File S1
TRD mapping algorithms.

Genotype probabilities: The crossing design (Fig.1) allows us to infer the population of origin of alleles at fully informative marker loci, and the origin of some alleles at partially informative loci. Hence, we can make inferences about the population of origin of the alleles at the remaining marker loci, i.e. we can infer haplotypes. Let us consider a marker with alleles abc and d where alleles a and c are from population 0 and alleles b and d from population 1. Instead of the name of the allele (a or b) we can use the population of origin as the "phase" of the maternal F_1 allele, i.e. 0 if allele a was inherited from the female F_1 parent, and 1 if allele b was inherited from the female F_1 parent. Similarly, we can write 0 if allele c was inherited from the male F_1 parent, and 1 if allele d was inherited from the male F_1 parent. Thus, we can re-write the four possible genotypes ac, ad, bc, and bd as "phases" 00, 01, 10, and 11, respectively. These haplotype phases correspond to genotypes at fully genotyped markers, but their advantage as compared to genotypes is that they are comparable between loci, so that they can be used to infer genotypes at pseudomarkers. For example, at another locus alleles b and c may originate from population 0, so that genotype bc is assigned haplotype phase 00.

Haplotype phases of flanking markers are generally used to infer haplotype phases of pseudomarkers in QTL mapping. Considering the phase of just the maternal allele, if both flanking markers are in phase 0, the pseudomarker is more likely to be in phase 0 than in phase 1. More precisely, if φ is the (unknown) phase of the pseudomarker, the probability that the pseudomarker is in phase 0 is:

Eq. S1 $$p(\varphi = 0) = \frac{p(\varphi = 0|\varphi_l)p(\varphi_r|\varphi = 0)}{p(\varphi_l|\varphi_r)}$$

where the phases of the left and right flanking markers are indicated with subscripts l and r. The alternative (φ=1) has probability 1-p(φ=0). The conditional probabilities in Eq. S1 depend on the distances between the pseudomarker and its flanking markers. Let us express the distance between the pseudomarker and the left flanking marker as a recombination fraction, d_l. Then, the probability of the pseudomarker phases is given by Haldane's map function:

Eq. S2 $$p(\varphi|\varphi_l) = \begin{cases} 0.5\exp(-2d_l) & \varphi \neq \varphi_l \\ 1-0.5\exp(-2d_l) & \varphi = \varphi_l \end{cases}$$

In the case a flanking marker is lacking (for example if we consider the first or last marker on a chromosome) the distance d on that side is infinite so that $p(\varphi|\varphi_l) = 0.5$ and the probability of the phase of the pseudomarker is influenced only by the remaining flanking marker. Thus, we can infer haplotype phases (and hence transmission ratios) using flanking markers.

The above method to infer pseudomarker phases is widely used, but requires modification for the present purpose, where we consider an experimental cross between natural, outcrossing populations. Consider for example a locus where allele a originates

from population 0 and allele b from population 1, and where both F_1 parents are have genotype ab, so that the possible F_2 genotypes are aa, ab, and bb. Genotypes aa and ab represent phases 00 and 11, respectively, while genotype ab is either 01 or 10. Hence, an individual with genotype ab at this locus provides no phase information (both alleles are equally likely to stem from both populations) even though it clearly provides information about transmission ratios (as it is neither 00 nor 11). In order to employ the information about transmission ratios provided by partly informative markers such as the example above above, we must extend Haldane's mapping function to incorporate both maternal and paternal alleles simultaneously, and to more than two flanking markers.

The extension of the mapping function to incorporate both alleles is rather straightforward. Let r denote the recombination rate on a very short distance, for example $r = 0.01$ per centimorgan (cM). Then, the probability that two flanking markers one cM apart are in phases 00 and 10 would equal r. These markers are in phases 00 and 11 only if recombination occurs twice, that is with probability r^2. We can conveniently write all the possible transitions between the 4 phases in the 4x4 transition matrix Q:

$$Q = \begin{matrix} -2r - r^2 & r & r & r^2 \\ r & -2r - r^2 & r^2 & r \\ r & r^2 & -2r - r^2 & r \\ r^2 & r & r & -2r - r^2 \end{matrix}$$

where rows and columns refer to 00, 01, 10, and 11, respectively. Entries on the off-diagonal are chosen so that rows sum to 0. We can also write the probability of the phases as a matrix, corresponding to the row and columns order of Q (i.e. 00, 01, 10, and 11). For example, at a genotyped marker the phase may be 00, which can be written:

$$P_A = \begin{matrix} 1 \\ 0 \\ 0 \\ 0 \end{matrix}$$

The probabilities of the phases at a flanking marker B at distance d_{AB} are then given by the Chapman-Kolmogorov equation:

Eq. S3 $$P_{B|A} = P_A \exp(d_{AB} Q)$$

where the exponent is a matrix exponent. For example, for P_A above and d_{AB}=5cM, $P_{B|A}$=[0.9066; 0.0453; 0.0453; 0.0027]. Like Haldane's mapping function, equation S3 accounts for multiple recombination events, assuming that these are independent, random events. In other words, equation S3 is Haldane's mapping function applied to both maternal and paternal alleles simultaneously, and written in matrix representation for mathematical convenience.

