Users Guide for New BC_sF_t Tools for R/qtl

Laura M. Shannon Brian S. Yandell Karl Broman

29 January 2013

Introduction

Historically QTL mapping studies have employed a variety of crossing schemes including: backcrosses [1], sib-mating [2], selfing [3], RI lines [4], and generations of random mating within mapping populations [5]. Different cross designs offer different advantages. Backcrossing allows for the isolation of limited regions of the donor parent genome in an otherwise recurrent parent background. Selfing and sib-mating in an intercross provide the opportunity to examine all genotype combinations and observe dominance. RI lines allow for multiple phenotype measures on a single line. Random mating increases recombination frequency. In order to use a combination of these cross types and access their various benefits, a more flexible analysis approach is needed.

This guide develops methods to analyze advanced backcrosses and lines created by repeated selfing by extending features of R/qtl [6]. Interval mapping requires estimating the probable genotype of a putative QTL based on the neighboring markers [7]. The probability that a loci between two genotyped markers is of a given genotype depends on the recombination history of the population, which depends on the type of cross. In Figure 1 we have two markers, A and B, each with two possible allele genotypes, capital or lower case. Let us assume that markers A and B are spaced such that double crossovers in a single generation are unlikely. Let Q be the position of the putative QTL between A and B. When the observed genotypes at A and B are both homozygous capital in an F_2 or BC_1 the genotype at Q is most likely homozygous capital (Figure 1 part B). However, in an F_3, Q might be heterozygous or homozygous for either parent (Figure 1 part C). Similarly, in a BC_2, Q might be homozygous capital or heterozygous, when the observered genotype is AB/AB. Each generation brings an additional opportunity for crossing over

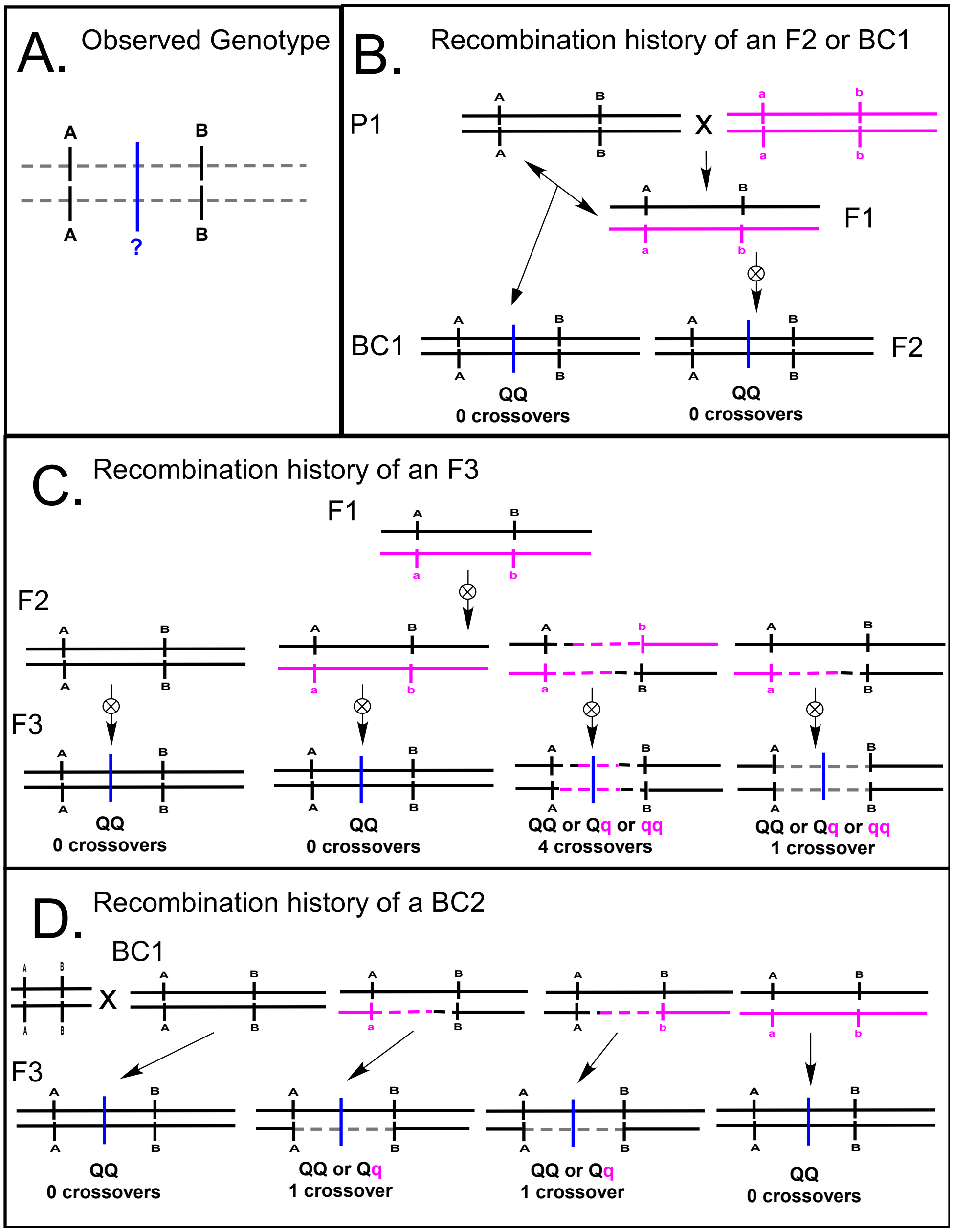

Figure 1: An illustration of QTL geneotype inference in populations created through different crossing structures. All images are of a chromosome section including 2 markers (A and B) and a putative QTL (Q). Chromosomal segments are pink when they share a genotype with the lower case parent and black when they share a genotype with the capital parent. Regions where the genotype cannot be observed are dashed. Regions where the genotype is unknown are gray.

