Bayesian QTL analysis with the BayesQTLBIC R package

Roderick D. Ball

October 27, 2009

Abstract

This package implements a Bayesian model selection approach to QTL mapping based on markers (Ball 2001), considering linear regression models where a model corresponds to a set of selected markers. These models correspond, within the resolution of the marker map to the number and location of QTL. All possible models are considered according to their posterior probabilities estimated by the BIC criterion. This proves to be accurate with commonly used sample sizes (Ball 2007). Information on QTL location within an interval is represented in our framework by models with one or more of the flanking markers selected. Missing values are accommodated by multiple imputation and adjustment is made for multiple imputation and selective genotyping. Functions are provided implementing a strategy for whole genome analysis of multiple chromosomes when the total number of markers is too large for brute force search of model space. Results, including most probable models, posterior probabilities for model size and marginal probabilities for markers are conveniently summarised by the summary method for the R fitted model object. Our approach does not require MCMC simulations so can be faster and easier to use than full Bayesian MCMC approaches.

1 Introduction

1.1 History of QTL mapping

In quantitative trait loci (QTL) mapping we would like to infer the location and effects of QTL, *i.e.* locations in the genome associated with variation in quantitative traits.

Traditional non-Bayesian methods such as single marker hypothesis tests or interval mapping (Lander and Botstein 1989) and generalisations test for QTL at single loci along the genome, and the LOD score graph plotted along the genome. This approach has several major limitations:

1. It does not give a probability for existence of QTL—in general there is no relation between the p-values given (whether comparison-wise or genome-wise from permutation tests) and the strength of evidence for an effect.

2. The genetic architecture (existence and number of QTL) cannot be inferred from the graph of the LOD score. The LOD score from interval mapping is approximately equivalent to the

log posterior probability for QTL location within a region, assuming a single QTL exists within the region (Sen and Churchill 2001). Having effectively made the assumption that there is a single QTL within a region, we cannot logically infer the number of QTL from the graph.

3. The estimates of QTL effects will be biased (selection bias) unless estimated in an independent population.

Any decision made with these methods will necessarily be *ad hoc*, *e.g.* based on some arbitrary choice of threshold, leading to problems with spurious associations being reported that were not subsequently verified, if an independent verification population was used.

1.2 QTL analysis based on markers—selecting a model

An increase in power was claimed for interval mapping, equivalent to an extra marker, however, this was based on comparisons that did not adjust for the differing type I error rates and the problem of 'nuisance parameters present only under the alternative hypothesis' (Davies 1977, 1987; Rebai *et al.* 1995).

Broman and Speed (Broman 1997; Broman and Speed 2002) suggested simply regressing the trait on marker genotypes, and choosing a (single, 'best') model based on a modified version of the BIC criterion. This approach has the advantage of simplicity and in Broman and Speed (2002) simulations performed as well as or better than alternatives such as interval mapping or using MCMC to find a single 'best' fitting model.

However there is no reason to select a single best model—in QTL mapping a single model is not unequivocally selected by the data, rather multiple models are generally consistent with the data. This was our motivation for considering multiple models—a Bayesian model selection approach.

1.3 QTL analysis based on markers—using multiple models

Traditionally, statisticians select a single model best fitting the data, perhaps by some criterion such as p-values, AIC or BIC, then proceed with inference and estimation as if this was the true model. This is often done with the same data used for selecting the model and subsequent inference and estimation. This would often not be a problem, where the choice of model is ancillary to the inference or estimation being made, or where the data strongly selects a single model.

For QTL mapping, if we select a single model, we still have the problems with inference and selection bias affecting traditional methods, (unless effects are estimated in an independent sample, which is an inefficient use of the data). The choice of model is not ancillary to the inference but is central to the inference being made, *e.g.* the number of markers in the model is related to the number of QTL; and, we shall see the data does not strongly select a single model but many alternative models are consistent with the data.

1.4 BayesQTLBIC approach—a Bayesian model selection approach to QTL mapping based on markers

This package implements a Bayesian model selection approach to QTL mapping based on markers (Ball 2001), considering linear regression models where a model corresponds to a set of selected markers. These models correspond, within the resolution of the marker map to the number and location of QTL. All possible models are considered according to their posterior probabilities estimated by the BIC criterion. This proves to be accurate with commonly used sample sizes (Ball 2007). More accurate estimates could be obtained with analytical calculations and/or the Laplace approximation. These methods may be incorporated in future versions.

