Tutorial on Bayesian 1-D and 2-D Genome Scans

Brian S. Yandell and W. Whipple Neely

October 23, 2006

Contents

Abstract

We present an introduction to QTL analysis based on new Bayesian scan routines available in the `R/qtlbim` package. These routines allow exploration of single and multiple QTL models. Additionally plotting routines allow visual exploration of the genetic architecure for a phenotypic trait. We present tutorial examples based on both experimental and simulated data.

1 Overview

This document describes 1-D and 2-D Bayesian genome scan routines available in the `R/qtlbim` package. In the present context, the term "scan" refers to methods based on constructing one or two dimensional profiles of QTL likelihoods or posterior distributions. These new scan routines in `R/qtlbim` are analogous to the routines `scanone` and `scantwo` from the `R/qtl` package. On a practical level, using `R/qtlbim` scan routines is very similar to using `R/qtl`'s `scanone` and `scantwo` methods. The key difference between the scan routines in `R/qtlbim` and the scan routines in `R/qtl` lies in the technique used for constructing QTL summaries. `R/qtlbim` extends `R/qtl` by providing the ability to generate Markov chain Monte Carlo (MCMC) samples from a posterior distribution for the genetic architecture of a trait. Furthermore the putative genetic architectures sampled can include an arbirary number of QTL.

The `R/qtlbim` package's scan routines are called `qb.scanone` and `qb.scantwo`. Because these scans are motivated by Bayesian MCMC techniques we refer to `qb.scanone` and `qb.scantwo` collectively as "qb.scans" or "qb.scan routines". The utility of the qb.scan routines lies in their ability to provide interpretable summaries of the high-dimensional MCMC samples. The scan summaries use ideas of Bayesian model averaging to explore the most probable models given the data. For example, in a one dimensional genome scan, we might consider the contribution of each potential locus averaging over all sampled models that include that locus. This allows us to adjust for the possible effects of all other loci by examining the marginal distributions. This has the advantage of reducing variation explained by other loci and reducing bias due to linked loci. Thus a one dimensional marginal scan can be informative about higher-order models directly without bias or variance inflation. Although the development of the `qb.scan` routines is motivated by Bayesian techniques, the interpretation of qb.scans involve a mix of frequentist and Bayesian ideas. In what follows we show the resolving power of low-dimensional scans that condition on the presence of other QTL using simulated data with one QTL and the `hyper` data set.

2 Using and Interpreting Qb.Scan Routines

This section illustrates the basic uses and interpretation of the qb.scan routines using both simulated data and experimental data. The value of the simulated data is that we have complete control over the model from which the data are generated, so that we can provide very simple examples with predictable outcomes. Our real data is the `hyper` data included in the `R/qtl` package.

2.1 Simulating Data Using R/qtl

As an initial illustration, we use simulated data with a modest sample size. For the simulation we set the number of individuals in the sample (`n.ind`) to be 100,

```
> n.ind <- 100
```

assign a single QTL at position (`qtl.pos`) 100 on a 200cM chromosome (`n.mark`) with markers at every 1cM (`by.mark`),

```
> n.mark <- 200
> by.mark <- 1
> qtl.pos <- n.mark/2
```

and set the substitution effect size (`qtl.effect`) to be 2.

```
> qtl.effect <- 2
```

In order to use these setting to simulate a data set requires four steps.

1. First, invoke the R/qtlbim library. Notice that when `R/qtlbim` is loaded the `R/qtl` library and any other required packages will be automatically loaded.

   ```
   > library(qtlbim)
   ```

2. Second, create a sequence of marker data (`markers`) and use `sim.cross` to simulate a data set with one QTL.

   ```
   > markers <- seq(0, n.mark, by = by.mark)
   > names(markers) <- paste("M", markers, sep = "")
   ```

3. Third, construct a marker map (`sim.map`) and a model (`sim.model`). The marker map specifies which markers appear on each chromosome. The model is a matrix specifying the chromosome, the qtl position and effect size.

   ```
   > sim.map <- list(ch1 = markers)
   > sim.model <- matrix(c(1, qtl.pos, qtl.effect/2),
   +     1, 3)
   > colnames(sim.model) <- c("chromosome", "qtl-position",
   +     "effect-size")
   ```

4. Fourth, simulate data using the `R/qtl` function `sim.cross`. Because `sim.cross` uses R's internal random number generator, setting R's random number seed prior to calling `sim.cross` guarantees that the simulation is repeatable. We only do 1000 iterations here for demonstration purposes.

   ```
   set.seed(1234)
   sim <- sim.cross(map = sim.map, model = sim.model,
     n.ind = n.ind, type = "bc")
   ```

The `summary` function can be used to give a quick synopsis of the simulated data.

```
> summary(sim)

    Backcross

    No. individuals:    100

    No. phenotypes:     1
```

```
Percent phenotyped: 100

No. chromosomes:    1
    Autosomes:      ch1

Total markers:      201
No. markers:        201
Percent genotyped:  100
Genotypes (%):      AA:47.9  AB:52.1
```

It is worth noting that creating our simulated data has only required functions available in the `R/qtl` package and that up to this point we have not called any functions from the `R/qtlbim` package. Later we simulate data from multiple QTL models. Simulating these multiple qtl models requires using the `qb.sim.cross` function in the `R/qtlbim` package.

2.2 Loading the `hyper` Data

The `hyper` data come from 250 backcross individuals in which the phenotype is blood pressure. To load `hyper` use the command

```
data(hyper)
```

A summary of `hyper` can be displayed with

```
> summary(hyper)

    Backcross

    No. individuals:    250

    No. phenotypes:     1
    Percent phenotyped: 100

    No. chromosomes:    19
        Autosomes:      1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

    Total markers:      170
    No. markers:        22 8 6 20 14 11 7 6 5 5 14 5 5 5 11 6 12 4 4
    Percent genotyped:  47.9
    Genotypes (%):      AA:50.1  AB:49.9
```

We need to exclude the X chromosome from our work with `hyper` as `R/qtlbim` does not yet handle this properly. To accomplish this we identify the chromosome named "`X`" and select all but this chromosome.

```
hyper <- subset(hyper,sapply(hyper$geno,class) != "X")
```

2.3 Running the MCMC Analysis on Simulated Data & The hyper Data

Now that we have data, we can begin to use the new methods available through R/qtlbim. The next step in using the R/qtl package would be to use the function calc.genoprob to create genotype probabilities based on a Hidden Markov model. For the Bayesian model selection, we replace calc.genoprob with the R/qtlbim function qb.genoprob followed by the MCMC sampler (qb.mcmc). The MCMC sampler has random number generator and does not use R's built-in random number generator, in order to make our simulations repeatable we pass an integer seed (qb.random.seed) directly to qb.mcmc.

```
sim <- qb.genoprob(sim,step=2)
qbSim <- qb.mcmc(sim, epistasis=FALSE, n.iter = 1000,
   seed=1616, verbose=FALSE, mydir="scanPDF")
```

