DOI: 10.1534/genetics.108.098913

Binary Trait Mapping in Experimental Crosses With Selective Genotyping

Ani Manichaikul*,[1] and Karl W. Broman†

**Department of Biomedical Engineering, University of Virginia, Charlottesville, Virginia 22908 and †Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wisconsin 53706*

Manuscript received November 19, 2008
Accepted for publication April 21, 2009

ABSTRACT

Selective genotyping is an efficient strategy for mapping quantitative trait loci. For binary traits, where there are only two distinct phenotypic values (*e.g.*, affected/unaffected or present/absent), one may consider selective genotyping of affected individuals, while genotyping none or only some of the unaffecteds. If selective genotyping of this sort is employed, the usual method for binary trait mapping, which considers phenotypes conditional on genotypes, cannot be used. We present an alternative approach, instead considering genotypes conditional on phenotypes, and compare this to the more standard method of analysis, both analytically and by example. For studies of rare binary phenotypes, we recommend performing an initial genome scan with all affected individuals and an equal number of unaffecteds, followed by genotyping the full cross in genomic regions of interest to confirm results from the initial screen.

We consider the problem of mapping genetic loci contributing to a binary trait in an experimental cross with selective genotyping. There are two clear approaches for linkage analysis with a binary trait. Typically, we compare the proportion of affected individuals across genotype groups (Xu and Atchley 1996). Alternatively, we can compare genotype frequencies between affected and unaffected individuals, similar to Henshall and Goddard (1999). Beyond these two basic approaches, binary trait mapping has seen fundamental advances in regression models (McIntyre *et al.* 2001; Deng *et al.* 2006), extensions to multiple-QTL mapping (Coffman *et al.* 2005; Chen and Liu 2009), and the development of Bayesian algorithms (Yi and Xu 2000; Huang *et al.* 2007). However, the original data structure and approach have remained intact. Existing methods for binary trait mapping largely require the availability of genotype and phenotype data for a representative sample of both affected and unaffected individuals, and we have not yet seen a well-developed framework for binary trait mapping in the presence of selective genotyping.

It is not uncommon to see genotype data on affected individuals only, in which case the above methods cannot be used. Instead, we can compare observed genotype frequencies to the expected segregation ratios given the cross type, in a test for segregation distortion (see Faris *et al.* 1998; Lambrides *et al.* 2004). For example, the expected segregation proportions for an intercross are 1:2:1. The observed genotypes can then be described by a multinomial model, and statistically significant deviation from the expected segregation ratios among the genotyped affected individuals would suggest genotype–phenotype association. Gene mapping approaches that model genotypes rather than phenotypes have been developed extensively in the analysis of affected human relative pairs (see, for example, Risch 1990; Holmans 1993; Hauser and Boehnke 1998). In the analysis of experimental crosses, however, this type of approach has been developed primarily for the identification of monogenic mutants (Moran *et al.* 2006).

Once all affected individuals are genotyped, an investigator may go on to genotype unaffected individuals. With this genotyping strategy in mind, we present several potential methods of analysis that might be applied in this context. First, we consider a standard analysis of the genotyped individuals, with disease proportions compared across genotype groups (Xu and Atchley 1996). Having omitted ungenotyped individuals, this method of analysis appears invalid because the estimated disease proportions are biased upward, reflecting an overrepresentation of affecteds in the set of genotyped individuals under consideration. As an alternative, we develop a reverse approach with genotype frequencies compared across phenotype groups. Because selective genotyping does provide a representative sample of genotypes for each phenotype group, this reverse approach does not face the bias in parameter estimation seen with the standard approach. We further extend the reverse approach to incorporate a segregation assumption, as is necessary for an affecteds only analysis. Finally, we present a full-likelihood analysis accounting for selective genotyping, similar to

[1]*Corresponding author:* Department of Biomedical Engineering, University of Virginia, Box 800759, Health System, Charlottesville, VA 22908. E-mail: amanicha@virginia.edu

TABLE 1

Data at a single genetic marker from a backcross experiment with binary phenotypes

	Genotype			
	AA	*AB*	Missing	Total
Phenotype				
Affected	$n_{AA,D}$	$n_{AB,D}$	$n_{\text{mis},D}(=0)$	n_D
Unaffected	$n_{AA,\bar{D}}$	$n_{AB,\bar{D}}$	$n_{\text{mis},\bar{D}}$	$n_{\bar{D}}$
Total	n_{AA}	n_{AB}	$n_{\text{mis}}\ (=n_{\text{mis},\bar{D}})$	N

All affecteds are genotyped, while some or all unaffecteds may be left ungenotyped.

that suggested by Lander and Botstein (1989) for quantitative traits. We develop the full-likelihood approach both with and without incorporating an assumption on the genotype segregation proportions.

Having put forth each of these methods, we derive analytic relationships among them. These relationships provide important insight regarding application of the presented methods under selective genotyping. Most notably, we find that making a segregation assumption can lead to spurious evidence of a QTL, but is necessary to treat the case of affecteds only genotyping. We demonstrate properties of the methods in an analysis of recovery from infection by *Listeria monocytogenes* in intercross mice and further compare power of the methods through computer simulations. Finally, we synthesize our analytical and simulation results to offer more general suggestions for the analysis of binary trait data with selective genotyping.

METHODS

For simplicity, we present methods for the case of a backcross. For analysis, we consider backcross data consisting of binary phenotypes (affected or unaffected) for all individuals and marker genotypes (*AA* or *AB*) on all affecteds and all, some, or none of the unaffecteds. We present methods of analysis to address these three genotyping strategies for binary trait data. The observed data are represented in Table 1.

Throughout our description of the methods, we use the following notation. Let N be the total number of individuals in the cross, with n_{obs} genotyped individuals and $n_{\text{mis}} = N - n_{\text{obs}}$ ungenotyped individuals. We assign the indexes $i = 1, \ldots, N$ such that individuals $i \in \{1, \ldots, n_{\text{obs}}\}$ are genotyped, and the remaining individuals are ungenotyped. Let $D_i = 1$ or 0 according to whether individual i is affected or unaffected. We take $G_i \in \{AA, AB\}$ to denote the underlying unobserved genotype at the putative QTL of interest, while $O_{mi} \in \{AA, AB, -\}$ denotes the observed genotype at marker m, with " $-$ " indicating a missing value.

Standard approach: In dealing with selective genotyping, one possible approach is to ignore individuals without genotype data, performing an analysis of genotyped individuals only. Once we have omitted individuals with missing genotypes from our analysis, we can use the approach of Xu and Atchley (1996) to look for genotype–phenotype association by testing for a difference in disease probability across genotypes.

Let $\pi_{AA} = \Pr(D_i = 1 \mid G_i = AA)$ and $\pi_{AB} = \Pr(D_i = 1 \mid G_i = AB)$ denote the penetrance values (the conditional phenotype probabilities given the genotype at a putative QTL), and let $\pi_\cdot = \Pr(D_i = 1)$ denote the marginal phenotype probability (*i.e.*, the prevalence of disease).

Similar to standard interval mapping for quantitative traits (Lander and Botstein 1989), the approach of Xu and Atchley (1996) makes use of the conditional QTL genotype probabilities given the full set of multipoint marker data for the ith individual, $p_{ig} = \Pr(G_i = g \mid \mathbf{O}_{\cdot i})$. By convention, evidence against the null hypothesis of genotype–phenotype independence, $\Pr(\mathbf{D} \mid \mathbf{G}) = \Pr(\mathbf{D})$, in favor of the alternative hypothesis of a QTL, $\Pr(\mathbf{D} \mid \mathbf{G}) \neq \Pr(\mathbf{D})$, is presented as the $\log_{10}$-likelihood ratio

$$\begin{aligned}\text{LOD}_{\text{F}} &= \log_{10}\left\{\frac{\max_{\pi_{AA},\pi_{AB}} \Pr(\mathbf{D} \mid \mathbf{O}_{\cdot\cdot}; \pi_{AA}, \pi_{AB})}{\max_{\pi_\cdot} \Pr(\mathbf{D}; \pi_\cdot)}\right\} \\ &= \log_{10}\left\{\frac{\prod_i \sum_{g\in\{AA,AB\}} [p_{ig} \cdot (\hat{\pi}_g)^{D_i} \cdot (1-\hat{\pi}_g)^{1-D_i}]}{\prod_i (\hat{\pi}_\cdot)^{D_i} \cdot (1-\hat{\pi}_\cdot)^{1-D_i}}\right\},\end{aligned}$$

where we model affection status as a Bernoulli random variable with a common probability under the null and with genotype-specific probabilities under the alternative hypothesis. Assuming no missing genotype data for reasons other than the selective genotyping, and no genotyping error, the maximum-likelihood estimates (MLEs) at the markers are simply sample proportions. Between markers, we can perform interval mapping by an EM algorithm (Dempster *et al.* 1977), which has been previously described for this application (Xu and Atchley 1996; Broman 2003).