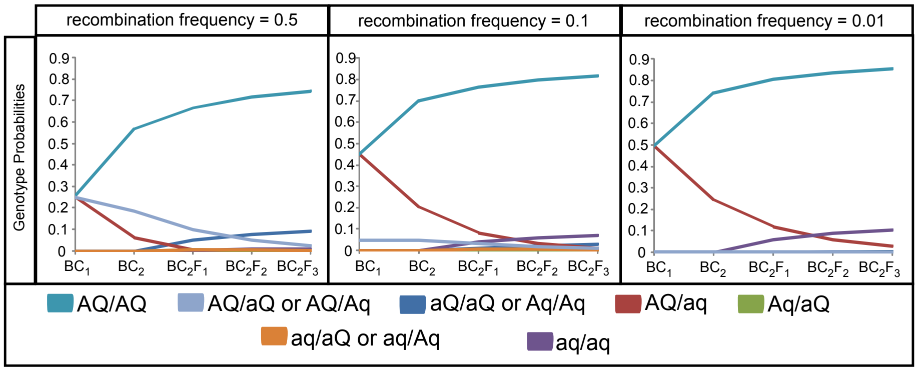

Figure 2: The probability that a pair of loci is of a given genotype based on the transition probabilities from the known genotype of marker one (As) to the unknown genotype of the putative QTL (Bs).

within the interval, increasing the likelihood that Q will not share a genotype with A and B. This has real consequences when determining genotype probabilities (Figure 2).

Genetic map creation is also based on recombination history. Assuming an F_2 or BC_1 and sufficiently close markers to make double crossovers in a single generation improbable, individuals which are homozygous for the recurrent parent allele at two adjacent markers exhibit no recombination events between those two markers (figure 1 part B). However, in an F_3 the state of being homozygous for the recurrent parent allele at neighboring markers can be accomplished with 0, 1, or 4 recombination events (figure 1 part C). If an F_3 is treated as an F_2, an individual with 2 adjacent markers homozygous for the same parent will be counted as having undergone 0 recombination events. However, the actual expected number of recombination events for the described individual is:

$$r^4 + \frac{r(1-r)}{8} \approx r/8 \ ,$$

where r is the recombination frequency.

Therefore, treating an F_3 as an F_2 would artificially shorten the map length (Figure 3). The number of recombination events between two markers depends on the recombination frequency and cross history, and the number of recombination events in agregate determines the map length.

In this guide we present our method for analyzing mapping populations with advanced cross histories while avoiding the pitfalls described above. Specifically, we address populations resulting from repeated backcrossing (BC_s), repeated selfing (F_t), and backcrossing

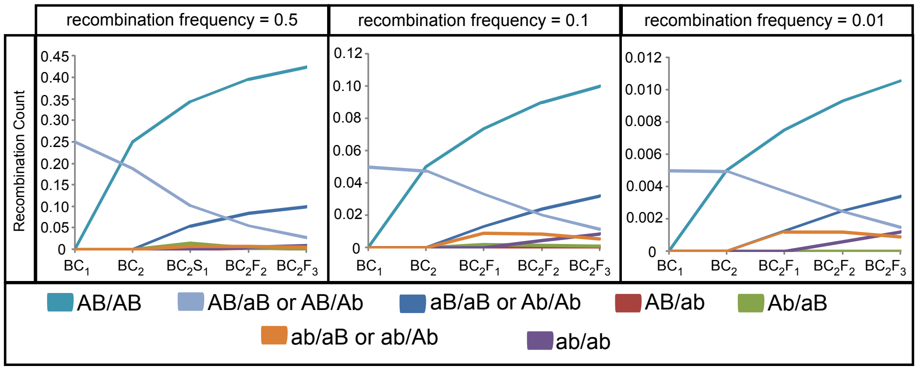

Figure 3: Estimated recombination counts between pairs of markers with observed genotypes.

followed by selfing (BC_sF_t). The first section is a tutorial on how to use the new tools. The second section lays out the way we derived the equations for probabilities and recombination counts, which allow for the analysis of advanced cross histories. The third section contains a technical description of the modifications to the code of the previous release of R/qtl [6].

Tutorial

These changes to R/qtl are mostly internal. The one thing that does change for the user is reading in the data. Data can be read in using *read.cross()* as for all other crosses. We will use the listeria sample data from R/qtl below.

```
> library(qtl)
> listeria.bc2s3<-read.cross(format="csv",
+   file=system.file(file.path("sampledata", "listeria.csv"), package = "qtl"),
+       BC.gen=2, F.gen=3)
```

```
 --Read the following data:
         120  individuals
         133  markers
         1  phenotypes
 --Cross type: bcsft
```

Here's another way to convert a cross. Suppose the R/qtl hyper data was really a BC_3 (or BC_3F_0). You can convert it as follows:

```
> data(hyper)
> hyper3 <- convert2bcsft(hyper, BC.gen = 3)
```

```
 --Estimating genetic map
```

We will briefly highlight the difference in results between crosses analyzed using the traditional program and those analyzed using our new tools. However, we do not discuss the entire process of QTL mapping. Please refer to the tutorials available through rqtl.org or A Guide to QTL Mapping with R/qtl by Karl Broman [8] for guidence on complete analysis.

