Chapter 23
Finding Quantitative Trait Loci Genes

Hui Wang, Sherri Rose, Mark J. van der Laan

The goal of quantitative trait loci (QTL) mapping is to identify genes underlying an observed trait in the genome using genetic markers. In experimental organisms, the QTL mapping experiment usually involves crossing two inbred lines with substantial differences in a trait, and then scoring the trait in the segregating progeny. A series of markers along the genome is genotyped in the segregating progeny, and associations between the trait and the QTL can be evaluated using the marker information. Of primary interest are the positions and effect sizes of QTL genes.

Early literature (Sax 1923; Thoday 1960) focused on directly analyzing a single marker using analysis of variance (ANOVA). The biggest disadvantage of such marker-based analysis is its inability to assess QTL genes between markers. In 1989, Lander and Botstein proposed the interval mapping (IM) method (Lander and Botstein 1989). With IM, the genotypic value of a QTL follows a multinomial distribution, determined by the distance of the QTL to its flanking markers and the genotypes of the flanking markers. The trait value is modeled as a Gaussian mixture with the mixing proportions being the multinomial probabilities of the QTL genotype. The significance of the QTL effect is then assessed using likelihood ratio test. By testing positions at small increments along the genome, a whole-genome finely scaled test statistic profile can be constructed. IM has greatly increased the accuracy of estimating QTL parameters, and it has gained wide popularity in the genetic mapping community. Later, Haley and Knott developed a regression method to approximate IM (Haley and Knott 1992). This method imputes the unobserved genotypic value of a putative QTL with its expected value.

IM methods unrealistically assume there is only one gene underlying the observed trait in the entire genome, represented as testing each potential position separately (Lander and Botstein 1989) or computing the univariate association between the expected genotypic value and the phenotypic trait in Haley–Kott regression. In other words, IM only considers the current QTL; all other QTL genes are ignored. When this assumption is violated, the effects of other QTL genes are contained within the residual variance, affecting the assessment of QTL parameters.

M.J. van der Laan and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics, DOI 10.1007/978-1-4419-9782-1_23,

To handle multiple QTL genes, Jansen (1993) and Zeng (1994) developed a composite interval mapping (CIM) approach. In CIM, background markers are added to a standard IM statistical model to reduce noise and increase the precision of QTL effect estimates. Thus, the CIM approach estimates QTL effects adjusted for confounding markers and can substantially improve the performance of IM when the background markers are properly chosen. Multiple interval mapping (MIM) was also developed to simultaneously estimate effects and positions of multiple QTL genes (Kao et al. 1999). MIM enjoys greater power but is computationally difficult. It also has a long-standing estimator selection problem: Which QTL genes are to be included? Bayesian approaches have also been studied and applied in QTL mapping (Satagopan et al. 1996; Heath 1997; Sillanpaa and Arjas 1998).

In recent years, with finely scaled single nucleotide polymorphism (SNP) markers replacing the traditional widely spaced microsatellite markers, identifying QTL genes between markers has become less concerning. Due to the high-dimensional nature of SNP data, the univariate marker-trait regression is widely used for its simplicity and computational feasibility despite its noisy results. Machine learning algorithms, such as random forests (Breiman 2001b), are also used to map QTL genes (Lee et al. 2008).

Most of these QTL methods are fully parametric and typically assume a Gaussian distribution for the phenotypic trait, as well as require specification of a parametric regression model. The estimation of QTL effects often relies on the method of maximum likelihood estimation. Maximum likelihood estimation based on such parametric regression models is widely used and well studied, with software available in many platforms. However, quite often, these parametric models represent an over-simplified description of the underlying genetic mechanism and leads to biased estimates. In addition, if the parametric model is data-adaptively selected among a set of candidate parametric regression models, then the reported standard errors and the p-values are not interpretable.

In this chapter, we address the QTL mapping problem through the use of a semiparametric regression model and the TMLE. The only assumption of the semiparametric regression model is that the phenotypic trait changes linearly with the QTL gene. We also define the C-TMLE, which is a particularly appealing estimator for the high-dimensional genomic data structures. Portions of this chapter were adapted from Wang et al. (2010).