The forward approach using LOD_{F} is appropriate in the case that we have genotyped all individuals. However, if we have done selective genotyping with regard to phenotypes, the approach will yield biased and inconsistent estimates of π_{AA} and π_{AB}. As a result, the validity of this approach for selective genotyping is not immediately apparent.

Reverse approach, conditioning on phenotypes: As an alternative, we can also look for genotype–phenotype association by reversing the standard approach and instead modeling genotypes conditional on phenotypes. This approach is technically quite similar to the logistic regression model presented by Henshall and Goddard (1999), but we present it in a framework that elucidates its relationship with the standard approach of Xu and Atchley (1996). Placing the reverse approach in this likelihood framework also allows it to be easily adapted for analysis of affecteds only, as will be seen

TABLE 2

Summary of forward and reverse approaches and likelihood functions

	Hypothesis	MLEs	Likelihood
Forward			
Alternative	$\pi_{AA} \neq \pi_{AB}$	$\hat{\pi}_{AA} = \frac{n_{D,AA}}{n_{AA}}$	$\text{lik}(\hat{\pi}_{AA}, \hat{\pi}_{AB} \mid \mathbf{O}_{..}) = \prod_i \sum_{g\in\{AA,AB\}} [p_{ig} \cdot (\hat{\pi}_g)^{D_i} \cdot (1-\hat{\pi}_g)^{1-D_i}]$
		$\hat{\pi}_{AB} = \frac{n_{D,AB}}{n_{AB}}$	
Null	$\pi_{\cdot} = \pi_{AA} = \pi_{AB}$	$\hat{\pi}_{\cdot} = \frac{n_{D,AA} + n_{D,AB}}{n_{\text{obs}}}$	$\text{lik}(\hat{\pi}_{\cdot}) = \prod_i (\hat{\pi}_{\cdot})^{D_i} \cdot (1-\hat{\pi}_{\cdot})^{1-D_i}$
Reverse			
Alternative	$\phi_D \neq \phi_{\tilde{D}}$	$\hat{\phi}_D = \frac{n_{D,AA}}{n_{D,AA} + n_{D,AB}}$	$\text{lik}(\hat{\phi}_D, \hat{\phi}_{\tilde{D}} \mid \mathbf{D}) = \prod_i \sum_{g\in\{AA,AB\}} [q_{ig} \cdot \Pr(G_i = g \mid D_i; \phi_D, \phi_{\tilde{D}})]$
		$\hat{\phi}_{\tilde{D}} = \frac{n_{\tilde{D},AA}}{n_{\tilde{D},AA} + n_{\tilde{D},AB}}$	
Null	$\phi_{\cdot} = \phi_D = \phi_{\tilde{D}}$	$\hat{\phi}_{\cdot} = \frac{n_{AA}}{n_{\text{obs}}}$	$\text{lik}(\hat{\phi}_{\cdot}) = \prod_i \sum_{g\in\{AA,AB\}} \cdot \Pr(G_i = g; \phi_{\cdot})$
Modified Null	$\phi_{\cdot} = \phi_D = \phi_{\tilde{D}}$	$\hat{\phi}_{\cdot} = \frac{1}{2}$	$\text{lik}(\hat{\phi}_{\cdot} = \frac{1}{2})$

Analytical maximum-likelihood estimates (MLEs) are stated for marker locations.

below. Again, we consider only genotyped individuals and omit the rest from our analysis.

Let $\phi_D = \Pr(G_i = AA \mid D_i = 1)$ and $\phi_{\tilde{D}} = \Pr(G_i = AA \mid D_i = 0)$ denote the affection-status-specific probabilities of the *AA* genotype at the putative QTL of interest, and let $\phi_{\cdot} = \Pr(G_i = AA)$ denote the marginal probability of the *AA* genotype. (For an intercross, we must handle the three possible genotypes {*AA*, *AB*, *BB*}, and so we would consider the vector $\phi_{\cdot} = [\Pr(G_i = AA)\ \Pr(G_i = AB)]^{\mathrm{T}}$ and analogous vectors for ϕ_D and $\phi_{\tilde{D}}$.)

We calculate the LOD score measuring support for a QTL as the $\log_{10}$-likelihood ratio comparing evidence for the alternative hypothesis, $\Pr(\mathbf{D} \mid \mathbf{G}) \neq \Pr(\mathbf{D})$ [or equivalently, $\Pr(\mathbf{G} \mid \mathbf{D}) \neq \Pr(\mathbf{G})$], in favor of the null hypothesis of independence. Here, we model genotypes at the putative QTL using a Bernoulli process (or multinomial for an intercross) with a common probability under the null and with disease-status-specific probabilities under the alternative hypothesis (see Table 2).

To allow analysis at both marker and nonmarker locations, we perform interval mapping using an EM algorithm to calculate the necessary MLEs. Analogous to the p_{ig} for standard interval mapping, we make use of the reverse quantities, $q_{ig} = \Pr(\mathbf{O}_{\cdot i} \mid G_i = g)$, probabilities of the full set of observed marker data given a specified value of the underlying QTL genotype, *g*. We have developed hidden Markov models to obtain the q_{ig}; details are provided in APPENDIX A.

Reverse approach, modified: Having presented the reverse approach above, a simple modification allows us to incorporate knowledge regarding the structure of the cross. In particular, we may specify the null hypothesis value of $\phi_{\cdot}$ to be $\frac{1}{2}$ for a backcross (or $\phi_{\cdot} = \left[\frac{1}{4}\ \frac{1}{2}\right]^{\mathrm{T}}$ for an intercross). This prior knowledge is crucial in the analysis of affecteds only, for which it is infeasible to simply check for a difference in genotype proportions across phenotype groups, as was done above.

Here, the modified LOD score, $\text{LOD}_{\text{R,seg}} = \log_{10}\left\{\max_{\phi_D,\phi_{\tilde{D}}} \Pr(\mathbf{O}_{..} \mid \mathbf{D};\ \phi_D,\ \phi_{\tilde{D}}) / \Pr(\mathbf{O}_{..};\ \phi_{\cdot} = \frac{1}{2})\right\}$, quantifies evidence for segregation distortion, *i.e.*, deviation of observed genotype counts from their expected distribution. In an affecteds only analysis, this view of evidence suggests genotype–phenotype association and so indicates the presence of a QTL.

Since the alternative hypothesis remains the same as in the original reverse approach, LOD_{R}, the MLEs for ϕ_D and $\phi_{\tilde{D}}$ may be obtained in the same way as described above. Note that a reasonable approach to take with this method is to constrain the conditional genotype probabilities such that their weighted average, $\phi_D \cdot \pi_{\cdot} + \phi_{\tilde{D}} \cdot [1 - \pi_{\cdot}]$, is equal to the marginal genotype probability, $\phi_{..}$ However, we use the unconstrained value in calculating $\text{LOD}_{\text{R,seg}}$, incorporating the constraint through a full-likelihood analysis developed further below.

Full-likelihood analysis: Performing a full-likelihood analysis allows us to forgo conditioning on either genotypes or phenotypes. Conceptually, this model makes complete use of the available data in assessing evidence of a QTL. An apparent advantage of this approach is that ungenotyped individuals can be included in the likelihood. However, careful examination in the RESULTS below shows that full-likelihood analysis yields results quite similar to those of both the forward and reverse approaches.