By default the qb.mcmc function will print out progress messages of the number of iterations completed. These progress messages can be suppressed by setting verbose=FALSE. To run the MCMC sampler on the hyper data we use the command

```
hyper <- qb.genoprob(hyper, step=2)
qbHyper <- qb.mcmc(hyper, genoupdate=TRUE, n.iter = n.iter,
  seed = qb.random.seed, verbose=FALSE,mydir="scanPDF")
```

2.4 Using qb.scanone

The object qbSim created above contains the results of the MCMC run. Each iteration of the Monte Carlo chain represents a single QTL model. The entire Monte Carlo chain represents a sample from the posterior distribution of all possible models. One simple summary of the MCMC sample is the posterior profile, or the posterior probability of a QTL at each locus. A summary and plot of these counts is carried out as follows.

```
> temp <- qb.scanone(qbSim, type = "posterior")
> summary(temp)

posterior of phenotype for main


          n.qtl pos m.pos  main
c1:c2.677  2.68  98    98 0.277
```

The plot of qb.scanone.counts shows that the overwhelming majority of models included just one QTL in the vicinity of marker 100. This is consistent with out simulated data since the true model was for a single QTL at marker 100.

In order to run qb.scanone on the qbHyper we follow the same procedure. Here we show 2log(BF), or log of the Bayes factor, measuring the strength of evidence (> 2.1 is high) for a QTL.

```
> temp <- qb.scanone(qbHyper, type = "2logBF")
```

The plot of qb.scanone shows noticable peaks on chromosomes 1, 4, 6 and 15. The blue lines in the plot indicate main effects, the purple indicate epistatic effects and black curves (where visible) represent the sum of main and epistatic effects. In order to

```
> plot(temp)
```

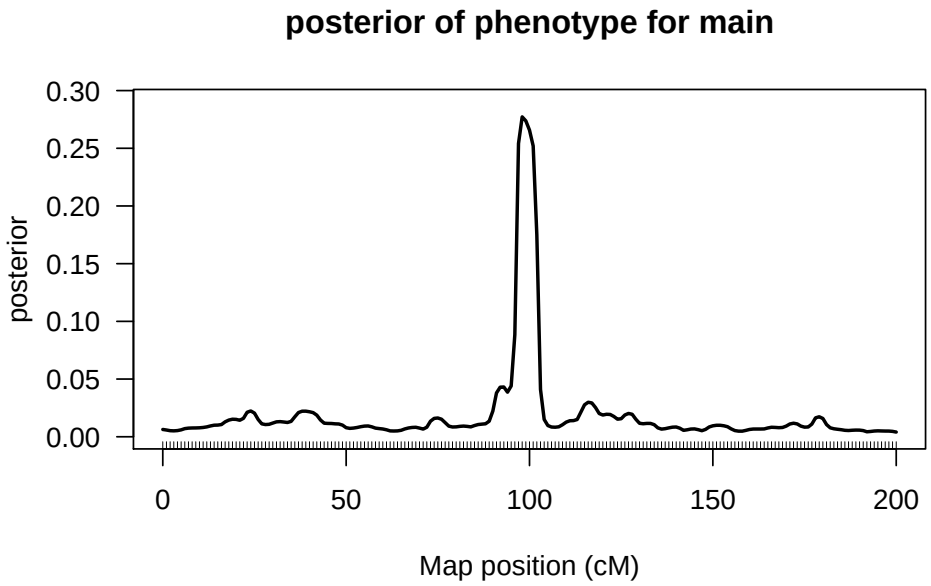

Figure 1: Plot of `qb.scanone` for posterior of simulated data. Notice that the overwhelming majority of the posterior is concentrated in the vicinity of marker 100.

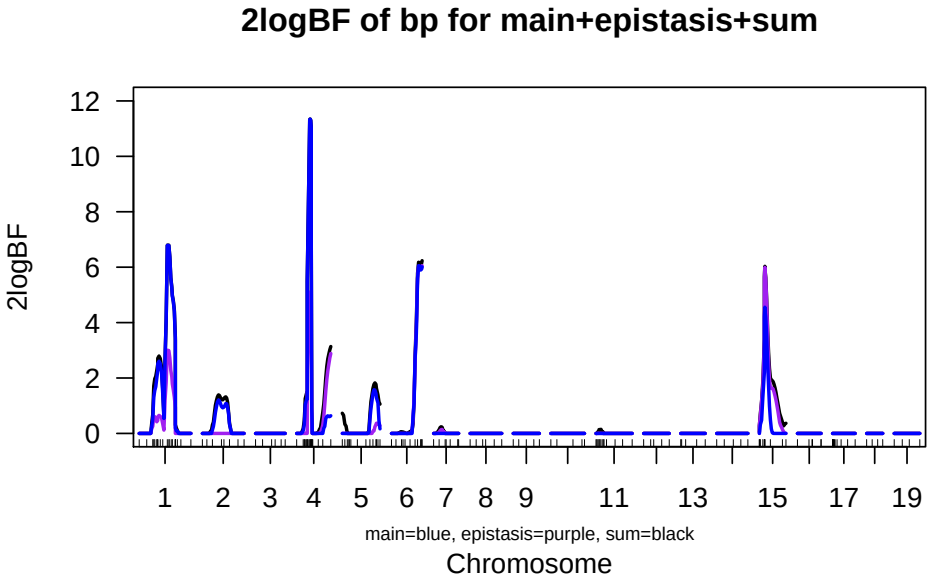

Figure 2: Plot of `qb.scanone` for 2log(Bayes factor) on blood pressure (`bp`) for the `hyper` data.

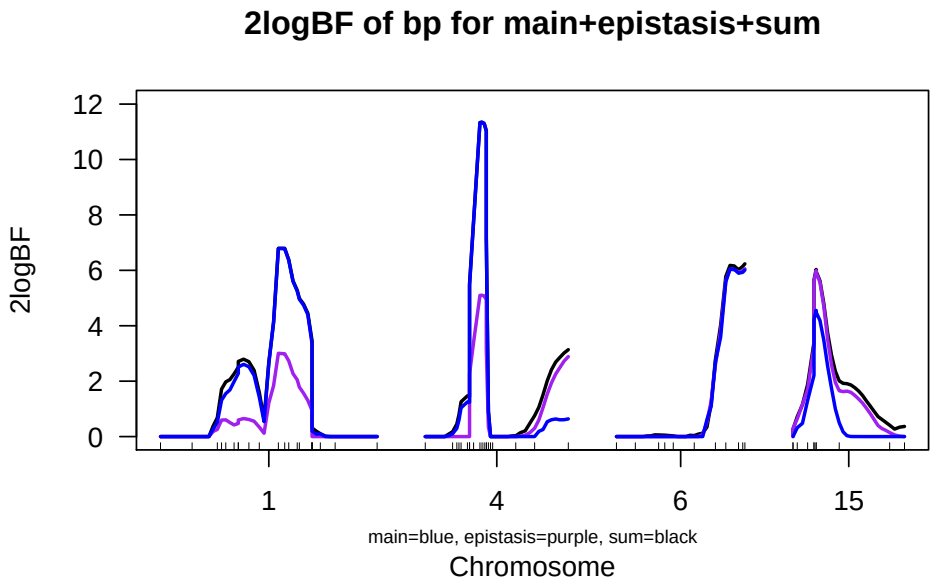

Figure 3: The `qb.scanone` results for the `hyper` data restricted to chomosomes 1,4,6 and 15.

examine the effects on 1, 4, 6 and 15 more closely, we can plot subsets of chromosomes (Figure 3) by using the plot command `plot(temp, chr=c(1,4,6,15)`.

