Diagnostic Plots for Radiation Hybrid Maps

Daniel P. Gaile[ab1], P. Fred Dahm[c], Clare Gill[d], David L. Adelson[d]

[a]Department of Biostatistics, University at Buffalo, Buffalo, NY 14214-3000, USA
[b]New York State Center of Excellence in Bioinformatics and Life Sciences, Buffalo, NY 14203-1199, USA
[c]Department of Statistics, Texas A&M University, College Station, Texas, USA
[d]Department of Animal Science, Texas A&M University, College Station, Texas, USA

Short title:

RH plots

[1]Corresponding Author. Department of Biostatistics, School of Public Health and Health Professions, 249 Farber Hall, University at Buffalo, 3435 Main Street, Buffalo NY 14214-3000, USA. Tel:+1(716) 881 8955. e-mail:dpgaile@buffalo.edu

Abstract

Background

Radiation hybrid (RH) mapping is a technique for mapping relative positions of markers on chromosomes with the goal of mapping animal genomes. In spite of the widespread usage of RH mapping, few diagnostic tools seem to have been developed to assess the goodness of fit for candidate maps. This paper presents graphical diagnostics of maximum likelihood estimates of relative marker positions and applies the diagnostics to simulated and actual data.

Results

Simulated data sets with and without map errors are utilized to confirm expected behaviors of the diagnostic plots. Diagnostics then are applied to a data set that is comprised of a published RH map for Bos Taurus Autosome 5 (BTA5) and an RH panel consisting of 90 bovine-hamster cell lines typed for 85 markers on BTA5. Our diagnostics suggest possible local errors in marker position, a significant error of typing in at least one marker and a block inversion in a specific region of the map. Possible map inflation due to mapping errors is also detected and discussed in both the simulated and the BTA5 data sets.

Conclusions

Graphical diagnostics presented in this paper appear to provide a useful tool for assessing RH maps, as illustrated by our examination of the RH map for BTA5. We believe that application of our proposed diagnostics can facilitate the construction of maps for agriculturally and scientifically important species.

Introduction

Radiation hybrid (RH) mapping (Goss and Harris, 1975; Cox et al., 1990) is a somatic cell technique for estimating the relative positions of markers along chromosomes which was useful in providing some of the first comprehensive human genome maps (Slonim et al., 1997). Although the usefulness of this technique with respect to the human genome has diminished, there is still significant demand for radiation hybrid mapping tools in agriculturally important species, which are not likely to have genomic sequence available soon.

In Radiation Hybrid experiments, cells are exposed to a dose of radiation that results in the fragmentation of the chromosomes within the cell. The irradiated cells can be fused with hamster cells and grown to form a hybrid cell line. Hybrid cells are plated at limiting dilution so that only individual clones are grown. Each clone is expanded to produce large quantities of DNA for typing and then assayed for the presence or absence of markers unique to the genome of the irradiated cell line. The probability that the radiation will cause a break between two marker loci is a function of the distance between the two loci; with smaller probabilities associated with closer markers. Hence, the relative merit of putative marker orderings can be infered from the patterns of marker coretention observed in the hybrid cell lines. Maximum likelihood (Lange et al., 1995; Slonim et al., 1997), non-parametric Boehnke (1992), graph theoretic(Ben-Dor and Chor, 1997; Bo et al., 2002) and Traveling Salesman Problem (Ben-Dor et al., 2000; Agarwala et al., 2000) based methods of inference have all been developed for the RH mapping problem.

Problem Definition

In this paper, we consider the problem of assessing the quality of a candidate marker ordering. We provide graph-based tools which allow for visualization of the overall quality of the map and for identification of error prone regions within the map. We present three RH diagnostic plots. The first plot was designed to assess whether or not the multi-point fit is consistent with all possible two-point fits. The remaining plots are designed to assess whether the multi-point fit is consistent with all possible three-point fits.

Summary of Previous Methods

Models Used in Recently Developed RH Mapping Algorithms

Recent advances in RH mapping have been in the area of translating the RH problem to a graph theoretic problem (Ben-Dor and Chor, 1997; Bo et al., 2002) and a Traveling Salesman Problem (Ben-Dor et al., 2000; Agarwala et al., 2000). These approaches are based on either minimum number of breaks (MNB) or two-point and three-point approximations to the multi-point likelihood. The parametric models adopted by these approaches commonly assume an error free model with equal retention probabilities consistent with Boehnke et al. (1991) and Lange et al. (1995).

Radiation Hybrid Mapping Diagnostics

Diagnostics to identify influential hybrids have been proposed by Lange et al. (1995) and Slonim et al. (1997). These diagnostics have been designed to identify hybrids which are inconsistent or improbable with respect to the optimal marker ordering. Lange et al. (1995) use the number of OCB as a measure of consistency and Slonim et al. (1997) employ a likelihood ratio based statistic. Both diagnostics assume that the estimated marker ordering is correct and seek to identify hybrid data that are in error. We are unaware of any previous efforts to develop

diagnostics that identify portions of a marker ordering that are least probable given the observed data. We are also unaware of any efforts to provide graph-based tools for RH Mapping.

Overview of the current method

In this chapter, we consider the problem of assessing the quality of a candidate marker ordering. We provide graph-based tools which allow for visualization of the overall quality of the map and for identification of error prone regions within the map. We present three RH diagnostic plots. The first plot was designed to assess whether or not the multi-point fit is consistent with all possible two-point fits. The remaining plots are designed to assess whether the multi-point fit is consistent with all possible three-point fits.

Methods

Radiation Hybrid Data

Let $X = (X_1, \ldots, X_m)$ denote the observed hybrid vector for a single hybrid cell line where m denotes the number of markers that have been typed. If no markers are present at locus k, then $X_k = 0$. If one or more markers are present, then $X_k = 1$. To control errors in marker typing, cell lines are often typed multiple times for each marker. If the marker could not be typed or a consensus call was not possible, $X_k = 2$ or $X_k =?$. We define θ_k as the probability of at least one break occurring between markers k and $k+1$. The breakage probability is then related to the scaled distance d_k

$$d_k = -\log(1 - \theta_k).$$

This distance is said to have the unit measure of Rays(R). Sometimes the distance $100 \times d_k$ is reported with unit measure centi-Rays(cR). The quantity d_k can be interpreted as the expected number of breaks that will occur between markers k and $k+1$.

Likelihoods

Our diagnostic plots require that we fit two-point, three-point and the multi-point model which is to be diagnosed. We consider the equal retention model which is a subset of the general class of Markovian models described in Lange et al. (1995). The two-point and three-point likelihoods can be evaluated under the multi-point model but we have opted to evaluate them differently for reasons of computational efficiency. The two-point model allows a closed form solution for the maximum likelihood estimates of the retention and breakage probabilities. The three-point model can be evaluated more directly than the multi-point method detailed by Lange et al. (1995). In this section we will provide an overview of the likelihood and maximization strategies for the two-point, three-point and multi-point cases.

Multi-point Likelihood

The multi-point model presented in Lange et al. (1995) is given by

$$\begin{aligned} P &= \Pr(X = (x_1, \ldots, x_m)) \\ &= \sum_{g_1} \cdots \sum_{g_m} \binom{c}{g_1} r_1^{g_1} (1 - r_1)^{c - g_1} \\ &\times \prod_{k=1}^{m-1} t_{c,k}(g_k, g_{k+1}) \prod_{k=1}^{m} \phi_k(x_k | g_k) \end{aligned} \tag{1}$$

where g_k is the number of copies of marker k present in the clone, r_1 is the retention probability of any irradiated chromosomal fragment containing the first marker, c is the ploidy, $\phi_k(x_k|g_k)$ is the penetrance (defined below) for marker k, and $t_{c,k}(g_k, g_{k+1})$ is the transition probability from state g_k at locus k to state g_{k+1} at locus $k+1$. The transition probability, $t_{c,k}$, can be expressed as

$$\begin{aligned} t_{c,k}(i,j) &= \sum_{l=\max\{0,i+j-c\}}^{\min\{i,j\}} \binom{i}{l} t_{1,k}(1,1)^l t_{1,k}(1,0)^{i-l} \\ &\quad \times \binom{c-i}{j-l} t_{1,k}(0,1)^{j-l} t_{1,k}(0,0)^{c-i-j+l}. \end{aligned}$$

where

$$\begin{aligned} t_{1,k}(0,0) &= 1-\theta_k r_{k+1} \\ t_{1,k}(0,1) &= \theta_k r_{k+1} \\ t_{1,k}(1,0) &= \theta_k(1-r_{k+1}) \\ t_{1,k}(1,1) &= 1-\theta_k(1-r_{k+1}) \end{aligned}$$

where r_k is the retention probability of any irradiated chromosomal fragment such that the k^{th} marker is the left-most marker. The likelihood in equation (1) can be evaluated rapidly using Baum's forward and backward algorithms (Baum, 1972) and can be maximized utilizing the EM algorithm (Dempster et al., 1977) in concert with a useful update formula from Weeks and Lange (1989). The equal retention model, $r = r_1 = \ldots = r_m$, is merely a special case of the general retention model given in equation (1). Details concerning the evaluation and maximization of equation (1) are provided in Lange et al. (1995).

Three-point Likelihood

Let n_{ijk} represent the number of hybrid clones with $X = (i, j, k)$ in the case where $m = 3$ markers are typed without error or missing data. Note that the vector $N = (n_{000}, n_{001}, \ldots, n_{011}, n_{111})$ is an observation on a multinomial distributed variable with cell probabilities $(p_{000}, p_{001}, p_{010}, p_{011}, p_{100}, p_{101}, p_{110}, p_{111})$ where each p_{ijk} is a function of r, θ_1, θ_2. The log-likelihood is given by

$$L(r, \theta_1, \theta_2|X = x) = K + \sum_i \sum_j \sum_k n_{ijk} \log p_{ijk}. \tag{2}$$

The likelihood in equation (2) is maximized via the EM algorithm with two-point based estimates used for initial parameter estimates. We employ a three-point equal-retention error-free model (Markovian) but we evaluate the likelihood in a different manner than Lange et al. (1995). Each p_{ijk} can be expressed as a polynomial function of r, θ_1, θ_2. These polynomial functions as well as their first and second derivatives were derived using a subroutine that was written in R (Ihaka and Gentleman, 1996). The polynomials were then passed to another R subroutine which generated Fortran code that was dynamically loaded back into R. The EM algorithm was coded in Fortran and also dynamically loaded in R. This approach coupled the quick evaluation and maximization of the likelihood via Fortran with the powerful object-oriented graphic front-end provided by R.

Two-point Likelihood

Let n_{ij} represent the number of hybrid clones with $X = (i, j)$ in the case where $m = 2$ markers are typed without error or missing data. Note that the vector $N = (n_{00}, n_{01}, n_{10}, n_{11}, n_{111})$ is an observation on a multinomial distributed variable with cell probabilities $(p_{00}, p_{01}, p_{10}, p_{11})$ where each p_{ij} is a function of r and θ. Specifically,

$$\begin{aligned} p_{00} &= [(1-r)(1-\theta r)]^c \\ p_{01} &= p_{10} \\ &= (1-r)^c - [(1-r)(1-\theta r)]^c \\ p_{11} &= 1 - p_{00} - 2p_{01} \end{aligned}$$

where c is the ploidy of the data.The log-likelihood is given by

$$L(r, \theta | X = x) = K + \sum_i \sum_j n_{ij} \log p_{ij}$$

where K is a constant. Lange et al. (1995) demonstrate that the likelihood is maximized by

$$\begin{aligned} \widehat{r} &= 1 - \left[\frac{1 - \widetilde{p}_{11} + \widetilde{p}_{00}}{2} \right]^{\frac{1}{c}} \\ \widehat{\theta} &= \frac{1 - r - [\widetilde{p}_{00}]^{\frac{1}{c}}}{r(1-r)} \end{aligned}$$

when $\widetilde{p}_{00}\widetilde{p}_{11} \geq \widetilde{p}_{10}^2$ where $\widetilde{p}_{ij} = \frac{n_{ij}}{\sum_i \sum_j n_{ij}}$. If the $\widetilde{p}_{ij}$'s do not satisfy $\widetilde{p}_{00}\widetilde{p}_{11} \geq \widetilde{p}_{10}^2$ then θ is set to some number close to 1 and the EM algorithm is run until the likelihood converges to within a given tolerance.

Missing Data

For the two-point and three-point models, hybrids with non-consensus or missing data are discarded from the analysis. In the multi-point case, discarding hybrids with non-consensus or missing data could result in substantial loss of data, so these hybrids are typically not discarded. For the multi-point model presented in Lange et al. (1995) inclusion of hybrids with non-consensus or missing data (i.e., with a score of $X_k =?$) is accomplished by setting $\phi(?|g_k) = 1$ for all hidden states g_k.

Definitions of LOD, Δ_{LOD}, and δ_{LOD}

The level of significance attached to a particular order for a subset of markers is often expressed in LODs. The LOD scores have become popular because of their ease of interpretation: a LOD j difference in likelihoods corresponds to a likelihood ratio of $10^j : 1$. We consider the analysis of sets of three markers which we may also refer to as "triples" or "triplets".