23.1 Semiparametric Regression Model and TMLE

Suppose the observed data are i.i.d. realizations of $O_i = (Y_i, M_i) \sim P_0, i = 1, \ldots, n$, where Y represents the phenotypic trait value, M represents the marker genotypic values, and i indexes the ith subject. Let A be the genotypic value of the putative QTL under consideration. When A lies on a marker, A is observed. When A lies between markers, it is unobserved. In this case, we impute A with its expected value from a multinomial distribution computed from the genotypes and the relative locations of its flanking markers. This is the same strategy used in Haley–Knott regression (Haley and Knott 1992), and we will thus only be estimating the effect of an imputed A. The semiparametric regression model for the effect of A at value $A = a$ relative to $A = 0$, adjusted for a user-supplied set of other markers M^-, is given by

$$E_0(Y \mid A = a, M^-) - E_0(Y \mid A = 0, M^-) = \beta_0 a.$$

Other parametric forms, such as $a \sum_{j=1}^{J} \beta_j V_j$ incorporating effect modification by other markers V_j, can be incorporated as well. We view β_0 as our parameter of interest, which also corresponds with a marginal average effect obtained by averaging this conditional effect over the distribution of M^-.

The TMLE of β_0 was presented in the previous chapter and involves an initial machine learning (e.g., super learner) fit of $E_0(Y \mid M)$, which yields a fit of $E_0(Y \mid A = 0, M^-)$, mapping the latter into an initial estimator of β_0 and thereby of $E_0(Y \mid A, M^-)$ in the semiparametric regression model. After obtaining this initial estimator of $E_0(Y \mid A, M^-)$ of the semiparametric form as enforced by the semiparametric regression model, we carry out a single targeted update step by adding an estimate of the clever covariate $A - E_0(A \mid M^-)$, and fitting the coefficient ϵ in front of this clever covariate with univariate regression, using the initial estimator of $E_0(Y \mid A, M^-)$ as offset. Note that the TMLE of β_0 is now simply $\beta_n^0 + \epsilon_n$.

The estimation of the clever covariate requires an estimator of $E_0(A \mid M^-)$. The latter can be carried out with a machine learning algorithm regressing A on M^-. In particular, one could decide to fit this regression of the marker of interest on two flanking markers, thereby dramatically simplifying the estimation problem, while potentially capturing most of the confounding by the total marker set M^-. The choice of how great the distance between the flanking markers will be is a delicate issue. If one selects the flanking markers right next to the marker of interest, the data might not allow the separation of the effect of interest from the effect of the flanking markers. That is, one is aiming to adjust for confounders that are too predictive of the marker of interest. On the other hand, if one selects the flanking markers too far away from the marker of interest, the flanking markers will not adjust well for the markers that are in between the marker of interest and the flanking markers. Simulations in the previous chapter suggest that the TMLE shows no sign of deterioration for correlations smaller than 0.7 between the marker of interest and the confounders. This could be used to set the window width defined by the two flanking markers. Subject matter considerations, such as that the scientist would be satisfied with a claim that the targeted effect of the marker can be due to other markers in a window of a particular size, could also be used to set this window width of the flanking markers.

An alternative approach is to let the data decide what other markers to include in the model for $E_0(A \mid M^-)$. For that purpose, we can employ the C-TMLE (using a linear regression working model for fluctuation of initial estimator), first presented in Chap. 19 for estimation of an additive effect $E_0(E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W))$ for the observed data structure $O = (W, A, Y)$ and nonparametric model for the probability distribution P_0 of O. This C-TMLE has also been implemented for this estimation problem, but, obviously, now in terms of TMLEs in this semiparametric regression model. Thus, this algorithm involves using forward selection of main terms to build a main term linear regression fit for $E_0(A \mid M^-)$, based on the sum of squared residuals (i.e., MSE) of the corresponding TMLE of $E_0(Y \mid A, M^-)$ that uses this main term regression fit of $E_0(A \mid M^-)$ in the clever covariate. Cross-validation is used to select the number of main terms (i.e., the number of forward selection steps that the algorithm carries out) that will actually be included in the fit of $E_0(A \mid M^-)$. The candidate main terms can include fits of $E_0(A \mid M^-)$ such as one based on two flanking markers defined by a choice of window width, across a number of possible window widths. In this manner the C-TMLE algorithm can data-adaptively decide how aggressive the targeting step should be in its effort to reduce bias due to residual confounding.