The full-likelihood function models the joint probability of disease status and observed genotypes. We write the full likelihood of a QTL at the putative site of interest in terms of parameters ϕ_D, $\phi_{\tilde{D}}$, and $\pi_{\cdot}$ as

$$\begin{aligned}\text{lik}(\phi_D, \phi_{\tilde{D}}, \pi_{\cdot}) &= \prod_{i=1}^{N} \Pr(D_i, \mathbf{O}_{\cdot i}; \phi_D, \phi_{\tilde{D}}, \pi_{\cdot}) \\ &= \text{lik}(\phi_D, \phi_{\tilde{D}} \mid \mathbf{D}) \cdot \text{lik}(\pi_{\cdot}). \qquad (1)\end{aligned}$$

Since the full likelihood can be decomposed into orthogonal components to separate ϕ_D and $\phi_{\tilde{D}}$ from $\pi_{\cdot}$

(see APPENDIX B for details), the resulting MLEs are simply those obtained by performing the maximization separately. At the markers, these are again the appropriate sample averages as specified in the previous sections above. Estimating the position of a QTL does not depend on the value $\pi_\cdot$ since this parameter estimate is fixed across the genome.

Constrained full likelihood: Analogous to the modified reverse approach that incorporates evidence for segregation distortion, we can also perform full-likelihood analysis under the assumption that marginal genotype probabilities should follow their null segregation values. The resulting full-likelihood function is the same as that specified above, but subject to the constraint that disease-status-specific genotype probabilities average to a marginal value of $\phi_\cdot = \frac{1}{2}$. We write this constraint in terms of the overall probability of disease, $\pi_\cdot$:

$$\phi_\cdot = \phi_D \cdot \pi_\cdot + \phi_{\tilde{D}} \cdot [1 - \pi_\cdot] = \frac{1}{2}. \qquad (2)$$

(For the case of an intercross, $\phi_\cdot = \left[\frac{1}{4}\ \frac{1}{2}\right]^{\mathrm{T}}$, so the equation above becomes a two-component constraint.) Analysis is performed using all N individuals in the cross, both genotyped and ungenotyped. The constrained full-likelihood LOD score is written as

$$\mathrm{LOD}^{N}_{\mathrm{Full,seg}} = \log_{10}\left\{\frac{\max_{\phi_D,\phi_{\tilde{D}},\pi_\cdot|\phi_\cdot=1/2}\mathrm{lik}(\phi_D,\phi_{\tilde{D}},\pi_\cdot)}{\max_{\pi_\cdot}\mathrm{lik}(\phi_\cdot=1/2,\pi_\cdot)}\right\}.$$

Under the null hypothesis, the MLE $\tilde{\pi}_\cdot = n_D/N$ is the same as in the absence of the constraint, while $\tilde{\phi}_\cdot = \frac{1}{2}$ according to the segregation assumption. To maximize the constrained likelihood under the alternative hypothesis, we use an EM algorithm, described in APPENDIX C.

Significance thresholds: After performing a genome scan using any of the methods presented above, we can make use of significance thresholds in reporting statistical significance for genomic regions of interest. The significance thresholds must account for multiple comparisons arising in the complete genome scan. A typical way to perform this adjustment while controlling the rate of detecting false positive QTL is to examine the distribution of the genomewide maximum LOD score under the global null hypothesis of no QTL.

For standard interval mapping, significance thresholds conditioning on observed genotypes and phenotypes may be obtained empirically by permutation (CHURCHILL and DOERGE 1994), shuffling phenotypes while keeping genotypes fixed to approximate the null distribution of the genomewide maximum LOD score.

In the case of methods incorporating a segregation assumption, such as $\mathrm{LOD}_{\mathrm{R,seg}}$, it may not make sense to condition on observed genotypes. For example, in an affecteds only analysis, the observed genotypes contain all of the information we use to test for linkage. If we condition on those observed genotypes in calculating the null distribution, we effectively condition out any evidence in the data. Put more simply, we cannot shuffle phenotypes in an affecteds only analysis because all individuals have the same phenotype. Instead, we can estimate the null distribution by simulation using a gene-dropping approach (MACCLUER *et al.* 1986) to simulate new genotypes preserving the cross structure and pattern of missing genotypes from the original data set. The resulting significance thresholds are reported conditional on observed phenotypes, while averaging across possible sets of genotypes given the cross used to generate the data. Since the simulation-based significance thresholds do not condition on the observed genotypes, they are appropriate for analyses in which we have incorporated evidence for segregation distortion or deviation from expected segregation ratios. We describe this form of evidence more explicitly in the RESULTS below.

RESULTS

We have presented approaches to calculating likelihoods and corresponding likelihood ratios to assess genotype–phenotype association for binary trait mapping in the presence of selective genotyping. Since all of these methods are likelihood based, they are closely related. In this section, we highlight key relationships between the various approaches. We summarize all presented relationships at the end of this section.

Reverse approach *vs.* modified reverse approach: There is a direct relationship between $\mathrm{LOD}_{\mathrm{R}}$, in which we use the MLE $\hat{\phi}_\cdot$ to calculate the null likelihood, and $\mathrm{LOD}_{\mathrm{R,seg}}$, in which we assume $\phi_\cdot = \frac{1}{2}$ according to the expected genotype frequencies in a backcross,

$$\begin{aligned}\mathrm{LOD}_{\mathrm{R,seg}} &= \log_{10}\left\{\frac{\mathrm{lik}(\hat{\phi}_D,\hat{\phi}_{\tilde{D}}\,|\,\mathbf{D},\mathbf{O}_{m\cdot})}{\mathrm{lik}(\phi_\cdot=\frac{1}{2}\,|\,\mathbf{O}_{m\cdot})}\right\}\\ &= \log_{10}\left\{\frac{\mathrm{lik}(\hat{\phi}_D,\hat{\phi}_{\tilde{D}}\,|\,\mathbf{D},\mathbf{O}_{m\cdot})}{\mathrm{lik}(\hat{\phi}_\cdot\,|\,\mathbf{O}_{m\cdot})}\times\frac{\mathrm{lik}(\hat{\phi}_\cdot\,|\,\mathbf{O}_{m\cdot})}{\mathrm{lik}(\phi_\cdot=\frac{1}{2};\mathbf{O}_{m\cdot})}\right\}\\ &= \mathrm{LOD}_{\mathrm{R}} + \mathrm{LOD}_{\mathrm{seg.dist.}},\end{aligned}$$

where $\mathrm{LOD}_{\mathrm{seg.dist.}} = \log_{10}\{\mathrm{lik}(\hat{\phi}_\cdot\,|\,\mathbf{O}_{m\cdot})/\mathrm{lik}(\phi_\cdot=\frac{1}{2}\,|\,\mathbf{O}_{m\cdot})\}$ quantifies evidence for segregation distortion or deviation of the observed genotypes from the assumed segregation proportion, $\phi_\cdot = \frac{1}{2}$.

Full likelihood *vs.* reverse approach: Let $\mathrm{LOD}^{n_{\mathrm{obs}}}_{\mathrm{Full}}$ be the full-likelihood LOD score based on genotyped individuals only and $\mathrm{LOD}^{N}_{\mathrm{Full}}$ be the corresponding LOD score obtained from all individuals whether genotyped or not. The LOD score to test for genotype–phenotype association based on the reverse approach is closely related to the corresponding LOD based on full-likelihood analysis. Recall that the full likelihood can be

decomposed as a product of the likelihood for parameters ϕ_D, $\phi_{\bar{D}}$ and the likelihood for $\pi_\cdot$, as shown in (1) above. So, we can factor out the likelihood for phenotype probability from the full likelihood to relate it back to the reverse approach. Because this factorization applies at both marker and nonmarker locations, the relationship $\mathrm{LOD}^{n_{\mathrm{obs}}}_{\mathrm{Full}} = \mathrm{LOD}^{N}_{\mathrm{Full}} = \mathrm{LOD}_{\mathrm{R}}$ holds quite generally even in the presence of missing genotype data or genotyping error (see APPENDIX D for details).