Suppose we label three markers M_a, M_b, and M_c. Under the equal retention model presented by Lange et al. (1995), the three unique orderings of these three markers are $\{(M_a, M_b, M_c)$, (M_a, M_c, M_b), $(M_c, M_a, M_b)\}$. Note that, under the equal retention model, the orders (M_a, M_b, M_c) and (M_c, M_b, M_a), the orders (M_a, M_c, M_b) and (M_b, M_c, M_a), and the orders $(M_c, M_a, M_b)\}$

and (M_b, M_a, M_c) provide likelihood functions that are identical to one another[2]. The LOD score of marker order (M_a, M_b, M_c) compared to marker order (M_a, M_c, M_b) is given by

$$LOD_{(M_a,M_b,M_c):(M_a,M_c,M_b)} = \log_{10} e \times \Big\{ L(\widehat{r}, \widehat{\theta}_1, \widehat{\theta}_2 | X = (x_a, x_b, x_c)) \quad (3)$$

$$- L(\widetilde{r}, \widetilde{\theta}_1, \widetilde{\theta}_2 | X = (x_a, x_c, x_b)) \Big\} \quad (4)$$

where $\widehat{r}, \widehat{\theta}_1, \widehat{\theta}_2$ are the parameter MLEs under marker order (M_a, M_b, M_c) and $\widetilde{r}, \widetilde{\theta}_1, \widetilde{\theta}_2$ are the parameter MLEs under marker order (M_a, M_c, M_b).

By definition, the LOD score in equation (3) is meaningful only in the context of comparing two different marker orderings. In practice, it is common to refer to the LOD score of a group of markers or of a single marker ordering. The LOD score of a group of markers is the LOD score of the best ordering of the group of markers compared to the second best ordering. The LOD score of a single marker ordering is the LOD score of that ordering compared to the best ordering from the set of possible alternative orderings. We define two distinct statistics to handle these cases. First, we define the Δ_{LOD} score of a set of three markers to be the value of the LOD of the ordering with the highest likelihood compared to the ordering with the second highest likelihood. Second, we define the δ_{LOD} score of a given marker ordering to be the LOD score of that marker ordering compared to the most likely of the remaining two marker orderings. Note that Δ_{LOD} is non-negative by definition while δ_{LOD} may assume both negative and positive values. A positive value of δ_{LOD} indicates that the marker ordering is superior to both remaining orderings. A negative δ_{LOD} score indicates that the marker ordering is less likely than at least one of the remaining orderings.

Results

Data-Sets

We apply our diagnostics to a real data-set which is comprised of a published RH map for *Bos taurus* Autosome 5 (BTA5) (Womack et al., 1997) and an RH panel consisting of 90 bovine-hamster hybrid cell lines typed for 85 markers on BTA5 (Womack et al., 1997). A copy of the panel data can be found in the supplemental meterials. We will refer to this data-set as W-BTA5.

We also apply our diagnostics to two simulated data sets which we label SDAT-N and SDAT-E. Each simulated data-set is comprised of simulated RH panel data and a candidate marker ordering (i.e., a candidate map or a candidate multi-point map). The simulated RH panel data consists of 90 hybrid cell lines typed for 85 markers and were generated under the assumption that the 85 markers are equally spaced along a chromosome of total length 2.5 Rays (i.e., markers are located every 2.97cR)[3]. The panel data in SDAT-N was generated under the assumption that markers were typed without error. The panel data in SDAT-E was generated under the assumption that markers were typed with false positive and false negative error rates fixed at 0.05 for all markers except M_{50} and M_{60}. M_{50} had 14 positive typing scores converted to negative (i.e., a 57% false negative error rate). Marker M_{60} had 12 negative typing scores were converted

[2]The probability of a break between markers M_a and M_b is the same for marker ordering $\{(M_a, M_b, M_c)$ and $\{(M_c, M_b, M_a)$. The same can be said for the probability of a break between markers M_b and M_c and the probability of a break between markers M_a and M_c. Therefore, under the equal retention model, the likelihoods must be identical.

[3]The sum of the estimated distances from the left-most marker to the center marker and from the center marker to the right-most marker in the candidate map for W-BTA5 is approximately 2.5 Rays. Simulated data-sets were designed to be similar to W-BTA5 in this respect.

to positive (i.e., a 16% false positive error rate). The simulated data-set SDAT4 contains RH panel data which was generated by applying simulated typing errors to the panel data used in SDAT-E. The candidate marker ordering corresponding to SDAT-N is the true marker ordering. The candidate marker ordering corresponding to SDAT-E was generated by subjecting the true marker ordering to 50 sequentially applied random marker flips[4]. An inversion was then placed in the map by inverting the order of the 21^{st} through 30^{th} markers. A complete listing of the simulated data-sets is provided in the supplemental materials.

For each data-set, we label the markers $M_1, M_2, \ldots, M_{85}$ in accordance with the order in which they appear in that data-set's candidate RH Map. For the real data-set, a table relating these labels to their published names can be found in the supplemental materials.

Calculations

In this section, we discuss the calculations that were performed on the real and simulated data-sets.

Multi-point Calculations

The RH panel data was analyzed with markers constrained to the ordering dictated by the candidate map. The retention model in equation (1) was fitted to the data but the retention parameters were constrained to be equal. Maximum likelihood estimates for r and $\theta_1, \theta_2, \ldots, \theta_{84}$ were obtained. The MLEs for distance, $\widehat{d}_k = -\log(1-\widehat{\theta}_k)$, $k = 1, 84$ were also calculated.

All possible inter-marker distances were estimated conditioned on the candidate map. Suppose, as mentioned above, that the markers are labeled $M_1, M_2, \ldots, M_{85}$ in accordance with the order in which they appear in the candidate map. Let $d_{i,j}$ denote the distance between marker i and marker j where $i < j$. The MLE for $d_{i,j}$ is calculated

$$\widehat{d}_{i,j} = \sum_{k=i}^{j-1} \widehat{d}_k.$$

All $\binom{85}{2} = 3570$ values of $\widehat{d}_{i,j}$ were calculated conditioned on the candidate map.

Three-point Calculations

Maximum likelihood estimates for $r_{3pt,i,j,k}, \theta_{3pt,i,j}$, and $\theta_{3pt,j,k}$, were calculated for each set of unique combinations of i, j, k under the three-point equal retention model described above. The MLEs for distance, $\widehat{d}_{3pt,i,j} = -\log(1-\widehat{\theta}_{3pt,i,j})$ and $\widehat{d}_{3pt,j,k} = -\log(1-\widehat{\theta}_{3pt,j,k})$ were also calculated. The maximum likelihoods were compared for the three unique marker orderings associated with each set of values of i, j, k. Using these likelihood values, the observed Δ_{LOD} and δ_{LOD} were calculated. The δ_{LOD} scores were calculated with respect to the triplet marker ordering that was consistent with the candidate RH Map. If we consider the three markers M_1, M_7, and M_{29} then, by definition, the marker ordering (M_1, M_7, M_{29}) is the unique ordering consistent with the candidate RH map. The δ_{LOD} is calculated as the difference between the maximum $\log_{10}$ likelihood associated with the marker ordering (M_1, M_7, M_{29}) and the greater of the maximum $\log_{10}$ likelihoods associated with the orderings of (M_1, M_{29}, M_7) and (M_{29}, M_1, M_7).

[4]The i^{th} marker is randomly selected and the map positions of the i^{th} and $(i+1)^{th}$ markers are flipped

Two-point Calculations

Maximum likelihood estimates for $r_{2pt,i,j}$ and $\theta_{2pt,i,j}$ were calculated for each of the $\binom{85}{2} = 3570$ combinations of markers M_i and M_j, under the two-point equal retention model described above. The MLEs for distance, $\widehat{d}_{2pt,i,j} = -\log(1 - \widehat{\theta}_{i,j})$ also were calculated.

The Two-point Diagnostic Panel

The first diagnostic consists of a collection of $m = 85$ plots that assess whether the two-point estimates for distance between markers, $\widehat{d}_{2pt,i,j}$, are consistent with their multi-point counterparts, $\widehat{d}_{i,j}$. In practice, we assess consistency of the two-point estimates for breakage probability $\widehat{\theta}_{2pt,i,j} = 1 - e^{-abs(\widehat{d}_{2pt,i,j})}$ with the multi-point distances $\widehat{d}_{i,j} = -\log(1 - \widehat{\theta}_{i,j})$ because scaling the y-axis of the graphs with respect to $\widehat{\theta}_{2pt,i,j}$ provides a graph that is easier to interpret than one scaled with respect to $\widehat{d}_{2pt,i,j}$. We call this collection of plots, a "Two-Point Diagnostic Panel".

Each plot within a Two-Point Diagnostic Panel corresponds to the selection of a single marker as a reference marker. The plot is designed to compare the two-point and multi-point estimates of distance between the reference marker and each of the $m - 1 = 84$ remaining non-reference markers. The plot consists of $m - 1$ points corresponding to each of the $m - 1$ non-reference markers. The plotted points have an x-coordinate corresponding to the position of the non-reference marker in the multi-point map[5] and a y-coordinate corresponding to the two-point estimates for the breakage probability with respect to the reference and non-reference marker. In each plot, a dotted vertical line indicates the position of the reference marker as estimated by the multi-point map and dashed lines indicate the two-point breakage probabilities consistent with the candidate multi-point map. If the candidate map order were correct and the RH panel were error free, we would expect the plotted points to be positioned close to the dashed lines. Therefore, if a marker ordering is generally correct (i.e., mis-orderings are confined to local regions of the chromosome) we expect the plot to display points in the general shape of a convex function in the neighborhood of the minimum with the minimum occurring at the position of the reference marker. If portions of the map have been inverted (block inversions) then we expect that to see sections of the map where the plotted points form patterns discordant with the expected convex pattern. Plotted points that are consistently positioned beneath the dashed lines suggest that some, if not all, multi-point marker distance estimates may be inflated.

For the remainder of this section, we apply our diagnostic to the simulated and real example data-sets. The simulated data-sets have been designed to illustrate how our diagnostic responds to plausible errors in the map ordering and in the RH panel marker typing data. To establish a point of reference, we begin by applying our diagnostic to a data-set (SDAT1) with a candidate map that is correct and RH panel data that are error-free.

Figure 1 contains nine of the 85 plots from the two-point diagnostic panel for the simulated data-set SDAT-E. In practice, all 85 plots would be generated and inspected. For clarity of presentation, we have opted to include only the subset of plots corresponding to markers $M_1, M_{11}, M_{22}, M_{33}, M_{44}, M_{55}, M_{66}, M_{77}$, and M_{85}. The points in each of the graphs in Figure 1 lie close to the dashed lines in the neighborhood of the minima; indicating that the estimated two-point breakage probabilities are consistent with the multi-point estimates.

[5]The position of a marker is quantified as the estimated distance in Rays from the left most marker in the multi-point map.

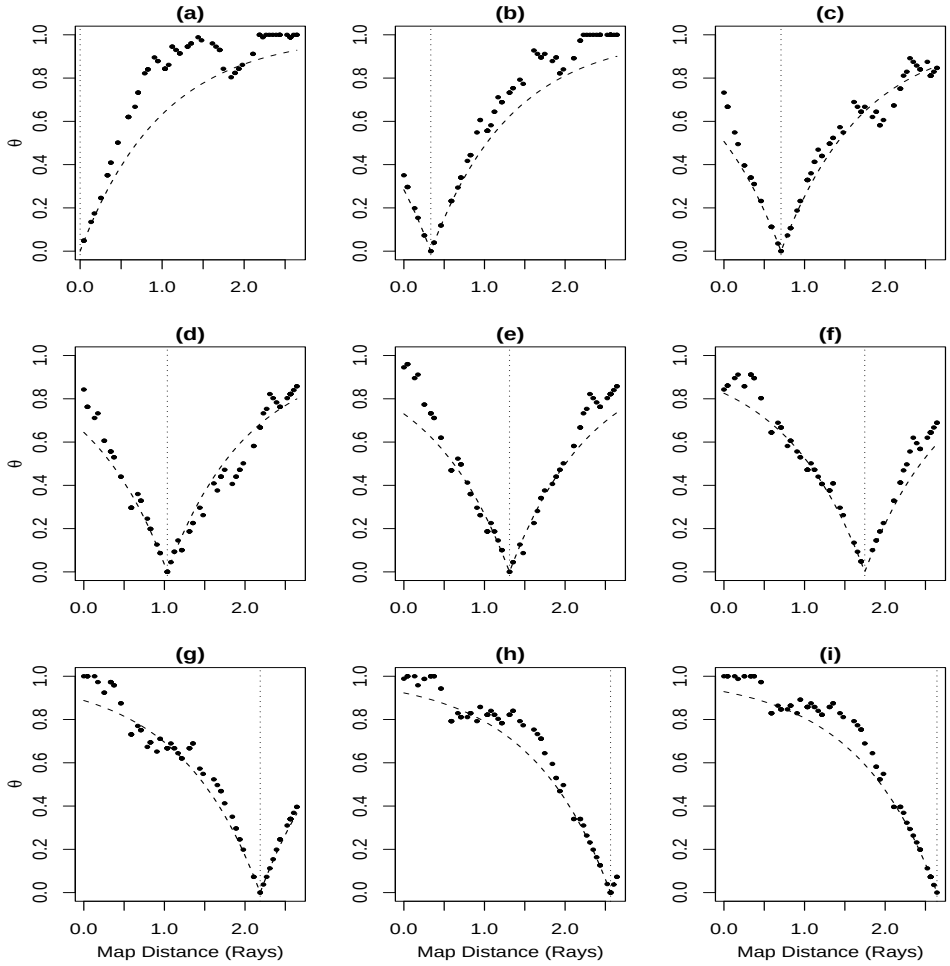

Figure 1: A two-point diagnostic panel with nine reference markers from simulated data-set SDAT-E. Horizontal axes represent the position of markers in Rays as estimated by the multi-point map. Vertical axes represent the estimated two-point breakage probability between the reference and all non-reference markers. A dotted vertical line indicates the position of the reference of marker as estimated by the multi-point map. Dashed lines indicate the two-point breakage probabilities consistent with the multi-point map. The estimated map length is approximately equal to the true value of 2.5 Rays. The plotted points lie in close proximity to the dashed lines; indicating that the estimated two-point breakage probabilities are consistent with the multi-point estimate