Information on QTL location within an interval is represented in our framework by models with one or more of the flanking markers selected. If we are willing to assume a single QTL within the interval, the interval mapping graph can also be plotted, so our approach is complementary to interval mapping. Our approach does not require MCMC simulations so can be faster and easier to use than full Bayesian MCMC approaches. Our approach is also complementary to full Bayesian MCMC approaches, *e.g.* can be used as a check on the latter, or to obtain starting values for MCMC runs.

1.5 QTL analysis based on markers versus modelling QTL location continuously

From a general modelling perspective, regression on markers amounts to choosing a reduced level of detail and complexity of modelling compared to modelling QTL positions continuously along the genome. Little is gained by a more detailed model if the information is not present in the data to reveal the level of detail modelled. In the case of QTL mapping, the extra level of detail (in interval mapping or full Bayesian methods discussed below) consists of QTL location modelled continuously on a distance scale less than the inter-marker spacing, and additional hypothetical QTL modes of action.

Broman and Speed (2002)'s simulations suggest that there is no major benefit from modelling QTL location continuously. Modelling QTL location continously has 2 drawbacks:

1. The implicit assumption, in the case of interval mapping, of the existence of one and only one QTL, and;

2. in the case of Bayesian multi-locus methods, the asymptotics of the BIC approximation to posterior probabilities for models may not apply (Geiger *et al.* 2001; Sillanpää and Corander 2002), and analytical calculations of marginal probabilities are not tractable, hence the need for mixture models and computationally intensive trans-dimensional MCMC samplers.

Certainly, QTL location can be estimated to an accuracy of less that the width of marker intervals if the sample size is sufficiently large and if a single QTL with given mode of action is assumed. The equivalent information is also contained in our output—the posterior probabilities for flanking markers to be selected will indicate how close the QTL is to one or other flanking markers. Moreover QTL can only be localised to smaller distances than marker spacing if the sample size is large enough, and the marker spacing small enough so that the prior probability for multiple QTL within an interval is small.

2 Single chromosome analysis

Here we demonstrate a single chromosome analysis, which can readily be applied to each chromosome separately, with the typical number of markers covering the chromosome in QTL mapping studies.

Since the BayesQTLBIC analysis works by exhaustively enumerating models there is a limit to the number of markers that can be considered simultaneously. Around 30 markers can be considered simultaneously but runs with around 10 markers (*e.g.* covering a single chromosome with 10–20cM spacing) will be faster.

Since chromosomes assort independently in meiosis, chromosomes can be analysed separately and the results combined. This is often a good approximation to the full joint model, although where a substantial proportion of the variation is explained by QTL on other chromosomes improvements in power can be obtained by various strategies (Section 3).

2.1 Example analysis

Data simulation: First we simulate some data. This example has 1200 backcross progeny, with 2 chromosomes, and 11 backcross markers per chromosome, and 5 QTL. (*cf.* the `bicreg.qtl()` help page):

```
> library(BayesQTLBIC)
> set.seed(1234)
> ex1.marker.pos <- seq(5,105,by=10)
> chrom <- rep(1:2,rep(length(ex1.marker.pos),2))
> ex1.qtldata <- sim.bc.progeny(n=1200,Vp=c(0.1,0.2,0.3,0.15,0.25)/2,
+            map.pos=list(chrom=rep(1:2,rep(length(ex1.marker.pos),2)),
+            pos=rep(ex1.marker.pos,2)),qtl.pos=list(chrom=rep(1:2,c(3,2)),
+                                    pos=c(40,50,80,30,55)))
chromosome 1
chromosome 2
> ex1.qtldata$x[1:5,chrom==1]
     c1m1 c1m2 c1m3 c1m4 c1m5 c1m6 c1m7 c1m8 c1m9 c1m10 c1m11
[1,]    2    2    2    2    2    2    2    2    2     2     2
[2,]    2    2    2    2    2    2    2    2    2     2     2
[3,]    2    2    2    2    2    2    2    2    1     1     1
[4,]    2    2    2    2    2    2    2    2    2     2     2
[5,]    2    2    2    2    2    2    2    2    2     2     2
```

Running the analysis: The QTL analysis for chromosome 1 is run using the function `bicreg.qtl()`:

```
> ex1n200c1.bicreg <- bicreg.qtl(x=ex1.qtldata$x[,chrom==1],
+      y=ex1.qtldata$y,OR=1000, nbest=10, nvmax=5,prior=0.2,
+      keep.size=1)
```

The main arguments are `x`, the matrix of marker values, with one row per individual and, `y`, the vector of trait values, and `prior`.