2.5 Using `qb.scantwo`

The function `qb.scantwo` gives a two dimensional scan that allows us to look for possible epistatic effects between putative QTL. Before using `qb.scantwo` on the `hyper` data set we will illustrate the the use and interpretation of `qb.scantwo` on extremely simple simulated data. After examining the results from the simulated data we will use `qb.scantwo` to explore the `hyper` data set.

2.5.1 Simulating Data with Epistatic Effects & No Main Effects

In order to simulate data for which there is an epistatic effect but no main effects, we use the `R/qtlbim` function `qb.sim.cross`. Through the rest of this document we *reuse* the names of objects `sim` and `qbSim` to simplify presentation, but their contents change depending on the example.

1. Determine the positions of qtl by specifying the `qtl.positions` parameter. This parameter gives a matrix with dimensions `(number of qtl)` x 2. Each row identifies a qtl, the first column's entries represent the chromosome's index, the second column's entries represent the location on the chromosome of the qtl. The order in which qtl are listed in this parameter is the index by which they are identified later on in the parameters `qtl.main` and `qtl.epi`.

```
> qtl.positions <- rbind(qtl1 = c(chromosome = 1,
+     locus = 5), qtl2 = c(chromosome = 1, locus = 50),
```

```
+     qtl3 = c(chromosome = 2, locus = 33))
> qtl.positions

     chromosome locus
qtl1          1     5
qtl2          1    50
qtl3          2    33
```

2. Specify the main effects of the qtl by the `qtl.main.model` parameter. The qtl indices listed here are the row indices of the `qtl.positions` (or `qtl.pos`) parameter.

```
> qtl.main.model <- rbind(qtl1.main.effect = c(qtl = 1,
+     main.effect.size = 0), qtl2.main.effect = c(qtl = 2,
+     main.effect.size = 0), qtl3.main.effect = c(qtl = 3,
+     main.effect.size = 0))
> qtl.main.model

                 qtl main.effect.size
qtl1.main.effect   1                0
qtl2.main.effect   2                0
qtl3.main.effect   3                0
```

3. Specify the epistatic effects using the `qtl.epi.model` parameter.

```
> qtl.epi.model <- rbind(qtl1.and.qtl3.epi.effect = c(qtl1 = 1,
+     qtl2 = 3, epi.effect.size = 10))
> qtl.epi.model

                          qtl1 qtl2 epi.effect.size
qtl1.and.qtl3.epi.effect    1    3              10
```

4. Call the `qb.sim.cross` function. The parameter `len` gives the lengths of each chromosome. Thus `len = c(80,90,44)` would represent a model with three chromosomes of lengths 80, 90, and 44 respectively. Similarly the parameter `n.mar` gives the number of markers on each chromosome. If a single number is entered for `n.mar` then all chromosomes will have the same number of markers.

```
set.seed(1234)
sim <- qb.sim.cross(len=rep(100,2), n.mar=10, eq.spacing=TRUE,
  n.ind=100, type="bc", missing.geno=0.03, qtl.pos=qtl.positions,
  qtl.main=qtl.main.model, qtl.epis=qtl.epi.model)
```

Finally we can run `qb.genoprob` and `qb.mcmc` on the simulated data, just as we would for data that arose from an actual experiment.

```
## Call qb.genoprob to fill in missing data.
sim <- qb.genoprob(sim)
## Call qb.mcmc and then analysis code.
qbSim <- qb.mcmc(sim,n.iter=n.iter,verbose=FALSE,
                    seed=qb.random.seed,mydir="scanPDF")
```

Next we use qb.scantwo to examine the heritability, or percent varianced explained, per pair of loci.

```
> temp <- qb.scantwo(qbSim)
> summary(temp, digits = 2)

upper: heritability of pheno.normal for epistasis
lower: heritability of pheno.normal for full

      n.qtl l.pos1 l.pos2 lower u.pos1 u.pos2 upper
c1:c1 1.237   53.3   64.4  16.0  48.89   66.7  19.4
c1:c2 2.851    0.0   33.3  84.5   2.22   24.4  84.5
c2:c2 0.667   20.0   60.0  21.7  62.22   88.9   8.7
```

The plot of the results from running `qb.scantwo` on the simulated data shows two isolated peaks each representing the interaction of the first and third QTL. This illustrates that under idealized circumstances we would expect a plot of `qb.scantwo` results to show evidence for epistasis in the form of a peak in the vicinity of the position corresponding to the loci of the two QTL.

2.5.2 Simulating Data with Main Effects But No Epistatic Effects.

In order to illustrate the extreme case where there are main effects but no epistatic effects we can modify the simulation parameters `qtl.main.model` and `qtl.epi.model`. Since `qtl.main.model` consisted entirely of zeros (indicating no main effects), we can add a main effect of size 10 for the first QTL as follows.

```
> qtl.main.model[1, "main.effect.size"] = 10
```

Since we have no epistatic component to the new model, we replace `qtl.epi.model` in the call to `qb.sim.cross` with `NULL`.

```
set.seed(1234)
sim <- qb.sim.cross(len=rep(100,2), n.mar=10, eq.spacing=TRUE,
  n.ind=100, type="bc",
  missing.geno=0.03, qtl.pos=qtl.positions,
  qtl.main=qtl.main.model, qtl.epis=NULL)
```

Running `qb.sim.cross`, `qb.mcmc` and plotting the results of `qb.scantwo` gives a plot of data in which there is no epistatic effect, but in which there is a main effect. This is indicated by the horizontal band at 5`cm` on chromosome 1.

```
sim <- qb.genoprob(sim)
qbSim <- qb.mcmc(sim,n.iter=n.iter,verbose=FALSE,
                  seed=qb.random.seed,mydir="scanPDF")
```

```
> temp <- qb.scantwo(qbSim)
> summary(temp, digits = 2)

upper: heritability of pheno.normal for epistasis
lower: heritability of pheno.normal for full

      n.qtl l.pos1 l.pos2 lower u.pos1 u.pos2 upper
c1:c1  3.40   2.22  100.0 95.08   22.2   64.4  7.78
c1:c2  1.20   6.67   71.1 94.95   84.4   57.8  6.77
c2:c2  0.15  28.89   84.4  8.93   28.9   84.4  9.90
```