As in Chap. 19, the C-TMLE implementation may also involve the selection of a penalty to be added to the MSE in order to make the procedure more robust in the context of having to adjust for highly correlated markers: for details we refer to the technical report (Wang et al. 2011). C-TMLE allows one to data-adaptively determine the markers to include in the fit of $E_0(A \mid W)$. For example, one may wish to only adjust for the two closest markers that are farther than δ-apart from the marker A, and one can use C-TMLE to data-adaptively select this choice δ based on the log-likelihood of the TMLE of the semiparametric regression fit. In our simulations and data analysis we have implemented both TMLEs as well as C-TMLEs.

23.2 The C-TMLE

Let $Q_n^0 = m(A, V \mid \beta_n^0) + r(M^-)$ be the initial estimate of Q_0 contained in the same semiparametric regression model that we also used in the TMLE. The C-TMLE is concerned with iteratively updating this initial estimate of Q_0. Firstly, we compute a set of K univariate covariates $W_1, \ldots, W_K$ from M^-, which we will refer to as main terms, even though a term could be an interaction term or a super learning fit of the regression of A on a subset of the components of M^-. Let's refer to M^- by $W = (W_1, \ldots, W_K)$. In this subsection we will suppress in the notation for estimates of Q_0 and g_0 their dependence on the sample size n. Let $\Omega = \{W_1, \ldots, W_K\}$ be the full collection of main terms. A linear regression model fit g^K of $g_0(W) = E_0(A \mid W)$ using all main terms in Ω is viewed as the most nonparametric estimate of g_0. For a given subset of main terms $\mathcal{S} \subset \Omega$, let $\mathcal{S}^c$ be its complement within Ω. For a given subset $\mathcal{S}^k$, we will define g^k as the least squares fit of the linear regression model for $E_0(A \mid W)$ that includes as main terms all the terms in $\mathcal{S}^k$. In the C-TMLE algorithm we use a forward selection algorithm that augments a given set $\mathcal{S}^k$ into a next set $\mathcal{S}^{k+1}$ obtained by adding the best main term among all main terms in the complement $\mathcal{S}^{k,c}$ of $\mathcal{S}^k$. In other words, the algorithm iteratively updates a current estimate g^k into a new estimate g^{k+1}, but the criterion for g does not measure how well g fits g_0; it measures how well the TMLE using this g fits Q_0.

Let $L(Q)(O) = (Y - Q(A, W))^2$ be the squared error loss function for the true regression function $Q_0 = E_0(Y \mid A, W) = \beta_0 A + E_0(Y \mid A = 0, W)$. For a given initial estimate Q, let $Q_g(\epsilon) = Q + \epsilon(A - g(W))$ be the parametric working fluctuation model used in the TMLE of Q_0 defined in the previous section. For a given estimate g of g_0 and initial Q of Q_0, the corresponding TMLE (as defined in the previous section) of Q_0 is given by $Q_g(\epsilon_n)$, where $\epsilon_n = \arg\min_\epsilon P_n L(Q_g(\epsilon))$ is the univariate least squares estimator of ϵ using the initial estimate Q as offset, and P_n denotes the empirical probability distribution of $O_1, \ldots, O_n$. Here we used the notation $Pf \equiv \int f(o)dP(o)$. That is, an initial estimate Q, an estimate g, and the data $O_1, \ldots, O_n$ are mapped into a new targeted maximum likelihood estimate $Q^* = Q_g(\epsilon_n)$. Let's refer to this mapping as $Q^* = \text{TMLE}(Q, g)$, suppressing its dependence on P_n.