First we compare the maps for the listeria data set (figure 4).

```
> listeria.f2<-read.cross(format="csv",
+    file=system.file(file.path("sampledata", "listeria.csv"), package = "qtl"))
```

```
 --Read the following data:
         120  individuals
         133  markers
         1  phenotypes
 --Cross type: f2
```

```
> map.bc2s3 <- est.map(listeria.bc2s3)
> map.f2<-est.map(listeria.f2)
```

Now, we will compare the maps for the hyper data (figure 5).

```
> map.bc1 <- est.map(hyper)
> map.bc3<-est.map(hyper3)
> plot(map.f2, map.bc2s3, label=FALSE, main="")
```

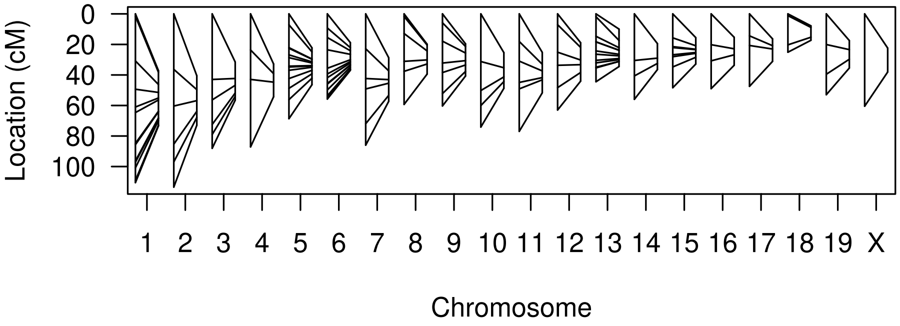

Figure 4: A comaprison of genetic maps of the listeria data set analyzed as though it were a F_2 (left) and as though it were a BC_2F_3 (right).

```
> plot(map.bc1, map.bc3, label=FALSE, main="")
```

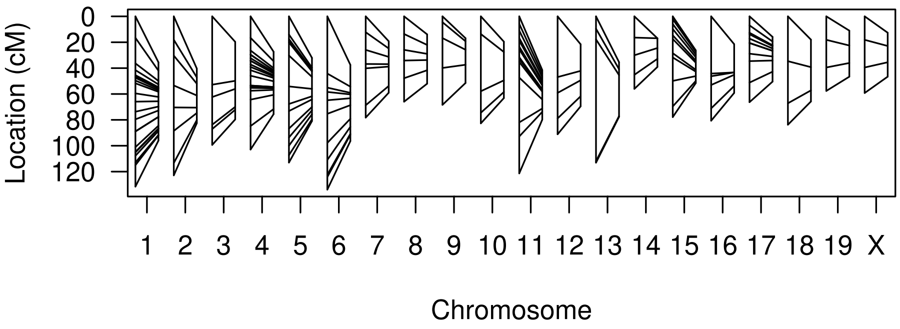

Figure 5: A comparison of genetic maps of the hyper data set analyzed as though it were a BC_1 (left) and as though it were a BC_3(right).

```
> plot(one.f2, one.bc2s3, col=c("red", "purple"))
```

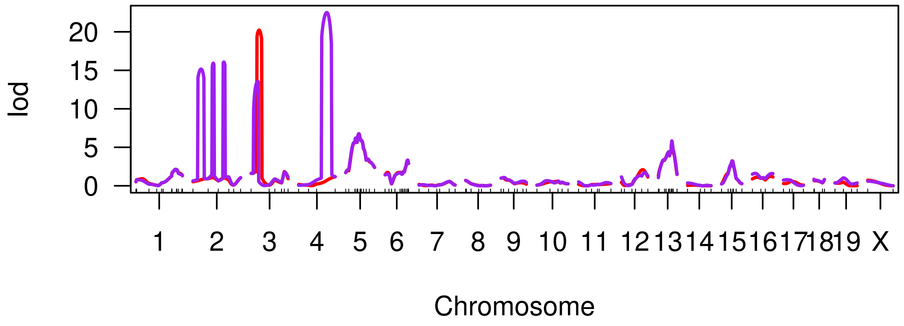

Figure 6: LOD plots for simple interval mapping with the listeria data set. The red curves are from analysis as though the population were a F_2. The purple curves are from analysis as though the population were a BC_2F_3. Both were analyzed using the same map distances to facilitate comparison

In both cases the map length is smaller when the cross is analyzed as a BC_sF_t because the same number of recombination events are attributed to multiple generations. In order to demonstrate that the cross history makes a real difference in outcome of a QTL analysis, we asign the same map to both cross objects regardless of cross history for direct comparisson. Comparing identical data sets with identical maps using the *scanone* command illustraits that position-wise LOD score also depends on cross history (figures 6 and 7) .

```
> listeria.bc2s3<-replace.map(listeria.bc2s3, map.f2)
> listeria.f2<-replace.map(listeria.f2, map.f2)
> listeria.f2<-calc.genoprob(listeria.f2, step=1 )
> one.f2<-scanone(listeria.f2, method="em",pheno.col=1)
> listeria.bc2s3<-calc.genoprob(listeria.bc2s3, step=1 )
> one.bc2s3<-scanone(listeria.bc2s3, method="em",pheno.col=1)
```

```
> hyper3<-replace.map(hyper3, map.bc1)
> hyper<-replace.map(hyper, map.bc1)
> hyper<-calc.genoprob(hyper, step=1 )
> plot(one.hyp, one.hyp3, col=c("red", "purple"))
```