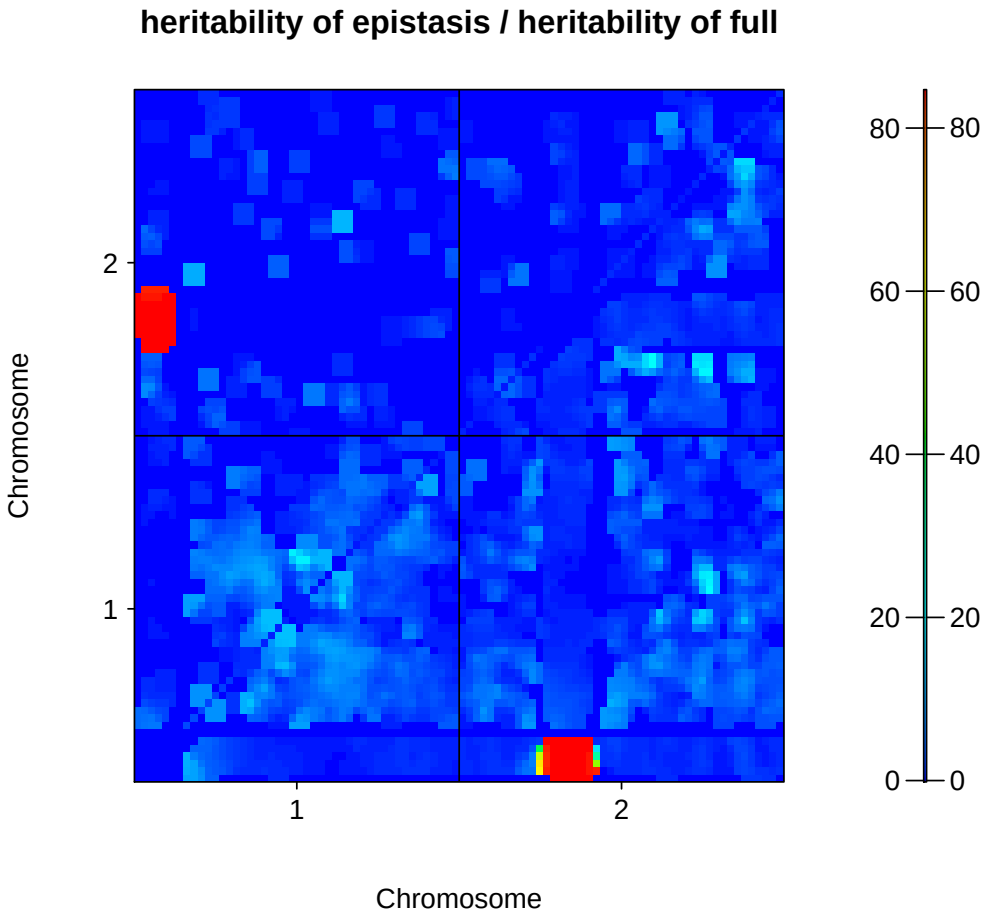

Figure 4: Plot of `qb.scantwo` results from simulated data with an epistatic effect but no main effects. Heritability is R-squared, or percent variation explained. Epistasis in the upper triangle is indicated by the two isolated peaks. Notice that the peaks occur at the locations expected from the simulation: around 5 `cm` on chromosome one and 33 `cm` on chromosome two.

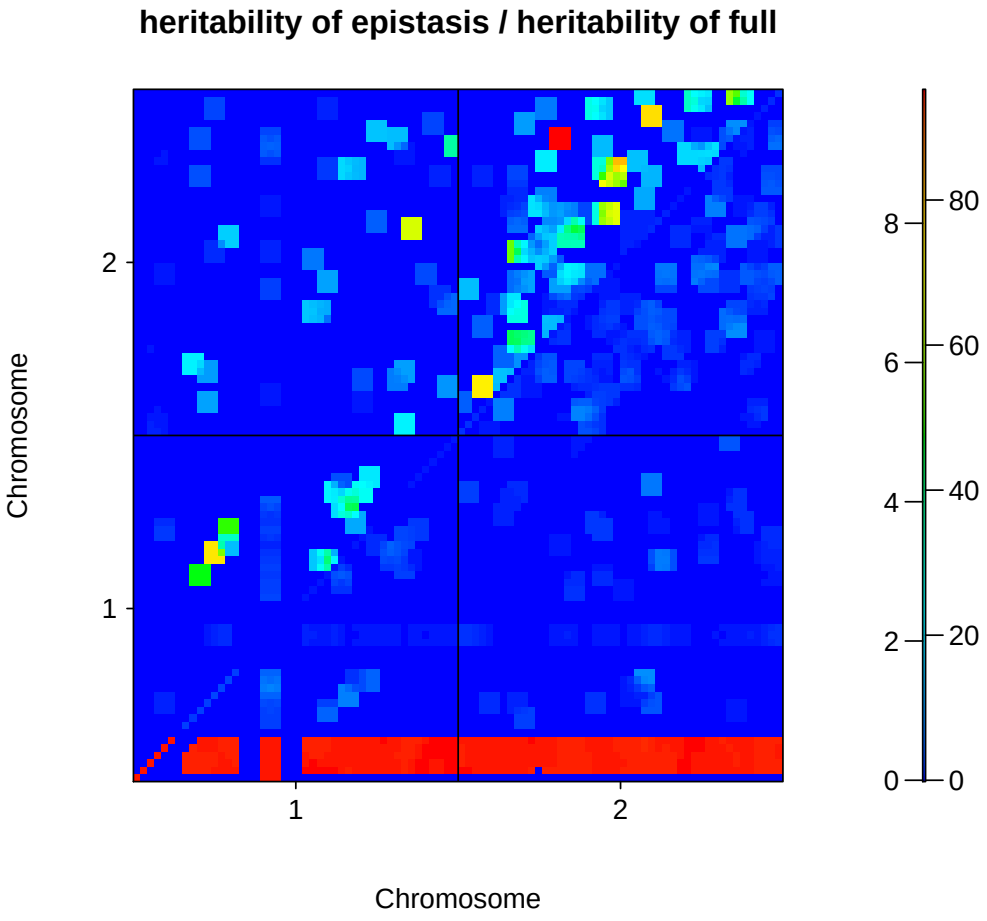

Figure 5: The plot of `qb.scantwo` with a main effect for a QTL at position 5 on chromosome 1 but no epistatic effect. Notice the long band in the lower triangle.