We have set `prior = 0.2` corresponding to a prior probability of 0.2 for a QTL to be in the vicinity of a marker. This corresponds to a prior probability of 0.02 per cM, or a prior expected number of QTL of 40 if the genome length is 2000cM.

Where marker spacings are not approximately uniform the `prior` can be set to a vector proportional to the average distance to flanking markers.

We have limited consideration to the 10 best models of each size with the option `nbest=10`, and models with at most 5 QTL with the option `nvmax=5`. This limits the computing time and storage required and number of models to store and print.

The option `OR=1000` specifies that models less than 1/1000 as likely as another model may be dropped by 'Occam's razor' (Cf the `bicreg` package; Raftery 1995).

The option `keep.size=1` ensures that the null model and models of size 1 are nevertheless kept.

Note:

1. We have run the analysis for chromosome 1 here. In general it is not possible to simultaneously consider all possible models for typically hundreds of markers covering the genome. However 10–20 markers in a single chromosome is manageable. A strategy to sample jointly from the models for each chromosome will be described below.

2. Markers are coded as 1,2 coding in the output from `sim.bicprogeny`. These can be used as is in the analysis. Any coding is possible but we prefer $\pm 1/2$ so that marker effects are approximately orthogonal to the intercept, and marker effects are approximately the effects of allelic substitution for a single QTL located at the marker position.

Summary output: The analysis is summarised by the summary function:

```
> summary(ex1n1200c1.bicreg,nbest=23)
R-squared, BIC, and approximate posterior probabilties for individual models:
   c1m1 c1m2 c1m3 c1m4 c1m5 c1m6 c1m7 c1m8 c1m9 c1m10 c1m11    R2    BIC postprob cumprob
1     0    0    0    0    1    0    0    1    1     0     0 43.91 -659.3 0.295776  0.2958
2     0    0    0    1    1    0    0    1    1     0     0 44.36 -659.2 0.285950  0.5817
3     0    0    0    1    1    0    1    0    1     0     0 44.24 -656.6 0.075194  0.6569
4     0    0    0    1    1    1    0    1    1     0     0 44.68 -656.2 0.064309  0.7212
5     0    0    0    0    1    1    0    1    1     0     0 44.22 -656.1 0.061292  0.7825
6     0    0    1    0    1    0    0    1    1     0     0 44.17 -655.1 0.035804  0.8183
7     0    0    0    0    1    0    1    0    1     0     0 43.71 -655.0 0.035331  0.8537
8     0    0    0    1    1    1    0    0    1     0     0 44.13 -654.2 0.023549  0.8772
9     0    0    0    1    1    0    1    1    1     0     0 44.55 -653.5 0.016071  0.8933
10    0    1    0    0    1    0    0    1    1     0     0 44.08 -653.2 0.014067  0.9073
11    0    0    0    0    1    0    1    1    1     0     0 44.08 -653.1 0.013621  0.9210
12    0    0    0    0    1    1    0    0    1     0     0 43.60 -652.8 0.011549  0.9325
13    0    0    0    0    1    0    0    1    0     1     0 43.60 -652.7 0.010951  0.9435
14    0    0    1    0    1    1    0    1    1     0     0 44.49 -652.1 0.008131  0.9516
15    0    0    0    1    1    0    0    1    0     1     0 44.03 -652.0 0.007715  0.9593
16    1    0    0    0    1    0    0    1    1     0     0 43.99 -651.2 0.005135  0.9644
17    0    0    0    0    1    0    0    1    1     1     0 43.98 -651.1 0.004815  0.9693
18    0    0    0    1    1    0    0    1    1     1     0 44.44 -650.9 0.004488  0.9737
19    0    0    0    1    1    1    1    0    1     0     0 44.43 -650.8 0.004207  0.9780
```

