BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 23 2010, pages 2990–2992
doi:10.1093/bioinformatics/btq565

Genetics and population analysis

Advance Access publication October 21, 2010

R/qtl: high-throughput multiple QTL mapping

Danny Arends[1,†], Pjotr Prins[1,2,†], Ritsert C. Jansen[1] and Karl W. Broman[3,*]

[1]Groningen Bioinformatics Centre, University of Groningen, Groningen, [2]Department of Nematology, Wageningen University, Wageningen, The Netherlands and [3]Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: R/qtl is free and powerful software for mapping and exploring quantitative trait loci (QTL). R/qtl provides a fully comprehensive range of methods for a wide range of experimental cross types. We recently added multiple QTL mapping (MQM) to R/qtl. MQM adds higher statistical power to detect and disentangle the effects of multiple linked and unlinked QTL compared with many other methods. MQM for R/qtl adds many new features including improved handling of missing data, analysis of 10 000 s of molecular traits, permutation for determining significance thresholds for QTL and QTL hot spots, and visualizations for *cis–trans* and QTL interaction effects. MQM for R/qtl is the first free and open source implementation of MQM that is multi-platform, scalable and suitable for automated procedures and large genetical genomics datasets.

Availability: R/qtl is free and open source multi-platform software for the statistical language R, and is made available under the GPLv3 license. R/qtl can be installed from http://www.rqtl.org/. R/qtl queries should be directed at the mailing list, see http://www.rqtl.org/list/.

Contact: kbroman@biostat.wisc.edu

Received on June 16, 2010; revised on September 15, 2010; accepted on September 30, 2010

1 INTRODUCTION

R/qtl is an extensible, interactive environment for the mapping of quantitative trait loci (QTL) in experimental crosses. It is implemented as an add-on package for the freely available and widely used statistical language/software R (R Development Core Team, 2010). Since its introduction, R/qtl (Broman *et al.*, 2003) has become a reference implementation with an extensive guide on QTL mapping (Broman and Sen, 2009). R/qtl development is continuous, with input from multiple collaborators and users. We have introduced a full testing environment with regression testing, updated the license to the GPL version 3 and hosted the source code repository on Github, which gives R/qtl software development high visibility and transparency. The development of R/qtl reflects trends in quantitative genetics, in particular the use of larger datasets, larger calculations and requirements for controlling the false discovery rate (FDR). These developments are partly driven by high-throughput genetical genomics—the name coined for the study of gene expression QTL (eQTL; Jansen and Nap, 2001), metabolite QTL (mQTL) and protein QTL (pQTL).

Multiple QTL Mapping (MQM) belongs to a family of QTL mapping methods, that include Haley–Knott regression (Haley and Knott, 1992) and composite interval mapping (CIM; Zeng, 1994). MQM combines the strengths of generalized linear model regression with those of interval mapping (Jansen, 1993; Jansen and Stam, 1994). Recent developments in QTL mapping include Bayesian modelling of multiple QTL [e.g. R/qtlbim package (Banerjee *et al.*, 2008; Yandell *et al.*, 2007)]. Bayesian modelling, however, is computationally expensive, and arguably has little additional power when applied to high density maps, and (nearly) complete genotype data (Jansen, 2007). Still, we intend to combine the strengths of the different methods in future versions of R/qtl.

MQM provides a practical, relevant and sensitive approach for mapping QTL in experimental populations. The theoretical framework of MQM was introduced and explored by one of us (Jansen, 1994) and explained in the 'Handbook of Statistical Genetics' (Jansen, 2007). MQM has one known commercial implementation (Van Ooijen *et al.*, 2002), which has been used effectively in practical research, resulting in hundreds of papers [e.g. in mouse, plant, and fish, respectively (de Mooij-van Malsen *et al.*, 2009; Jeuken *et al.*, 2009; Kitano *et al.*, 2009)]. Now, with MQM for R/qtl, we present the first free and open source implementation of MQM, that is multi-platform, scalable and suitable for automated procedures and large datasets.

2 FEATURES

MQM for R/qtl is an *automated* three-stage procedure in which, in the first stage, missing genotype data are 'augmented'. (In other words, rather than guessing one likely genotype, multiple genotypes are modelled with their estimated probabilities.) In the second stage, important marker cofactors are selected by multiple regression and backward elimination. In the third stage, a QTL is moved along the chromosomes using these preselected markers as cofactors. QTL are interval mapped using the most informative model through maximum likelihood. A refined and automated procedure for cases with large numbers of marker cofactors is included. The method lets users test different QTL models by elimination of non-significant cofactors. MQM for R/qtl brings the following advantages to QTL mapping: (1) higher power, as long as the QTL explain a reasonable amount of variation; (2) protection against over-fitting, because MQM fixes the residual variance from the full model, which allows the use of more cofactors than may be used in, for example, CIM (Zeng, 1994); (3) prevention of ghost QTL detection (between two QTL in coupling phase); and (4) detection of negating QTL (QTL in repulsion phase).