The C-TMLE algorithm defined below generates a sequence $(Q^k, \mathcal{S}^k)$ and corresponding TMLEs Q^{k*}, $k = 0, \ldots, K$, where Q^k represents an initial estimate, $\mathcal{S}^k$ a subset of main terms that defines g^k, and Q^{k*} the corresponding TMLE that updates Q^k using g^k. These TMLEs Q^{k*} represent subsequent updates of the initial estimator Q_n^0, and the corresponding main term set $\mathcal{S}^k$, as used to define g^k in this k-specific TMLE, increases in k, one unit at a time: $\mathcal{S}^0$ is empty, $\mid \mathcal{S}^{k+1} \mid = \mid \mathcal{S}^k \mid +1$, $\mathcal{S}^K = \Omega$. The C-TMLE uses cross-validation to select k, and thereby to select the TMLE Q^{k*} that yields the best fit of Q_0 among the $K + 1$ k-specific TMLEs that are increasingly aggressive in their bias-reduction effort. This C-TMLE algorithm is defined as follows:

Initiate algorithm: Set initial TMLE. Let $k = 0$. $Q^k = Q_n^0$ is the initial estimate of Q_0, and $\mathcal{S}^k$ is the empty set so that g^k is the empirical mean of A. Thus, Q^{k*} is the TMLE updating this initial estimate Q^k using as clever covariate $A - g^k$.

Determine next TMLE. Determine the next best main term to add to the linear regression working model for $g_0(W) = E_0(A \mid W)$:

$$\mathcal{S}^{k+1,cand} = \arg \min_{\mathcal{S}^k \cup W_j : W_j \in \mathcal{S}^{k,c}} P_n L(\text{TMLE}(Q^k, \mathcal{S}^k \cup W_j)).$$

If

$$P_n L(\text{TMLE}(Q^k, \mathcal{S}^{k+1,cand})) \leq P_n L(\text{TMLE}(Q^{k*})),$$

then $(\mathcal{S}^{k+1} = \mathcal{S}^{k+1,cand}, Q^{k+1} = Q^k)$, else $Q^{k+1} = Q^{k*}$, and

$$\mathcal{S}^{k+1} = \arg \min_{\mathcal{S}^k \cup W_j : W_j \in \mathcal{S}^{k,c}} P_n L(\text{TMLE}(Q^{k*}, \mathcal{S}^k \cup W_j)).$$

[In words: If the next best main term added to the fit of $E_0(A \mid W)$ yields a TMLE of $E_0(Y \mid A, W)$ that improves upon the previous TMLE Q^{k*}, then we accept this best main term, and we have our next TMLE Q^{k+1*}, g^{k+1} (which still uses the same initial estimate as Q^{k*} uses). Otherwise, reject this best main term, update the initial estimate in the candidate TMLEs to the previous TMLE Q^{k*} of $E_0(Y \mid A, W)$, and determine the best main term to add again. This best main term will now always result in an improved fit of the corresponding TMLE of Q_0, so that we now have our next TMLE Q^{k+1*}, g^{k+1} (which now uses a different initial estimate than Q^{k*} used).]

Iterate. Run this from $k = 1$ to K at which point $\mathcal{S}^K = \Omega$. This yields a sequence (Q^k, g^k) and corresponding TMLE Q^{k*}, $k = 0, \ldots, K$.

This sequence of candidate TMLEs Q^{k*} of Q_0 has the following property: the estimates g^k are increasingly nonparametric in k and $P_n L(Q^{k*})$ is decreasing in k, $k = 0, \ldots, K$. It remains to select k. For that purpose we use V-fold cross-validation. That is, for each of the V splits of the sample in a training and validation sample, we apply the above algorithm for generating a sequence of candidate estimates $(Q^{k*} : k)$ to a training sample, and we evaluate the empirical mean of the loss function at the resulting Q^{k*} over the validation sample, for each $k = 0, \ldots, K$. For each k we take the average over the V-splits of the k-specific performance measure over the validation sample, which is called the cross-validated risk of the k-specific TMLE. We select the k that has the best cross-validated risk, which we denote with k_n. Our final C-TMLE of Q_0 is now defined as Q^{k_n*}, and the corresponding updated regression coefficient is our TMLE β_n^* of β_0.

Remark. The candidate main terms can also include fits of $E_0(A \mid M^-)$ such as one based on two flanking markers defined by a choice of window width, across a number of possible window widths. In this manner, the above C-TMLE algorithm data-adaptively decides which window width yields effective bias reduction. C-TMLE implementation in the following data analysis involved a penalized mean squared error as a measure of fit instead of the mean squared error, where the penalty is defined as a variance estimator of the corresponding TMLE of β_0.