Figure 2 contains nine of the 85 plots from the two-point diagnostic panel for the simulated data-set SDAT-E. The candidate marker ordering corresponding to SDAT-E was generated by subjecting the true marker ordering to 50 sequentially applied random marker flips (i.e., the i^{th} marker is randomly selected and the map positions of the i^{th} and $(i+1)^{st}$ markers are flipped) and by inverting the order of the 21^{st} through 30^{th} markers. Introducing the random marker flips into the candidate map yields a map in which markers are positioned in the correct regions of the map but not the correct order within regions. With the exception of 2(a) each of the plots in Figure 2 contain regions where plotted points fall consistently below the dashed lines. This indicates that the two-point estimates for distance (and breakage probability) are less than their multi-point counterparts. The simulated map length for the data used in all simulated data-sets was 2.5 Rays. A comparison of the x-axis scale in 2 reveals that the estimated map length has been inflated by errors in the candidate map. Inverted markers are denoted by open circles in Figure 2. Plotted points that correspond to inverted markers are discordant with the convex shape expected under correct ordering in plots in Figure 2(a), 2(b), 2(d), 2(e), and 2(f). Marker M_{50} had 14 positive typing scores converted to negative (i.e., a 57% false negative error rate). Marker M_{60} had 12 negative typing scores were converted to positive (i.e., a 16% false positive error rate). Plotted points for markers M_{50} and M_{60} are highlighted in Figure 2 with gray squares and gray inverted triangles, respectively. Plotted points for markers M_{50} and M_{60} appear as distinct outliers in Figure 2(f). We consider these points to be outliers based on comparisons of their vertical positions with the vertical positions of the plotted points in the regions of the map surrounding markers M_{50} and M_{60}.

Figure 3 contains nine of the 85 plots from the two-point diagnostic panel of the example data set W-BTA5. The plotted points lie beneath the dashed lines across all nine plots. The general trend of the points between 0 to 4 Rays in Figure 3 (d) is similar to the patterns found in Figure 2 for points corresponding to markers that were inverted in the simulated data-set. In Figure 3, the plotted points for markers M_{36}, M_{57}, and M_{75} have been highlighted with open circles, triangles and diamonds respectively. Figures 3(d), 3(g) and 3(h) all contain an outlying plotted point corresponding to marker M_{36} (highlighted with open circles). In each figure, the estimated two-point breakage probabilities associated with M_{36} is much higher than the estimated breakage probabilities associated with most, if not all, other non-reference markers. The behavior of marker M_{36} in these plots is similar to the behavior of the mistyped markers M_{50} and M_{60} from SDAT4 in Figure 2.

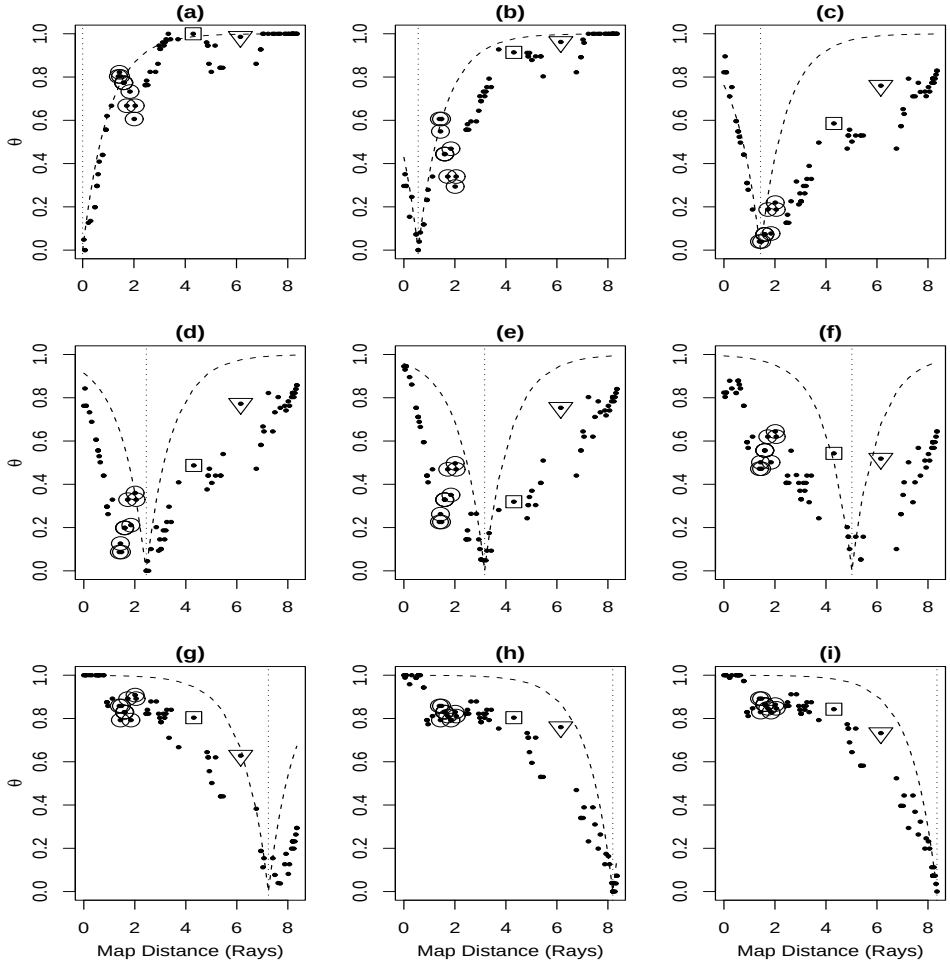

Figure 2: A two-point diagnostic panel with nine reference markers from simulated data-set SDAT-E. Horizontal axes represent the position of markers in Rays as estimated by the multi-point map. Vertical axes represent the estimated two-point breakage probability between the reference and all non-reference markers. A dotted vertical line indicates the position of the reference of marker as estimated by the multi-point map. Dashed lines indicate the two-point breakage probabilities consistent with the multi-point map. Plotted points for markers M_{50} and M_{60} are highlighted with gray squares and gray inverted triangles, respectively. Inverted markers are denoted by open circles.

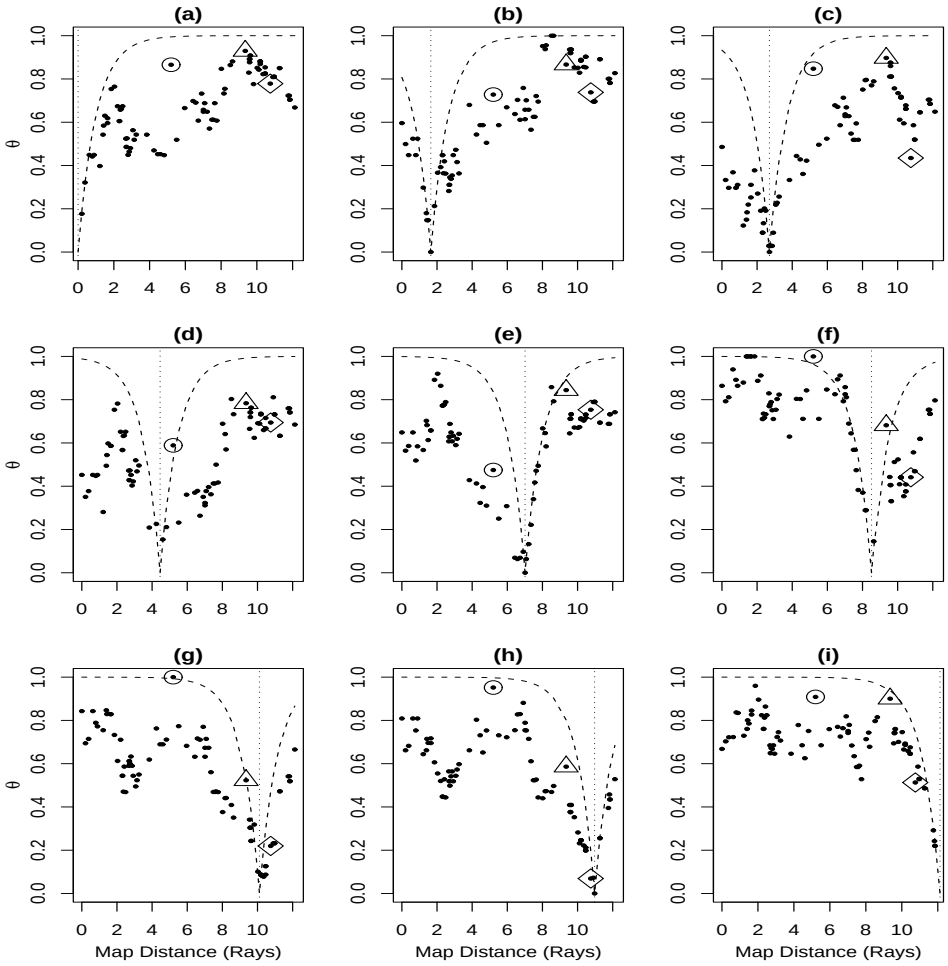

Figure 3: A two-point diagnostic panel with nine reference markers from the example data-set W-BTA5. Horizontal axes represent the position of markers in Rays as estimated by the multi-point map. Vertical axes represent the estimated two-point breakage probability between the reference and all non-non-reference markers. A dotted vertical line indicates the position of the reference of marker as estimated by the multi-point map. Dashed lines indicate the two-point breakage probabilities consistent with the multi-point map. The plotted points for markers M_{36}, M_{57}, and M_{75} have been highlighted with open circles, triangles and diamonds respectively.

Our two-point diagnostic plot also can be useful in estimating the locations of markers that are not included in the candidate RH map. Figure 4 contains two-point diagnostic plots for four positional candidate genes for meat tenderness. The dashed lines included in the previous figures are not included in Figure 4 as the locations of the four markers are not specified by the candidate marker map. We expect that these plots should be consistent with the the plots in Figure 3 . For example, all plots in Figure 4 suggest a block inversion in the neighborhood of 0R to 4R. The plot in Figure 4(b) corresponds to the marker MMP19 and exhibits provides strong evidence for the placement of the marker in this putative region. We consider the evidence to be strong because (1) the minimum estimated breakage probability is relatively small (which indicates a close association with markers that have already been mapped) and, (2) the plotted points form a convex pattern in the area of the minimum. The plot in Figure 4(c) corresponds to marker MYF5TX. The possible block inversion may explain the discordant points in the area of 0 Rays to 4 Rays. Otherwise Figure 4(c) indicates that MYF5TX maps to a position on the map consistent with prior knowledge of this marker. The plots in Figure 4(a) and Figure 4(d) correspond to markers WNT10B and WIF1, respectively. The patterns found in Figures 4(a) 4(d) may be explained by some or all of the following: (1) the possibility that the markers are positioned is a sparse region of the map; (2) the possibility that the markers have significant typing errors; and (3) the possibility that the markers are positioned in a region of the map populated by a marker (or markers) with significant typing errors in the RH panel data.

Table 1: Information for selected points in Figure 6.