Forward *vs.* reverse approach: The forward and reverse methods of analysis are closely related as they are both likelihood-ratio-based tests of independence. For the case of no missing genotypes or genotyping error, we can examine the relationship analytically at marker locations

$$\begin{aligned}\mathrm{LOD}_{\mathrm{F}} &= \log_{10}\left\{\frac{\Pr(\mathbf{D}\mid\mathbf{O}_{\cdot\cdot};\hat{\pi}_{AA},\hat{\pi}_{AB})}{\Pr(\mathbf{D};\hat{\pi}_{\cdot})}\right\}\\ &= \log_{10}\left\{\frac{\Pr(\mathbf{D}\mid\mathbf{O}_{\cdot\cdot};\hat{\pi}_{AA},\hat{\pi}_{AB})\cdot\Pr(\mathbf{O}_{\cdot\cdot};\hat{\phi}_{\cdot})}{\Pr(\mathbf{D};\hat{\pi}_{\cdot})\cdot\Pr(\mathbf{O}_{\cdot\cdot};\hat{\phi}_{\cdot})}\right\}(=\mathrm{LOD}^{n_{\mathrm{obs}}}_{\mathrm{Full}})\\ &= \mathrm{LOD}_{\mathrm{R}},\end{aligned}$$

where the estimated parameters $\hat{\pi}_{AA}$, $\hat{\pi}_{AB}$, and $\hat{\phi}_{\cdot}$ are the MLEs as specified in the METHODS section above. Note that these relationships apply only at marker locations, in the case of no missing genotypes. It is in this special case that all MLEs are obtained as sample averages, and so the computed likelihood ratio is the same whether conditional on genotypes or phenotypes. At nonmarker locations, we must employ the relationships in (A6) to obtain the full-likelihood MLEs, $\hat{\pi}_{AA}$, $\hat{\pi}_{AB}$, and $\hat{\phi}_{\cdot}$, so these values could differ from those obtained by the standard approach, $\mathrm{LOD}_{\mathrm{F}}$.

Overall relationships: Here, we summarize the relationships among the proposed methods presented here, together with those shown in APPENDIX D. At the markers we have the following relationships:

$$\mathrm{LOD}^{n_{\mathrm{obs}}}_{\mathrm{Full}} = \mathrm{LOD}_{\mathrm{F}} = \mathrm{LOD}_{\mathrm{R}}$$

$$\mathrm{LOD}_{\mathrm{R,seg}} = \mathrm{LOD}_{\mathrm{F}} + \mathrm{LOD}_{\mathrm{seg.dist.}}.$$

The following relationships hold more generally at both marker and nonmarker locations:

$$\mathrm{LOD}^{n_{\mathrm{obs}}}_{\mathrm{Full}} = \mathrm{LOD}^{N}_{\mathrm{Full}} = \mathrm{LOD}_{R}$$

$$\mathrm{LOD}^{N}_{\mathrm{Full,seg}} \le \mathrm{LOD}_{\mathrm{R,seg}} = \mathrm{LOD}_{\mathrm{R}} + \mathrm{LOD}_{\mathrm{seg.dist.}}$$

That $\mathrm{LOD}_{\mathrm{R}}$ agrees with full-likelihood analysis, whether or not we include ungenotyped individuals in the analysis, suggests it is unnecessary to perform full-likelihood analysis, since we can get the exact same results using the simpler reverse approach. However, we also see that the modified reverse approach incorporates evidence for segregation distortion, which can be irrelevant to linkage if we have genotyped both affecteds and unaffecteds. Hence, full-likelihood analysis may still be necessary to incorporate the segregation assumption while avoiding spurious evidence, as in $\mathrm{LOD}^{N}_{\mathrm{Full,seg}}$.

APPLICATION

To demonstrate features of our proposed methods, we perform analysis of recovery from *L. monocytogenes* infection in 116 mice from an intercross of the resistant strain C57BL/6ByJ and the susceptible strain BALB/cByJ (BOYARTCHUK *et al.* 2001), using the data set available in the R/qtl package (BROMAN *et al.* 2003). In our analysis, we make use of genotypes at 131 genetic markers on the 19 autosomes. Although phenotypes were recorded as survival times in the original study, we converted them to binary values to demonstrate application to our proposed methods. (Note that analyzing survival data as binary values is only one possible strategy in handling survival data; see BROMAN 2003 for a more complete treatment of this particular data set.) Accordingly, binary phenotypes were calculated to indicate whether or not mice survived to 264 hr following infection. Among the 116 phenotyped mice in this data set, 35 survived and 81 died within 264 hr. Since survival is the rarer phenotype in this cross, we refer to the 35 survivors as affected individuals. With the full data, an appropriate analysis would use the standard approach, $\mathrm{LOD}_{\mathrm{F}}$, with the available full genotypes. To explore the set of possible genotyping strategies using real data, we subset the available genotypes to create two additional versions of this data set for analysis: one data set with genotypes on affecteds only and another with equal numbers of affecteds and unaffecteds genotyped. In each case, we apply methods of analysis appropriate for the genotyping strategy at hand.

Since some of the presented methods are sensitive to segregation distortion, we note in advance that chromosome 13 showed the strongest evidence of overall deviation from the expected genotype proportions. In particular, the marker D13M233 had genotype segregation proportions of 40:41:21 rather than the expected 1:2:1, giving a segregation distortion LOD score of 2.97 ($P = 0.097$ by gene-dropping simulation with 10,000 replicates). In contrast, the marker D2M365 on chromosome 2 showed little segregation distortion, with segregation proportions of 24:59:27 and a segregation distortion LOD score of 0.12.

Standard analysis: We first consider a standard analysis with full genotypes. In this case, it is appropriate to perform standard interval mapping using the forward approach, $\mathrm{LOD}_{\mathrm{F}}$, conditioning phenotypes on genotypes. In calculating our LOD curve, we use an EM algorithm, as implemented in R/qtl (BROMAN *et al.* 2003), with genotype probabilities calculated every 1 cM. We use a permutation test (CHURCHILL and

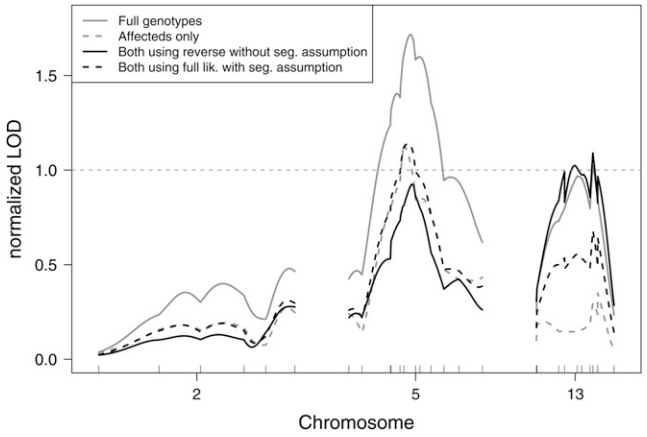

Figure 1.—Analysis of intercross data from Boyartchuk *et al.* (2001) with significant results on chromosomes 5 and 13, and chromosome 2 shown for comparison. LOD curves are generated using four different methods and normalized by their corresponding 5% significance thresholds for comparison: (i) The LOD with full genotypes is calculated by standard interval mapping according to the method of Xu and Atchley (1996) (shaded line), with a 5% permutation threshold of 3.57; (ii) with genotypes on affecteds only, the LOD curve is calculated by the reverse approach using an intercross segregation assumption, $\mathrm{LOD}_{\mathrm{R,seg}}$ (dashed shaded line), and the appropriate 5% simulation threshold is 3.57; (iii) using genotypes on 35 unaffecteds and all 35 affecteds, the LOD curve is calculated by the reverse approach, $\mathrm{LOD}_{\mathrm{R}}$ (solid line), with a corresponding 5% permutation threshold of 3.65; and (iv) using genotypes on 35 unaffecteds and all 35 affecteds, the LOD curve is calculated by the full likelihood with the segregation assumption, $\mathrm{LOD}^{N}_{\mathrm{Full,seg}}$ (dashed solid line), using constrained maximum likelihood. The corresponding 5% permutation threshold is 3.56.

Doerge 1994) to assign statistical significance and obtain (with 10,000 permutation replicates) a 5% significance threshold of 3.57.

The LOD curve (Figure 1) has statistically significant peaks on chromosomes 5 and 13, with corresponding *P*-values of <0.001 and 0.042, respectively. The evidence seen on chromosome 13 is not influenced by segregation distortion and shows that genotype proportions differ significantly by affection status.