Statistical Properties of the C-TMLE

To understand the appeal of the C-TMLE, we make the following remarks. Including a main term in the fit of the clever covariate that has no effect on the outcome will only harm the TMLE of β_0 both with respect to bias and mean squared error. If one uses the log-likelihood (i.e., MSE) of the regression of A on M^- as a criterion for selection of the main terms, then one will easily select main terms that have a weak effect on the outcome, while truly important main terms are not included. Therefore, it is crucial to use a main term selection criterion for $E_0(A \mid M^-)$ that actually measures the fit of the resulting TMLE of the outcome regression. In addition, one can formally prove that the TMLE achieves the full bias reduction with respect to β_0 if the clever covariate uses a true regression, $E_0(A \mid M^s)$, with M^s being a reduction of M^- that is rich enough so that $E_0(Y \mid A = 0, M^-)$ is captured. In fact, the result is stronger, since M^s only needs to capture the function of M^- that is obtained by taking the difference between the true $E_0(Y \mid A = 0, M^-)$ and its initial estimator $E_n(Y \mid A = 0, M^-)$ (van der Laan and Gruber 2010). Thus, theory indeed fully supports that we should be selecting main terms in the clever covariate that are predictive of residual bias of the initial estimator of $E_0(Y \mid A = 0, M^-)$, and the C-TMLE algorithm presented above indeed targets such main terms.

23.3 Simulation

A single chromosome of 100 markers was simulated on 600 backcross subjects. Markers were evenly spaced at 2 centimorgan (cM). A single QTL main effect was generated at marker position 100 cM, denoted by $M_{(100)}$. Here, the number in the subscript of M indicates the position of the marker. There were also four epistatic effects on markers $M_{(60)}$, $M_{(90)}$, $M_{(120)}$, and $M_{(150)}$. Phenotypic values were generated from the data-generating distribution: $Y = 5 + 1.2M_{(100)} - 0.8M_{(60)}M_{(90)} - 0.8M_{(90)}M_{(120)} - 0.8M_{(120)}M_{(150)} - 0.8M_{(150)}M_{(60)} + U$, where U is the error term drawn from an exponential distribution scaled to have a variance of 10. We generated 500 simulated data sets of this type.

In this simulation, the density of markers is fairly high, the phenotypic outcome follows a nonnormal distribution, and there are strong counteracting epistatic effects in linked markers. A univariate regression effect estimate of the effect of, for example, $M_{(100)}$ will be biased due to the lack of adjustment for the effect of the highly correlated markers. Indeed, the CIM estimate for the effect of $M_{(100)}$ is negative, far away from the true value 1.2. On the other hand, taking the CIM prediction function as the initial estimator $\bar{Q}_n^{(0)}$, TMLE was then able to recover some of the signal and hence improved on the CIM estimates. In TMLE, the true regression of A on the other 99 markers, M^-, was estimated with a main terms linear regression including two flanking markers with a prespecified distance to A. We used two distances, 20 cM and 40 cM, and denote the estimators by TMLE$_{(20)}$ and TMLE$_{(40)}$. The CIM analysis was carried out using QTL Cartographer (Basten et al. 2001), with default settings. We analyzed markers without considering positions between them. For CIM, the mean effect estimate for $M_{(100)}$ is −0.2731 and is dominated by the epistatic effects from its nearby markers. TMLE$_{(40)}$ is able to correct some of the bias, and its effect estimate is 0.5365. TMLE$_{(20)}$ utilizes an estimator of $E_0(A \mid M^-)$ with more predictive power than TMLE$_{(40)}$ and produced an estimate closest to the truth. We list the averages of the effect estimates for $M_{(100)}$ across 500 simulations in Table 23.1 along with their standard errors for CIM, TMLE$_{(20)}$, and TMLE$_{(40)}$.