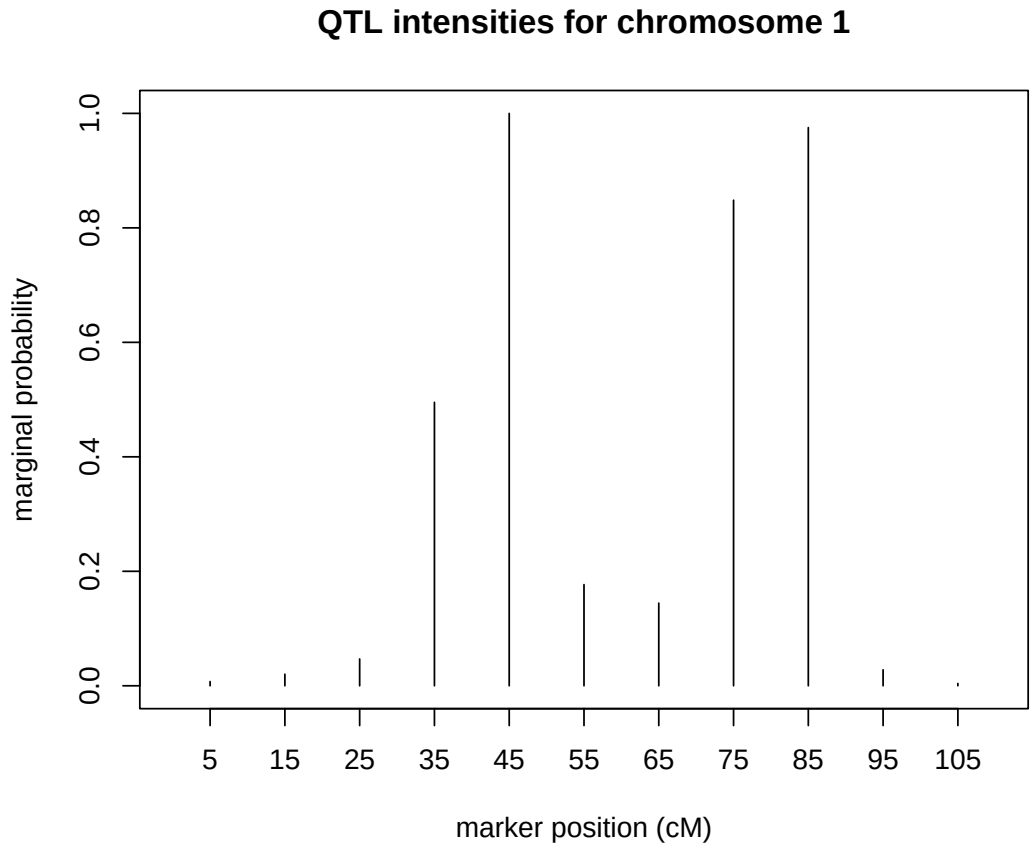
```
20    0    1    0    0    1    1    0    1    1     0     0 44.41 -650.4 0.003390  0.9813
21    0    0    0    1    1    0    0    1    1     0     1 44.41 -650.3 0.003353  0.9847
22    0    0    0    0    1    0    0    1    0     0     0 43.00 -650.0 0.002826  0.9875
23    0    1    0    1    1    0    0    1    1     0     0 44.38 -649.8 0.002616  0.9901
marginal probabilities for model sizes
         0          1          2          3          4          5
1.742e-145  2.155e-36  3.386e-03  3.582e-01  5.271e-01  1.113e-01
marginal probabilities for individual variables
  c1m4   c1m5   c1m6   c1m7   c1m8   c1m9
0.4875 0.9901 0.1764 0.1444 0.8403 0.9686
attr,"prior")
[1] 0.2
attr,"intercept")
[1] TRUE
```

Note:

1. The top 23 models account for 99% of the probability.
2. The probability of model size 0 is very low, indicating strong evidence for one or more QTL. Model sizes 3 to 5 have combined probability 99.7%, indicating strong evidence for 3 or more QTL
3. The most probable model had markers 5,8,9 selected. Hence the simulated QTL at positions 40, 50 cM could not be distinguished.
4. There is evidence for 1 or 2 QTL in the vicinity of markers 8,9—further calculation shows that markers 8 and/or 9 are selected in each of the top 23 models. The marginal posterior probability for marker 9 is higher than for marker 8, hence the estimated QTL position from interval mapping (*i.e.* assuming a single QTL) would be closer to marker 9.

More detailed information can be obtained by examining the fitted model object. For example it may be useful to examine graphs of posterior probability for a QTL to be in an interval:

```
> qprobs <- lapply(as.list(seq(along=ex1.marker.probs)),
+      function(i,w,p){sum(p[w[,i]])},
+      w=ex1n1200c1.bicreg$which,
+      p=ex1n1200c1.bicreg$postprob)
> plot(c(0,110),c(0,1),type='n',xaxt='n',
+      main='QTL intensities for chromosome 1'
+      xlab='marker position (cM)',
+      ylab='marginal probability')
> segments(ex1.marker.pos,rep(0,11),ex1.marker.pos,unlist(qprobs))
> axis(1,at=ex1.marker.pos)
```



2.2 Choice of prior

Choice of prior is an important consideration. Of course, higher prior probabilities *will* result in higher posterior probabilities. The choice of prior *is* subjective. We make no apology for using subjective priors, in fact this is vital in real world applications. Ideally, prior probabilities should be elicited from 'experts' who may be geneticists or statisticians involved with analysing previous similar data. Considerations may include general knowledge, number of QTL found previously in other species or related traits in the same or other species. How to combine these various sorts of information is inherently subjective, but this is unavoidable, in fact underlies any prediction or decision based on statistical data. One is interested in predicting some future or unobserved event based on a model or data from similar or related events. One rarely or never observes repeated events of the hypothetical type considered in frequentist theory. For further information on eliciting priors see O'Hagan *et al.* (2006).