Affecteds only analysis: If only the 35 affected individuals were genotyped, the forward and reverse approaches, $\mathrm{LOD}_{\mathrm{F}}$ and $\mathrm{LOD}_{\mathrm{R}}$, respectively, are not appropriate because they will always produce LOD scores of strictly zero. Instead, we calculate the LOD curve using the reverse approach, $\mathrm{LOD}_{\mathrm{R,seg}}$, making use of the intercross segregation assumption, with genotypes *CC*, *CB*, and *BB* segregating according to the ratios 1:2:1. A 5% significance threshold is calculated by simulation to be 3.57.

The LOD peak on chromosome 5 (Figure 1) is statistically significant with a LOD score of 4.07 ($P = 0.019$). To follow up on this result, it is appropriate to obtain genotypes for the full cross at the peak position on chromosome 5. In our follow-up analysis of D5M357, we test for a difference in genotype proportions across affected and unaffected individuals. Among affected individuals, the observed genotype distribution was 18:16:1, compared to 12:39:30 among unaffecteds, yielding an exact pointwise *P*-value of 2.9×10^{-6}. On the basis of this analysis, we conclude there is evidence for a QTL on chromosome 5, and the significant result obtained using affecteds only was not driven by spurious segregation distortion.

Although there was good evidence of a QTL on chromosome 13 using full genotypes, we detected no evidence of a QTL on chromosome 13 using affecteds only (maximum LOD score of 1.25, $P > 0.99$). The results on chromosome 13 demonstrate that the reverse approach, $\mathrm{LOD}_{\mathrm{R,seg}}$, provides reliable results only when there is little segregation distortion in the overall set of genotypes. In this particular data set, there was strong segregation distortion among the pooled set of affecteds and unaffecteds, while affected individuals alone had genotype proportions close to their null values.

Equal numbers of affecteds and unaffecteds genotyped: By genotyping both affecteds and unaffecteds, we no longer require the use of a segregation assumption and so can avoid the problem of distorted evidence that we encountered in our affecteds only analysis. Here, we consider the same intercross as above, but now we have genotyped 35 unaffected mice (selected at random), in addition to the 35 affected individuals. In this case, we perform analysis using the reverse approach, $\mathrm{LOD}_{\mathrm{R}}$, and use permutation to get a 5% significance threshold of 3.65.

The profile of our observed LOD curves (Figure 1) is very similar to that obtained by standard analysis with complete genotyping. The overall strength of the signal obtained by partial genotyping is somewhat attenuated, but we still have reasonable evidence of QTL on chromosome 5 and 13, with *P*-values 0.085 and 0.026, respectively. After the initial genome scan using this portion of the cross, we recommend following up with genotypes on the full cross in genomic regions of interest. For example, on chromosome 5 we find that the remaining unaffected individuals show D5M357 genotype proportions of 5:21:20, compared to 7:18:10 among the first 35 genotyped unaffecteds. These results confirm that the unaffecteds overall have a relatively larger proportion of homozygote *BB* individuals and smaller proportion of *CC* individuals. Further, the peak positions obtained by our partial genotyping strategy are identical to those obtained by analysis of the full cross. Thus, genotyping an equal number of affected and unaffected individuals combined with follow-up using the full cross provides an effective way to locate QTL while vastly reducing the amount of genotyping required.

To examine sensitivity of our results to randomness in the set of unaffected individuals selected for genotyping, we repeated the analysis with 100 different sets of 35

randomly selected unaffected mice. We saw qualitatively similar results across this set of replicates, with reasonable evidence of QTL on chromosomes 5 and 13 in the majority of samples (results not shown). This investigation suggests genotyping all affecteds and an equal number of unaffecteds is an effective way to capture evidence of QTL in the full cross, while genotyping only a fraction of the individuals.

Some unaffecteds genotyped using constrained maximum likelihood: Incorporating a segregation assumption is most useful for affecteds only analysis, where we have zero power to map QTL without such an assumption. Here, we consider incorporating this assumption in the more moderate case of selective genotyping with equal numbers of affecteds and unaffecteds genotyping.

We examine the same set of genotyped individuals as above, with genotypes on 35 unaffecteds and all 35 affected individuals, and perform analysis with the inclusion of a segregation assumption. Toward this end, both the reverse approach and the full likelihood are possible options to accommodate the assumption. However, as shown in RESULTS above, $LOD_{R,seg}$ is equivalent to a standard analysis, plus the LOD for segregation distortion. In mapping QTL, we are generally not interested in overall segregation distortion. We perform analysis by $LOD^{N}_{Full,seg}$ to eliminate the possibility of spurious evidence. Since we have eliminated evidence for segregation distortion, we use permutation, rather than simulation, to obtain a 5% significance threshold of 3.56.

The overall shape of the LOD curve produced by this approach agrees quite closely with the full and partial genotyping results shown in Figure 1. Still, we do note differences that reflect properties of the methods. The peak on chromosome 5 shows a LOD score of 4.06 ($P = 0.018$), greater than the value of 3.37 using the reverse approach. On the other hand, the peak evidence on chromosome 13 using $LOD^{N}_{Full,seg}$ was only 2.42 ($P = 0.449$), which is considerably less than the reverse LOD of 3.98. These results show that constrained maximum likelihood can improve the strength of our signal, particularly when the segregation assumption matches the observed data well. At the same time, using a segregation assumption can also attenuate the strength of evidence when there is deviation from this assumption, as seen on chromosome 13.

POWER STUDIES

Simulations were performed varying cross type, heritability, and expected proportion of affecteds, to investigate the impact of these factors on power to detect a QTL across four approaches to binary trait mapping. Data were generated for backcrosses of 250 individuals and intercrosses of 500 individuals, using a marker map based on the mouse genome with markers about every 10 cM [the full map is included with the R/qtl package (BROMAN *et al.* 2003)]. In all simulations, a single QTL was placed between the sixth and seventh markers on chromosome 1 with heritability of the continuous liability phenotype (XU and ATCHLEY 1996) set at either 5 or 10%. Binary traits were generated from the continuous liability values, with thresholds set such that the expected proportion of affecteds was either 10 or 25%.

The four mapping strategies assessed in the simulations are the same as those presented in the APPLICATION: (1) full genotyping of the cross using the standard approach of XU and ATCHLEY (1996), (2) affecteds only analysis using $LOD_{R,seg}$, (3) genotypes on all affecteds and an equal number of unaffecteds using the reverse approach, LOD_R, without the segregation assumption, and (4) genotypes on all affecteds and an equal number of unaffecteds using constrained full-likelihood analysis, $LOD^{N}_{Full,seg}$. For each set of parameter values and each of the four mapping strategies, we obtained a 5% significance threshold as the 95% quantile of the distribution of genomewide maximum LOD scores under the null, as estimated by 10,000 simulation replicates.

Power for all combinations of parameters based on 10,000 simulation replicates is shown in Table 3. For each of the scenarios considered, the highest power was obtained by full genotyping with standard analysis. Affecteds only analysis and constrained full-likelihood analysis with partial genotyping had comparable power to detect a QTL, with slightly lower power in the affecteds only analysis under all investigated scenarios. Finally, reverse analysis with partial genotyping showed notably lower power to detect a QTL compared to the other three approaches.

When the affected phenotype is rare, an affecteds only analysis can provide power comparable to analysis of the full cross and requires only a small fraction of the genotyping. To check for spurious evidence due to segregation distortion, further genotyping of unaffecteds can serve as a useful supplement to affecteds only analysis, particularly when incorporated by constrained full-likelihood analysis, and with greater improvements seen when the affected phenotype is more common. Although analysis of affecteds and unaffecteds using the reverse approach, LOD_R, has lower power than other approaches, we should keep in mind that this approach is more robust to segregation distortion than the constrained full-likelihood analysis, which can suffer from reduced power in the presence of segregation distortion. The relatively weaker performance of the reverse approach, LOD_R, suggests this robust strategy is more suitable as a follow-up check for segregation distortion, rather than as a genomewide QTL mapping strategy.