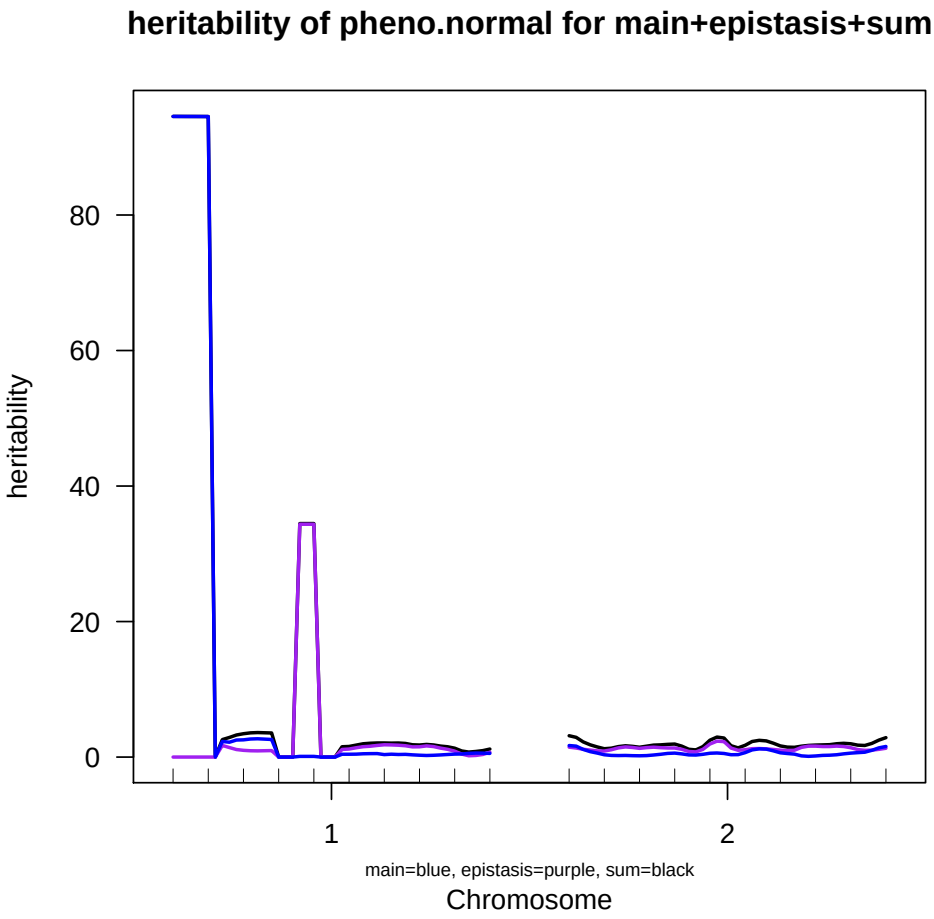

Figure 6: Heritability for simulated data with main effect but no epistasis examined under `qb.scanone`. As expected from the simulate there is a peak in the blue (main effects) curve near position five on chromosome one.

Notice that in Figure 5 the main effect is represented by the horizontal band at `5cm`. The corresponding `qb.scanone` results for `qbSim` are shown in Figure 6.

2.5.3 Simulating Data with Main Effects & Epistatic Effects.

As a final example of `qb.scantwo` operating on simulated data, we can take a look at data that involves both an epistatic effect and a main effect. We use the value of `qtl.main.model` above, which specifies a main effect and the original value of `qtl.epi.model`.

```
set.seed(1234)
sim <- qb.sim.cross(len=rep(100,2), n.mar=10, eq.spacing=TRUE,
  n.ind=100, type="bc", missing.geno=0.03, qtl.pos=qtl.positions,
  qtl.main=qtl.main.model, qtl.epis=qtl.epi.model)
```

Running `qb.sim.cross`, `qb.mcmc` and plotting the results of `qb.scantwo` gives an idealized plot of data with both main and epistatic effects. It is essentially an overlay of the previous two plots. We leave the details to the reader.

2.5.4 Running `qb.scantwo` on the `hyper` Data

To run qb.scan two on the hyper data set, we can use our previous results of the MCMC algorithm running on the `hyper` data.

```
> temp <- qb.scantwo(qbHyper, chr = c(4, 6, 15))
> summary(temp, digits = 2)

upper: heritability of bp for epistasis
lower: heritability of bp for full

        n.qtl l.pos1 l.pos2 lower u.pos1 u.pos2 upper
c4 :c4  0.298   28.4   31.7  22.7   14.2   47.0  5.04
c4 :c6  1.561   31.7   61.2  24.1   54.4   47.0  7.41
c4 :c15 0.446   32.8   27.5  23.9   14.2   51.7 20.13
c6 :c6  1.214   19.9   31.7  16.4   21.9   59.0  1.06
c6 :c15 1.080   61.2   19.5  18.0   59.0   19.5 10.27
c15:c15 0.105   21.5   23.5   6.4   13.1   31.5  3.18
```

Using the results from the two-dimensional qb.scans of the simple simulated data as a guide, the plot of `qb.scantwo` shows a main effect from a QTL on chromosome 4 and epistatic effects between the pairs of QTLs on chromosomes 4 and 15 and 6 and 15.

3 Types of Scan Summaries

We have created several types of scan summaries, illustrated below. These include the following LPD, heritability, variance components, parameter estimates, cell means, posterior probabilities and Bayes factors. Below we detail what these are and how they are calculated.

For each type, we can provide a summary scan, and in addition provide detail broken down by main effects, epistatic effects, and/or GxE (genotype by environment, or genotype by covariate) interactions. These breakdowns can be further divided into Cockerham (1954; see Kao and Zeng 2002) type effects (additive and dominance for main effects, or the four epistatic interactions of aa, ad, da, dd) if desired.

- `count` gives the count of the number of MCMC samples including this locus. Currently this can be viewed on a log scale using type `log10`.
- `posterior` is the Bayesian posterior probability, basically the `count` divided by the total number of MCMC samples.
- `BF` provides the Bayes factor comparing the model with and without this locus. It is more easily viewed as `2logBF`.
- `estimate` gives model parameter estimates for main effects, epistasis, and GxE interactions.