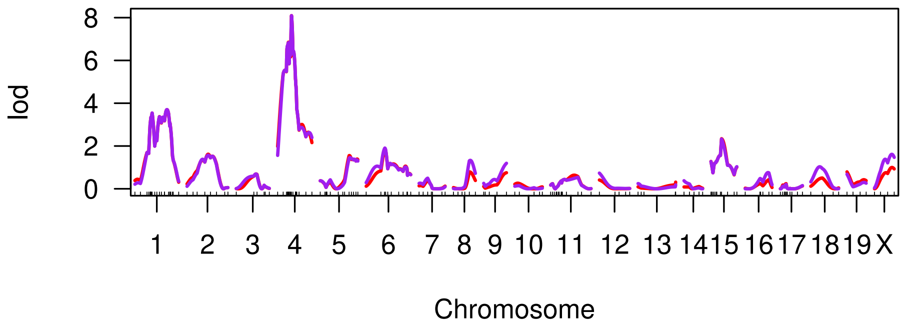

Figure 7: LOD plots for simple interval mapping with the hyper data set. The red curves are from analysis as though the population were a BC_1. The purple curves are from analysis as though the population were a BC_3. Both were analyzed using the same map distances to facilitate comparison

```
> one.hyp<-scanone(hyper, method="em",pheno.col=1)
> hyper3<-calc.genoprob(hyper3, step=1 )
> one.hyp3<-scanone(hyper3, method="em",pheno.col=1)
```

Calculations

Allowing for the analysis of BC_SF_T crosses in R/qtl required two new sets of calculations: genotype probabilities for different cross histories and recombination counts for these cross histories. The genotype probabilities were derived based on Jiang and Zeng's [9] calculations and the recombination counts are estimated using a golden section search.

Genotype Probabilities

Jiang and Zeng [9] provide a guide for calculating genotype frequencies resulting from several types of crosses of inbred lines. Although they examine many cases (F_2, selfed F_t, random mating F_t, backcross from selfed F_t, and BC_s) they do not address all possible cross structures. Most notably, they do not discuss BC_sF_t crosses. In this section we derive the equations for calculating genotype probabilities for a BC_sF_t cross. The equations we arrived at are heavily based on those of Jiang and Zeng. However they have been modified both to address BC_sF_t cross histories and to function within the context of the existing R/qtl program. We include all the implemented equations, both new and modified, below.

QTL mapping requires estimating the putative QTL genotype based on the observed genotypes of flanking markers. In all cases there are 2 parental inbred lines. Line 1 will be indicated by capital letters, while line 2 will be indicated by lower case letters. A particular descendant of these lines has a known genotype at locus A (indicated with A or a), however the genotype at locus B (indicated with B or b), the putative QTL, has not been observed. The genotype at locus B is dependent on the genotype at locus A, the recombination rate between locus A and locus B (r), and the cross history.

Backcross BC_S

The simplest case is a BC_1 with line 1 as the reccurrent parent. Let q be a vector of the frequency of all possible genotypes of loci A and B

$$q = \left[\begin{array}{cccc} freq(\frac{AB}{AB}) & freq(\frac{Ab}{AB}) & freq(\frac{aB}{AB}) & freq(\frac{ab}{AB}) \end{array}\right] = \left[\begin{array}{cccc} \frac{w}{2} & \frac{r}{2} & \frac{r}{2} & \frac{w}{2} \end{array}\right]$$

where $w = 1 - r$.

After a subsequent generation of backcrossing the genotype frequencies will change based on the probability that a pair of loci with a particular genotype will produce offspring of each genotype when backcrossed to the recurrent parent. We will call this the transition probability. In order to calculate q for a BC_2 we will need transition probabilities for all possible genotype combinations. Let M be the matrix of transition probabilities.

$$M = \left[\begin{array}{cccc}
P\left(\frac{AB}{AB}|\frac{AB}{AB}\right) & P\left(\frac{AB}{AB}|\frac{AB}{Ab}\right) & P\left(\frac{AB}{AB}|\frac{AB}{aB}\right) & P\left(\frac{AB}{AB}|\frac{AB}{ab}\right) \\
P\left(\frac{AB}{Ab}|\frac{AB}{AB}\right) & P\left((\frac{AB}{Ab}|\frac{AB}{Ab}\right) & P\left(\frac{AB}{Ab}|\frac{AB}{aB}\right) & P\left(\frac{AB}{Ab}|\frac{AB}{ab}\right) \\
P\left(\frac{AB}{aB}|\frac{AB}{AB}\right) & P\left(\frac{AB}{aB}|\frac{AB}{Ab}\right) & P\left(\frac{AB}{aB}|\frac{AB}{aB}\right) & P\left(\frac{AB}{aB}|\frac{AB}{ab}\right) \\
P\left(\frac{AB}{ab}|\frac{AB}{AB}\right) & P\left(\frac{AB}{ab}|\frac{AB}{Ab}\right) & P\left(\frac{AB}{ab}|\frac{AB}{aB}\right) & P\left(\frac{AB}{ab}|\frac{AB}{ab}\right)
\end{array}\right] = \left[\begin{array}{cccc}
\frac{w}{2} & \frac{r}{2} & \frac{r}{2} & \frac{w}{2} \\
0 & \frac{1}{2} & 0 & \frac{1}{2} \\
0 & 0 & \frac{1}{2} & \frac{1}{2} \\
0 & 0 & 0 & 1
\end{array}\right]$$

