R/qtlbim: QTL with Bayesian Interval Mapping in Experimental Crosses

Brian S. Yandell[1,2], **Tapan Mehta**[3], **Samprit Banerjee**[3], **Daniel Shriner**[3], **Ramprasad Venkataraman**[3], **Jee Young Moon**[1], **W. Whipple Neely**[1], **Hao Wu**[4], **Randy von Smith**[5], and **Nengjun Yi**[3]

Brian S. Yandell: byandell@wisc.edu; Nengjun Yi: nyi@ms.soph.uab.edu

[1]Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53706, USA

[2]Department of Horticulture, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53706, USA

[3]Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294, USA

[4]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD 21205 USA

[5]The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA

Summary

R/qtlbim is an extensible, interactive environment for the Bayesian Interval Mapping of QTL, built on top of R/qtl (Broman *et al.* 2003), providing Bayesian analysis of multiple interacting quantitative trait loci (QTL) models for continuous, binary and ordinal traits in experimental crosses. It includes several efficient Markov chain Monte Carlo (MCMC) algorithms for evaluating the posterior of genetic architectures, *i.e.* the number and locations of QTL, their main and epistatic effects, and gene-environment interactions. R/qtlbim provides extensive informative graphical and numerical summaries, and model selection and convergence diagnostics of the MCMC output, illustrated through the vignette, example and demo capabilities of R (R Development Core Team 2006).

1 INTRODUCTION

The freely available QTL mapping package, R/qtlbim (www.rqtlbim.org), provides a comprehensive framework for Bayesian model selection of the genetic architecture of complex traits in experimental crosses. Classical approaches to model selection in QTL mapping, such as multiple interval mapping in QTL Cartographer (Basten, Weir and Zeng 2002), largely rely on stepwise model selection with separate fits to each possible model. The Bayesian approach has the advantage of sampling across the more probable models, and

providing graphical summaries that can compare many models at once. R/qtlbim can infer multiple QTL in the presence of epistasis (gene-gene interaction) and gene-environment interactions.

R/qtlbim is built on the widely used R/qtl framework (Broman *et al.* 2003), which provides many graphical tools for data checking and classical model selection. R/qtlbim shares the extensibility features of R/qtl. Computationally intensive algorithms are written in C, with data manipulation and graphics in R. R/qtlbim is available across Window, Linux and Mac OSX platforms and accepts a variety of input formats via R/qtl.

2 MCMC TECHNOLOGY

Central to R/qtlbim is the Markov chain Monte Carlo (MCMC) technology (Yi 2004; Yi *et al.* 2004; Yi *et al.* 2005). MCMC samples are drawn from the posterior distribution of genetic architecture, including number and location of genetic loci, gene action effects at all loci, epistatic interactions between pairs of loci, fixed and random covariates, and gene-environment interactions. These MCMC samples are then summarized and interpreted with graphs to infer key aspects of the genetic architecture.

MCMC provides a mechanism to study the full Bayesian posterior distribution for any particular genetic architecture. Further, using a prior to allow uncertainty of genetic architecture leads to MCMC samples across multiple genetic architectures. This brings model selection formally into a Bayesian framework in which the genetic architecture is just another parameter to be estimated.

3 FEATURES

R/qtlbim contains several efficient MCMC algorithms to search for genetic architectures that are most probable. It includes graphical and tabular summaries that assess the contribution of individual loci and pairs of loci while adjusting for effects of all other possible loci and covariates via model averaging.

Graphical tools allow the user to examine the MCMC samples by loci and genotypic effects in a variety of ways. They include estimation of Bayes factors for model selection on the number of QTL, the pattern of QTL across chromosomes, and the patterns of epistatic and gene-environment interactions. High posterior density (HPD) regions provide estimates of QTL analogous to LOD support intervals. Suggestions for model selection are provided in vignettes, examples and demos using the R (R Development Core Team 2006) interactive documentation system. The primary vignette, *qtlbim.pdf*, gives an overview of the package. Each command has a help page and corresponding example, which can be adapted to the user's own data. The command *qb.demo* leads the user through several demonstrations of the package capabilities.

The *scan.pdf* vignette provides more detail on R/qtlbim extensions of typical interval mapping scans. The R/qtl package (Broman *et al.* 2003) offers genome scans of the classical log odds (LOD) of Lander and Botstein (1989) and the Bayesian log posterior density (LPD) of Sen and Churchill (2001). These only consider the effect of one (*scanone*) or two

(*scantwo*) QTL on a phenotype. R/qtlbim's scan routines, principally *qb.scanone* and *qb.scantwo*, use R/qtl's generic plot routines for 1-D and 2-D scans, respectively. However, our philosophy differs in important ways beyond changes in the calling sequence. Our scans consider the contribution of a given locus, or pair of loci, to the LPD after adjusting for all other possible QTL and covariates by model averaging over all genetic architectures that contain the QTL(s) being examined. These marginal scans are partitioned into contributions from main effects, epistatic interactions and gene-environment interactions. Another important distinction is that we provide the facility to estimate marginal heritability, Bayes factors, means by genotypes, and other features in addition to LPD.

A third vignette, *hyperslide.pdf*, offers a way to automate model selection for genetic architecture of a complex trait. By default, it analyzes the hypertension data of Sugiyama *et al.* (2001) illustrated below. However, this vignette is an interactive object that can be reconfigured with another dataset using the *qb.sweave* command, via the Sweave framework (Leisch 2002). Sweave is an advanced feature using and the LaTeX (www.latex-project.org) type-setting system, which must be separately installed.