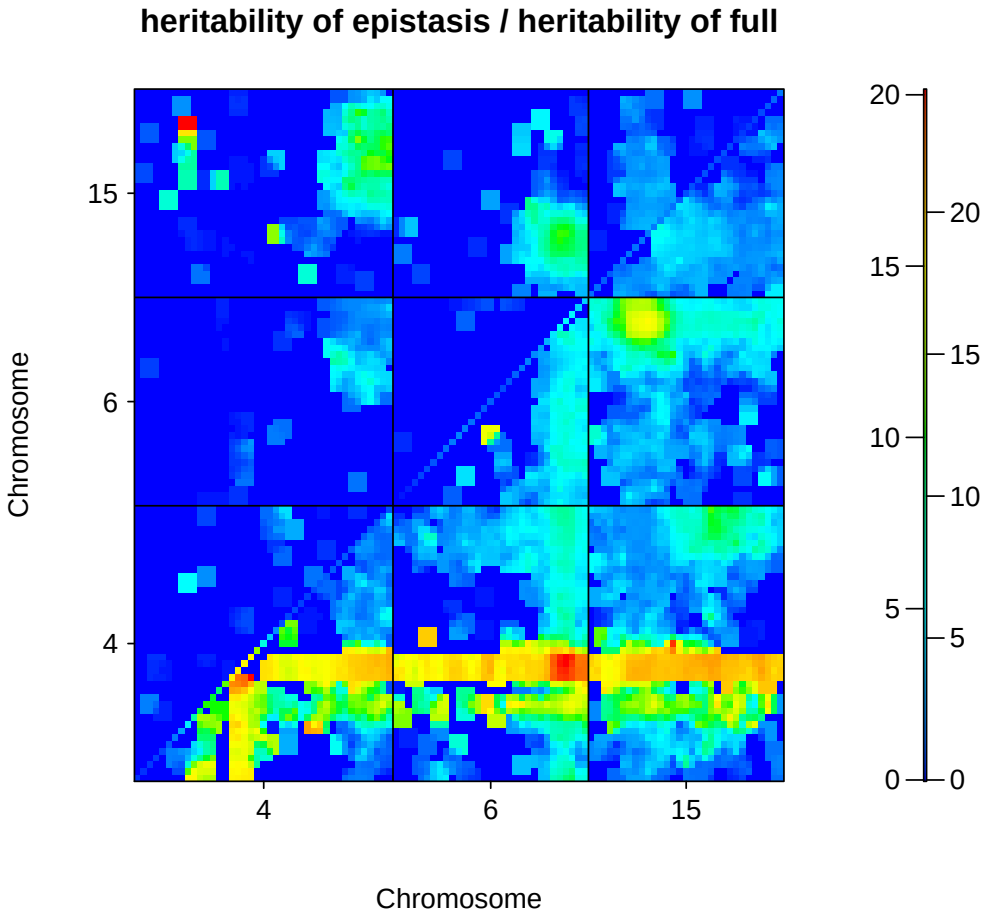

Figure 7: A plot of a `qb.scantwo` scan of the `hyper` data showing results for chromosomes 1, 4,6, and 15. Note the main effect from the QTL on chromosome 4 and the epistatic effect between the pairs of QTLs on chromosomes 4 and 15 and 6 and 15.

- `cellmean` provides marginal means at a locus, adjusted for all other model effects from other QTL and covariates.
- `variance` yields the variance components for QTL effects associated with a particular locus.
- `heritability` is actually at this point explained variation. In a future release we may distinguish Rsquared and idealized heritability.
- `LPD` is the log posterior density, adapted from Morton's (1995) log odds ratio (LOD) used in human genetics to LOD maps by Lander and Botstein (1989). The LPD for QTLs was introduced by Sen and Churchill (2001). It tests presence or absence of a QTL at a locus, adjusting for all other possible model effects (other QTL, epistasis and GxE). The `LPD`, the `LR` or likelihood ratio, and the `deviance` are detailed in the next section.
- `detection` is the posterior probability of detection of a QTL at a locus.

4 Theoretical Development

This section could be skipped. It is aimed at those quantitative folks who have read Yi et al. (2005) for the math and want to know more. Here we leave out details concerning covariates to simplify presentation.

Given complete data on genotypes for all individuals across the genome, we could consider a model relating phenotype y to genotype g through a design matrix X,

$$y = \mu + X\Gamma\beta + e \ .$$

The unknown effect parameters are the grand mean, μ, the effect parameters, β, and the unexplained variance, $\sigma^2 = V(e)$, which for convenience, we bring together as $\theta = (\mu, \beta, \sigma^2)$. The genetic architecture is specified by $\Gamma = \text{diag}\gamma$, which has values of 1 or 0 to indicate presence or absence, respectively, of the corresponding model effect. The QTL model could thus be written as $p(y|\gamma, X, \theta)$.

This genetic architecture specified by a 0-1 vector γ allows us to consider models of different dimensions, e.g. one vs. two QTL, without resorting to a more complicated (reversible jump) sampling scheme. The unknown values γ are the key device in sampling over many different possible genetic architectures, in terms of what loci λ are included and what gene action is important. There is some redundancy between γ and λ: a locus is in the model only if at least one γ associated with that locus is 1. Technically, we consider probabilities $p(\lambda|\gamma)$ that can only be 0 or 1 to indicate whether the loci, λ, are compatible with the genetic architecture, γ. While the loci are determined by the genetic architecture, γ is not completely determined by λ. We exploit this to make more efficient code and to build diagnostic summaries.

Recall from Yi et al. (2005) that the whole-genome genotype information, g, and the design matrix, X, are 1-1 mappings. In other words, $p(X|g)$ is either 0 or 1, depending on whether, for instance, the design is compatible with the genotypes.

4.1 Likelihood and posterior

In a classical setting, the full likelihood augmented by genotypes, g, over the genome is

$$p(y, g|m, \gamma, \theta) = p(y|\gamma, X, \theta)p(X|g)p(g|m, \lambda)p(\lambda|\gamma)\,,$$

with m the marker genotypes across the genome and $p(g|m, \lambda)$ the map function. At most loci, we do not fully know genotypes g, hence the likelihood given observable data is averaged over g,

$$L(\gamma, \theta|y, m) = \sum_g p(y, g|m, \gamma, \theta)\,.$$

With no QTL, we write $L(\mu|y)$ for the null likelihood.

In a Bayesian perspective, a prior $p(\gamma, \theta)$ is placed on the unknowns, and we study the posterior,

$$p(g, \gamma, \theta|y, m) \propto p(y, g|m, \gamma, \theta)p(\gamma, \theta)\,.$$

To study the unknown parameters of interest, (γ, θ), we average the posterior over the genotypes, or equivalently, form a weighted average of the augmented likelihood with weights proportional to the prior on (γ, θ),

$$p(\gamma, \theta|y, m) = \sum_g p(g, \gamma, \theta|y, m) \propto \sum_g p(y, g|m, \gamma, \theta)p(\gamma, \theta)\,.$$

4.2 Parameter estimation

Classically, the parameters of interest, (λ, θ), are estimated by maximizing the likelihood. This is usually done in a QTL setting by profiling the likelihood, or LOD (see below), with respect to one locus or two loci over the genome. We think of that here as profiling with respect to a given genetic architecture, γ, to find the maximum likelihood estimate (MLE) for β ,

$$\hat{\beta} = V\Gamma X^T y\,,$$

with $V = (\Gamma X^T X\Gamma)^{-1}$ and $\sigma^2 V$ the variance-covariance matrix for $\hat{\beta}$. Here we assume the columns of X are centered on zero, so the MLE for the reference is $\hat{\mu} = \bar{y}$.