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.

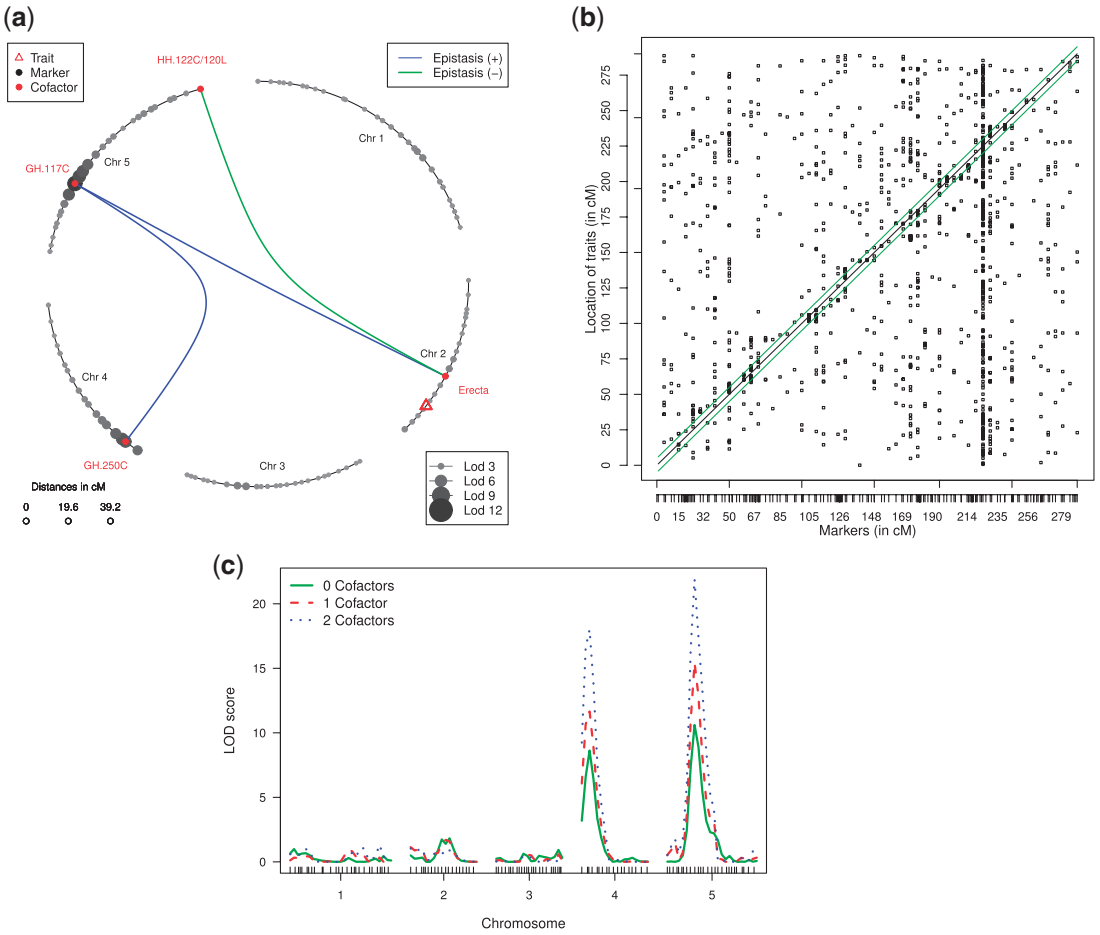

Fig. 1. Three examples of MQM plots included in R/qtl. (**a**) Circular genome interaction plot of the *Arabidopsis thaliana* glucosinolate pathway (Fu *et al.*, 2007). Logarithm of odds (LOD) scores shown at marker positions are scaled (grey circles), with selected cofactors (red circles) and epistasis between multiple cofactors (green and blue splines). (**b**) *Cis–trans* plot of significant QTL (squares) showing *cis*-acting QTL (diagonal) and a *trans*-band (vertical, chromosome 5) in *Caenorhabditis elegans* (Li *et al.*, 2006). (**c**) Three-way comparison of MQM performance in *Arabidopsis thaliana* (Fu *et al.*, 2007). LOD score increases when cofactors are added manually to the model. Here, adding more than two cofactors does not improve the model any further (as discussed in the online MQM tutorial).

