BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 5 2011, pages 727–729
doi:10.1093/bioinformatics/btq719

Genetics and population analysis

Advance Access publication January 8, 2011

# R/mpMap: a computational platform for the genetic analysis of multiparent recombinant inbred lines

B. Emma Huang[1,2,*] and Andrew W. George[1,2]

[1]Division of Mathematics, Informatics and Statistics and [2]Food Futures National Research Flagship, CSIRO, St Lucia, QLD 4067, Australia

Associate Editor: Alex Bateman

## ABSTRACT

**Summary:** Multiparent crosses of recombinant inbred lines provide opportunity to map markers and quantitative trait loci (QTL) with much greater resolution than is possible in biparental crosses. Realizing the full potential of these crosses requires computational tools capable of handling the increased statistical complexity of the analyses. R/mpMap provides a flexible and extensible environment, which interfaces easily with other packages to satisfy this demand. Functions in the package encompass simulation, marker map construction, haplotype reconstruction and QTL mapping. We demonstrate the easy-to-use features of mpMap through a simulated data example.

**Availability:** www.cmis.csiro.au/mpMap.

**Contact:** emma.huang@csiro.au

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

Received on September 16, 2010; revised on December 9, 2010; accepted on December 14, 2010

## 1 INTRODUCTION

Multiparent recombinant inbred line (RIL) crosses are a new type of experimental design. Already, they are being implemented in mice (Complex Trait Consortium, 2004), Arabidopsis (Kover *et al.*, 2009), rice and wheat. Their growing popularity stems from their potential for increased mapping resolution and their ability to capture a wide range of genetic and phenotypic diversity from a breeding population. As yet, however, there are few computational tools for the analysis of these data.

Existing analysis software for biparental crosses cannot be applied directly or extended easily to these new designs. Analysis software for biparental crosses relies on being able to unambiguously infer the parental origin of allelic information from the observed marker data. Since this does not hold for multiparent crosses, analysis software for multiparent crosses must contend with all patterns of allelic segregation, which are consistent with the observed marker data and their associated probabilities.

Two software packages have been developed for the analysis of data from multiparent RIL crosses—R/qtl (Broman *et al.*, 2003) and R/happy (Mott *et al.*, 2000). Both packages focus primarily on quantitative trait loci (QTL) mapping. In contrast, we have developed mpMap as a comprehensive suite of functions for multiparent designs in close consultation with plant breeders and geneticists. In addition to native functions, mpMap interfaces with both R/qtl and R/happy and is easily extensible. This versatile software helps meet the urgent need for analysis tools capable of handling the complexities associated with multiparent inbred line crosses.

*To whom correspondence should be addressed.

## 2 IMPLEMENTATION

mpMap is implemented as a software package in *R* consisting of functions for a range of genetic analyses in multiparent crosses. Functionality in four main areas is described below. The (base) data structure for a multiparent cross is defined as an *R* object of class mpcross. This object includes the observed marker data from the founders and final inbred lines in addition to the cross pedigree and trait data. By calling the summary function, information on the percentage of missing data, potential genotyping errors and segregation ratios can be accessed and used to preprocess the data. The mpcross object can either be created within *R* or imported from files and can be exported in R/qtl and R/happy formats.

### 2.1 Simulation

Simulation is a valuable tool for both assessing the performance of analysis methods and determining the statistical limitations of a proposed genetic study. We provide functions for the simulation of $2^n$-parent inbred line crosses. Markers and quantitative trait data are simulated on the experimental cross under a user-specified map. A quantitative trait value is calculated for each observed offspring by summing its QTL effects with a random deviate. The random deviate is drawn from a normal distribution with mean 0 and environmental variance $\sigma_e^2$.

### 2.2 Marker map construction

High resolution, high-quality linkage maps are critical to an understanding of genomic structure. We provide a semiautomated approach to marker map construction using the following basic algorithm. The algorithm consists of three steps, and we provide functions to assist with each step.

In Step 1, the maximum likelihood estimate of the recombination between a pair of loci is calculated for each loci pair. The maximized (two-point) likelihood is constructed from haplotype probabilities given in Broman (2005).