We also used a univariate regression (UR) fit for $\bar{Q}_n^{(0)}$ within TMLE, and these results can be found in Table 23.1. The UR initial estimate was even more biased than that of CIM. TMLE$_{(20)}$, using UR as $\bar{Q}_n^{(0)}$, produced very similar estimates to TMLE$_{(20)}$ using CIM as initial estimator. On the other hand, TMLE$_{(40)}$ using the CIM as initial estimator produced a better estimator than TMLE$_{(40)}$ using the

Table 23.1 Mean effect estimates of $M_{(100)}$ over 500 simulations

	$\bar{Q}_n^{(0)}$=CIM		$\bar{Q}_n^{(0)}$=UR	
	Estimate	SE	Estimate	SE
Initial Estimate	−0.2731	0.3273	−0.6248	0.2684
TMLE$_{(40)}$	0.5365	0.4538	0.2705	0.3135
TMLE$_{(20)}$	0.8478	0.4508	0.8093	0.4079

univariate regression as initial estimator. This demonstrates the robustness of TMLE with respect to misspecification of the initial estimator, which predicts that the more predictive the regression of A on M^{-}, the more robust TMLE will be to the choice of its initial estimator. A closer look at Table 23.1 also reveals that compared to $\mathrm{TMLE}_{(40)}$, the additional bias reduction of $\mathrm{TMLE}_{(20)}$, using univariate regression as initial estimator, comes with an increase in standard error.

23.4 Wound-Healing Application

In this section, we analyze a data set published in Masinde et al. (2001). The original study was designed to identify QTL genes involved in the wound-healing process. A genomewide scan of 119 codominant markers was performed using 633 F2 (MRL/MP x SJL/J) mice. Each mouse was punctured with a 2-mm hole in its ear, and the phenotypic trait was the hole closure measurement at day 21. The marginal distribution of the phenotypic trait is bell-shaped.

We analyzed this data set with TMLE (results not shown; see Wang et al. 2011), C-TMLE, and CIM. Based on the evaluation of a discrete super learner (Chap. 3) that included both DSA and random forests, the DSA machine learning algorithm was selected as initial estimator of $E_0(Y \mid M)$, and subsequently mapped into the desired initial estimator for $E_0(Y \mid A, M^{-})$ satisfying the semiparametric regression model. To lessen the computational load, we first screened additive and dominant effects of all markers with univariate regression and supplied to this machine learning algorithm the markers with a p-value less than 0.10. In the TMLE, the conditional mean of A, given M^{-} is fitted with a main terms linear regression model with main

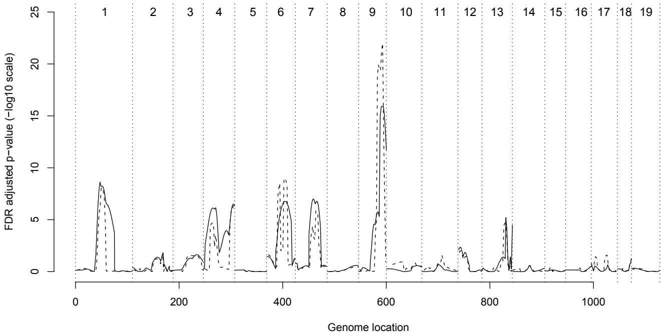

Fig. 23.1 —The genomewide FDR-adjusted p-value profile for the additive effects in the wound-healing data set. The *solid line* represents CIM, and the *dashed line* represents C-TMLE. Chromosome numbers are superimposed on top of the picture

terms A_c, W_1^a, W_1^d, W_2^a, W_2^d, where A_c denotes the dominant effect of A when A is additive and the additive effect of A when A is dominant, W_1 and W_2 are the closest flanking markers 20 cM away from A, and the superscript a denotes the additive effect and d the dominant effect.

Four hundred putative QTL positions were tested at 2-cM increments for both the additive and dominant effects. The p-values were adjusted using FDR. The TMLE and C-TMLE produced similar results, and we only present C-TMLE results in this chapter. Figure 23.1 displays the genomewide FDR-adjusted p-value profile for the additive effect at each tested position. The CIM p-values were computed from the asymptotic χ^2 distribution. No significant dominant effect was detected in this data set. The (C)-TMLE essentially identified the same QTL genes as CIM, albeit with an improved resolution. Many of these genes were also reported in Masinde et al. (2001). However, on chromosome 6, the (C-)TMLE suggests two linked QTL genes instead of one, as indicated by CIM.