Uncertainty in prior probabilities for QTL can be incorporated into the prior. Although we specify a single prior probability per marker, different values can be used in separate runs and

the results combined, equivalent to a mixture prior.

We do not recommend the use of 'non-informative' priors like `prior=0.5` (saying all models are equally likely). As with any Bayesian analysis, others are free to re-run the analysis with *their* prior. The function `recalc.bicprobs()` is useful in this regard.

Information relevant to the choice of prior are: the expected number of QTL per genome, the genome length. The expected number of QTL per genome can be elicited in discussions with biologists, based on knowledge of the trait, the species and QTL that have or have not been detected in previous experiments, including experiments on related traits and related species.

When considering the prior probability to use it is best to first consider the prior expected number of detectable QTL genome-wide. Since very small QTL of undetectable size are 'academic' we limit consideration to those over some practical threshold. This threshold itself may be a moving target when larger and larger datasets become available, but generally should be set at a size that can be detected with reasonable power. If we expect 40 detectable QTL, this means the experiment should be powerful enough to detect QTL explaining $1/40 = 2.5\%$ of the genetic variance of the trait.

2.3 Backcross or pseudo-backcross markers

The above example uses a simulated backcross. For an outcross, pseudo-backcross markers can also be used. In this case a 'phase alignment' is needed, *i.e.* to specify for each marker genotype a code (*e.g.* 1 or 2 for grandparents 1 or 2). This is presently up to the user. Adjacent marker pairs can be examined to check if the recombination fraction is as expected (a binomial distribution with probability predicted from the map distance using the mapping function). Otherwise the phase is reversed, *e.g.* the phase alignment might be as follows:

```
--M----n---o---
--m----N---O---
```

in which case the recombination rate would be high between `M` and `N` if we incorrectly assumed `Mm` and `Nn` were in phase. Given this alignment we then code `M,m,o` as *e.g.* -1/2 and `m,N,O` as *e.g.* +1/2.

For outcrosses, with co-dominant, dominant or mixed informative markers proceed as above constructing and aligning pseudo-backcross markers informative for each parent, then analyse each chromosome for each parent separately, or combine the X-matrices for each parent for a given chromosome and analyse each of these, optionally adding interaction terms corresponding to dominance.

Prior probabilities for dominance could be set reasonably high conditional on the markers for one or both parents being selected

Interactions between loci (epistasis) can be accommodated by adding interaction terms to the X-matrix, and specifying prior probabilities for interactions between loci as discussed in Ball (2001) and implemented by Bogdan *et al.* 2004. Note, however, prior probabilities for interactions necessarily have to be sufficiently low or the method (or any other) will not work.

Bogdan *et al.* (2004) specify a given prior for all interactions which fits nicely with the `bicreg.qtl()` function. However a more natural prior is that the prior for interactions is conditional on the pair of markers being selected, *i.e.* the prior for an interaction between `M` and `N` is higher if `M` and `N` are both selected, than if only one is selected, and the prior if one of `M` or

N is selected is higher that if none are. This can be accommodated later, *e.g.* by importance sampling.

2.4 Inference of the number of QTL

Posterior probabilities for the number of QTL is obtained by summing probabilities of models with each given size, with the caveat that it is not possible to distinguish 1 QTL from 2 QTL within a marker interval. Two adjacent markers with moderate to high probabilities could result from 1 or 2 QTL.

2.5 Marker effects

Marker effects can be estimated in each model. Conditional estimates (effects estimated in the models where a marker are selected) are subject to selection bias, a well known effect in statistics and the explanation for the 'Beavis effect' where QTL estimates tend to be smaller or non-existent when re-estimated in independent populations (Miller 1990; Beavis 1994; Ball 2001).

Unconditional or model averaged effects are obtained by averaging over models, where the effect of a marker is assumed to be zero in models where the marker is not selected. These effects are free of selection bias, because they do not come from a selected set of models.

These are the 'true' marker effects. We can also estimate the 'effects of allelic substitution' or the difference in trait values between marker classes if a given single marker is used for selection.