MQM for R/qtl brings additional advantages to genetical genomics datasets with hundreds to millions of traits: (5) a pragmatic permutation strategy for controlling the FDR and prevention of locating false QTL hot spots, as discussed in Breitling *et al*. (2008). Marker data are permuted, while keeping the correlation structure in the trait data; (6) high-performance computing by scaling on multi-CPU computers, as well as clustered computers, by calculating phenotypes in parallel, through the message passing interface (MPI) of the SNOW package for R (Tierney *et al*., 2009); (7) visualizations for exploring interactions in a genomic circle plot (Fig. 1a) and *cis*- and *trans*-regulation (Fig. 1b).

A 40-page tutorial for MQM explores, both the automated procedure, and the manual procedure of adding and removing cofactors, in an *Arabidopsis thaliana* recombinant inbred line (RIL) metabolite (mQTL) dataset with 24 metabolites as phenotypes (Fu *et al.*, 2007). In addition, the tutorial visually explains the effects of data augmentation, cofactor selection, model selection and tweaking of input parameters, such as cofactor significance. Genetic interactions (epistasis) are explored through effect plots, and an example is given of parallel computation. The tutorial is part of the software distribution of R/qtl and is available online.

3 CONCLUSION

MQM for R/qtl is a significant addition to the QTL mapper's toolbox. R/qtl provides the user with the most frequently used statistical analysis methods: single-marker analysis, interval mapping, Haley–Knott regression (Haley and Knott, 1992), CIM (Zeng, 1994) and MQM (Jansen, 1994). MQM has improved handling of missing data and allows more powerful and precise detection of QTL, compared with many other methods. Not only is this new implementation of MQM available in the statistical R environment, which allows scripting for pipe-lined setups, but it is also highly scalable through parallelization and paves the way for high-throughput QTL analysis. With MQM, R/qtl is a free and high-performance comprehensive QTL mapping toolbox for the analysis of experimental populations. R/qtl now includes permutation strategies for determining thresholds

of significance relevant for QTL and QTL hot spots, the first step towards causal inference and network analysis.

Funding: National Institutes of Health (GM074244 to K.W.B.); the Netherlands Organisation for Scientific Research/TTI Green Genetics (1CC029RP to P.P.); the Centre for BioSystems Genomics (CBSG) and the Netherlands Consortium of Systems Biology (NCSB), both of which are part of the Netherlands Genomics Initiative / Netherlands Organisation for Scientific Research [to DA]; the EU 7th Framework Programme under the Research Project PANACEA (222936 to R.J.).

Conflict of Interest: none declared.

REFERENCES

Banerjee,S. *et al*. (2008) Bayesian quantitative trait loci mapping for multiple traits. *Genetics*, **179**, 2275–2289.

Broman,K.W. and Sen,S. (2009) *A Guide to QTL Mapping with R/qtl*. Springer, New York.

Broman,K.W. *et al*. (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, **19**, 889–890.

de Mooij-van Malsen,J.G. *et al*. (2009) Evidence for epigenetic interactions for loci on mouse chromosome 1 regulating open field activity. *Behav. Genet.*, **39**, 176–182.

Fu,J. *et al*. (2007) MetaNetwork: a computational tool for the genetic study of metabolism. *Nat. Protocols*, **2**, 685–694.

Haley,C.S. and Knott,S.A. (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, **69**, 315–324.

Jansen,R.C. (1993) Interval mapping of multiple quantitative trait loci. *Genetics*, **135**, 205–211.

Jansen,R.C. (1994) Controlling the type I and type II errors in mapping quantitative trait loci. *Genetics*, **138**, 871–881.

Jansen,R.C. (2007) Quantitative trait loci in inbred lines. In Balding,D.J. *et al*. (eds) *Handbook of Statistical Genetics*, John Wiley & Sons, Ltd., New York, pp. 589–622.

Jansen,R.C. and Nap,J.P. (2001) Genetical genomics: the added value from segregation. *Trends Genet.*, **17**, 388–391.

Jansen,R.C. and Stam,P. (1994) High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, **136**, 1447–1455.

Jeuken,M.J. *et al*. (2009) Rin4 causes hybrid necrosis and race-specific resistance in an interspecific lettuce hybrid. *Plant Cell*, **21**, 3368–3378.

Kitano,J. *et al*. (2009) A role for a neo-sex chromosome in stickleback speciation. *Nature*, **461**, 1079–1083.

Li,Y. *et al*. (2006) Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. *PLoS Genet.*, **2**, e222.

R Development Core Team (2010) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Van Ooijen,J.W. *et al*. (2002) MapQTL 4.0, Software for the Calculation of QTL Position on Genetic Maps. Availabe at http://www.kyazma.nl/index.php/mc.MapQTL/.

Yandell,B.S. *et al*. (2007) R/qtlbim: QTL with Bayesian interval mapping in experimental crosses. *Bioinformatics*, **23**, 641–643.

Zeng,Z.B. (1994) Precision mapping of quantitative trait loci. *Genetics*, **136**, 1457–1468.