The Set-up:

Suppose we have information at 10 SNPs for two homozygous parental strains. Suppose the haplotypes for the parents are known to be as follows:

Parent 1: AAAAAAAAAA. Denote this by P1
Parent 2: TTTTTTTTTT. Denote this by P2.

Suppose we create an RIL from those two strains, run our favourite base-calling algorithm, and get the results shown in Table 1. Each row corresponds to a SNP (1st row is the leftmost SNP; last row is the rightmost SNP). Elements show the probability that each SNP is of each of four possible types.

Table 1: Result of base-calling algorithm for the RIL (i.e. the 'prior')

A	C	G	T
0.1	0.1	0.1	0.7
0.1	0.1	0.1	0.7
0.25	0.25	0.25	0.25
0.1	0.1	0.1	0.7
0.1	0.1	0.1	0.7
0.7	0.1	0.1	0.1
0.7	0.1	0.1	0.1
0.7	0.1	0.1	0.1
0.7	0.1	0.1	0.1
0.7	0.1	0.1	0.1

In summary, this prior suggests that SNPs 1,2,4, and 5 are inherited from P2, while SNPs 6,7,8,9 and 10 are inherited from P1. The identity of SNP 3 is unknown. Common sense suggests that SNP 3 should also most likely be inherited from P2 (and that's what this algorithm is designed to tell us).

The Method:

We treat the results in table 1 as a 'prior' distribution for the haplotype, *H*, of the RIL at each locus. We then add a model for recombination to produce a posterior for the H. We do this by calculating *M*, the minimum number of recombinations necessary to construct H from the two parental lines. We then model the number of recombinations as a Poisson random variable. In other words, we add a piece to the model of the form:
P(H | P1, P1)=P(H | M)=Pr(X=M)
Where *X* is a Poisson(*Rho*) random variable and *Rho* is a parameter chosen to reflect our belief about the recombination rate.

NB. This is not the 'correct' model, in some sense. What we would really need to calculate is the prob. that we get an odd number of recombinations between SNPs at which the inherited parental line appears to change, etc. These details can be added later, but the model we give here is a good approximation when recombination is rare.

We then construct an MCMC algorithm to explore the space of possible haplotypes *H,* conditional on *P1, P2,* and *M.*
In Tables 2-4, we present three, example analyses, for differing recombination parameters *Rh*o. Again, each row corresponds to a SNP. Each element shows the posterior prob. that the given SNP is of one of the four possible types.

Table 2: Posterior for RIL with Poisson recombination rate of Rho=1e-1 (a small rho)

A	C	G	T
0.01260500	0.00000000	0.00000000	0.98739000
0.0070341	0.0000000	0.0000000	0.9929700
0.0231460	0.0000000	0.0000000	0.9768500
0.03915800	0.00000000	0.00000000	0.96084000
0.14673000	0.00000000	0.00000000	0.85327000
0.88695000	0.00000000	0.00000000	0.11305000
0.982970	0.000000	0.000000	0.017034
0.99545000	0.00000000	0.00000000	0.00454910
0.9936300	0.0000000	0.0000000	0.0063727
0.986090	0.000000	0.000000	0.013908

Table 3: Posterior for RIL with Poisson recombination rate of 1 (a higher rho)

A	C	G	T
0.07044100	0.00000000	0.00000000	0.92956000
0.0604610	0.0000000	0.0000000	0.9395400
0.1738300	0.0000000	0.0000000	0.8261700
0.05845700	0.00000000	0.00000000	0.94154000
0.13158000	0.00000000	0.00000000	0.86842000
0.88407000	0.00000000	0.00000000	0.11593000
0.966390	0.000000	0.000000	0.033607
0.96898000	0.00000000	0.00000000	0.03102200
0.9706200	0.0000000	0.0000000	0.0293790
0.932650	0.000000	0.000000	0.067355

Table 4: Posterior for RIL with Poisson recombination rate of 1e-5 (a tiny rho)

A	C	G	T
0.99928000	0.00000000	0.00000000	0.00072144
0.9994000	0.0000000	0.0000000	0.0006012
0.9997000	0.0000000	0.0000000	0.0003006
0.99972000	0.00000000	0.00000000	0.00028056
0.99982000	0.00000000	0.00000000	0.00018036
0.99998000	0.00000000	0.00000000	0.00002004
1.000000	0.000000	0.000000	0.000000
0.99994000	0.00000000	0.00000000	0.00006012
1.0000000	0.0000000	0.0000000	0.0000000
1.000000	0.000000	0.000000	0.000000

The algorithm performs as expected (row 3 of the tables is the key thing to look at): When Rho is small, the method predicts that SNP 3 is inherited from P2 (and gives less certainty for SNPs near the obvious recombination location that for those far from it. When Rho is somewhat larger, the degree of certainty about the identity of SNP 3 decreases (since we could now explain it as being inherited from P1 with two flanking recombinations). When Rho is tiny, and we are forced to construct a solution with no recombinations, the posterior picks one of the parents and sets the RIL equal to that parent. [In fact, if I wrote an algorithm that mixed correctly it would pick the two parents with equal probability, whereas the current version just picks one parent and gets stuck there.]

QED, I think, as far as a proof of principle goes. Obviously there are many more layers that could be added.