Unconditional and conditional marker effects can be obtained from the fitted model object *e.g.* :

```
> ex1n1200c1.bicreg$postmean
 [1] -2.8796504  0.0004358  0.0021863  0.0071707  0.1392444  0.8068102
 [7]  0.0535064  0.0487821  0.3955589  0.4494677  0.0058909  0.0003383
> ex1n1200c1.bicreg$condpostmean
 [1] -2.87965  0.05921  0.10801  0.15310  0.28120  0.80681  0.30275
 [8]  0.33777  0.46627  0.46091  0.21062  0.08389
```

2.6 Imputation of missing values

Marker values can be missing *e.g.* due to PCR failure. Several missing marker values are often present in a given individual, so it would be highly inefficient to discard individuals with missing values. To make best use of the data from an individual we impute missing marker values conditional on the flanking non-missing values, but ignoring the trait. Since most QTL are small these are approximately independent of the trait. Where multiple markers are missing these can be imputed sequentially in random order, conditional on flanking markers which are non-missing or previously imputed.

To reduce bias or sampling variation due to imputation we use multiple imputations. Multiple imputation means that the missing data are imputed multiple times and the resulting datasets combined into one larger dataset with the response repeated.

Missing marker values can be imputed using the function `impute.marker()`. To allow for multiple imputations `bicreg.qtl()` is called with `num.imputations` set to the number of imputations used.

Note:

1. Some authors use expected values, and believe this works well. We have not examined this possibility. However this theoretically underestimates uncertainty in the imputed values. Therefore it is incumbent on those authors to show whether this uncertainty affects the results in a given situation.

2. 'Virtual markers', with all data missing, interspersed between actual markers can also be imputed (Cf. Sen and Churchill 2001). As with interval mapping this can give a smoother plot of probabilities for QTL presence, but cannot give new information on recombinations within marker intervals.

3. Regression on markers estimates the genetic architecture to within the resolution of the marker map. More physical markers are needed to estimate the genetic architecture to a finer level of resolution.

Exercise. Simulate 'virtual markers' in the vicinity of and between of markers 4 and 5 and in the vicinity of and between markers 8 and 9, and apply `bicreg.qtl()` to the imputed dataset, and re-generate the above plot. (Cf. section below on imputation).

3 Multiple chromosome analysis

For multiple chromosomes, proceed initially by analysing each chromosome separately, and using these samples to generate multi-chromosome samples, viz:-

1. Analyse each chromosome separately.

2. Optionally, restrict to a subset of most probable models (*e.g.* accounting for a fixed percentage of the posterior probability, *e.g.* 99% or 99.9% , for the chromosome) for each chromosome

3. Sample jointly from the Cartesian product of these sets by sampling with replacement models for each chromosome according to its within chromosome posterior probability. Generate a sufficiently large sample, so that a high proportion of the probability is accounted for. This may require some experimentation.

4 . Re-evaluate approximate posterior probabilities for the multi-chromosome models using the BIC criterion.

5. Use this set of models together with their probabilities.

This process is automated using the functions `sample.bicreg.qtl.models()` for step (3.) and `bicreg.models()` for steps (4.) and (5.).