Bayesian parameter estimates are typically found as the posterior means, which shrink $\hat{\mu}$ toward its prior mean μ_0 and $\hat{\beta}$ toward the prior mean of 0, leading to posteriors

$$\mu \sim N\left((1-b)\mu_0 + b\bar{y}, b\sigma^2/n.ind\right)\ ,$$

and

$$\beta \sim N\left(B\hat{\beta}, B\sigma^2 V\right)\ ,$$

with b and B being Bayesian shrinkage factors. As we gather more data, the Bayesian priors focus on the MLEs, i.e. b and B tend to 1. The likelihood and the posterior are both fairly symmetric around the maximum, for any given γ. Thus, the posterior mean and the MLE for β are very close in practice. This is less apparent from the summaries in the previous section, as the Bayesian estimates are attenuated by the putative effects of other QTL along the genome. This is a technical post-processing issue of properly sorting out the effects of multiple linked loci, which we intend to address in the next freeze. Notice in the second plot how closely the qb.scanone (solid red line) and scanone (dashed black line) profiles of the substitution effect agree near 100cM.

4.3 Variance components

Variance components can also be estimated in both approaches. The classical unbiased estimate for environmental variance is $\hat{\sigma}^2 = RSS(\hat{\theta})/df$, with $RSS(\theta) = \sum(y - \mu - X\Gamma\beta)^2$ and $df = n.ind - 1 - \sum\gamma$.

A Bayesian posterior estimate of σ^2 is its posterior mean, which is a weighted average of $RSS(\theta)/n.ind$ and its prior mean. Its empirical estimate can be found by averaging the posterior samples,

```
> summary(qb.scanone(qbSim, type = "variance", scan = "env"))

variance of pheno.normal for env

          n.qtl  pos m.pos  env
c1:c3.124  3.12 42.2  4.44 1.45
c2:c0.37   0.37 75.6 60.00 1.44
```

Heritability is computed as the percent of explained variation, $h^2 = 100(TSS - RSS(\theta))/TSS$, with $TSS = \sum(y - \bar{y})^2$ the total sum of squares. [The idealized variation would substitute expected fractions for the X^2 terms based on the type of cross.] We can find the posterior estimate of variability as the `main` entry below:

```
> summary(qb.scanone(qbSim, type = "heritability"))

heritability of pheno.normal for main,epistasis,sum

          n.qtl  pos m.pos e.pos  main epistasis   sum
c1:c3.124  3.12 42.2  4.44  42.2 94.54     34.36 34.46
c2:c0.37   0.37 75.6 60.00  75.6  1.21      1.64  1.72
```

4.4 LOD, LPD and BF

The classical approach introduced by Lander and Botstein (1989) profiles the likelihood only along the ridge of maximum β for each λ. That is, at each λ, find β that maximizes the LOD. The LOD map is a plot of this profile. The LOD statistic to assess QTL is

$$LOD(\lambda) = c + \log_{10}\left(\max_{\theta} L(\gamma, \theta|y, m)p(\lambda|\gamma)\right) ,$$

with the constant being $c = -\log_{10}(\max_{\mu} L(\mu|y))$. The likelihood ratio is $LR = 10^{LOD}$, and deviance is $D = 2\log(10)LOD$.

The Bayesian approach provides a direct estimate of the posterior as the histogram of the samples from the Markov chain Monte Carlo. Sen and Churchill (2002) proposed profiling the log posterior density, LPD, which involves averaging over the unknown parameters θ,

$$LPD(\lambda) = C + \log_{10}\left(\sum_{\theta} p(\gamma, \theta|y, m)p(\lambda|\gamma)\right) .$$

[The sum over θ is actually an multidimensional integral, but we ignore those details here.] Here the constant C would involve averaging over the null likelihood with respect to the prior on μ. In practice, LOD and LPD are often pretty close to each other and can be used interchangeably.

One advantage of sampling a large set of possible models by MCMC is that Bayes factors are easily computed. We do not have to resort to fancy harmonic means as in Newton and Raftery (199x). Instead, we construct marginal posterior histograms for models to be compared, and rescale by their priors. For instance, to compare two genetic architectures, we construct

$$BF = \frac{p(\gamma|y,m)/p(\gamma)}{p(0|y)/p(0)} \ ,$$

in which $p(0)$ is the prior on γ being all zero (no QTL at all) and $p(0|y)$ is the posterior. Actually, $p(0|y)/p(0) \propto p(y) = \sum_\mu p(y|\mu)p(\mu)$, with the sum really an integral over the real line. Often this is more interpretable on a log scale as $2\log(BF)$, which we can compute as

```
> summary(qb.scanone(qbSim, type = "2logBF"))

2logBF of pheno.normal for main,epistasis,sum

          n.qtl  pos m.pos e.pos main epistasis sum
c1:c3.124  3.12 42.2  4.44  42.2 4.86      15.9   0
c2:c0.37   0.37 75.6 60.00  75.6 0.00       0.0   0
```

4.5 Marginal Summaries

Our primary interest here is in marginal statistics. Consider that the model has genetic architecture γ that include loci λ . We want to ask what is the contribution to the model of some subset of indicators, γ_2 , associated with a locus, or a set of loci, λ_2 . We might ask this in a variety of ways, looking at evidence in terms of LOD or a related statistics, or the contribution in terms of variance components, heritability, or parameter effects. We can think of partitioning the genetic architecture into two components, $\gamma = (\gamma_1, \gamma_2)$, with a corresponding partition of the effect parameters,

$$\Gamma\beta = (\Gamma_1 + \Gamma_2)\beta \ .$$

The subset of effect parameters, $\beta_2 = \Gamma_2\beta$, may include, for instance, the main effects for locus λ_2 plus some or all epistatic effects that involve this locus. We can then ask questions about β_2, or about γ_2 and λ_2, adjusting for the presence of effects $\beta_1 = \Gamma_1\beta$. Note that β_1 could include some model parameters for λ_2.

4.5.1 Variance components

Here and through the rest of this document, we argue that we can characterize important diagnostic summaries using marginal properties of MCMC samples. The key technical argument is in the next paragraph. Namely, we can use the marginal variance components of our model fit, ignoring covariances, to construct approximate statistics.