DISCUSSION

We have presented methods for linkage analysis of binary phenotypes in the presence of selective genotyp-

TABLE 3

Power to detect QTL using four different strategies for binary trait mapping

Expected proportion of affecteds (%)	Cross type	Heritability (%)	Power (%) of methods			
			Full genotypes	Affecteds only	Both with reverse	Both with full likelihood
10	Backcross	5	9.5	7.7	4.2	7.7
		10	28.6	24.2	12.9	24.8
	Intercross	5	19.9	16.5	7.1	16.7
		10	59.6	52.0	25.5	53.0
25	Backcross	5	20.2	12.5	10.4	14.8
		10	58.3	39.9	34.6	46.4
	Intercross	5	41.7	26.0	22.1	31.3
		10	89.5	72.4	64.2	79.3

Estimates are based on 10,000 simulation replicates for each combination of parameter values, in backcrosses of 250 individuals and intercrosses of 500 individuals. The segregation assumption is incorporated for affecteds only analysis, as well as for full-likelihood analysis of affecteds and some unaffecteds.

ing. As alternatives to standard interval mapping, we presented a reverse approach of modeling genotypes conditional on phenotypes and also a full-likelihood approach. Our suggested modifications to the standard approach of Xu and Atchley (1996) are developed in terms of fundamental likelihood modeling strategies. Accordingly, a key contribution here is our presentation of approaches to binary trait analysis using a cohesive likelihood framework, elucidating fundamental relationships among the methods. Our formal development of allele sharing methods presented as the reverse approach, LOD_R, led to the use of hidden Markov models to allow interval mapping in the reverse and full-likelihood approaches (appendixes a and c). Through analytical comparisons, we found that our reverse approach, LOD_R, is identical to a full-likelihood approach at both marker and nonmarker locations.

We also proposed another version of the reverse approach, $LOD_{R,seg}$, which incorporates a segregation assumption of the expected genotype proportions based on the type of cross that was performed. This approach formalizes a natural method of analysis for dealing with genotypes on affecteds only and presents it in a more general form that can be applied with genotypes on both affecteds and unaffecteds. We found the $\log_{10}$-likelihood ratio, $LOD_{R,seg}$, could be decomposed as the sum of two types of evidence: (1) deviation of genotype proportions by phenotype group and (2) segregation distortion. For the case of genotypes on affecteds only, incorporating the segregation assumption was especially crucial, as it provided a view of evidence for a QTL where none was available by LOD_R or LOD_F.

The inclusion of evidence for segregation distortion was deemed inappropriate in dealing with data having genotypes available on both affecteds and unaffecteds. For this case, we proposed incorporating the segregation assumption using a constrained full-likelihood approach. In this way, the segregation assumption was imposed under both the null and alternative hypotheses, so that the resulting test statistic, $LOD^N_{Full,seg}$, did not contain evidence for segregation distortion. Eliminating evidence for segregation distortion helps ensure that a large LOD score indicates evidence of a QTL, as segregation distortion can arise simply by random chance, systematically as a result of genotyping error, or as a result of embryonic lethal alleles that are unrelated to the trait of interest.

An understanding of these approaches as they relate to one another helps us to decide which method to use on the basis of the existing pattern of selective genotyping. In the case that we have genotyped everybody in our sample, we are not interested in overall evidence for segregation distortion and would choose a standard approach LOD_F. On the other hand, if we have genotyped affecteds only, the standard approach does not allow us to detect association. In this case, all evidence of association will be captured as evidence for segregation distortion, which shows up only with use of $LOD_{R,seg}$ as $LOD_{seg.dist.}$.

Since full genotyping of a cross can be costly, while affecteds only analysis can be prone to spurious evidence, a reasonable balance is to genotype some affecteds and some unaffecteds. Specifically, we recommend an initial screen with genotypes on all affected individuals and an equal number of unaffecteds, followed by analysis of the full cross in genomic regions of interest. As demonstrated in the application, this economical strategy can be an effective way to characterize QTL from the full cross and requires only a fraction of the genotyping. The ideal selective genotyping approach for any particular study may of course vary from this recommendation and could be studied as a function of animal rearing and phenotyping costs relative to genotyping cost.

The reverse approach, LOD_R, with no segregation assumption is a natural method of analysis for data with genotypes on affecteds and some unaffecteds. Although this approach was shown to be quite similar to the standard approach LOD_F at marker locations, we prefer LOD_R as it is not susceptible to biased parameter estimates and so produces more reliable results in between markers.

A full-likelihood analysis with constrained maximization under both the null and alternative hypotheses, presented as $LOD^{N}_{Full,seg}$, is another reasonable way to approach selective genotyping data with both affecteds and unaffecteds. Incorporating the segregation assumption in this setting is a practical compromise to preserve the evidence reported in an affecteds only analysis while bringing in unaffected individuals. The drawback, as seen in the Listeria example (Boyartchuk *et al.* 2001), is that evidence can be attenuated when there is overall segregation distortion in the data. Still, our computer simulation studies indicate constrained full-likelihood analysis offers notably higher power than the reverse approach, LOD_R, across a variety of parameter values. Thus, our power studies suggest constrained full-likelihood analysis is preferable as long as there is no pervasive segregation distortion in the cross.

A further limitation of the reverse approach lies in the treatment of multiple-QTL models. While single-QTL models may be set up quite naturally by conditioning genotypes at a single locus on the observed phenotypes, modeling genotypes at multiple loci can be much more cumbersome. Instead, when exploring multiple-QTL models with data on both affecteds and unaffecteds available, the standard approach of conditioning phenotypes on genotypes is more natural. The close relationship between the forward and reverse approaches in a single-QTL scan makes it quite reasonable to go ahead with the forward approach for the consideration of multiple-QTL models in the presence of selective genotyping. When using the forward approach for multiple-QTL mapping under selective genotyping, inferences at nonmarker positions may still be somewhat unreliable, while entirely valid results will be produced for models involving marker positions only.

After performing an analysis of a cross experiment under selective genotyping, we may always follow up by genotyping all individuals in genomic regions of interest identified from the initial scan. Such follow-up will be especially important if the initial scan is performed with affecteds only, since this strategy is most sensitive to spurious evidence due to segregation distortion.

We thank William Pu at the Children's Hospital in Boston for presenting us with data motivating this research. This work was supported in part by National Institutes of Health grant GM074244 (to K.W.B.) and by a National Science Foundation Graduate Research Fellowship (to A.M.).

LITERATURE CITED

Boyartchuk, V. L., K. W. Broman, R. E. Mosher, S. E. D'Orazio, M. N. Starnbach *et al.*, 2001 Multigenic control of *Listeria monocytogenes* susceptibility in mice. Nat. Genet. **27:** 259–260.

Broman, K. W., 2003 Mapping quantitative trait loci in the case of a spike in the phenotype distribution. Genetics **163:** 1169–1175.

Broman, K. W., H. Wu, S. Sen and G. A. Churchill, 2003 R/qtl: QTL mapping in experimental crosses. Bioinformatics **19:** 889–890.

Chen, Z., and J. Liu, 2009 Mixture generalized linear models for multiple interval mapping of quantitative trait loci in experimental crosses. Biometrics (in press).

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

Coffman, C. J., R. W. Doerge, K. L. Simonsen, K. M. Nichols, C. K. Duarte *et al.*, 2005 Model selection in binary trait locus mapping. Genetics **170:** 1281–1297.

Dempster, A., N. Laird and D. Rubin, 1977 Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Stat. Soc. Ser. B **39:** 1–38.

Deng, W., H. Chen and Z. Li, 2006 A logistic regression mixture model for interval mapping of genetic trait loci affecting binary phenotypes. Genetics **172:** 1349–1358.

Faris, J. D., B. Laddomada and B. S. Gill, 1998 Molecular mapping of segregation distortion loci in *Aegilops tauschii*. Genetics **149:** 319–327.

Hauser, E. R., and M. Boehnke, 1998 Genetic linkage analysis of complex genetic traits by using affected sibling pairs. Biometrics **54:** 1238–1246.

Henshall, J. M., and M. E. Goddard, 1999 Multiple-trait mapping of quantitative trait loci after selective genotyping using logistic regression. Genetics **151:** 885–894.