3.1 Example multi-chromosome analysis

```
> data(ex3n300a.data)
> chrom <- rep(1:12,rep(16,12))
> marker <- rep(1:16,12)
> x <- sapply(ex3n300a.data$Markers,c)
> y <- ex3n300a.data$Trait$t1
```

```
> nchrom <- length(sort(chrom.levels <- unique(chrom)))
> quick.demo <- TRUE
> if(quick.demo){
+   nc <- 2; nsim <- 20;x <- x[,chrom %in% 1:2];
+   chrom <- chrom[chrom %in% 1:2]
+ }else{nc <- 12; nsim <- 200}
> chrom.fits <- list()
> for(ii in seq(along=chrom.levels[1:nc])){
+   cat(paste("*** chromosome",ii,"***","\n"))
+   ci <- chrom.levels[ii]
+   chrom.sel <- chrom==ci
+   chrom.fits[[ii]] <- bicreg.qtl(x[,chrom.sel],y, prior=0.1,nbest=20,nvmax=3)
+ }
*** chromosome 1 ***
*** chromosome 2 ***
> mWhich <- sample.bicreg.qtl.models(chrom.fits,nsim=nsim)
> mres <- bicreg.models(x=x,y=y,which=mWhich,prior=0.1)
> summary(mres,nbest=38,min.marker.prob=0.05)
R-squared, BIC, and approximate posterior probabilties for individual models:
   c1m1 c1m2 c1m3 c1m4 c1m5 c1m6 c1m7 c1m8 c1m9 c1m10 c1m11 c1m12 c1m13 c1m14 c1m15 c1m16 c2m1
1     0    0    0    0    0    0    0    0    0     0     0     0     0     0     0     0    0
2     0    0    0    0    0    0    0    0    0     0     0     0     0     0     0     0    0
3     0    0    0    0    0    0    0    1    0     0     0     0     0     0     0     0    0
4     0    0    0    0    0    0    0    0    0     0     0     0     0     0     0     0    0
5     0    0    0    0    0    0    0    0    0     0     0     0     0     0     0     0    0
6     0    0    0    0    0    1    0    0    0     0     0     0     0     0     0     0    0
7     0    0    0    0    0    0    0    0    0     0     0     0     0     0     0     0    0
8     0    0    0    0    1    0    0    0    0     0     0     0     0     0     0     0    0
9     0    0    0    0    0    0    0    0    0     0     0     0     0     0     0     0    0
10    0    0    0    0    0    0    0    0    0     0     0     0     0     0     1     0    0
11    1    0    0    0    0    0    0    0    0     0     0     0     0     0     0     0    0
12    0    0    0    0    0    0    0    0    0     0     0     0     0     0     0     0    0
13    0    0    0    1    0    0    0    0    0     0     0     0     0     0     0     0    0
14    0    0    0    0    0    0    0    0    0     0     0     0     0     0     0     0    0
15    0    0    0    0    0    0    0    0    0     0     0     0     0     0     0     0    1
16    0    0    0    0    0    1    0    0    0     0     0     0     0     0     0     0    0
17    0    0    0    0    0    0    0    0    0     0     0     0     0     0     0     0    0
18    0    0    0    0    0    0    0    1    0     0     0     0     0     0     0     0    1
19    0    0    0    1    0    0    0    0    0     0     0     0     0     0     0     0    0
20    0    0    0    0    0    0    0    0    0     0     1     0     0     0     0     0    0
   c2m2 c2m3 c2m4 c2m5 c2m6 c2m7 c2m8 c2m9 c2m10 c2m11 c2m12 c2m13 c2m14 c2m15 c2m16   R2
1     0    0    0    0    0    0    0    0     0     0     0     0     0     0     0 0.00
2     0    0    0    0    0    0    0    1     0     0     0     0     0     0     0 2.05
3     0    0    0    0    0    0    0    0     0     0     0     0     0     0     0 1.92
4     0    0    0    0    0    0    1    0     0     0     0     0     0     0     0 1.73
5     0    0    0    0    0    0    0    0     0     0     0     0     0     0     1 1.72
6     0    0    0    0    0    0    0    0     0     0     0     0     0     0     0 1.31
7     0    0    0    0    1    0    0    0     0     0     0     0     0     0     0 0.89
8     0    0    0    0    0    0    0    0     0     0     0     0     0     0     0 0.81
9     0    0    1    0    0    0    0    0     0     0     0     0     0     0     0 0.72
```

```
10    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0 0.45
11    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0 0.37
12    0    0    0    0    0    0    0    1    0    0    1    0    0    0    0 3.64
13    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0 0.20
14    0    0    0    0    0    0    0    0    0    0    0    0    1    0    0 0.00
15    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0 0.00
16    0    0    0    0    0    0    0    1    0    0    0    0    0    0    0 3.20
17    0    0    0    0    0    0    0    1    1    0    0    0    0    0    0 2.51
18    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0 1.94
19    0    0    0    1    0    0    0    0    0    0    0    0    0    0    0 1.17
20    0    0    1    0    0    0    0    1    0    0    0    0    0    0    0 2.95
      BIC  postprob cumprob
1   6.743 6.206e-01  0.6206
2  10.627 8.899e-02  0.7096
3  11.025 7.294e-02  0.7825
4  11.606 5.456e-02  0.8371
5  11.636 5.373e-02  0.8908
6  12.885 2.878e-02  0.9196
7  14.159 1.522e-02  0.9348
8  14.401 1.348e-02  0.9483
9  14.673 1.177e-02  0.9601
10 15.488 7.831e-03  0.9679
11 15.729 6.942e-03  0.9749
12 15.816 6.648e-03  0.9815
13 16.241 5.376e-03  0.9869
14 16.841 3.981e-03  0.9909
15 16.841 3.981e-03  0.9948
16 17.183 3.357e-03  0.9982
17 19.313 1.157e-03  0.9994
18 21.062 4.824e-04  0.9998
19 23.409 1.492e-04  1.0000
20 28.055 1.462e-05  1.0000
marginal probabilities for model sizes
        0         1         2         3
6.206e-01 3.676e-01 1.179e-02 1.462e-05
marginal probabilities for individual variables
   c1m8    c2m8    c2m9   c2m16
0.07342 0.05456 0.10017 0.05373
attr,"prior")
[1] 0.1
attr,"intercept")
[1] TRUE
```