K	# triplets with $\Delta_{LOD} \geq K$	Percent in Map
0	98770(100%)	57.303%
1	33375(33.791%)	65.567%
2	11005(11.142%)	70.023%
3	2861(2.897%)	66.32%
4	571(0.578%)	59.691%
5	80(0.081%)	42.5%
6	3(0.003%)	66.667%

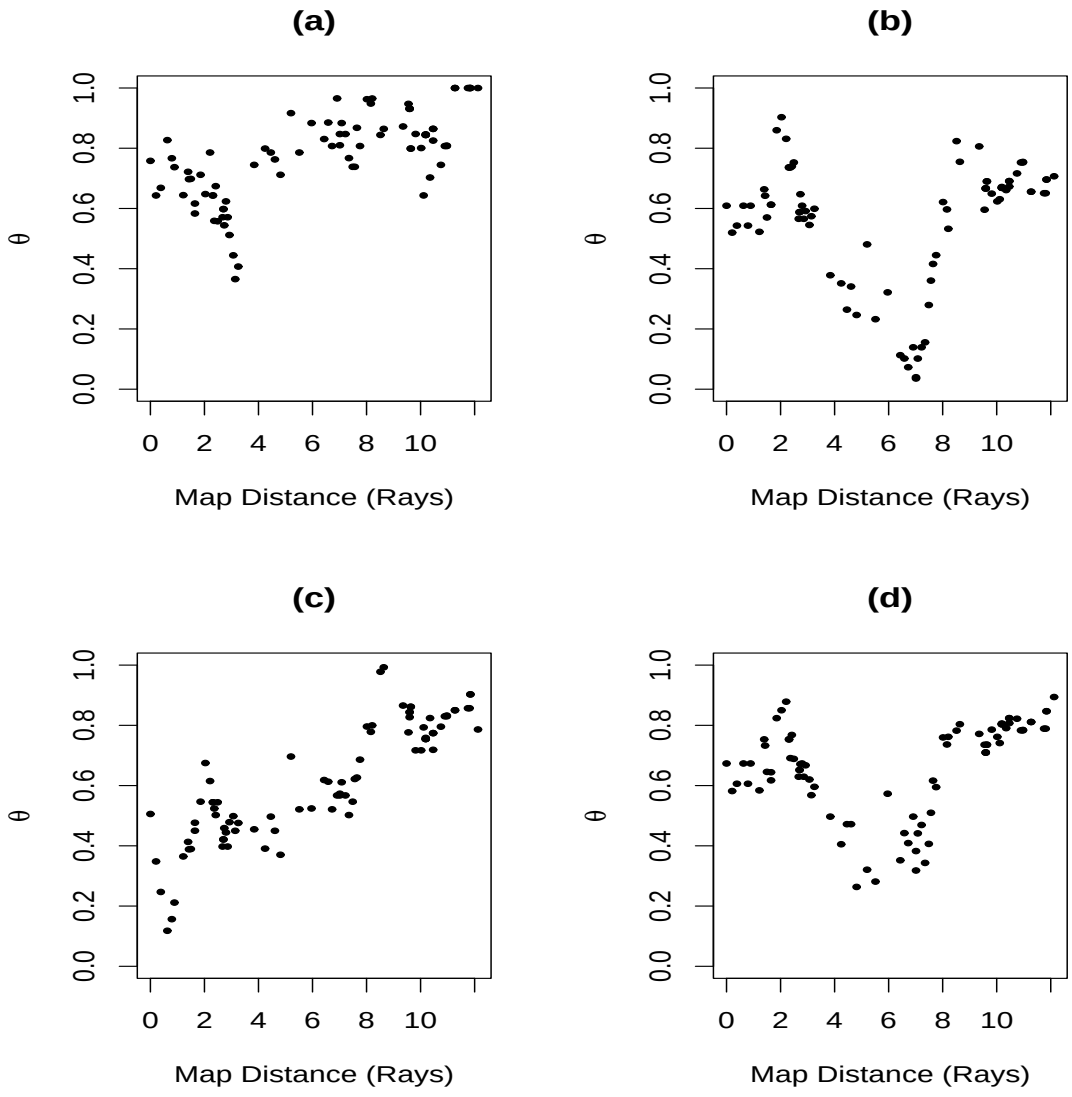

Figure 4: Two-point diagnostic plots for four positional candidate genes for meat tenderness. Horizontal axes represent the position of markers in Rays as estimated by the multi-point map. Vertical axes represent the estimated two-point breakage probability between the reference marker and non-reference markers. (a) Two-point diagnostic plot for candidate gene WNT10B.(b) Two-point diagnostic plot for candidate gene MMP19. (c) Two-point diagnostic plot for candidate gene MYF5.(d) Two-point diagnostic plot for candidate gene WIF1.

The Three-point δ_{LOD} Diagnostic Plot

The remaining diagnostics were designed to assess whether the multi-point fit is consistent with all possible three-point fits. The first such diagnostic utilizes the δ_{LOD} score which measures the likelihood of the candidate map triplet ordering map relative to the best alternative ordering. A negative δ_{LOD} score indicates inconsistency with the candidate map ordering. If a marker is well positioned in the candidate map then we would expect that the number of triplets with negative δ_{LOD} scores will not be significantly larger than the numbers associated with other well placed markers. If the marker position is in error, then we expect the number of triplets with negative δ_{LOD} scores will be larger than the numbers associated with well placed markers. For our diagnostic, we calculate for each marker the number of triplets with δ_{LOD} less than or equal to some number $-K$ and compare these values across markers. We recommend re-examining the placement of markers (or regions of markers) with the largest numbers of triplets having $\delta_{LOD} \leq -K$. Since selection of an optimal value for K appears to be data dependent, we recommend performing these calculations for several values of K.

Figure 5 contains three-point δ_{LOD} diagnostic plots for the eighty five markers in the example data-set W-BTA5. Diagnostic plots are presented for values of $K \in \{1, 3, 5, 7\}$. The horizontal axes represents the position of markers in Rays as estimated by the multi-point map. The vertical axes represent a range of values for the number of the marker triplets with $\delta_{LOD} \leq -K$. Many of the markers in the region of 0 Rays to 4 Rays are included in relatively large numbers of triplets with negative δ_{LOD} scores. This is consistent with Figure 3 which suggests a possible block inversion in that region of the candidate map. The plotted points for markers M_{36}, M_{57}, and M_{75} have been highlighted as in Figure 3. Markers M_{57}, and M_{75} appear as outliers across many of the graphs in Figure 5; most notably marker M_{75} in plots 5(a) through 5(c). Marker M_{36}, which may have significant typing errors, does not behave as an outlier in Figure 5. It is possible that marker M_{36} is relatively well placed within the candidate map despite having significant typing errors.

The Three-point Δ_{LOD} Diagnostic Plot

The Δ_{LOD} score measures the likelihood of the optimal triplet ordering relative to the next best alternative ordering. We calculate the proportion of triplets with $\Delta_{LOD} \geq K$ that are consistent with the candidate map. By consistent we mean that the most likely ordering of a given triplet is the ordering indicated by the candidate map. Assuming adequacy of the panel typing data, then we expect that the proposition of triplets with $\Delta_{LOD} \geq K$ that are consistent with the candidate map, given that $\Delta_{LOD} \geq K$, will be close to 1 for all values of K. Figure **??** contains a three-point Δ_{LOD} diagnostic plot for eighty five markers in the simulated data-set SDAT1. The horizontal axis represents range of values for Δ_{LOD} across all $\frac{85!}{82!3!} = 98770$ unique sets of triplets.The vertical axis represents a range of values for the proportion of triplets included in the multi-point map, given a Δ_{LOD} greater than or equal to values specified by the horizontal axis. Plotted vertical lines indicate observed quantiles of Δ_{LOD} across all unique sets of triplets.

Figure 6 contains a three-point Δ_{LOD} diagnostic plot for eighty five markers in the example data-set W-BTA5 and Table 1 summarizes information for selected points. Table 1 reveals that 57.303% of the total set of triplets and only 42.5% (34 of 80) of the triplets with $\Delta_{LOD} \geq 4$ were consistent with the candidate map. Implications of these observations are discussed in Section D.

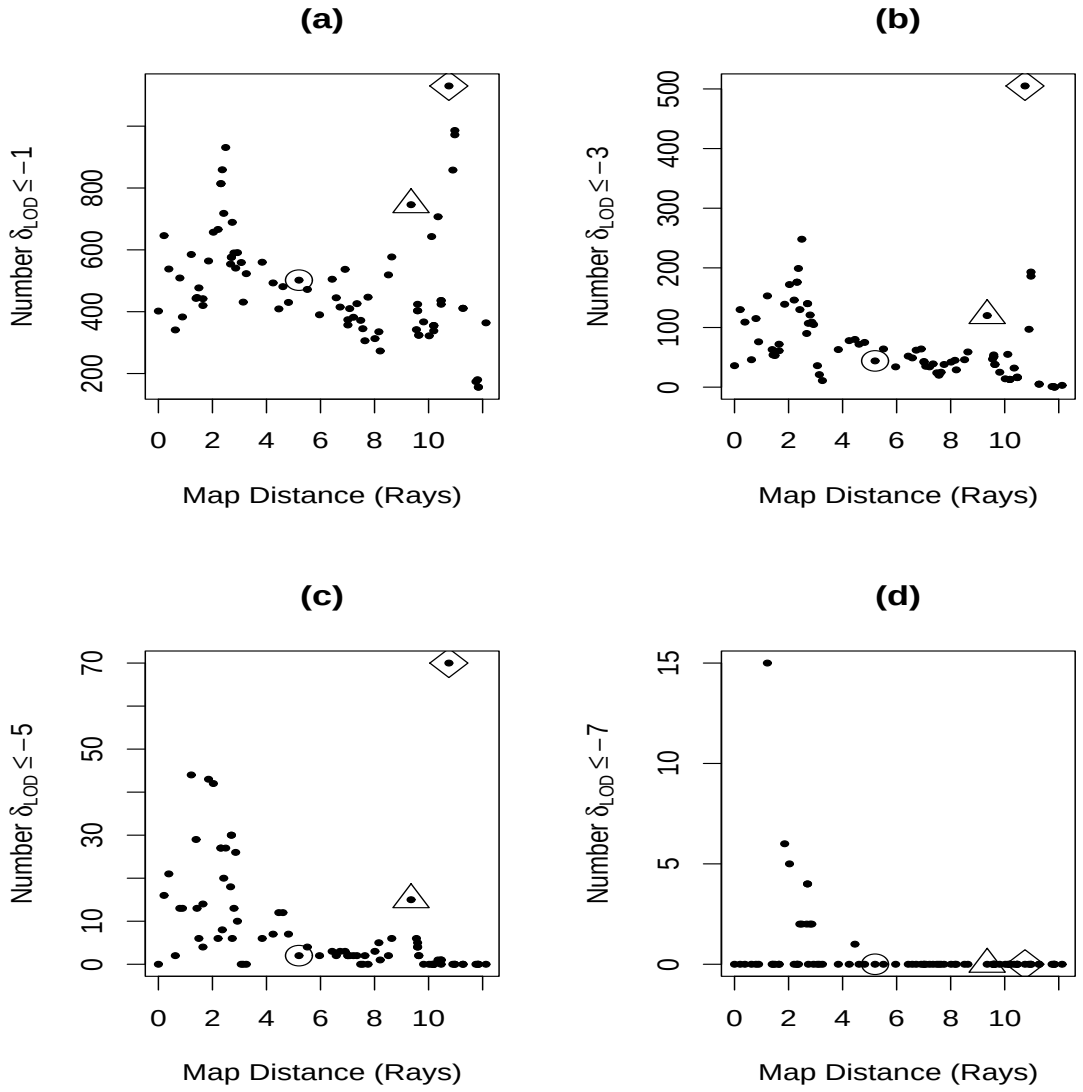

Figure 5: A δ_{LOD} diagnostic plot for eighty five markers in the example data-set W-BTA5. The horizontal axis represents the position of markers in Rays as estimated by the multi-point map. The vertical axis represents a range of values for the number of triplets with $\delta_{LOD} \leq -K$ for each marker. The plotted points for markers M_{36}, M_{57}, and M_{75} have been highlighted with open circles, triangles and diamonds respectively. (a) $K = 1$. (b) $K = 3$. (c) $K = 5$. (d) $K = 7$.