An example of R/qtlbim is provided in Figure 1 using the salt-induced hypertension data of Sugiyama *et al.* (2001) that is available in R/qtl. Note the improvement in LPD over R/qtl scanone by (1) adjusting for affects of all other possible QTL and (2) providing marginal evidence for epistasis.

4 FUTURE DEVELOPMENT

R/qtlbim is under continual development. Future plans include proper treatment of the X chromosome (Broman *et al.* 2006) and extensions to correlated traits and experimental crosses derived from multiple inbred lines and outbred populations. We are also investigating ways to assess false discovery and check the fit of a model to data and prior assumptions. More extensive graphics for gene-environment interactions and for ordinal traits are planned. We intend to build on the graphical user interface for R/qtl that is under development, and we are in close communication with the R/qtl development team.

Acknowledgments

References

Basten, CJ.; Weir, BS.; Zeng, ZB. QTL Cartographer, Version 1.16. Department of Statistics, North Carolina State University; Raleigh, NC: 2002.

Broman KW, Sen Œ, Owens SE, Manichaikul A, Southard-Smith EM, Churchill GA. The X chromosome in quantitative trait locus mapping. Genetics. 2006; 00:000–000. dx.doi.org/10.1534/genetics.106.064311.

Broman KW, Wu H, Sen Œ, Churchill GA. R/qtl: QTL mapping in experimental crosses. Bioinformatics. 2003; 19:889–890. www.rqtl.org. [PubMed: 12724300]

Leisch, F. Dynamic generation of statistical reports using literate data analysis. In: Härdle, W.; Rönz, B., editors. Compstat 2002 - Proceedings in Computational Statistics; Heidelberg, Germany: Physika Verlag; 2002. p. 575-580.

Lander ES, Botstein D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics. 1989; 121:185–199. [PubMed: 2563713]

R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2006. www.R-project.org

Sen Œ, Churchill GA. A statistical framework for quantitative trait mapping. Genetics. 2001; 159:371–387. [PubMed: 11560912]

Sugiyama F, Churchill GA, Higgens DC, Johns C, Makaritsis KP, Gavras H, Paigen B. Concordance of murine quantitative trait loci for salt-induced hypertension with rat and human loci. Genomics. 2001; 71:70–77. [PubMed: 11161799]

Yi N. A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. Genetics. 2004; 167:967–975. [PubMed: 15238545]

Yi N, Xu S, George V, Allison DB. Mapping multiple quantitative trait loci for ordinal traits. Behavior Genetics. 2004; 34:3–15. [PubMed: 14739693]

Yi N, Yandell BS, Churchill GA, Allison DB, Eisen EJ, Pomp D. Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. Genetics. 2005; 170:1333–1344. [PubMed: 15911579]
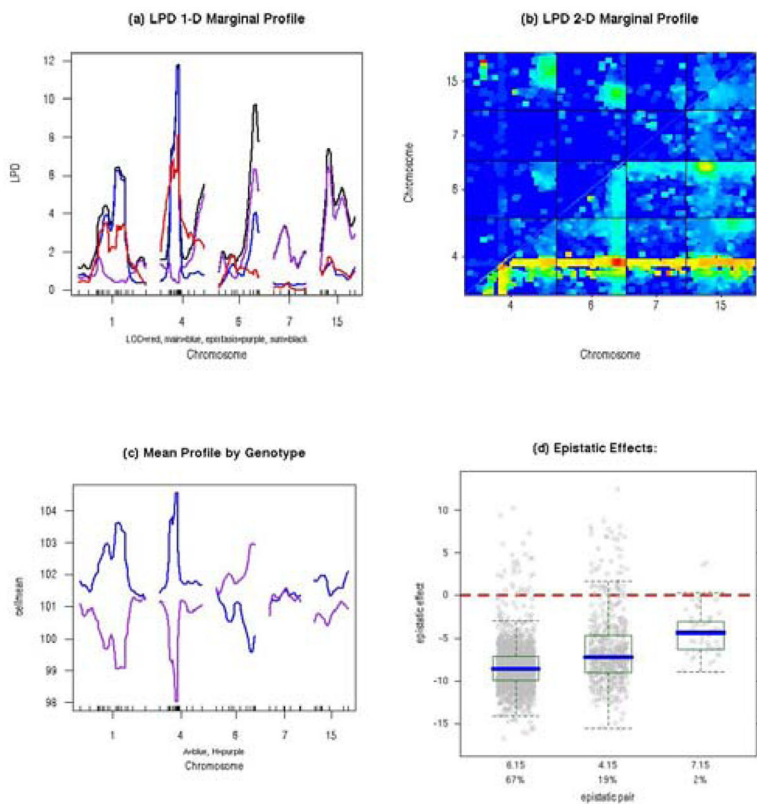
Figure 1.
Example graphs from R/qtlbim based on the backcross of Sugiyama *et al.* (2001). **(a)**
Marginal LPD one-dimensional profile scan on chromosomes 1, 4, 6, 7, 15 (red = 1-QTL
LPD scan from R/qtl, blue = LPD marginal scan for main effects, purple = LPD marginal
scan for epistasis, black = LPD marginal for all effects of QTL); (b) marginal 2-D genome
scan for LPD on chromosomes 4, 6, 7 and 15; (c) marginal means by genotype profiled
across chromosomes; (d) epistatic effects for chromosome pairs 6.15 (chr 6 and 15, with
67% of MCMC samples included this pair), 4.15 (19%) and 7.15 (2%).