The frequency vector from the BC_1 can then be multiplied by the transition matrix to arrive at a frequency vector for a BC_2:

$$q_{BC_2} = qM \ .$$

With each subsequent generation of backcrossing it is necessary to multiply by the transtion matrix again. The equation for determining genotype frequencies based on any number of backcross generations (s) is:

$$q_{BC_s} = qM^{s-1} \ .$$

This can be further simplified [10]. $P(s,0)$ is a set of probabilities for all genotype combinations at two loci in a BC_s population. It is equivalent to q_{BC_s}, but organized differently to make it easier to read.

$$P(s,0) = \begin{array}{ccc} & BB & Bb \\ AA & A_{11} & A_{12} \\ Aa & A_{12} & A_{22} \end{array}$$

When $s = 1$

$$A_{11} = A_{12} = \frac{w}{2}$$
$$A_{12} = \frac{r}{2}$$

For any value of s

$$A_{11} = \frac{2^s - 2 + w^s}{2^s}$$
$$A_{12} = \frac{1 - w^s}{2^s}$$
$$A_{22} = \frac{w^s}{2^s}$$

Note the symmetry on the diagonal of recombinant alleles (Ab/AB and aB/AB) but not on the diagonal with only non-recombinant alleles (AB/AB and ab/AB). This asymmetry is due to the fact that ab alleles are only introduced in the F_1 and therefore all such alleles remaining in the population have never recombined where as AB alleles are introduced every generation. Genotype frequencies can be calculated for all types of crosses using a vector of initial frequencies and a transition matrix.

Repeated Selfing F_t

Next, we will discuss the calculations for genotype frequencies from an F_t population resulting from repeated selfing. This crossing structure is also sometimes refered to as an S_t, but we are using F_t to be consistant with the notation used by R/qtl. The major difference between the calculations for an F_t and a BC_s is that while in a backcross one allele is always AB, so there are only 4 genotype possibilities, in an F_t there are 10 genotype possibilities.

$$q_{F_1} = \left[\begin{array}{cccccccccc} \frac{AB}{AB} & \frac{AB}{Ab} & \frac{Ab}{Ab} & \frac{AB}{aB} & \frac{AB}{ab} & \frac{Ab}{aB} & \frac{Ab}{ab} & \frac{aB}{aB} & \frac{aB}{ab} & \frac{ab}{ab} \end{array}\right] = \left[\begin{array}{cccccccccc} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

The transition matrix for an F_t is the same as Jiang and Zeng [9]:

$$N = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 \\ \frac{w^2}{4} & \frac{rw}{2} & \frac{r^2}{4} & \frac{rw}{2} & \frac{w^2}{2} & \frac{r^2}{2} & \frac{rw}{2} & \frac{r^2}{4} & \frac{rw}{2} & \frac{w^2}{4} \\ \frac{r^2}{4} & \frac{rw}{2} & \frac{w^2}{4} & \frac{rw}{2} & \frac{r^2}{2} & \frac{w^2}{2} & \frac{rw}{2} & \frac{w^2}{4} & \frac{rw}{2} & \frac{r^2}{4} \\ 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Again, these can be multiplied to arrive at the probability of all genotypes in the F_t.

$$q_{F_t} = q_{F_1} N^{t-1}$$

This can be simplified. $P(0,t)$ contains the probabilities for all genotype combinations for two loci in an F_t population (once again it is equivalent to q_{F_t} reorganizes).

$$P(0,t) = \begin{array}{ccccc} & BB & Bb & bB & bb \\ AA & B_{11} & B_{12} & B_{12} & B_{14} \\ Aa & B_{12} & B_{22} & B_{23} & B_{12} \\ aA & B_{12} & B_{23} & B_{22} & B_{12} \\ aa & B_{14} & B_{12} & B_{12} & B_{11} \end{array}$$

The probabilities B_{ij} of ending up in a particular genotype after t generations can be modeled in terms of generations spent in the double heterozygous stage (at least 1, as F_1 is a double heterozygote), the probability of moving from that genotype to either one of the intermediate stages or to a double homozygote, the time spent at an intermediate stage (could be 0), and the probability of moving from an intermediate stage to a double homozygote. There are four transient states (double heterozygotes), 8 intermediate states (single heterozygotes) and 4 absorbing states (double heterozygotes).