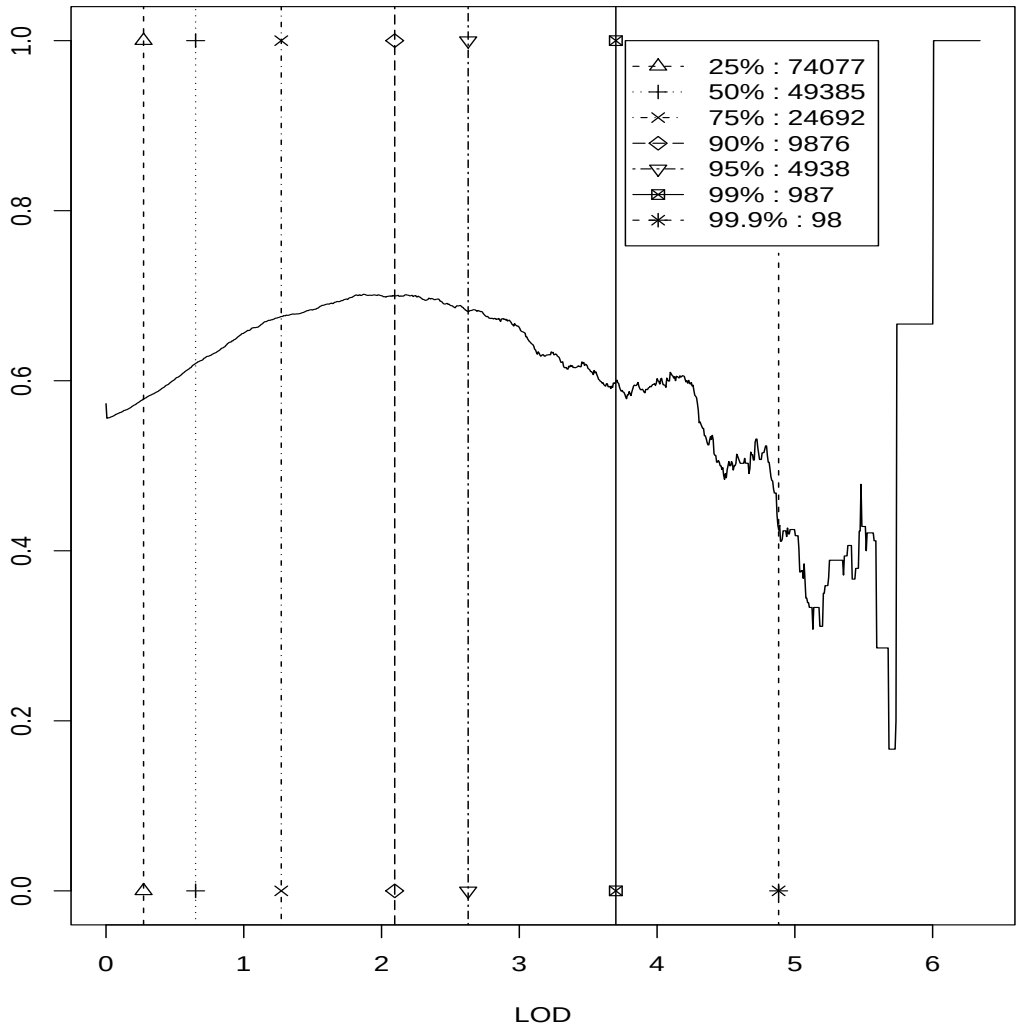

Figure 6: A three-point Δ_{LOD} diagnostic plot for eighty five markers in the example data-set W-BTA5. The horizontal axis represents range of values for Δ_{LOD} across all $\frac{85!}{82!3!} = 98770$ unique sets of triplets.The vertical axis represents a range of values for the proportion of triplets included in the multi-point map, given a Δ_{LOD} greater than or equal to values specified by the horizontal axis. Plotted vertical lines indicate observed quantiles of Δ_{LOD} across all unique sets of triplets.

Discussion

We provided three diagnostic plots to assess the quality of a candidate Radiation Hybrid map. The philosophy of our approach is to assess whether the multi-point fit is consistent with two-point and three-point fits. We have applied our diagnostics to real and simulated data-sets. The simulated data-sets were designed to illustrate how our diagnostics respond to realistic errors in the map ordering and in the RH panel marker typing data. Our battery of diagnostic plots responded informatively to each type of simulated error and demonstrated that our diagnostics can be effective under realistic conditions. For example, the two-point diagnostic panel for the real data-set W-BTA5 (Figure 3) exhibited many of the error-specific characteristics found in the the two-point diagnostic panels for the simulated data-sets; lending credibility to the hypothesis that the real data-set has local errors in marker position, significant errors of typing in at least one marker and a block inversion in the region of 0 Rays to 4 Rays.

Map inflation is an expected consequence of introducing error into the RH map order and the typing data. Although data-sets SDAT-N and SDAT-E were simulated with a map length of 2.5 Rays, the estimated map length for SDAT-E (the data-set with simulated errors) was severely inflated. Application of our two-point diagnostic to SDAT-E (Figure 2) detected the presence of this inflation. Inspection of the two-point diagnostic plots in Figure 3 supports the hypothesis that the candidate map for W-BTA5 is significantly inflated. Since we have not established that map inflation occurs only in presence of errors in the RH data-set (either with the candidate map or with the marker typing data), we cannot interpret the inflation of the W-BTA5 map as definitive evidence for errors in the candidate map. However, we can can conclude that evidence of inflation is consistent with the presence of such errors and we recommend that the candidate map and typing data be re-examined with this possibility in mind.

Block inversions of the type simulated in SDAT-E and presented in Figure 2 can generate patterns that are clearly discordant with those expected in the two-point diagnostic plots. Figures 1(a) and 1(b), in which our diagnostic is applied to error-free data, each exhibit discordant patterns in the region of 1.5 to 2.0 Rays. These patterns appear for points in which the estimated two-point breakage probabilities are very large but do not appear in plots 1(d) through 1(i) which correspond to reference markers that are closer to the region of interest. For the simulated data-set SDAT-E, in which a region of the candidate map was inverted, we see that the discordant pattern associated with this region is clearly visible in the plots 2(d) and 2(e) which correspond to reference markers that are relatively close to the region of interest. Therefore, we recommend affording more credibility to discordant patterns that manifest themselves in the two-point diagnostic plots for reference markers positioned close to, and possibly within, the region of interest.

Inspection of the plots in Figure 3 reveals a pattern in the region of 0 Rays to 4 Rays which is consistent with the pattern associated with our simulated block inversion in Figure 2. This pattern appears in all plots except (perhaps) Figure 2 (b), which corresponds to a reference marker in the middle of the possibly inverted region. We feel that this pattern indicates a possible block inversion in the candidate map since the pattern is formed by many points and is clearly defined in plots 2(c) and 2(d) which correspond to reference markers that are close to the possibly inverted region.

The δ_{LOD} diagnostic plots in Figure 5 also provide strong evidence for a possible block inversion between 0 Rays and 4 Rays, as many of the markers in this region are included in a relatively large numbers of triplets with negative δ_{LOD} scores. Under the assumption that the candidate map is correct one would expect that the markers involved in relatively large numbers

of triplets with negative δ_{LOD} scores[6] would be distributed approximately uniformly across the map. Therefore, observing the majority of these markers in the area of 0 to 4 Rays is compelling evidence that the candidate map is flawed in that region.

The plotted points for markers M_{57} have been highlighted with open triangles and the plotted points for marker M_{77} have been highlighted with open diamonds in Figures 3 and 5. The candidate map requires that marker M_{57} is positioned closer than M_{75} to the reference markers in plots 3(a) through 3(f). In each instance, however, the two point estimates indicate that the opposite is true (i.e., that M_{75} than is closer marker M_{57} to the reference marker). The δ_{LOD} diagnostic plots in Figure 5 suggest that markers M_{57} and M_{75} are included in relatively large numbers of triplets with negative δ_{LOD} scores. This evidence, in concert with the observations from plots 3(a) through 3(f), indicate that the relative positions of these markers in the candidate map is incorrect.

When a marker with significant levels of typing error is forced into the multi-point map, the estimates of inter-marker distances with respect to that marker may be inflated. The end result is that the portion of the map which is local to the mis-typed marker will appear to be sparsely populated by markers and the marker will not appear to be close (i.e., the minimum two-point breakage probability estimate across all $m-1$ remaining markers will be large) to any of the markers in the map.

Inspection of Figure 3 for M_{36} marker raises the question: "Does the marker M_{36} map to a sparse part of the candidate map, or is the map sparse because a marker (or markers) in that region of map was (were) typed with significant error?" Since the points corresponding to marker M_{36} (highlighted with open circles) are clearly outlying in plots 3(c), 3(d), 3(f), 3(g), and 3(h), we suspect that the latter is true and would recommend that the typing data for M_{36} be inspected. We note that marker M_{36} can not be characterized as an outlier in any of the δ_{LOD} diagnostic plots in Figure 5. The δ_{LOD} diagnostic is designed primarily for the detection of errors in the candidate map while the two-point diagnostic is designed to detect both errors in candidate map ordering and marker typing data.

Whereas the two-point and δ_{LOD} diagnostic plots are designed to identify specific markers or regions of markers that are in error, the Δ_{LOD} diagnostic is designed to provide an indication of the overall goodness of fit of the candidate map to the RH panel data. Application of the Δ_{LOD} diagnostic to the W-BTA5 data-set (presented in Figure 6 and summarized in Table 1) reveals that more than half of the triplets with $\Delta_{LOD} \geq 5$ are inconsistent with the map. An examination of the entire set of optimal triplet orderings reveals that only 57% are consistent with the multi-point map. Clearly, Figure 6 and Table 1 support the contention that the candidate map is not consistent with the RH panel data.

We have applied our diagnostics to data-sets with 85 markers. For this number of markers, it is possible to evaluate δ_{LOD} and Δ_{LOD} scores across the entire set of 98770 unique marker triplets. In cases where the number of markers is so large that it is impractical to evaluate the the δ_{LOD} and Δ_{LOD} scores across all unique marker orderings, we recommend partitioning the candidate map into overlapping regions containing manageable numbers of markers and then applying our diagnostics in each of these regions.

We have employed parametric equal retention models equivalent to those presented in Lange et al. (1995). Our diagnostics can be used in conjunction with other parametric models, provided that they do not require more degrees of freedom than are available in the two and three-point cases. It is unclear how our diagnostics are effected by model mis-specification (e.g., assuming an equal retention probability model when is is not true), but is the subject of future research.

[6] We say 'relatively large' since the number is large relative to the distribution of numbers across the markers in the map.

References

Agarwala, R., Applegate, D. L., Maglott, D., Schuler, G. D., and Schaffer, A. A. (2000), "A Fast Scalable Radiation Hybrid Map Construction and Integration Strategy," *Genome Research*, 10, 350–364.

Baum, L. (1972), "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, 3, 1–8.

Ben-Dor, A. and Chor, B. (1997), "On Constructing Radiation Hybrid Maps," *Journal of Computational Biology*, 4, 517–533.

Ben-Dor, A., Chor, B., and Pelleg, D. (2000), "RHO - Radiation Hybrid Ordering," *Genome Research*, 10, 365–378.

Bo, T. H., Jonassen, I., Eidhammer, I., and Helgesen, C. (2002), "A fast top-down method for constructing reliable radiation hybrid frameworks," *Bioinformatics*, 18, 11–18.

Boehnke, M. (1992), "Radiation hybrid mapping by minimization of the number of obligate chromosome breaks," *Cytogenet Cell Genet*, 59, 96–98.

Boehnke, M., Lange, K., and Cox, D. R. (1991), "Statistical Methods for Multipoint Radiation Hybrid Mapping," *American Journal of Human Genetics*, 49, 1174–1188.

Cox, D. R., Burmeister, M., Price, E. R., Suwon, K., and Myers, R. M. (1990), "Radiation Hybrid Mapping: A Somatic Cell Genetic Method for Constructing High-Resolution Maps of Mammalian Chromosomes," *Science*, 250, 245–250.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society, Series B, Methodological*, 39, 1–22.

Goss, S. J. and Harris, H. (1975), "New method for mapping genes in human chromosomes," *Nature*, 255, 680–684.

Hansen, G. R. (2003)*Evaluating Six Positional Candidate Genes for Effects on Meat Tenderness*. Ph.D. dissertation, Texas A&M University.

Ihaka, R. and Gentleman, R. (1996), "R: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics*, 5, 299–314.

Lange, K., Boehnke, M., Cox, D. R., and Lunetta, K. L. (1995), "Statistical Methods for Polyploid Radiation Hybrid Mapping," *Genome Research*, 5, 136–150.

Slonim, D., Kruglyak, L., Stein, L., and Lander, E. (1997), "Building Human Genome Maps with Radiation Hybrids," *Journal Comput Biol.*, 4, 487–504.

Weeks, D. E. and Lange, K. (1989), "Trials, Tribulations and Triumphs of the EM Algorithm in Pedigree Analysis," *IMA Journal of Mathematics Applied in Medicine & Biology*, 6, 209–232.

Womack, J. E., Johnson, J. S., Owens, E. K., Rexroad III, C. E., Schlapfer, J., and Yang, Y.-P. (1997), "A whole-genome radiation hybrid panel for bovine gene mapping," *Mammalian Genome*, 8, 854–856.

