A brief tour of R/qtlbim

Brian S. Yandell
Departments of Statistics and Horticulture, UW–Madison
www.qtlbim.org

June 7, 2007

Abstract

Bayesian interval mapping of QTL library R/qtlbim provides Bayesian analysis of multiple quantitative trait loci (QTL) models. This includes posterior estimates of the number and location of QTL, and of their main and epistatic effects. This tutorial assumes the reader has read *A brief tour of R/qtl* by Karl Broman, available at www.rqtl.org. We extend his hypertension example by analyzing the same data with Bayesian methods. Some familiarity with Bayesian methods is helpful but not required.

1 Overview of R/qtlbim

R/qtlbim is an extensible, interactive environment for mapping quantitative trait loci (QTL) in experimental crosses using Bayesian methods. It builds on R/qtl (www.rqtl.org), which in turn builds on the widely used statistical language system R (www.r-project.org). R/qtlbim is distributed in the same manner as R/qtl, and can be installed similarly.

This tutorial describes the MCMC sampling routines and some of the plotting facilities available through the `R/qtlbim` package. The purpose of these plots is to provide graphical tools for

1. exploring putative single and multiple QTL,
2. producing interpretable graphics of the relative evidence in favor of a set of putative QTL,
3. visual diagnostics of the MCMC model selection algorithm.

The package provides graphical diagnostics that can help investigate several ”better” models. It also provides a 1-D and 2-D genome scan. The `R/qtlbim` package provides plotting facilities for results generated by the analytical tools in the `R/qtlbim` package. These plotting facilities include time series plots of QTL model charactacteristics as basic MCMC diagnostic plots, visual tools for comparison of putative QTL models and exploratory plots whose purpose is the aid in the identification of likely QTL.

This package is currently in ”beta” release. That is, most of the basic features are stable, but we expect a learning curve. We would like feedback from experienced QTL mappers and R users especially. Please note that the command `qb.mcmc` that creates the MCMC samples produces external files in an output directory. These files are tens of Mb large. They are integral to `R/qtlbim` diagnostics. The proper way to remove a `qb` object created by `qb.mcmc` is to use the `qb.remove` routine, as indicated below.

This document walks through the `R/qtlbim` package by demonstrating the following major functions: creation of Bayesian samples from the posterior using MCMC sampling; use of plot and summary tools to examine genetic architecture; data management in R/qtlbim.

2 Citation of R/qtlbim

To cite R/qtl in publications, use the following:

> Yandell BS, Mehta T, Banerjee S, Shriner D, Venkataraman R, Moon JY, Neely WW, Wu H, von Smith R, Yi N (2007) R/qtlbim: QTL with Bayesian interval mapping in experimental crosses. *Bioinformatics 23*: 641-643.

The methodology is described in the following paper:

> Yi N, Yandell BS, Churchill GA, Allison DB, Eisen EJ, Pomp D (2005) Bayesian model selection for genome-wide epistatic QTL analysis. *Genetics 170*: 1333–1344.

3 Preliminaries

The Preliminaries of Broman's brief tour, as well as steps 1, 16 and 23 of his Example 1, provide important information on careful use of R.

This tutorial focuses on the `hyper` dataset from `R/qtl`. Please complete the R/qtl Tutorial for Hypertension in *A brief tour of R/qtl* available at (www.rqtl.org). Steps 1-4, 11-14 and 17-20 of Example 1 provide an overview of the core analysis in R/qtl.

Some other steps and examples might be skipped in the interest of time. Steps 15 and 21 show how to estimate permutation thresholds, which can take considerable time on slower machines. Step 22 of Example 1 and Example 5 develop a strategy for multiple QTL mapping. Example 4 shows how to incorporate covariates into R/qtl analysis.

The other skipped steps of Example 1 (5-10) concern further investigation of the marker genotypes and map construction. In addition, Example 2 provides further detail on marker order. Example 6 shows the internal data construct for cross objects for those familiar with R who want to dig deeper.

All of the code for this tutorial is available in a file. You can view this as

```
> url.show("http://www.stat.wisc.edu/~yandell/qtl/software/qtlbim/rqtlbimtour.R")
```

4 Hypertension Example

1. Run steps 1-4, 11-14 and 17-20 of Example 1 of Broman's brief tour. This provides an overview of R/qtl.
2. Load R/qtlbim package.

   ```
   > library(qtlbim)
   ```

3. Remove the X chromosome. R/qtlbim does not currently handle the X chromosome properly.

   ```
   > data(hyper)
   > hyper <- subset(hyper, chr = 1:19)
   ```

4. Calculate genotype probabilities.

   ```
   > hyper <- qb.genoprob(hyper, step = 2)
   ```

   This is essentially `calc.genoprob` of Broman's step 11, but with variable step width required for R/qtlbim.

5. The time-consuming part of R/qtlbim involves creating the MCMC samples. We will NOT do this step in the tutorial. The random seed of 1616 is included to allow reproducible samples. To obtain different MCMC samples, simply use a different seed or drop the seed argument.

   ```
   ## The following command is commented out.
   ## qbHyper <- qb.mcmc(hyper, pheno.col = 1, seed = 1616)
   ```

   Note that this step creates a uniquely named directory containing flat (text) files with the MCMC samples, as well as constructing the `qb` object.

6. Alternatively, we can load already prepared MCMC samples.

   ```
   > qb.load(hyper, qbHyper)
   ```

   This step actually loads the `hyper` dataset with the X chromosome removed and genotype probabilities properly calculated, as well as the `qb` object `qbHyper`.