Holmans, P., 1993 Asymptotic properties of affected-sib-pair linkage analysis. Am. J. Hum. Genet. **52:** 362–374.

Huang, H., C. D. Eversley, D. W. Threadgill and F. Zou, 2007 Bayesian multiple quantitative trait loci mapping for complex traits using markers of the entire genome. Genetics **176:** 2529–2540.

Lambrides, C. J., I. D. Godwin, R. J. Lawn and B. C. Imrie, 2004 Segregation distortion for seed testa color in Mungbean (*Vigna radiata* L. Wilcek). J. Hered. **95:** 532–535.

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

Lander, E. S., P. Green, J. Abrahamson, A. Barlow, M. J. Daly *et al.*, 1987 MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics **1:** 174–181.

MacCluer, J., J. Vandeburg, B. Read and O. Ryder, 1986 Pedigree analysis by computer simulation. Zoo Biol. **5:** 149–160.

McIntyre, L. M., C. J. Coffman and R. W. Doerge, 2001 Detection and localization of a single binary trait locus in experimental populations. Genet. Res. **78:** 79–92.

Moran, J. L., A. D. Bolton, P. V. Tran, A. Brown, N. D. Dwyer *et al.*, 2006 Utilization of a whole genome SNP panel for efficient genetic mapping in the mouse. Genome Res. **16:** 436–440.

Nelder, J., and R. Mead, 1965 A simplex method for function minimization. Comput. J. **7:** 308–313.

Risch, N., 1990 Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am. J. Hum. Genet. **46:** 229–241.

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

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

Communicating editor: R. W. Doerge

APPENDIX A: THE REVERSE APPROACH AT NONMARKER LOCATIONS

We describe the algorithm for obtaining $\hat{\phi}_\cdot$ here. Analogous methods for obtaining $\hat{\phi}_D$ and $\hat{\phi}_{\tilde{D}}$ follow directly by applying the same algorithm within each of the two phenotype groups.

Given the full set of observed genotype data $\mathbf{O}_{\cdot i} = (O_{1i}, \ldots, O_{pi})$ for a single individual at the p putative QTL positions to be considered, let G_i denote the underlying genotype at the mth putative site of interest. We can expand the probability of the observed marker data given the underlying genotype at the putative QTL of interest as

$$\begin{aligned} q_{ig} &= \Pr(O_{1i}, \ldots, O_{(m-1)i} \mid G_i = g) \times \Pr(O_{mi} \mid G_i = g) \times \Pr(O_{(m+1)i}, \ldots, O_{mi} \mid G_i = g) \\ &= \beta^{\mathrm{l}}_{mi}(g) \times e(g,\ O_{mi}) \times \beta^{\mathrm{r}}_{mi}(g), \end{aligned}$$

where the conditional probabilities of observed marker data, $\beta^{\mathrm{l}}_{mi}(g)$ and $\beta^{\mathrm{r}}_{mi}(g)$, to the left and right of putative QTL may be obtained inductively using the backward equations in the context of hidden Markov models (HMMs) (Lander *et al.* 1987). Here, $e(g,\ O_{mi})$ is the corresponding emission probability at the mth genetic position of interest for individual i, which can also be interpreted as the genotyping error rate.

The likelihood function for the parameter $\phi_\cdot$ based on the observed genotype data $\mathbf{O}_{\cdot i}$ on individuals $i \in \{1, \ldots, n_{\mathrm{obs}}\}$ is

$$\begin{aligned} \mathrm{lik}(\phi_\cdot) &= \prod_{i=1}^{n_{\mathrm{obs}}} \Pr(\mathbf{O}_{\cdot i}; \phi_\cdot) \\ &= \prod_{i=1}^{n_{\mathrm{obs}}} \sum_{g \in \{AA, AB\}} [q_{ig} \cdot \Pr(G_i = g; \phi_\cdot)]. \end{aligned}$$

At iteration $s + 1$, we have the parameter estimate, $\hat{\phi}_\cdot^{(s)}$. In the E-step, we calculate the expected number of individuals with genotype AA at the putative QTL of interest as

$$\begin{aligned} \hat{n}_{AA}^{(s+1)} &= \sum_{i=1}^{n_{\mathrm{obs}}} \frac{\Pr(G_i = AA, \mathbf{O}_{\cdot i}; \hat{\phi}_\cdot^{(s)})}{\Pr(\mathbf{O}_{\cdot i}; \hat{\phi}_\cdot^{(s)})} \\ &= \sum_{i=1}^{n_{\mathrm{obs}}} \frac{\hat{\phi}_\cdot^{(s)} \cdot q_{i,AA}}{\hat{\phi}_\cdot^{(s)} \cdot q_{i,AA} + (1 - \hat{\phi}_\cdot^{(s)}) \cdot q_{i,AB}}. \end{aligned} \tag{A1}$$

In the M-step, the updated parameter estimate is simply $\hat{\phi}_\cdot^{(s+1)} = \hat{n}_{AA}^{(s+1)} / n_{\mathrm{obs}}$. A reasonable initial estimate of $\phi_\cdot$ is the sample average of conditional genotype probabilities, $\hat{\phi}_\cdot^{(0)} = (1/n_{\mathrm{obs}}) \sum_{i=1}^{n_{\mathrm{obs}}} p_{i,AA}$.

APPENDIX B: ALTERNATE REPRESENTATIONS OF THE FULL LIKELIHOOD

We may expand our presentation of the full likelihood from Equation 1 as follows:

$$\begin{aligned} \mathrm{lik}(\phi_D, \phi_{\tilde{D}}, \pi_\cdot) &= \prod_{i=1}^{N} \Pr(D_i, \mathbf{O}_{\cdot i}; \phi_D, \phi_{\tilde{D}}, \pi_\cdot) \\ &= \left\{ \prod_{i=1}^{n_{\mathrm{obs}}} \Pr(\mathbf{O}_{\cdot i} \mid D_i; \phi_D, \phi_{\tilde{D}}) \right\} \cdot \left\{ \prod_{i=1}^{N} \Pr(D_i;\ \pi_\cdot) \right\} \qquad \text{(B1)} \\ &= \mathrm{lik}(\phi_D,\ \phi_{\tilde{D}} \mid \mathbf{D}) \cdot \mathrm{lik}(\pi_\cdot). \qquad \text{(B2)} \end{aligned}$$

Here, the pattern of missing genotypes generated by selective genotyping depends only on the observed phenotypes, **D**, and is conditionally independent of the underlying genotypes, **G**, given **D**. Hence, the model implicitly conditions on the pattern of selective genotyping, with $\Pr(\mathbf{O}_{\cdot i} \mid D_i; \phi_D, \phi_{\tilde{D}}) = 1$ for all ungenotyped individuals, $i \in \{n_{\mathrm{obs}} + 1, \ldots, N\}$.

We may reparameterize the full likelihood in terms of parameters π_{AA}, π_{AB}, and $\phi_\cdot$ as

$$\begin{aligned}\text{lik}(\pi_{AA}, \pi_{AB}, \phi_\cdot) &= \prod_{i=1}^{N} \Pr(D_i, \mathbf{O}_{\cdot i}; \pi_{AA}, \pi_{AB}, \phi_\cdot) \\ &= \prod_{i=1}^{N} \Pr(D_i | \mathbf{O}_{\cdot i}; \pi_{AA}, \pi_{AB}) \cdot \Pr(\mathbf{O}_{\cdot i}; \phi_\cdot), \end{aligned} \tag{B3}$$

where ungenotyped individuals, $i = n_{\text{obs}} + 1, \ldots, N$, are incorporated by applying the marginal genotype probabilities as mixing proportions in modeling disease status using a binomial model with $\pi_\cdot = \pi_{AA} \cdot \phi_\cdot + \pi_{AB} \cdot (1 - \phi_\cdot)$. To find the appropriate MLEs, note that the likelihood in (B3) is a reparameterization of (B1). Specifically, the parameters of interest π_{AA}, π_{AB}, and $\phi_\cdot$ can be written in terms of ϕ_D, $\phi_{\bar{D}}$, and $\pi_\cdot$:

$$\begin{aligned}\phi_\cdot &= \phi_D \cdot \pi_\cdot + \phi_{\bar{D}} \cdot (1 - \pi_\cdot) \\ \pi_{AA} &= \frac{\phi_D \cdot \pi_\cdot}{\phi_\cdot} \\ \pi_{AB} &= \frac{(1 - \phi_D) \cdot \pi_\cdot}{1 - \phi_\cdot}. \end{aligned} \tag{B4}$$

Then, the appropriate MLEs for π_{AA}, π_{AB}, and $\phi_\cdot$ in (B3) can be obtained as plug-in estimates using the relationships in (B4). We have provided these MLEs for completeness, but if our primary aim is the appropriate likelihood-ratio statistic, then calculating these MLEs is unnecessary.