A Appendix

A.1 Calculating the Polyploid Radiation Hybrid Likelihood and Its Derivatives

The likelihood presented in Lange et al. (1995) is given by

$$\begin{aligned} P &= \Pr(X = (x_1, \ldots, x_m)) \\ &= \sum_{g_1} \cdots \sum_{g_m} \binom{c}{g_1} r_1^{g_1} (1 - r_1)^{c - g_1} \\ &\quad \prod_{k=1}^{m-1} t_{c,k}(g_k, g_{k+1}) \prod_{k=1}^{m} \phi_k(x_k | g_k) \end{aligned}$$

where g_k is the number of copies of marker k present in the clone, r_1 is the retention probability of any irradiated chromosomal fragment containing the first marker, c is the ploidy, $\phi_k(x_k|g_k)$ is the penetrance (defined below) for marker k, and $t_{c,k}(g_k, g_{k+1})$ is the transition probability from state g_k at locus k to state g_{k+1} at locus $k+1$. The transition probability, $t_{c,k}$, can be expressed as

$$\begin{aligned} t_{c,k}(i,j) &= \sum_{l=\max\{0, i+j-c\}}^{\min\{i,j\}} \binom{i}{l} t_{1,k}(1,1)^l t_{1,k}(1,0)^{i-l} \\ &\quad \times \binom{c-i}{j-l} t_{1,k}(0,1)^{j-l} t_{1,k}(0,0)^{c-i-j+l}. \end{aligned}$$

with

$$\begin{aligned} t_{1,k}(0,0) &= 1 - \theta_k r_{k+1} \\ t_{1,k}(0,1) &= \theta_k r_{k+1} \\ t_{1,k}(1,0) &= \theta_k (1 - r_{k+1}) \\ t_{1,k}(1,1) &= 1 - \theta_k (1 - r_{k+1}) \end{aligned}$$

The likelihood is evaluated recursively using Baum's forward algorithm to evaluate the probabilities

$$f_k(g_k) = Pr(X_1 = x_1, \ldots, X_{k-1} = x_{k-1}, G_k = g_k)$$

and the backwards algorithm to evaluate the probabilities.

$$b_k(g_k) = Pr(X_{k+1} = x_{k+1}, \ldots, X_m = x_m | G_k = g_k)$$

The forward update of $f_k(g_k)$ is

$$f_{k+1}(g_{k+1}) = \sum_{g_k} f_k(g_k) \phi_k(x_k|g_k) t_{c,k}(g_k, g_{k+1})$$

with initial condition

$$f_1(g_1) = Pr(G_1 = g_1) = \binom{c}{g_1} r_1^{g_1} (1 - r_1)^{c - g_1}.$$

The backward update of $b_k(g_k)$ is

$$b_{k-1}(g_{k-1}) = \sum_{g_k} t_{c,k-1}(g_{k-1}, g_k) \phi_k(x_k|g_k) b_k(g_k)$$

with initial condition $b_m(g_m) = 1$. The likelihood can be constructed from the probabilities of the RH vector, P,

$$\begin{aligned} P &= \sum_{g_m} f_m(g_m)\phi_m(x_m|g_m) \\ &\text{or} \\ P &= \sum_{g_1} f_1(g_1)\phi_1(x_1|g_1)b_1(g_1) \end{aligned}$$

A. Evaluation of First Derivatives

The first derivatives are of three forms. Each form can be evaluated as follows.

First Derivatives of the form $\frac{d}{d\theta_k}P$:

$$\begin{aligned} \frac{d}{d\theta_k}P = &\sum_{g_k}\sum_{g_{k+1}} f_k(g_k)\phi_k(x_k|g_k)\frac{d}{d\theta_k}t_{c,k}(g_k, g_{k+1}) \\ &\times \phi_{k+1}(x_{k+1}|g_{k+1})b_{k+1}(g_{k+1}). \end{aligned}$$

First Derivatives of the form: $\frac{d}{dr_k}P$,$(k > 1)$

$$\begin{aligned} \frac{d}{dr_k}P = &\sum_{g_{k-1}}\sum_{g_k} f_{k-1}(g_{k-1})\phi_{k-1}(x_{k-1}|g_{k-1}) \\ &\times \frac{d}{dr_k}t_{c,k-1}(g_{k-1}, g_k)\phi_k(x_k|g_k)b_k(g_k). \end{aligned}$$

First Derivatives of the form: $\frac{d}{dr_1}P$

$$\frac{d}{dr_1}P = \sum_{g_1}\frac{d}{dr_1}f_1(g_1)\phi_1(x_1|g_1)b_1(g_1)$$

If an equal retention model is assumed, the derivative of the likelihood with respect to the retention probability $r = r_1 = \ldots = r_m$ can be calculated by applying the chain rule. Specifically,

$$\frac{d}{dr}P = \sum_k \frac{d}{dr_k}P.$$

B. Evaluation of Second Derivatives

We extend work presented in Lange et al. (1995) to provide for the evaluation of the second derivatives of P. To compute these second derivatives, we first define:

$$f^*_{k,k+j}(g_k, g_{k+j}) = \sum_{g_{k+j-1}} f^*_{k,k+j-1}(g_k, g_{k+j-1})t_{c,k+j-1}(g_{k+j-1}, g_{k+j})\phi_{k+j}(x_{k+j}|g_{k+j})$$

with initial condition

$$f^*_{k,k+1}(g_k, g_{k+1}) = t_{c,k}(g_k, g_{k+1})\phi_{k+1}(x_{k+1}|g_{k+1}).$$

The equations for the evaluation of each of the second derivatives of P with respect to $\theta_k, (k = 1, \ldots, m-1)$ and $r_{k'}, (k' = 1, \ldots, m)$ can be categorized as belonging to one of sixteen distinct cases.

Evaluation of second derivatives of the form $\frac{d^2}{d\theta_k d\theta_{k'}}$:

case 1: $k' = k$

$$\frac{d^2}{d\theta_k^2}P = \sum_{g_k}\sum_{g_{k+1}} f_k(g_k)\phi_k(x_k|g_k)\frac{d^2}{d\theta_k^2}t_{c,k}(g_k, g_{k+1}) \times \phi_{k+1}(x_{k+1}|g_{k+1})b_{k+1}(g_{k+1}).$$

case 2: $k' = k + 1$

$$\frac{d^2}{d\theta_k d\theta_{k+1}}P = \sum_{g_k}\sum_{g_{k+1}}\left\{ f_k(g_k)\phi_k(x_k|g_k)\frac{d}{d\theta_k}t_{c,k}(g_k, g_{k+1})\phi_{k+1}(x_{k+1}|g_{k+1}) \times \sum_{g_{k+2}}\frac{d}{d\theta_{k+1}}t_{c,k+1}(g_{k+1}, g_{k+2})\phi_{k+2}(x_{k+2}|g_{k+2})b_{k+2}(g_{k+2})\right\}$$

case 3: $k' = k + j, (j \geq 2)$

$$\frac{d^2}{d\theta_k d\theta_{k+j}}P = \sum_{g_k}\sum_{g_{k+1}}\left\{ f_k(g_k)\phi_k(x_k|g_k)\frac{d}{d\theta_k}t_{c,k}(g_k, g_{k+1})\phi_{k+1}(x_{k+1}|g_{k+1}) \times \sum_{g_{k+j}}\sum_{g_{k+j+1}} f^*_{k+1,k+j}(g_{k+1}, g_{k+j})\frac{d}{d\theta_{k+j}}t_{c,k+j}(g_{k+j}, g_{k+j+1}) \times \phi_{k+j+1}(x_{k+j+1}|g_{k+j+1})b_{k+j+1}(g_{k+j+1})\right\}$$

Evaluation of second derivatives of the form $\frac{d^2}{dr_k dr_{k'}}$:

case 4: $k' = k = 1$

$$\frac{d^2}{d^2 r_1}P = \sum_{g_1}\frac{d^2}{d^2 r_1}f_1(g_1)\phi_1(x_1|g_1)b_1(g_1)$$

case 5: $k = 1, k' = 2$

$$\frac{d^2}{dr_1 dr_2}P = \sum_{g_1}\sum_{g_2}\frac{d}{dr_1}f_1(g_1)\phi_1(x_1|g_1) \times \frac{d}{dr_2}t_{c,1}(g_1, g_2)\phi_2(x_2|g_2)b_2(g_2)$$

case 6: $k=1, k'=k+j, (j\geq 2)$

$$\frac{d^2}{dr_1 dr_{j+1}}P = \sum_{g_1}\Bigg\{\frac{d}{dr_1}f_1(g_1)\phi_1(x_1|g_1) \times \sum_{g_{k+j-1}}\sum_{g_{k+j}} f^*_{1,k+j-1}(g_1,g_{k+j-1}) \times \frac{d}{dr_{k+j}}t_{c,k+j-1}(g_{k+j-1},g_{k+j})\phi_{k+j}(x_{k+j}|g_{k+j})b_{k+j}(g_{k+j})\Bigg\}$$

case 7: $k'=k, (k\geq 2)$

$$\frac{d^2}{dr_k^2}P = \sum_{g_{k-1}}\sum_{g_k} f_{k-1}(g_{k-1})\phi_{k-1}(x_{k-1}|g_{k-1})\frac{d^2}{dr_k^2}t_{c,k-1}(g_{k-1},g_k) \times \phi_k(x_k|g_k)b_k(g_k).$$

case 8: $k, k'=k+1, (k\geq 2)$

$$\frac{d^2}{dr_k dr_{k+1}}P = \sum_{g_{k-1}}\sum_{g_k}\Bigg\{f_{k-1}(g_{k-1})\phi_{k-1}(x_{k-1}|g_{k-1})\frac{d}{dr_k}t_{c,k-1}(g_{k-1},g_k) \times \phi_k(x_k|g_k)b_k(g_k)\sum_{g_{k+1}}\frac{d}{dr_{k+1}}t_{c,k}(g_k,g_{k+1}) \times \phi_{k+1}(x_{k+1}|g_{k+1})b_{k+1}(g_{k+1})\Bigg\}$$

case 9: $k, k'=k+j, (k\geq 2, j\geq 2)$

$$\frac{d}{dr_k}P = \sum_{g_{k-1}}\sum_{g_k}\Bigg\{f_{k-1}(g_{k-1})\phi_{k-1}(x_{k-1}|g_{k-1})\frac{d}{dr_k}t_{c,k-1}(g_{k-1},g_k) \times \phi_k(x_k|g_k)\sum_{g_{k+j-1}}\sum_{g_{k+j}} f^*_{k,k+j-1}(g_k,g_{k+j-1}) \times \frac{d}{dr_{k+j}}t_{c,k+j-1}(g_{k+j-1},g_{k+j})\phi_{k+j}(x_{k+j}|g_{k+j})b_{k+j}(g_{k+j})\Bigg\}$$

Evaluation of cross derivatives of the form $\frac{d^2}{dr_{k'}d\theta_k}$ **:**

case 10: $k=1, k'=1$

$$\frac{d^2}{dr_1 d\theta_1}P = \sum_{g_1}\sum_{g_2}\frac{d}{dr_1}f_1(g_1)\phi_1(x_1|g_1)\frac{d}{d\theta_1}t_{c,1}(g_1,g_2) \times \phi_2(x_2|g_2)b_2(g_2).$$

case 11: $k \geq 2, k' = 1$

$$\frac{d^2}{dr_1 d\theta_k} P = \sum_{g_1} \Bigg\{ \frac{d}{dr_1} f_1(g_1)\phi_1(x_1|g_1) \times \sum_{g_k} \sum_{g_{k+1}} f^*_{1,k}(g_1, g_k) \times \frac{d}{d\theta_k} t_{c,k}(g_k, g_{k+1})\phi_{k+1}(x_{k+1}|g_{k+1})b_{k+1}(g_{k+1}) \Bigg\}$$

case 12: $k, k' = k+1, (k \geq 1)$

$$\frac{d^2}{dr_{k+1} d\theta_k} P = \sum_{g_k} \sum_{g_{k+1}} f_k(g_k)\phi_k(x_k|g_k) \frac{d^2}{dr_{k+1} d\theta_k} t_{c,k}(g_k, g_{k+1}) \times \phi_{k+1}(x_{k+1}|g_{k+1})b_{k+1}(g_{k+1}).$$

case 13: $k, k' = k+2, (k \geq 1)$

$$\frac{d^2}{dr_{k+2} d\theta_k} P = \sum_{g_k} \sum_{g_{k+1}} \Bigg\{ f_k(g_k)\phi_k(x_k|g_k) \frac{d}{d\theta_k} t_{c,k}(g_k, g_{k+1})\phi_{k+1}(x_{k+1}|g_{k+1}) \times \sum_{g_{k+2}} \frac{d}{dr_{k+2}} t_{c,k+1}(g_{k+1}, g_{k+2})\phi_{k+2}(x_{k+2}|g_{k+2})b_{k+2}(g_{k+2}) \Bigg\}$$

case 14: $k, k' = k+j, (j \geq 3)$

$$\frac{d^2}{dr_{k+j} d\theta_k} P = \sum_{g_k} \sum_{g_{k+1}} \Bigg\{ f_k(g_k)\phi_k(x_k|g_k) \frac{d}{d\theta_k} t_{c,k}(g_k, g_{k+1})\phi_{k+1}(x_{k+1}|g_{k+1}) \times \sum_{g_{k+j}} \sum_{g_{k+j+1}} f^*_{k+1,k+j}(g_{k+1}, g_{k+j}) \frac{d}{dr_{k+j+1}} t_{c,k+j}(g_{k+j}, g_{k+j+1}) \times \phi_{k+j+1}(x_{k+j+1}|g_{k+j+1})b_{k+j+1}(g_{k+j+1}) \Bigg\}$$

case 15: $k, k' = k, (2 \leq k \leq (m-1))$

$$\frac{d^2}{dr_k d\theta_k} P = \sum_{g_{k-1}} \sum_{g_k} \Bigg\{ f_{k-1}(g_{k-1})\phi_{k-1}(x_{k-1}|g_{k-1}) \frac{d}{dr_k} t_{c,k-1}(g_{k-1}, g_k)\phi_k(x_k|g_k) \sum_{g_{k+1}} \frac{d}{d\theta_k} t_{c,k}(g_k, g_{k+1})\phi_{k+1}(x_{k+1}|g_{k+1})b_{k+1}(g_{k+1}) \Bigg\}$$

case 16: $k, k', (3 \leq k \leq (m-1), 2 \leq k' \leq (k-1))$

$$\frac{d^2}{dr_{k'}d\theta_k}P = \sum_{g_{k-j-1}}\sum_{g_{k-j}}\Bigg\{f_{k-j-1}(g_{k-j-1})\phi_{k-j-1}(x_{k-j-1}|g_{k-j-1}) \times \frac{d}{dr_{k-j}}t_{c,k-j-1}(g_{k-j-1}, g_{k-j})\phi_{k-j}(x_{k-j}|g_{k-j}) \times \sum_{g_k}\sum_{g_{k+1}} f^*_{k-j,k}(g_{k-j}, g_k)\frac{d}{d\theta_k}t_{c,k}(g_k, g_{k+1}) \times \phi_{k+1}(x_{k+1}|g_{k+1})b_{k+1}(g_{k+1})\Bigg\}$$

C. Evaluation of Second Derivatives for Six Markers

Let γ represent a vector of length p which contains the retention and breakage probability parameters. For m markers and the retention model given in equation A-1,

$$\gamma = [\gamma_1 \quad \gamma_2 \quad \dots \quad \gamma_p]^T = [r_1 \quad \dots \quad r_m \quad \theta_1 \quad \theta_1 \quad \dots \quad \theta_{m-1}]^T .$$

where $p = 2m - 1$ in this case. Table 2 contains, as entries, the case numbers described in the previous section which correspond to the evaluation of all possible second derivatives, $\frac{d^2}{d\gamma_i d\gamma_j}P$.