The only genotypes which can produce all other genotypes are the transient double heterozygotes (B_{22} and B_{23}). Therefore with each generation there is an exponential decay in the probability of remaining in the double heterozygous state. In order to remain in the double heterozygous state there either has to be no recombination (w^2) or a double recombination event (r^2) in every generation. In order to model this we reparameterize w^2 and r^2 as β and γ, specifically $\beta + \gamma = w^2$ while $\beta - \gamma = r^2$. β is also the probability of remaining in a double heterozygous state given that the line started in one of the two double heterozygous states in a single generation.

$$B_{22} = \frac{\beta^{t-1} + \gamma^{t-1}}{2}$$

$$B_{23} = \frac{\beta^{t-1} - \gamma^{t-1}}{2}$$

$$\beta = \frac{w^2 + r^2}{2}$$

$$\gamma = \frac{w^2 - r^2}{2}$$

The 8 intermediate states, with one locus homozygous and one heterozygous. During one of the previous generations, one locus was fixed while the other remained heterozygous. There are two exponential decays, with the transition point unknown. After some simplification, this can be expressed asB_{12}:

$$B_{12} = \frac{rw\left(\frac{1}{2^{t-1}} - \beta^{t-1}\right)}{1-2\beta}$$

Finally, the four absorbing states, heterozygous at both loci, can be reached from a number of paths, involving simultaneous or separate fixation of both loci. The calculations are more involved, but simplify ty B_{11} or B_{14}:

$$B_{11} = f(w,r) = \frac{1}{8}\left[w^2\left(g\left(\beta,t\right) + g\left(\gamma,t\right)\right) + r^2\left(g\left(\beta,t\right) - g\left(\gamma,t\right)\right)\right] + \frac{rw}{5}\left[g\left(\beta,t\right) + g\left(2\beta,t-1\right)\right]$$

$$B_{14} = f(r,w)$$

With:

$$g\left(\beta,t\right) = (1-\beta^{t-1})/(1-\beta)\ .$$

Unlike P(s,0), P(0,t) is symmetric on both diagonals because both parental alleles are equally present in the F_1 and never introgressed again.

One major difference between working with a backcross and an F_t is that while in a backcross phase is always known, in an F_t phase cannot be observed. When dealing with phase unknown data the two heterozygote cases can be collapsed as follows:

$$\begin{array}{cccc} & BB & Bb & bb \\ AA & B_{11} & 2B_{12} & B_{14} \\ Aa & 2B_{12} & 2\left(B_{22}+B_{23}\right) & 2B_{12} \\ aa & B_{14} & 2B_{12} & B_{11} \end{array}$$

Since we cannot distinguish between the two heterozygote classes we add them and report the frequency of both.

The final difference between backcross and F_t calculations is that for F_t populations it is possible to have partially informative markers. Partially informative markers can only be interpreted as not belonging to a particular homozygous class. For instance if a marker were measured using the presence or absence of a band on a gel, heterozygotes would be indistinguishable from the homozygous present class. We will refer to partially informative markers as either "not AA" or "not aa". In order to calculate the probability of partially informative markers we add the probabilities of the genotypes we cannot distinguish between, much like the phase unknown case above. For example, *not* AA/BB could be Aa/BB, aA/BB, or aa/BB so we sum all of those probabilties to get $2B_{12} + B_{14}$. All other genotypes with partially informative markers can be calculated similarly.

Backcrossing followed by selfing BC_sF_t

The described equations for the BC_s and the F_t form the basis for the BC_sF_t. The two types of crosses can be thought of sequentially. The BC_s that forms the first steps of the BC_SF_t is exactly the same as the BC_S on it's own. The difference between calculating an F_t which follows several genetations of backcrossing and one which follows an F_1 is the vector of starting genotype frequencies. In this case the starting genotype frequencies can be supplied by $q_{BC_S} = qM^{s-1}$. The six genotypes not represented all have starting frequency 0.

$$q_{BC_SF_0} = \left[\begin{array}{cccccccccc} \frac{2^s-2+w^s}{2^s} & \frac{1-w^s}{2^s} & 0 & \frac{1-w^s}{2^s} & \frac{w^s}{2^s} & 0 & 0 & 0 & 0 & 0 \end{array} \right]$$

The q resulting from this modification of q_{BC_S} can be multiplied by the F_t transition matrix. Much like in the previous cases this can be simplified. $P(s,t)$ contains the probabilities of all possible genotype combinations at two loci for a BC_sF_t. Below, we explicitly identify the parts of the equations from the backcross (A) and selfing (B) probablities.

$$P(s,t) = \begin{array}{ccccc} & BB & Bb & bB & bb \\ AA & C_{11} & C_{12} & C_{12} & C_{14} \\ Aa & C_{12} & C_{22} & C_{23} & C_{24} \\ aA & C_{12} & C_{23} & C_{22} & C_{24} \\ aa & C_{14} & C_{24} & C_{24} & C_{44} \end{array}$$

Where:

$$
\begin{aligned}
C_{22} &= A_{22}(s)B_{22}(t) \\
C_{23} &= A_{22}(s)B_{23}(t) \\
C_{12} &= A_{22}(s)B_{12}(t) + A_{12}(s)\left(\frac{1}{2}\right)^t \\
C_{24} &= A_{22}(s)B_{12}(t) \\
C_{11} &= A_{22}(s)B_{11}(t) + A_{12}(s)\left(1 - \left(\frac{1}{2}\right)^t\right) + A_{11}(s) \\
C_{14} &= A_{22}(s)B_{11}(t) + A_{12}(s)\left(1 - \left(\frac{1}{2}\right)^t\right) \\
C_{44} &= A_{22}(s)B_{11}(t)
\end{aligned}
$$

Because these probabilities depend on the backcross probabilities there is only symmetry on one diagonal when $s > 0$. Partially informative markers and phase unknown data can be treated the same way as an F_t.