4 Selective genotyping

Selective genotyping (Darvasi and Soller 1992) is an experimental design where only the tails of the phenotypic distribution are sampled. For a given number of progeny genotyped, selective

genotyping is more powerful for identifying QTL, because there is little information in the centre of the distribution. Darvasi and Soller gave an adjustment factor γ_p by which naively estimated effects need to be reduced.

Notation: Let $\Phi(\cdot)$, $\varphi(\cdot)$ denote the standard normal cumulative distribution function and density function, and let Z_p denote the standard normal deviate corresponding to p, *i.e.* $Z_p = \Phi^{-1}(p)$.

The appropriate adjustment to the BIC criterion is given in Ball, (2009, submitted). The adjustment is as follows.

Assuming the model matrix of marker genotypes X, is full rank with k columns, the BIC criterion is adjusted by adding $k \log \gamma_p$ to the random sample BIC criterion giving:

$$\mathrm{BIC}_{\mathrm{sg}} = n \log(1 - R^2) + k \log(n\gamma_p)\,. \tag{1}$$

and γ_p is the Darvasi and Soller adjustment factor:

$$\gamma_p \approx 1 + Z_{1-p/2} \frac{\varphi(Z_{p/2})}{p/2}\,, \tag{2}$$

The adjustment for selective genotyping is applied by specifying the value of the argument `p.sg` (proportion `p.sg/2` genotyped per tail of the phenotypic distribution) to `bicreg.qtl`.

5 References

Ball, R. D. 2001: Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian Information Criterion. Genetics 159: 1351–1364.

Ball, R. D. 2007: Quantifying evidence for candidate gene polymorphisms—Bayesian analysis combining sequence-specific and QTL co-location information. Genetics 177: 2399-2416.

Ball, R. D. 2009: Bayes factors and the BIC criterion for QTL mapping in selectively genotyped families. Submitted.

Beavis, W. D. 1994: The power and deceit of QTL experiments: lessons from comparative QTL studies. Proc 49th Ann. Corn and Sorghum Indus. Res. Conf.

Bogdan M., Ghosh J. K., and Doerge R. W. 2004: Modifying the Schwarz Bayesian Information Criterion to Locate Multiple Interacting Quantitative Trait Loci. Genetics 167: 989–999.

Broman, K. W. 1997: Identifying quantitative trait loci in experimental crosses. PhD. thesis, University of California, Berkeley.

Broman K. W. and Speed T. P. 2002: A model selection approach for the identification of quantitative trait loci in experimental crosses (with discussion). J Roy Stat Soc B 64:641–656, 731–775.

Darvasi, A. and Soller, M. 1992: Selective genotyping for determination of linkage between a locus and a quantitative trait locus. Theoretical and Applied Genetics 85: 353–359.

Davies, R. B. 1977: Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 64: 247–254.

Davies, R. B. 1987: Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 74: 33–43.

Geiger, D., Heckerman, D., King, H., and Meek, C. 2001: Stratified exponential families: Graphical models and model selection. Annals of Statistics, 29 (2), pp. 505-529.

Lander, E. S. and Botstein, E. 1989: Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121: 185–199.

Miller, A. J., 1990: *Subset selection in regression*, Monographs on Statistics and Applied Probability 40, Chapman and Hall, London.

O'Hagan, A. Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. 2006: Uncertain judgements: eliciting experts' probabilities. Hoboken, NJ: Wiley, xiii+321pp. ISBN: 978-0-470-02999-2.

Raftery, A. E. 1995: Bayesian model selection in social research (with Discussion). Sociological Methodology 1995 (Peter V. Marsden, ed.), pp. 111-196, Cambridge, Mass.: Blackwells.

Rebai, A., Goffinet, B., and Mangin, B. 1995: Comparing power of different methods for QTL detection. *Biometrics* 51: 87–99.

Sen, S., and Churchill, G. A. 2001: A statistical framework for quantitative trait mapping. Genetics 159: 371–387.

Sillanpää, M. J. and Corander, J. 2002: Model choice in gene mapping: what and why. Trends in Genetics 18: 301–307.