Table 2: Cases for the evaluation of second derivatives for six marker. Row and column headings denote the parameters γ_i and γ_j, respectively. Tabled numbers refer to the second derivative case (described in the text) which corresponds to the evaluation of $\frac{d^2}{d\gamma_i d\gamma_j}P$.

						γ_j					
γ_i	θ_1	θ_2	θ_3	θ_4	θ_5	r_1	r_2	r_3	r_4	r_5	r_6
θ_1	1	2	3	3	3	10	12	13	14	14	14
θ_2	2	1	2	3	3	11	15	12	13	14	14
θ_3	3	2	1	2	3	11	16	15	12	13	14
θ_4	3	3	2	1	2	11	16	16	15	12	13
θ_5	3	3	3	2	1	11	16	16	16	15	12
r_1	10	11	11	11	11	4	5	6	6	6	6
r_2	12	15	16	16	16	5	7	8	9	9	9
r_3	13	12	15	16	16	6	8	7	8	9	9
r_4	14	13	12	15	16	6	9	8	7	8	9
r_5	14	14	13	12	15	6	9	9	8	7	8
r_6	14	14	14	13	12	6	9	9	9	8	7

D. Maximization of the Likelihood

Given the that the first derivatives $\frac{d}{d\gamma_k}P, (k = 1, \ldots, 2m-1)$, the corresponding contribution to the score function is equal to

$$\frac{d}{d\gamma_k}\log P = \frac{\frac{d}{d\gamma_k}P}{P}.$$

The log-likelihood can then can be maximized utilizing the EM algorithm (Dempster et al., 1977) in concert with the following useful update formula from Weeks and Lange (1989):

$$\begin{aligned}\gamma_k^{n+1} &= \frac{E(\#success|X,\gamma^n)}{E(\#trials|X,\gamma^n)} \\ &= \gamma_k^n + \frac{\gamma_k^n(1-\gamma_k^n)\frac{dL(\gamma^n)}{d\gamma_k}}{E(\#trials|X,\gamma^n)}.\end{aligned}$$

If γ_k corresponds to a breakage probability or to the retention probability r_1, then $E(\#trials|X,\gamma^n) = n_{hyb}c$. If γ_k corresponds to the retention probability r_j where $j \neq 1$, then $E(\#trials|X,\gamma^n) = n_{hyb}c\theta_j^{n+1}$.

A.2 Three-point RH Vector Probabilities Expressed as Polynomial Functions of θ_1, θ_2, and r.

The parameter θ_k is the probability of at least one break occurring between markers k and $k+1$. The breakage probability is related to the scaled distance,d_k, by:

$$\theta_k = 1 - e^{-d_k} \tag{A-1}$$

Let b_{lm} represent the breakage state of an irradiated chromosome with respect to three marker loci. The index l equals zero if no breaks have occurred between the first and second marker, and one otherwise. The index m equals zero if no breaks have occurred between the second and third marker, and one otherwise. Let f_{ijk} represent the retention fingerprint of an irradiated chromosome for a given clone and with respect to three marker loci. The indices i, j and k equal one if the i^{th}, j^{th}, and k^{th} marker loci, respectively, is retained in the clone, and zero otherwise. The probability of observing a given retention fingerprint for a irradiated chromosome depends upon the underlying breakage state. Table 3 provides the conditional probabilities of all possible retention fingerprints given the underlying breakage state.

Table 3: The probabilities of observing retention fingerprints conditioned upon the underlying breakage state. The breakage state of an irradiated chromosome with respect to three marker loci is labeled b_{lm}. The indices l and m indicate the breakage states between the first and second markers, and the second and third markers, respectively. A breakage state index value of one indicates a break and is zero otherwise. The retention fingerprint of an irradiated chromosome for a given clone and with respect to three marker loci is labeled f_{ijk}. The indices i, j and k equal one if the i^{th}, j^{th}, and k^{th} marker loci, respectively, is retained in the clone, and zero otherwise. The last column provides values for the conditional probabilities, $P(f = f_{ijk}|b = b_{lm})$, of all possible retention fingerprints given the underlying breakage state. the parameters θ_1 and θ_2 are the breakage probabilities the first and second markers, and the second and third markers, respectively. The parameter r is the retention probability.

breakage state		retention fingerprint	
b_{lm}	$P(b = b_{lm})$	f_{ijk}	$P(f = f_{ijk}\|b = b_{lm})$
b_{00}	$(1-\theta_1)(1-\theta_2)$	f_{000}	$(1-r)$
		f_{111}	r
b_{01}	$(1-\theta_1)\theta_2$	f_{000}	$(1-r)^2$
		f_{001}	$(1-r)r$
		f_{110}	$r(1-r)$
		f_{111}	r^2
b_{10}	$\theta_1(1-\theta_2)$	f_{000}	$(1-r)^2$
		f_{011}	$(1-r)r$
		f_{100}	$r(1-r)$
		f_{111}	r^2
b_{11}	$\theta_1\theta_2$	f_{000}	$(1-r)^3$
		f_{001}	$r(1-r)^2$
		f_{010}	$r(1-r)^2$
		f_{100}	$r(1-r)^2$
		f_{011}	$r^2(1-r)$
		f_{110}	$r^2(1-r)$
		f_{101}	$r^2(1-r)$
		f_{111}	r^3

Let $X = (X_1, X_2, X_3)$ denote the RH vector. If no markers are present at locus k, then $X_k = 0$. If one or more markers are present, then $X_k = 1$. Let h_{ijk} represent the probability of observing RH vector $X = (i, j, k)$ in an RH experiment using haploid cells, so that:

$$h_{ijk} = \sum_{l=0}^{1} \sum_{m=0}^{1} P(f = f_{ijk}|b = b_{lm}) P(b = b_{lm}) \tag{A-2}$$

Specifically, the haploid probabilities are given by

$$\begin{aligned}
h_{000} &= 1 - r - r\theta_2 - r\theta_1 + r^2\theta_2 + r^2\theta_1 + r^2\theta_1\theta_2 - r^3\theta_1\theta_2 \\
h_{001} &= r\theta_2 - r^2\theta_2 - r^2\theta_1\theta_2 + r^3\theta_1\theta_2 \\
h_{010} &= r\theta_1\theta_2 - 2r^2\theta_1\theta_2 + r^3\theta_1\theta_2 \\
h_{011} &= r\theta_1 - r\theta_1\theta_2 - r^2\theta_1 + 2r^2\theta_1\theta_2 - r^3\theta_1\theta_2 \\
h_{100} &= r\theta_1 - r^2\theta_1 - r^2\theta_1\theta_2 + r^3\theta_1\theta_2 \\
h_{101} &= r^2\theta_1\theta_2 - r^3\theta_1\theta_2 \\
h_{110} &= r\theta_2 - r\theta_1\theta_2 - r^2\theta_2 + 2r^2\theta_1\theta_2 - r^3\theta_1\theta_2 \\
h_{111} &= r - r\theta_2 - r\theta_1 + r\theta_1\theta_2 + r^2\theta_2 + r^2\theta_1 - 2r^2\theta_1\theta_2 + r^3\theta_1\theta_2.
\end{aligned}$$

In the case of diploid cells we must define probabilities as follows. Let p_{ijk} represent the probability of observing $X = (i, j, k)$ in an RH experiment using diploid cells. These diploid probabilities are related to haploid probabilities as:

$$p_{ijk} = \sum_{l}\sum_{m}\sum_{n}\sum_{l'}\sum_{m'}\sum_{n'} h_{lmn}h_{l'm'n'}\phi_{i,l,l'}\phi_{j,m,m'}\phi_{k,n,n'} \tag{A-3}$$

where

$$\phi_{a,b,b'} = \begin{cases} 1 & \text{iff } a = \min(1, b + b') \\ 0 & \text{otherwise} \end{cases} \tag{A-4}$$

if measurements are made without error. The diploid probabilities are given by

$$
\begin{aligned}
p_{000} &= 1 - 2r - 2r\theta_2 - 2r\theta_1 + r^2 + 4r^2\theta_2 + r^2\theta_2^2 + 4r^2\theta_1 + 4r^2\theta_1\theta_2 + r^2\theta_1^2 - 2r^3\theta_2 \\
&\quad -2r^3\theta_2^2 - 2r^3\theta_1 - 8r^3\theta_1\theta_2 - 2r^3\theta_1\theta_2^2 - 2r^3\theta_1^2 - 2r^3\theta_1^2\theta_2 + r^4\theta_2^2 \\
&\quad +4r^4\theta_1\theta_2 + 4r^4\theta_1\theta_2^2 + r^4\theta_1^2 + 4r^4\theta_1^2\theta_2 + r^4\theta_1^2\theta_2^2 - 2r^5\theta_1\theta_2^2 \\
&\quad -2r^5\theta_1^2\theta_2 - 2r^5\theta_1^2\theta_2^2 + r^6\theta_1^2\theta_2^2 \\
p_{001} &= 2r\theta_2 - 4r^2\theta_2 - r^2\theta_2^2 - 4r^2\theta_1\theta_2 + 2r^3\theta_2 + 2r^3\theta_2^2 + 8r^3\theta_1\theta_2 + 2r^3\theta_1\theta_2^2 \\
&\quad +2r^3\theta_1^2\theta_2 - r^4\theta_2^2 - 4r^4\theta_1\theta_2 - 4r^4\theta_1\theta_2^2 - 4r^4\theta_1^2\theta_2 - r^4\theta_1^2\theta_2^2 \\
&\quad +2r^5\theta_1\theta_2^2 + 2r^5\theta_1^2\theta_2 + 2r^5\theta_1^2\theta_2^2 - r^6\theta_1^2\theta_2^2 \\
p_{010} &= 2r\theta_1\theta_2 - 6r^2\theta_1\theta_2 - 2r^2\theta_1\theta_2^2 - 2r^2\theta_1^2\theta_2 + r^2\theta_1^2\theta_2^2 + 6r^3\theta_1\theta_2 \\
&\quad +6r^3\theta_1\theta_2^2 + 6r^3\theta_1^2\theta_2 - 2r^3\theta_1^2\theta_2^2 - 2r^4\theta_1\theta_2 - 6r^4\theta_1\theta_2^2 \\
&\quad -6r^4\theta_1^2\theta_2 + 2r^5\theta_1\theta_2^2 + 2r^5\theta_1^2\theta_2 + 2r^5\theta_1^2\theta_2^2 - r^6\theta_1^2\theta_2^2 \\
p_{011} &= 2r\theta_1 - 2r\theta_1\theta_2 - 4r^2\theta_1 + 6r^2\theta_1\theta_2 + 2r^2\theta_1\theta_2^2 - r^2\theta_1^2 + 2r^2\theta_1^2\theta_2 \\
&\quad -r^2\theta_1^2\theta_2^2 + 2r^3\theta_1 - 6r^3\theta_1\theta_2 - 6r^3\theta_1\theta_2^2 + 2r^3\theta_1^2 - 6r^3\theta_1^2\theta_2 \\
&\quad +2r^3\theta_1^2\theta_2^2 + 2r^4\theta_1\theta_2 + 6r^4\theta_1\theta_2^2 - r^4\theta_1^2 + 6r^4\theta_1^2\theta_2 - 2r^5\theta_1\theta_2^2 \\
&\quad -2r^5\theta_1^2\theta_2 - 2r^5\theta_1^2\theta_2^2 + r^6\theta_1^2\theta_2^2
\end{aligned}
$$