Recombination Counts

In the previous implementation of R/qtl recombination counts were calculated, however for advanced crossing schemes there is no direct analytic solution. Instead we implemented a hill climbing algorithm using a golden section search [11] which determines the most probable recombination frequency, rather than calculating an actual value. The search space starts between 0 and 0.5 (all possible recombination frequencies). The golden section search relies on comparing three points (figure 8). To start with the points are $r = 0$, $r = 0.5$, and $r = r_1$, where the value of r_1 is determined so that the ratio of a to a+b is equal to the ratio of a to b. Then a new point ($r = r_2$) is added in the larger interval so that the ratio of d to a is equal to the ratio of c to d. The set of 3 r values containing the highest maximum likelihood (as compared to the null model of unlinked markers $r = 0.5$) are kept, and the remaining value is dropped (in this case $r = 0.5$). The search algorithm starts again with 0, r_1, and r_2 as the three points. This process repeats until tolerance for the minimum improvement in likelihood is reached, then the r value with the highest likelihood is reported as the maximum likelihodd estimate used in the map. This provides an accurate estimate of recombination frequency.

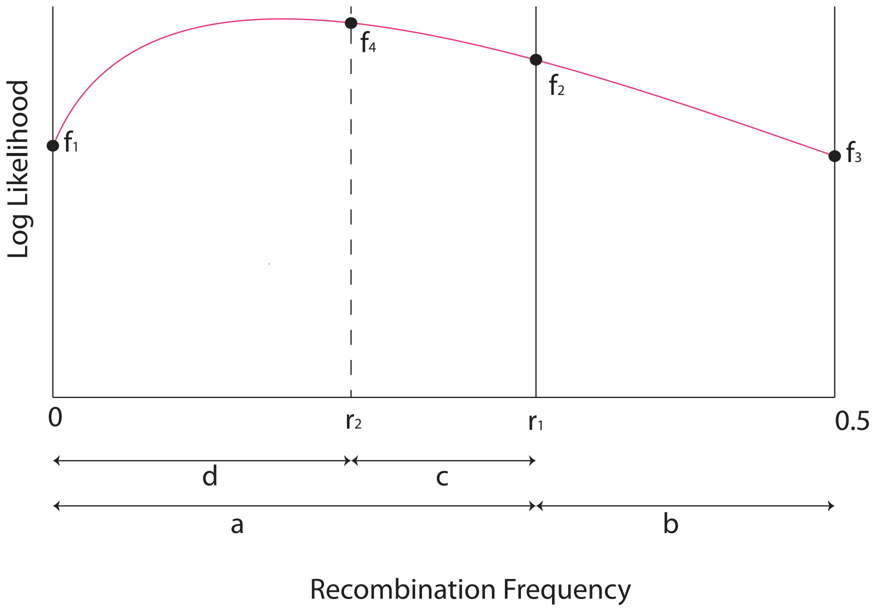

Figure 8: An illustration of the golden section search

A Note on Intercrosses and Random Matings

These equations are accurate for BC_sF_t when F_t refers to any number of selfed generations. We have not implemented code to address advanced intercross lines resulting from sib mating or random mating within an advanced cross. Below we sketch ideas to develop these algorithms.

In an F_2, selfing and sib-mating are interchangeable in terms of calculations because the entire population has an identical F_1 genotype. However after the F_2, calculations get more complicated for sib-mated populations. Each F_2 is sib-mated to create an F_3. Sib-mating brings the added complication that we need to think about families instead of individuals. There are 10 possible F_2 genotypes leading to 55 possible combinations of cross parents and their next generation families. A transition matrix (L) analagous to the one for selfing (N) would have to be 55 x 55 and account for the probability that a family that resulted from a cross between a particular set of parents in the F_{t-2} would yield a cross between another set of parents in the F_{t-1}. This would be multiplied by a vector of 55 starting probabilities for the F_1 $(q^*_{F_1})$, these being probabilities of genotypes for specific crosses rather than for individuals. Of course, for the F_1, there is only one type of cross $AB/abXAB/ab$ which can be the parents of the F_2, and the probability of the other 54 types of crosses is 0. Multiplying these successively will result in the probabilities of the crosses that produce the F_t, since our actual question is the probability of the genotypes of the F_t where one individual is selected from each family, we will need a second matrix, K. This matrix will give the probability for each of the 10 possible genotypes results based on the 55 possible crosses. The final result will be the probability of the 10 genotypes (q_{F_t}). The equation for the genotype probabilities after t generations of intercrossing, then is:

$$q_{F_t} = (q_{F_1} L^{t-2})K \ .$$

Note that the selfed F_t is a special case with only the 10 selfings of the 55 possible crosses being non-zero. In general, this formal equation can be simplified substantially by using symmetry arguments, and implemented in a similar manner to the selfing case. Transient and absorbing states can be handled in an analogous but somewhat more complicated manner to the selfed case. However, the devil is in the details!