If the columns of X are nearly orthogonal to each other, then the variance-covariance matrix for the effect parameter MLEs, $\mathrm{var}(\hat{\beta}) = \sigma^2 V$, would be *diagonally dominant*. That is, we suppose the variances along the diagonal are larger than the sum of the absolute covariances. Formally, with $v = \mathrm{diag}(V)$ and $V_{(j)}$ the j column of V ,

$$2v_{(j)} \geq \sum |V_{(j)}| \ .$$

In other words, we assume the covariances among effect estimates are negligible, and the diagonal values are approximately $v_{(j)} \approx \gamma_{(j)} / \sum X_{(j)}^2$, with $X_{(j)}$ the jth column of X . In this case we can approximate V by its diagonal, $D = \text{diag}(v)$, and get a good approximation of V^{-1} using D^{-1}:

$$V^{-1} = D^{-1}[I + O]^{-1} \ ,$$

with O being on the order of $(V - D)D^{-1}$. As long as the diagonal entries of D are large, then this approximation is good. Where these variances are small, the approximation is not so useful.

Since we are interested in learning about effects with larger variance components, this approximation seems quite workable in the present setting. It should a pretty reasonable between terms for unlinked loci, and under conditions of Hardy-Weinberg equilibrium among alleles at each locus. Note also that epistatic effects between linked loci will be addressed directly by construction of columns of X. [I believe the discrepancy of the diagonal can be readily checked under H-W by adding another `type` to the `qb.scan` routines–next freeze.]

With this approximation the explained variation can be approximated as

$$TSS - RSS(\theta) = \sum (X\Gamma\beta)^2 \approx \gamma^T r \, ,$$

with $r_{(j)} = \beta_{(j)}^2 \sum X_{(j)}^2$ being the variance explained by the jth component of the genetic architecture. Then the difference, $RSS(\theta_1) - RSS(\theta) \approx \gamma_2^T r = \sum r_2$, is simply the sum of variance components, which are readily stored for each MCMC iteration. Here, r_2 contains the elements of r corresponding to $\gamma_2 = 1$, and $\theta_1 = (\mu, \beta_1, \sigma^2)$.

Marginal heritability is computed as the additional variation explained by the genetic architecture γ_2 given γ_1 ,

$$h^2 = \frac{RSS(\theta_1) - RSS(\theta)}{TSS} = \frac{\gamma_2^T r}{TSS} \ .$$

4.5.2 LOD, LPD and BF

The adjusted LOD to compare the full model to the reduced model with $\gamma_2 = 0$ is

$$LOD(\gamma_2|\gamma_1) = \log_{10} \left(\frac{\max\limits_{\theta} L(\gamma, \theta|y, m)}{\max\limits_{\theta_1} L(\gamma_1, \theta_1|y, m)} \right) \ .$$

The adjusted LPD is similarly,

$$LPD(\gamma_2|\gamma_1) = \log_{10} \left(\sum_{\theta} \frac{p(\gamma, \theta|y, m)}{p(\gamma_1, \theta_1|y, m)} \right) \ ,$$

with again the sum actually being an integral over θ .

In the case of normal data and complete marker information, the LOD reduces to

$$LOD(\gamma_2|\gamma_1) = \frac{n.ind}{2} \log_{10} \left(\frac{\min\limits_{\theta_1} RSS(\theta_1)/df_1}{\min\limits_{\theta} RSS(\theta)/df} \right) \ ,$$

with degrees of freedom, $df = n.ind - 1 - \sum \gamma$, and $df_1 = n.ind - 1 - \sum \gamma_1$. The LPD follows a similar form, but involving an average (or really, integral) over θ,

$$LPD(\gamma_2|\gamma_1) = \frac{n.ind}{2} \log_{10} \left(\sum_{\theta} \frac{RSS(\theta_1)/df_1}{RSS(\theta)/df} \right) .$$

The Bayes factors are easily computed, as noted earlier. To compare the two genetic architectures γ and γ_1, we construct

$$BF = \frac{p(\gamma|y,m)/p(\gamma)}{p(\gamma_1|y,m)/p(\gamma_1)} .$$

Often this is more interpretable on a log scale as $2\log(BF)$, which we can compute as

4.6 Model Averaging Algorithm

Here we briefly describe the model averaging idea. The MCMC samples include a wide variety of models, indexed by γ. The 1-D and 2-D scans first compile a selected diagnostic for each sample (also known as an iteration). That is, at each genome position, or pair of positions, we average the values for samples that include that position, i.e. have $\gamma = 1$ at that position. The posterior is simply an average of the γ samples at each position.

These samples are kept for each model component, either in terms of the unaggregated Cockerham (1954) partition or in terms of `main` effects and `epistasis`, and for the `sum` of these components. There are some mechanics involved. For instance, for 1-D averages involving epistasis, we want to count each pair for both loci, and for 2-D averages, we want to count epistatic effects separately at each locus. But these are details that can be found by looking at the code if interested.

Chromosome summaries, or summaries within regions of chromosomes, are found as weighted averages of these per-position summaries. The weights are naturally the number of MCMC samples per position. At present the code does not separate out multiple loci on a chromosome [next freeze].

With small or moderate MCMC sample sizes, the 1-D and 2-D scans can be rather rough, or jagged. We have found nearest neighbor smoothing to be helpful. That is, a position is equally weighted against the sum of its neighbors, accounting for number of MCMC samples. This can be repeated several times (e.g. `smooth = 3`) to further local smoothing.

5 Summary

In this tutorial we have explored the use of the Bayesian scan routines `qb.scanone` and `qb.scantwo` as techniques for exploring the genetic architecture for a phenotypic trait. Through examples using both simulated and experimental data we have demonstrated the key steps in identifying both main and epistatic effects. Further information on using using `R/qtlbim` to explore the `hyper` data set can be found in the *hyperpaper* vignette. Inorder to view the vignette you can simply type

```
vignette(topic="hyperpaper",package="qtlbim")
```

at the R prompt. A demo for a simple analysis of `hyper` can be accessed by typing `demo(qb.hyper.tour)` after the `R/qtlbim` library has been loaded.

```
[1] TRUE
```

References

Cockerham CC (1954) An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics 39: 859-882.

Kao CH, Zeng ZB (2002) Modeling epistasis of quantitative trait loci using Cockerham's model. Genetics 160: 1243-1261.

Morton NE (1995) LODs past and present. Genetics 140: 7-12.

Newton MA, Raftery AE (1994) Approximate Bayesian inference by the weighted likelihood bootstrap (with Discussion). Journal of the Royal Statistical Society, series B, 56, 3-48.

Sen S, Churchill GA (2001) A statistical framework for quantitative trait mapping. Genetics 159: 371-387.

Sugiyama F, Churchill GA, Higgens DC, Johns C, Makaritsis KP, Gavras H, Paigen B (2001) Concordance of murine quantitative trait loci for salt-induced hypertension with rat and human loci. Genomics 71: 70-77.

Wright FA, Kong A (1997) Linkage mapping in experimental crosses: the robustness of single-gene models. Genetics 146: 417-425.