$$
\begin{aligned}
p_{100} &= 2r\theta_1 - 4r^2\theta_1 - 4r^2\theta_1\theta_2 - r^2\theta_1^2 + 2r^3\theta_1 + 8r^3\theta_1\theta_2 + 2r^3\theta_1\theta_2^2 + 2r^3\theta_1^2 \\
&\quad +2r^3\theta_1^2\theta_2 - 4r^4\theta_1\theta_2 - 4r^4\theta_1\theta_2^2 - r^4\theta_1^2 - 4r^4\theta_1^2\theta_2 - r^4\theta_1^2\theta_2^2 \\
&\quad +2r^5\theta_1\theta_2^2 + 2r^5\theta_1^2\theta_2 + 2r^5\theta_1^2\theta_2^2 - r^6\theta_1^2\theta_2^2 \\
p_{101} &= 4r^2\theta_1\theta_2 - 8r^3\theta_1\theta_2 - 2r^3\theta_1\theta_2^2 - 2r^3\theta_1^2\theta_2 + 4r^4\theta_1\theta_2 + 4r^4\theta_1\theta_2^2 \\
&\quad +4r^4\theta_1^2\theta_2 + r^4\theta_1^2\theta_2^2 - 2r^5\theta_1\theta_2^2 - 2r^5\theta_1^2\theta_2 - 2r^5\theta_1^2\theta_2^2 \\
&\quad +r^6\theta_1^2\theta_2^2 \\
p_{110} &= 2r\theta_2 - 2r\theta_1\theta_2 - 4r^2\theta_2 - r^2\theta_2^2 + 6r^2\theta_1\theta_2 + 2r^2\theta_1\theta_2^2 + 2r^2\theta_1^2\theta_2 \\
&\quad -r^2\theta_1^2\theta_2^2 + 2r^3\theta_2 + 2r^3\theta_2^2 - 6r^3\theta_1\theta_2 - 6r^3\theta_1\theta_2^2 - 6r^3\theta_1^2\theta_2 \\
&\quad +2r^3\theta_1^2\theta_2^2 - r^4\theta_2^2 + 2r^4\theta_1\theta_2 + 6r^4\theta_1\theta_2^2 + 6r^4\theta_1^2\theta_2 - 2r^5\theta_1\theta_2^2 \\
&\quad -2r^5\theta_1^2\theta_2 - 2r^5\theta_1^2\theta_2^2 + r^6\theta_1^2\theta_2^2 \\
p_{111} &= 2r - 2r\theta_2 - 2r\theta_1 + 2r\theta_1\theta_2 - r^2 + 4r^2\theta_2 + r^2\theta_2^2 + 4r^2\theta_1 - 6r^2\theta_1\theta_2 \\
&\quad -2r^2\theta_1\theta_2^2 + r^2\theta_1^2 - 2r^2\theta_1^2\theta_2 + r^2\theta_1^2\theta_2^2 - 2r^3\theta_2 - 2r^3\theta_2^2 - 2r^3\theta_1 \\
&\quad +6r^3\theta_1\theta_2 + 6r^3\theta_1\theta_2^2 - 2r^3\theta_1^2 + 6r^3\theta_1^2\theta_2 - 2r^3\theta_1^2\theta_2^2 + r^4\theta_2^2 \\
&\quad -2r^4\theta_1\theta_2 - 6r^4\theta_1\theta_2^2 + r^4\theta_1^2 - 6r^4\theta_1^2\theta_2 + 2r^5\theta_1\theta_2^2 + 2r^5\theta_1^2\theta_2 \\
&\quad +2r^5\theta_1^2\theta_2^2 - r^6\theta_1^2\theta_2^2.
\end{aligned}
$$

A.3 Radiation Hybrid Panel Data-set W-BTA5

Table 4: Marker names for data-set W-BTA5. Data-set W-BTA5 is comprised of a published RH map for *Bos taurus* Autosome 5 (BTA5) (Womack et al., 1997) and an RH panel consisting of 90 bovine-hamster hybrid cell lines typed for 85 markers on BTA5 (Womack et al., 1997).The 90 hybrid cell lines in our example also were typed for four positional candidate genes for meat tenderness which were developed in Hansen (2003). For each dataset, we label the markers $M_1, M_2, \ldots, M_{85}$ in accordance with the order in which they appear in that data-set's candidate RH Map. Markers labeled M_{G1}, M_{G2}, M_{G3}, and M_{G4} correspond to the four positional candidate genes.

label	Name	label	Name	label	Name	label	Name	label	Name
M_1	BMS1095	M_{19}	EST0110	M_{37}	RM500	M_{55}	BM1248	M_{73}	NOL1
M_2	ILSTS042	M_{20}	PDES1B	M_{38}	BR2936	M_{56}	BM8230	M_{74}	REA
M_3	BM6026	M_{21}	ILSTS022	M_{39}	SILV	M_{57}	EST1034	M_{75}	BM2830
M_4	MYF5	M_{22}	BMC1009	M_{40}	RDH5	M_{58}	M6PR	M_{76}	MAF48
M_5	BP1	M_{23}	K03534	M_{41}	AGLA254	M_{59}	OLR1	M_{77}	BM733
M_6	BMS610	M_{24}	EST0012	M_{42}	K02818	M_{60}	BM315	M_{78}	EST0260
M_7	NTS	M_{25}	SP1	M_{43}	EST1396	M_{61}	BMS1658	M_{79}	ACO2
M_8	BL23	M_{26}	TEGT	M_{44}	ETH10	M_{62}	GUCY2C	M_{80}	IDVGA9
M_9	BTG1	M_{27}	CSSM034	M_{45}	CDK2	M_{63}	URB052	M_{81}	ETH152
M_{10}	DCN	M_{28}	COL2A1	M_{46}	ILSTS066	M_{64}	MAGP2	M_{82}	BMS597
M_{11}	KERA	M_{29}	U63110	M_{47}	PHC	M_{65}	TNFRSF1A	M_{83}	URB060
M_{12}	LUM	M_{30}	K-ALPHA-1	M_{48}	IGF1	M_{66}	ETH2	M_{84}	BM8126
M_{13}	UBE2N	M_{31}	LYZ	M_{49}	TIMP3	M_{67}	EST1389	M_{85}	ACR
M_{14}	EST1320	M_{32}	AF016589	M_{50}	MB	M_{68}	SCNN1A	M_{G1}	WNT10B
M_{15}	BMS1315	M_{33}	BL4	M_{51}	BM1819	M_{69}	BMS772	M_{G2}	MMP19
M_{16}	OARFCB5	M_{34}	BL37	M_{52}	TST	M_{70}	CD9	M_{G3}	MYF5TX
M_{17}	X75935	M_{35}	IFNG	M_{53}	EST0328	M_{71}	CCND2	M_{G4}	WIF1
M_{18}	EST0062	M_{36}	EST0373	M_{54}	EST1290	M_{72}	EST0179		

The panel data for Data-set W-BTA5 is listed below. Each row consists of an RH vector for a radiation hybrid clone. The RH vectors are scored in the following order: $M_1, M_2, \ldots, M_{85}$, $M_{G1}, M_{G2}, M_{G3}, M_{G4}$.

```
hybrid
  1     00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
  2     00000000000000000?00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
  3     00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
  4     00000000000000000?00000000000000000010000000000000000000000000000000000000000000000000000000000000000000000000000000
  5     1111110000?00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000010
  6     00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
  7     00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
  8     000000000000000000000000000000000000000?111?11100000000000000000000000?00000000000000100
  9     00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
 10     00000000000000000000000001000000000011111111100?00000000000000000000000000000000000000100
 11     00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
 12     00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
 13     0000000000?00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000010000
 14     11010100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000111110000
 15     11000000000000000000000000000000000000000000000000000000000000000000000000000000000?00000000000000000000000000
 16     00000000000000000?00000000000000001010000001000000000000000000000000000000000000111111100000000
 17     00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
 18     00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
 19     00000000000000000000000000000000000000010000000000000000000000000000000000000000000000000000000000000000000000000000
 20     00010110000000000000000000000010100000000100000000000000000000100000000000000011000000?0
 21     11100011111111111101111111111111011101111101111111010000000000000100001000111111111111101
 22     00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
 23     00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
 24     00000011111111111111111?11111100000001000000000000000000000000000000000000000000000000000000000000000000000000000?000
 25     00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
 26     00000000000000000000000000000000000000000000000000000000?00000000000000000000000000000000000000000000000000000000000
 27     11111111111111111111111101111111111111111111111111111111111111111111111111111111111111111111111111111111111110111
 28     00000000000011111111111111110010000000000000000000111110?1100000000?11111111111111111111111111110000
 29     00000000000000000000000?00000000000000?00000011111111111111111111?11111111111111111111110000
 30     00000000000000000000000000000000000000001000000?00000000000000000000000000000000000000000000000000000
 31     00000000000000000000000000000010000000001000000000000000000000000000000000000000000000000000011110000
 32     00000000000000000000000000000000000011101111111111111101111111111111111111111111111110100
 33     00000000000000000000000000000000000000?00000000000000000000000000000000000000000000000000000000000000
 34     11111111100011111111111111111111111001000000000000000000000?00000000000000000000000000000010
 35     10000000000000000000000010000000000000000000000000000000000000000000000000000000010011000000000
 36     00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000010
 37     00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000111100000
 38     01111?1110011111?11111111111100100010000000000000000000000011111111111111111111?010
 39     00000000000000000000000000000000000000000000000000000000000000000000000000?00000000000000000
 40     00011111111111111111111111100000000000000000000000000?00000000000000000000000000000000000010
 41     00000000000000000000000000000000000000000000000000000000000000001111111111111111111111111110000
 42     10000000000000000000000000000000000000000000000000000000000000000000000000000000000000000111110000
 43     00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
 44     00000000000000000000000000000010110111111111110000001000000000000000000000000000011111110100
 45     11111111111111111111111111111100000000000000000000000000000000011111111111111100000001010
 46     11111111111111111111111111111100000000000000000000000000000000000?00000000000001010
 47     11111111111000000001111111111111111111111111111111000000000000000000000000000000011111
 48     0000000000?00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
 49     00000000000000000000000000000001000011111111100000000001000000001111111111000000000000100
 50     000000000000?00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
 51     11111111111111111111111110000111111111111111111000000000000000000000000000000000000000111
 52     100000?00000000000000000000000000000000000000000000000000000000000?00000000000000000111110000
 53     10000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
 54     00000000000000000000000000000001001001100010011?00010001000000000000000?00000000000000000
 55     00000000000000000000000000000000000000000000000?00111111101111111111111111111111111111110000
 56     1000000000000?00000000000000000000000000000000000000000000000000000000000000000000000000000111110000
 57     000000000000?00000000000000000000011011011110000000000000000000000000000000000000000?01
 58     00000001111111111111111111111110000000000000000000000000000000000000000000111111111110000
 59     00000000000000000000000000000000000010000000000000000000000000000000000000000000011111110000
 60     00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000010000
 61     00000000000000000000000000000000000000000000000000000000000000000000000000000000000?00000000000
 62     00000100000?01011111111111111111000000000000000000000011000001100100001000100000000001000
 63     0000000000?00000000000000000000000000000000000000000000000000000000000000000?00000000000000000
 64     00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
 65     00000000000000?00000000000000000000000000000000?00000000000000000000000000000000111110000
 66     00000000000001000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
 67     00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
 68     11111111111111111111111111111111111?11111111111111111100111111111111111111111111111110110
 69     11111110?000?11100001111110001011001000000000000010??1100000000000001?11101111000010000
 70     00111101111110000000000000000000000000000000000000000000000000000000000??00000000000000010
 71     00000000000000000000000000000?00000000000000000011110000000000000000000000000000000000000
 72     00001100000000000000000000000000000000000011111111100000000000000000?00000000000000000000?0
 73     00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000010000
```

```
74    11111111111000000001111111111111111111111111111100000000000000000000000000000011111?111
75    11111111111000000001111111001111101111111111111110000000000000000000000000000000111110110
76    11100000000000000000000000000000000000000000000000000000001111000000000000000000000000000000
77    00000000000000000000000000000000000000000000000000000000000000000000000?00000000000000000000
78    00000000000000000000000000000000000000010000000000?00000000000000000000000000000000000000
79    1111111111O?00111?10111?11111011111?110111?11111111111000100001000000000100011000010110
80    0011111111O?00111?00111111110000000000000000000000000000000000000000000000000000000000000000
81    11111110000000011111111111111111111?111111111111111111111111111111111111111111111111111111
82    00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
83    00000011111111111111111111100000000000000000000000000000000000?00000000001111111111110000
84    00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
85    00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
86    00010000000000000000000000000000000000000000000000000000000000000000000000000000000000010010
87    00110000000000000000000000000000000?00000000000000000000000000000000000000000000000010010
88    00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
89    00011100010000000000000000000010001111111111111100000000000000000000000000000011110000111
90    1111111000?00001111111111111111111111011111111111111111111111011111111111111111111111111110111
```