We consider the case of random mating after s generations of backcross and t generations of selfing. We begin with the $q_{BC_sF_t}$ 10-vector of phase-known genotype frequencies, and multiply by a 10×4 matrix (J) to convert these frequencies into the four possible two-locus alleles (u). These allele frequencies are cross multiplied $(u_T u)$ to create a 4×4 matrix of random mating frequencis of genotypes, which are then reduced to the 10-vector format of phase-known genotype frequencies ($q_{BC_sF_tR}$). Eight of the rows of the matrix J are simple (0, 0.5 or 1 values), while the middle two involve the possible recombinants and non-recombinants:

$$J = \begin{array}{c|cccc} & AB & Ab & aB & ab \\ \frac{AB}{AB} & 1 & 0 & 0 & 0 \\ \frac{AB}{Ab} & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{Ab}{Ab} & 0 & 1 & 0 & 0 \\ \frac{AB}{aB} & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{AB}{ab} & \frac{w}{2} & \frac{r}{2} & \frac{r}{2} & \frac{w}{2} \\ \frac{aB}{Ab} & \frac{r}{2} & \frac{w}{2} & \frac{w}{2} & \frac{r}{2} \\ \frac{ab}{Ab} & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{aB}{aB} & 0 & 0 & 1 & 0 \\ \frac{ab}{aB} & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{ab}{ab} & 0 & 0 & 0 & 1 \end{array}$$

Extension to sib-mating instead of selfing follows directly. Extension to selfing or sib-mating after random mating follows as for the F_t approach outlined earlier. Again, the devil is in the details.

Code modifications in R/qtl

Our goal was to make R/qtl fully functional for a wider variety of cross types. Ideally these changes will be minimally visible to the user after inputing the cross type. In order to create an identical user experience we had to modify all R routines so that they would recognize the `bcsft` cross type. Cross objects in R/qtl all have the attribute "class" consisting of 2 parts: one which identifies it as a cross object and one which specifies the cross type (`bc`, `f2`, `riself`, `risib`, etc.). We added an additional option, `bcsft`. A major difference between the previous cross types and `bcsft`, all other cross types are specific. In that there are no options for types of backcrosses, it's just a backcross. With the BC_sF_t we intentionally created a more flexible cross type, where the generation number can be set by the user. This means that we don't have to go back and add a cross type every time we want to analyze a population with a different history. The way we have created this flexibity is by adding the attribute `cross.scheme` to cross objects. The `cross.scheme` consists of two numbers, the first is the generations of backcrossing (s), the second is the generations of selfing (t). The addition of a cross type and an attribute allow all R routines to recognize all types of BC_sF_t crosses.

The previous sections detailed the way genotype probabilities and recombination counts are calculated for BC_sF_t crosses. These calculations are contained within the specific `init`, `emit`, and `step` functions for `bcsft` within the C code. All three of these functions are used in the Hidden Markov Model (HMM). The `init` function determines the probability of true genotypes. The `emit` function determines the probability of observed genotypes given the true genotypes, while the `step` function determines the probability of a genotype at a particular locus given the genotype at a linked locus as described in the previous section. We created BC_sF_t versions of all of these functions which follow the same format and work the same way as the existing versions, except for when they are called by `est.map`. We did not find a closed form solution for calculating the number of recombination events between pairs of markers in a BC_sF_t and so we implemented a golden section search as part of `est.map` instead.

There is a second difference between the way the HMM is implemented for BC_sF_t and all other types of crosses, leading to an improvement in efficiency with no effect on the estimates. Previously the probabilities for each pair of markers for each individual were calculated independently given the recombination rate between those markers in the entire data set. For the BC_sF_t the entire set of probabilities is calculated once for a set of markers given the recombination rate and then applied to all individuals. In a population with 100 lines, this is the difference between 10 calculations (1 for each possible genotype combination) and 100 calculations. This method could be readily expanded to analyze populations with mixed cross histories (where some lines have undergone more generations of selfing or backcrossing than others). Recombination rates could be calculated across all individuals and then probabilities would be calculated separately for each cross history and applied to pairs of markers in an individual according to cross history. However, record keeping about cross histories for each individual line would need to be implemented in the package.

Genotype probabilities differ for the autosomes and sex chromosomes. While this is not an issue for selfed populations it could be an issue in an advanced backcross or advanced intercross populations. We have arranged for proper handling of the X chromosome. Basically in an F_t the X chromosome is treated as though it were the product of a BC_t. The only real change here is that we created the capacity to keep track of t. All changes to the program have been unit tested and that code is included in the package.

References

[1] Tanksley SD, Nelson JC (1996) Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. Theoretical and Applied Genetics 92: 191–203.

[2] Xie C, Gessler DDG, Xu S (1998) Sib mating designs for quantitative trait loci. Genetica 104: 9–19.

[3] Collard BCY, Jahufer MZZ, Brouwer JB, Pang ECK (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. Euphytica 142: 169–196.

[4] King EG, Macdonald SJ, Long AD (2012) Properties and power of the Drosophilia Synthetic Population Resource for the routine dissection of complex traits. Genetics 191: 935–949.

[5] Darvasi A, Soller M (1995) Advanced intercrossing lines, an experimental population for fine genetic mapping. Genetics 141: 1199–1207.

[6] Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19: 889–890.

[7] Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121: 185–199.

[8] Broman KW, Sen S (2009) A guide to QTL mapping with R/qtl. Springer.

[9] Jiang C, Zeng ZB (1997) Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. Gentica 101: 47–58.

[10] Bulmer MG (1985) The Mathmatical Theory of Quantitative Genetics. Oxford University Press.

[11] Kiefer J (1953) Sequential minimax search for a maximum. Proceedings of the American Mathmatical Society 4: 502–506.