7. Show detailed summary of MCMC samples. This includes how the MCMC samples were constructed, where they were stored, etc.

   ```
   > summary(qbHyper)
   ```

   The diagnostic summaries characterize the number of QTL samples (`nqtl`), the posteriors for the `mean` and environmental variance (`envvar`), the explained variance components (`varadd` and `varaa`) and the total variance (`var`). In addition, the percentages of samples for number of QTL, number of epistatic pairs, and the most common epistatic pairs are shown.

8. A collection of diagnostic plots and summaries can be shown with the `plot` command:

   ```
   > plot(qbHyper)
   ```

   These include the following, which are identified by the separate routine that can be used to get that particular plot.

- Time series of mcmc runs. R/coda trace of MCMC samples to assess the Markov chain mixing.

  ```
  > tmp <- qb.coda(qbHyper)
  > summary(tmp)
  > plot(tmp)
  ```

- Jittered plot of quantitative trait loci by chromosome. A plot of samples loci across chromosomes (separated by main loci, epistatic loci and any GxE loci).

  ```
  > tmp <- qb.loci(qbHyper)
  > summary(tmp)
  > plot(tmp)
  ```

- Bayes Factor selection plots. Posteriors and Bayes factor ratios for number of QTL, pattern of QTL across chromosomes, chromosomes and epistatic pairs.

  ```
  > tmp <- qb.BayesFactor(qbHyper)
  > summary(tmp)
  > plot(tmp)
  ```

- HPD regions and best estimates. One dimensional scan of major QTL for test statistic (2logBF) and means by genotype.

  ```
  > tmp <- qb.hpdone(qbHyper)
  > summary(tmp)
  > plot(tmp)
  ```

- Epistatic effects. Size of epistatic effects for most common pairs of chromosomes.

  ```
  > tmp <- qb.epistasis(qbHyper)
  > summary(tmp)
  > plot(tmp)
  ```

- Summary diagnostics as histograms and boxplots by number of QTL. Posterior distribution overall and separately by number of QTL sampled for the overall mean, environmental variance, explained variance and heritability.

  ```
  > tmp <- qb.diag(qbHyper)
  > summary(tmp)
  > plot(tmp)
  ```

9. Perform log posterior density (LPD) scan of entire genome. This is analogous to R/qtl's `scanone`, which produces the LOD. However there are marginal LPD, adjusting for all other possible QTL, rather than one QTL summaries.

   ```
   > one <- qb.scanone(qbHyper, type = "LPD")
   ```

10. The plot for `qb.scanone` has separate LPD curves for overall (black), main effects (blue), epistatic effects (purple) and QTL by environment (dark red).

    ```
    > plot(one)
    ```

11. The summary shows the estimated peak by chromosome. There are two positions, `m.pos` for position of main effect peak and `e.pos` for position of epistatic effect peak.

    ```
    > summary(one)
    ```

12. We can filter the summary to only pick up chromosomes with large main effects and/or epistasis. We can then save those chromosome IDs.

    ```
    > sum.one <- summary(one, sort = "sum", threshold = c(sum = 4,
    +     epistasis = 4))
    > sum.one
    > chrs <- sort(sum.one$chr)
    > chrs
    ```

13. Now we can show a plot with this subset of chromosomes.

    ```
    > plot(one, chr = chrs)
    ```

14. Now look at cell means by genotype. We restrict attention to the key chromosomes.

    ```
    > onemean <- qb.scanone(qbHyper, chr = chrs, type = "cellmean")
    > plot(onemean)
    > summary(onemean)
    ```

15. An alternative way to filter the chromosomes is to use the highest posterior density (HPD) region. Here we ask for an LPD profile, rather than the default `2logBF`.

    ```
    > hpd <- qb.hpdone(qbHyper, profile = "LPD")
    > summary(hpd)
    > plot(hpd)
    ```

    The summary includes the limits of the HPD interval for each chromosome. The HPD region is computed across the entire genome.

16. Perform a two-dimensional scan on the key chromosomes.

    ```
    > two <- qb.scantwo(qbHyper, chr = chrs, type = "LPD")
    ```

17. Summarize the 2-D scan, sorting by the upper triangle, which contains epistasis by default. Threshold to include only values above 4.

    ```
    > summary(two, sort = "upper", threshold = c(upper = 4))
    ```

18. Plot to visualize epistatic chromosome pairs.

    ```
    > plot(two)
    ```

19. Slice along ridge relative to chromosome 15.

    ```
    > plot(two, chr = c(4, 6, 7), slice = 15)
    ```

20. Slice to examine cell mean for epistasis with chr 15. Plot shows profile of means for chromosome 6 and 7 when genotype on chr 15 is A (top) and H (bottom).

    ```
    > slice <- qb.sliceone(qbHyper, type = "cellmean", chr = c(4, 6,
    +     7), slice = 15)
    > summary(slice)
    > plot(slice, chr = 6:7)
    ```

21. Perform detailed slice at peak on chr 6 and 15. Rightmost plots are from R/qtl at nearest marker to peak.

    ```
    > slice = qb.slicetwo(qbHyper, c(6, 15), c(59, 19.5))
    > plot(slice)
    > summary(slice)
    ```

5 Hypertension Demo

An alternative demo of R/qtlbim run on the hypertension data can be run as

```
> library(qtlbim)
> demo(qb.hyper.tour)
```