APPENDIX C: CONSTRAINED FULL-LIKELIHOOD ANALYSIS AT NONMARKER LOCATIONS

To start, we incorporate the constraint by transforming the expression in (2) to yield $\phi_{\bar{D}} = (\phi_\cdot - \phi_D \cdot \pi_\cdot)/(1 - \pi_\cdot)$. Hence, our constrained likelihood is equivalent to a two-parameter model in which each individual has the following contribution to the likelihood,

$$\text{lik}(\phi_D, \pi_\cdot; \mathbf{O}_{\cdot i}, D_i) = \Pr(\mathbf{O}_{\cdot i} \,|\, D_i; \phi_D, \pi_\cdot) \cdot \Pr(\mathrm{D}_i; \pi_\cdot),$$

with observed genotypes modeled according to disease status as

$$\Pr(\mathbf{O}_{\cdot i} \,|\, D_i; \phi_D, \pi_\cdot) = \begin{cases} \sum_{g \in \{AA, AB\}} \Pr(G_i = g; \phi_D) \cdot q_{ig}, & D_i = 1 \\ \sum_{g \in \{AA, AB\}} \Pr(G_i = g; \phi_{\bar{D}} = \frac{\phi_\cdot - \phi_D \cdot \pi_\cdot}{1 - \pi_\cdot}) \cdot q_{ig}, & D_i = 0 \end{cases} \tag{C1}$$

for genotyped individuals and identically equal to one for ungenotyped individuals.

At iteration $s + 1$, we have the parameter estimate, $\hat{\phi}_D^{(s)}$. In the E-step, we calculate the expected number of affected and unaffected individuals with genotype AA at the putative QTL as shown in (A1).

In the M-step, the updated parameter estimates are obtained by maximizing the likelihood function in (C1), using a numerical optimization approach such as that of Nelder and Mead (1965). Similar to the EM for the reverse approach above, we use the sample average of conditional genotype probabilities, p_{ig}, among diseased individuals as an initial guess for ϕ_D and take the sample average n_D/N as the initial estimate for $\pi_\cdot$.

APPENDIX D: MORE RELATIONSHIPS BETWEEN THE FULL AND REVERSE APPROACHES

Full likelihood *vs.* reverse approach: Let $\text{LOD}_{\text{Full}}^{n_{\text{obs}}}$ be the LOD score based on full likelihood for genotyped individuals $i = 1, \ldots, n_{\text{obs}}$ only and $\text{LOD}_{\text{Full}}^{N}$ be the LOD score from full-likelihood analysis using all individuals $i = 1, \ldots, N$, whether genotyped or not. We see the following analytic results comparing $\text{LOD}_{\text{Full}}^{n_{\text{obs}}}$ to LOD_{R} from the reverse approach:

$$\begin{aligned}\text{LOD}_{\text{Full}}^{n_{\text{obs}}} &= \log_{10}\left\{\frac{\Pr(\mathbf{O}_{\cdot\cdot} \,|\, \mathbf{D}; \hat{\phi}_D, \hat{\phi}_{\bar{D}}) \cdot \Pr(\mathbf{D}; \hat{\pi}_\cdot)}{\Pr(\mathbf{O}_{\cdot\cdot}; \hat{\phi}_\cdot) \cdot \Pr(\mathbf{D}; \hat{\pi}_\cdot)}\right\} \\ &= \log_{10}\left\{\frac{\Pr(\mathbf{O}_{\cdot\cdot} \,|\, \mathbf{D}; \hat{\phi}_D, \hat{\phi}_{\bar{D}})}{\Pr(\mathbf{O}_{\cdot\cdot}; \hat{\phi}_\cdot)}\right\} = \text{LOD}_{\text{R}}. \end{aligned}$$

Likewise, we can compare $\text{LOD}_{\text{Full}}^{N}$ from a full-likelihood analysis with all N individuals to LOD_{R}, which uses genotyped individuals only. Recall that our full-likelihood analysis conditions on the pattern of missing genotype data,

with the conditional multipoint marker probabilities identically equal to one for all ungenotyped individuals. Using Equation B2 above, we see that

$$\begin{aligned}\mathrm{LOD}^N_{\mathrm{Full}} &= \log_{10}\left\{\frac{\max_{\phi_D,\phi_{\tilde{D}}}\prod_{i=1}^{n_{\mathrm{obs}}}\Pr(\mathbf{O}_{\cdot i}\,|\,D_i;\phi_D,\phi_{\tilde{D}})}{\max_{\phi_\cdot}\prod_{i=1}^{n_{\mathrm{obs}}}\Pr(\mathbf{O}_{\cdot i};\phi_\cdot)}\right\}\times\left\{\frac{\max_{\pi_\cdot}\prod_{i=1}^{N}\Pr(D_i;\pi_\cdot)}{\max_{\pi_\cdot}\prod_{i=1}^{N}\Pr(D_i;\pi_\cdot)}\right\}\\ &= \log_{10}\left\{\frac{\Pr(\mathbf{O}_{\cdot\cdot}\,|\,\mathbf{D};\hat{\phi}_D,\hat{\phi}_{\tilde{D}})}{\Pr(\mathbf{O}_{\cdot\cdot};\hat{\phi}_\cdot)}\right\} = \mathrm{LOD}_{\mathrm{R}},\end{aligned}$$

where the estimated parameters $\hat{\phi}_D$ and $\hat{\phi}_{\tilde{D}}$ are the MLEs specified in the METHODS section.

Constrained full likelihood *vs.* modified reverse approach: It is difficult to work with the constrained full likelihood analytically. However, we can derive the following inequality to put an upper bound on $\mathrm{LOD}^N_{\mathrm{Full,seg}}$. First note that constrained likelihood must be bounded above by the unconstrained likelihood so

$$\begin{aligned}\max_{\phi_D,\phi_{\tilde{D}},\pi_\cdot|\phi_\cdot=1/2}\mathrm{lik}(\phi_D,\phi_{\tilde{D}},\pi_\cdot) &\le \max_{\phi_D,\phi_{\tilde{D}},\pi_\cdot}\mathrm{lik}(\phi_D,\phi_{\tilde{D}},\pi_\cdot)\\ &= \mathrm{lik}(\hat{\phi}_D,\hat{\phi}_{\tilde{D}})\cdot\mathrm{lik}(\hat{\pi}_\cdot),\end{aligned}$$

where $\hat{\phi}_D$, $\hat{\phi}_{\tilde{D}}$, and $\hat{\pi}_\cdot$ are the unconstrained MLEs.

Plugging into $\mathrm{LOD}^N_{\mathrm{Full,seg}}$, we find

$$\begin{aligned}\mathrm{LOD}^N_{\mathrm{Full,seg}} &= \log_{10}\left\{\frac{\max_{\phi_D,\phi_{\tilde{D}},\pi_\cdot|\phi_\cdot=1/2}\mathrm{lik}(\phi_D,\phi_{\tilde{D}},\pi_\cdot)}{\max_{\pi_\cdot}\mathrm{lik}(\phi_\cdot=1/2,\pi_\cdot)}\right\}\\ &\le \log_{10}\left\{\frac{\mathrm{lik}(\hat{\phi}_D,\hat{\phi}_{\tilde{D}})\cdot\mathrm{lik}(\hat{\pi}_\cdot)}{\mathrm{lik}(\phi_\cdot=1/2)\cdot\mathrm{lik}(\hat{\pi}_\cdot)}\right\} = \mathrm{LOD}_{\mathrm{R,seg}}.\end{aligned}$$

Hence, the LOD score obtained by constrained full likelihood is bounded above by the LOD score from the modified reverse approach.