23.5 Listeria Application

Boyartchuk et al. (2001) published a data set on the survival time of 116 age-matched female mice following infection with *Listeria monocytogenes*, a Gram-positive bacteria causing a wide range of diseases. The mice were an F2 intercross population derived from susceptible BALB/cByJ and resistant C57BL/6ByJ strains, and the goal of the study was to map genetic factors of susceptibility to *L. monocytogenes*. The phenotypic trait is the recorded time to death for each mouse upon infection with *L. monocytogenes*. One hundred and thirty-one codominant markers were genotyped on the autosomal chromosomes. When a mouse survived beyond 240 h, it was considered recovered. About 30% of the mice recovered, and we refer to them as survivors and the remaining mice as nonsurvivors. This creates a spike in the phenotypic trait distribution, violating the normality assumption in traditional approaches of QTL mapping.

The outcome Y was defined as the logarithm of the phenotypic trait. The first step of TMLE is to obtain an initial estimator of $E_0(Y \mid M)$, which can then be mapped into an initial estimator of $E_0(Y \mid A, M^-)$, satisfying the semiparametric regression model. Y can be decomposed into a binary trait of survival or nonsurvival and a continuous trait of survival time among nonsurvivors (Broman 2003). We denote this binary trait of survival by $Z = I(Y = \log 264)$. Then, the expected value of Y given the marker data M can be represented as

$$E_0(Y \mid M) = P_0(Z = 1 \mid M) \log 264 + P_0(Z = 0 \mid M) E_0(Y \mid Z = 0, M).$$

In the above formula, $P_0(Z = 1 \mid M)$ and $P_0(Z = 0 \mid M)$ are conditional probabilities of whether a mouse has survived ($Z = 1$) or died ($Z = 0$) given the marker data M. We fit this with a super learning algorithm for binary outcomes. $E_0(Y \mid Z = 0, M)$ is the conditional expectation of Y on M given that the mouse has died, which can

Table 23.2 Mean risk of candidate initial regressions in discrete super learner from the Listeria data set

	DSA	RF	SL	2-part SL
CV risk	0.2212	0.1581	0.1589	0.1463

be obtained by applying super learning on nonsurvivors. We refer to this machine learning algorithm as the 2-part super learner.

The collection of algorithms in the super learner included DSA and random forests. As before, the machine learning algorithms were only provided the additive and dominant markers that had a significant univariate effect based on a p-value threshold of 0.10. Since we wished to evaluate if this 2-part super learner provided a better fit than a regular super learner, we implemented a discrete super learner whose library consisted of a total of four algorithms for estimation of $E_0(Y \mid M)$: DSA, random forests, super learner, and a 2-part super learner. In Table 23.2, we report the honest cross-validated risk of DSA, random forests, super learner, and the 2-part super learner. In the super learning fits, more than 95% of the weight was put on random forests, thereby strongly favoring a fit that allows for complex interactions.

The 2-part super learner had the smallest honest cross-validated risk and was therefore selected as the estimator of $E_0(Y \mid M)$. In the TMLE, we fitted the conditional mean of A, given M^-, with a main term linear regression model including the main terms used A_c, W_1^a, W_1^d, W_2^a, W_2^d, where A_c denotes the dominant effect of A when A is additive and the additive effect of A when A is dominant, W_1 and W_2 are the closest flanking markers 20 cM away from A, and the superscript a denotes the additive effect and d the dominant effect.

When inspecting Fig. 23.2, TMLE displays much less noise than the parametric CIM. Three additive genes on chromosomes 1, 5, and 13 are clearly identified. Two additive effects on chromosomes 15 and 18 are borderline significant. In addition, TMLE also detected dominant effects on chromosomes 12, 13, and 15. The chromosome 15 QTL gene is identified as carrying both the additive and dominant effects. The literature suggests that the chromosome 1 QTL gene has an effect on how long a mouse can live given it will eventually die, the chromosome 5 gene has an effect on a mouse's chance of survival, and the genes on chromosomes 13 and 15 are involved in both (Boyartchuk et al. 2001; Broman 2003; Jin et al. 2007). We detected all of these genes and, in addition, an additive gene on chromosome 18 and a dominant gene on chromosome 12. CIM also identified those major genes, however, with less significance and many more suspicious positives. See Table 23.3. This data analysis was first published in Wang et al. (2010).