To employ the information provided by partly informative markers we must also extend phase inference to multiple flanking loci. Consider for example a pseudomarker flanked by a partly informative marker, which in turn is flanked at close distance by a

fully informative marker. The partly informative marker itself provides little information to infer the phase of a pseudomarker next to it, while (due to close linkage) we could be rather certain that it is in the same phase as the fully informative marker. To make use of all the information provided by the genotypes, we must therefore use all fully and partially informative markers to infer phase probabilities at any (pseudo)marker locus.

Just as expressed in equation S1 for a single flanking marker, the phase probabilities of a (pseudo)marker are determined by two components: all the markers to the left, and all the markers to the right. Let us denote the phase probabilities of the *i*-th marker given the markers to the left as $P_{i|l}$, and the phase probabilities given the markers to the right as $P_{i|r}$. We calculate the phase probabilities as:

Eq.S4 $$P_i = \frac{P_{i|l} P_{i|r}}{\sum P_{i|l} P_{i|r}}$$

where the product is element-wise, and the summation over phases. The divisor, just like in equation S1, assures that P sums to unity.

In order to obtain $P_{i|l}$ we calculate sequentially, starting from the leftmost marker on the chromosome and proceeding to the right (using equation A3) $P_{i|l} = P_{i-1} \exp(d_{i(i-1)}Q)$. (For the first marker on the chromosome P_{i-1} =[0.25 0.25 0.25 0.25], reflecting that all phases are equally likely a priori.) If the *i*-th marker is (partly) informative, we can set some elements of $P_{i|l}$ to zero, and re-scale the remaining probabilities so that $P_{i|l}$ sums to unity. Thus, $P_{i|l}$ can be regarded as the phase probabilities of the *i*-th marker if these were determined only by the markers to the left of it. Starting from the rightmost marker, we can similarly calculate $P_{i|r}$ as the phase probabilities of the *i*-th marker if these were only determined by the markers to the right. Finally, we use equation S5 to calculate the phase probabilities P for every marker.

Likelihood maximization As explained above, (pseudo)marker phases can be analyzed for TRD as genotypes. At fully informative markers, phases are known with certainty, but at partly informative markers and pseudomarkers phases can only be assigned probabilities. This has implications for calculating the likelihood of genotype frequencies (Eq. S1): At partially informative loci the likelihoods L_f under different hypotheses depend on the assignment of phases. Thus, we should still maximize the likelihood, but the likelihood will now consist of two components: L_f, and a component representing the likelihoods of the phase assignments. Let $L_{\varphi,j}$ denote the log-likelihood of the phases of the *j*-th F_2 individual. If the phase is known with certainty (e.g. at a fully informative marker) this likelihood will be $L_{\varphi,j}$ =log(1)=0. If the phase is not certain, but for example P=[0.9066; 0.0453; 0.0453; 0.0027] then $L_{\varphi,j}$ = log 0.0453 if the individual is assigned phase 01. Maximizing the likelihood now involves choosing the phases for the *n* F_2 individuals in such a way that it maximizes the likelihood:

Eq. S5 $$L = L_f + \sum_{j=1}^{n} L_\varphi$$

It should be noted that the number of unknown phases is a property of the data that is independent of the hypothesis that is being evaluated. Therefore unknown phases do not

affect the difference in the numbers of estimated parameters between hypotheses, i.e. the number of degrees of freedom of the χ^2-distribution used to compare likelihoods.

Maximizing the likelihood is not an easy task (except at fully informative markers). If all individuals are assigned the phase that is most likely, L_φ is maximized, but this may lead to genotype frequencies that render L_f sub-optimal. In an F_2 of n individuals with k possible phases, there are k^n possible phase assignments across individuals. It is clear that for realistic n, the number of assignments is truly large, and exhaustive search for the assignment that maximizes L is prohibitive. Therefore, we use an iterative algorithm to attempt to find the phase assignment that maximizes the likelihood.

1. The algorithm used to maximize the likelihood starts by assigning every individual a plausible phase. For every individual, the initial probability of the i-th phase was calculated as p_i*P, where p_i is the expected frequency of the i-th phase (Eq. 1) and P the probability of this phase according to the flanking markers (Eq. S4). The individual was then assigned the most probable phase. (That is the phase suggested by the flanking markers (i.e. suggested by P), except when that genotype is not expected to be observed (i.e. $p_i = 0$) based on the TRD hypothesis being evaluated.) The likelihoods (L_f and L_φ) of the initial assignment are then calculated using equation S5.

2. For every individual, and for every possible alternative phase, it is calculated how the likelihood (both L_f and L_φ) would change. For example, if an individual is currently assigned phase 10, there are three alternatives, 00, 01, and 11, each of which may have a different effect on L_f as well as L_φ.

3. The single individual and alternative phase is selected that results in the greatest increase in likelihood L.

Steps 2 and 3 are repeated until no further improvement of the likelihood can be achieved. It should perhaps be emphasized that in every iteration, only one individual is assigned a different haplotype in step 3.