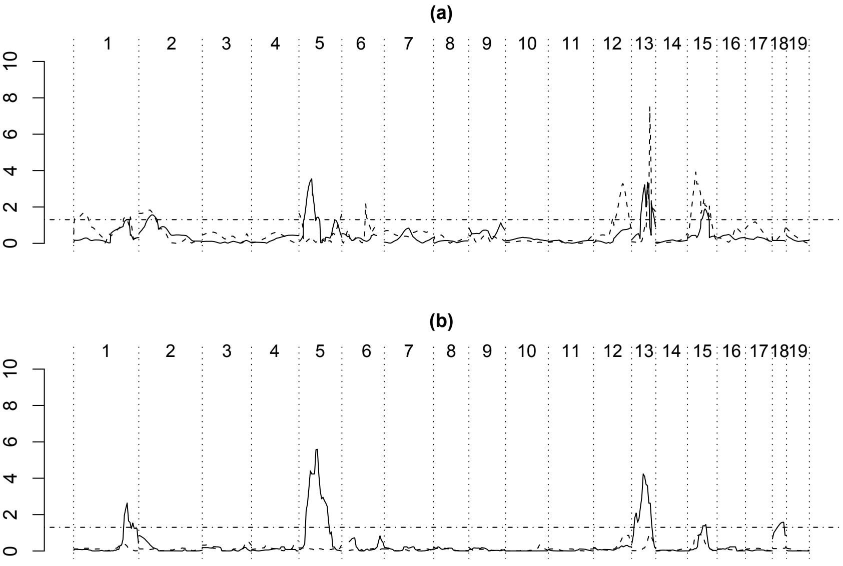

Fig. 23.2 The genomewide p-value profile for the additive and dominant effects in the Listeria data set. The p-values are FDR adjusted and on a negative log10 scale. (a) p-value profile from the CIM. (b) p-value profile from TMLE. In both panels, the *solid line* represents additive effects, and the *dashed line* represents dominant effect. The *dash-dot* line indicates the 0.05 p-value threshold. Chromosome numbers are superimposed on top of each panel

Table 23.3 The estimates of effect sizes and positions of QTL genes from CIM and TMLE in Listeria data set. QTL genes with FDR-adjusted p-values smaller than 0.05 are reported

		CIM			C-TMLE		
QTL ID	Type	Chr	cM	Effect size	Chr	cM	Effect size
1	dom	1	15.0	−0.2351	–	–	–
2	dom	1	72.8	0.1606	–	–	–
3	add	1	78.8	−0.1349	1	78.1	−0.1074
4	dom	2	14.0	−0.2623	–	–	–
5	add	2	18.0	−0.1744	–	–	–
6	dom	5	0.0	−0.1468	–	–	–
7	dom	5	61.0	−0.1693	–	–	–
8	add	5	18.1	0.2764	5	26.1	0.1960
9	dom	6	33.8	−0.1235	–	–	–
10	dom	12	41.8	−0.2352	12	40.1	−0.1372
11	add	13	22.7	−0.3409	13	14.4	−0.1668
12	dom	13	25.9	0.3525	13	26.4	0.1458
13	add	15	25.1	0.1540	15	22.1	0.0678
14	dom	15	12.0	0.2042	15	22.1	0.1438
15	add	18	–	–	18	14.1	−0.0692

23.6 Discussion

Current practice for assessing the effects of genes on a phenotype involves the utilization of parametric regression models. One of the advantages of parametric regression models is that they also provide a p-value, allowing one to rank the different estimated effects and assess their significance. However, both the effect estimates as well as the reported statistical significance are subject to bias due to model misspecification. On the other hand, machine learning algorithms such as random forests, are not sufficient when used alone since these algorithms are tailored for prediction, report generally poor effect estimates, and do not provide a measure of significance. TMLE allows us to incorporate the state of the art in machine learning, without significant computational burden (the targeting step is relatively trivial, although it needs to be carried out for each effect), while still providing an estimate tailored for the effect of interest and CLT-based statistical inference.