In Step 2, loci are grouped and ordered automatically. First, markers are assigned to the same linkage group if their recombination fractions are $<\theta^*$ and the associated logarithm of odds (LOD) scores are $>L^*$. Default values for $\theta^*$ and $L^*$ are 0.15

and 5.0. Second, linked loci are ordered using R/seriation (Hahsler *et al.*, 2009). This package uses techniques such as simulated annealing, multidimensional scaling, hierarchical clustering and a Travelling Salesman Solver to order a distance matrix. We calculate the order using each technique and select that which minimizes the total chromosome length. An alternate criterion would be to maximize the sum of the adjacent two-point likelihoods; this can be done by selecting the option of 'lkhdsum' for ordering criterion.

In Step 3, we refine the marker map generated from Step 2 using both three-point and multipoint probabilities (Broman, 2005). The three-point ordering permits the inclusion of markers which might have been set aside at an early stage due to segregation distortion or missing data. Given a marker and a framework map, it selects the maximum likelihood estimator of the marker position based on the probability of that marker lying between the two flanking marker loci. Multipoint ordering borrows functionality from R/qtl to compute log likelihoods for all possible orderings of markers within a specified window. The best order is that with the largest log likelihood. The computational cost of this process increases dramatically with window size, so that in practice it is only feasible to consider small windows. This restricts the search for the best order to local permutations of the input order.

### 2.3 Haplotype reconstruction

Knowing the population's haplotype structure allows us to identify highly recombinant progenies, detect recombination hot spots and assess mapping resolution. Haplotypes are constructed from marker data by identifying the parental origin of the marker alleles. However, the presence of multiple founders makes direct observation of parental origin difficult. Note that these haplotype probabilities are also the basis for imputing missing marker data and for mapping QTL. We assign a putative parental origin at a given locus if the probability of that founder exceeds a given threshold (such as 80%).

Three different approaches for estimating haplotype probabilities are possible. First, three-point haplotype probabilities can be computed conditionally on the observed marker data at flanking markers. This is done directly within mpMap and is based on the three-point probabilities described in Broman (2005). Second, multipoint haplotype probabilities can be computed from information on all available linked marker loci. A front end to this computation in R/qtl is provided in mpMap. Third, multipoint haplotype probabilities can be computed using dynamic programming (Mott *et al.*, 2000). The independence from knowledge of the pedigree makes this method the most general, but may impact performance in cases where the additional information from the pedigree is available. We provide a front end to R/happy to compute these probabilities.

### 2.4 QTL mapping

Detection of gene–trait relationships is a primary goal for most breeding experiments, but most methods for QTL mapping have not or cannot be generalized to accommodate multiparent designs. Currently, we have implemented (composite) interval mapping in mpMap using the haplotype probabilities described above. For each position at which probabilities are computed, we use a regression approach to estimate QTL effects for each founder. The Wald statistic for the joint significance of all founder QTL effects is used to detect QTL. Marker covariates are selected for inclusion in the linear model using the Akaike Information Criterion (AIC) in a forward selection process. Significance thresholds for the composite interval mapping results are obtained empirically by simulating the null distribution.

We have limited the QTL mapping methods in R/mpMap to interval mapping-based approaches. However, when combined with the ASReml-R (Gilmour *et al.*, 2009) package, functions exist to perform interval mapping in a mixed model context and thereby include environmental and pedigree effects in the analysis. QTL mapping in R/happy is also limited to a regression on founder alleles; while R/qtl contains functions for more sophisticated QTL mapping, these have not yet been extended to multiparent RIL designs. Unlike R/mpMap, neither R/qtl nor R/happy has the capacity to accommodate linear mixed models.

## 3 SIMULATED DATA

We use the simulation features of mpMap to compare various approaches and provide an analyzed example of a single replicate. The data contains multilocus marker and quantitative trait data on a four-way multiparent advanced generation intercross (Cavanagh *et al.*, 2008). The experimental population contains 1000 individuals, which have been selfed for six generations. We assume 100 markers are equally spaced and evenly distributed along two chromosomes, with inter-marker distances varying between 1 and 10 cM. The population size and density of markers are chosen to reflect current studies and the theoretical resolution of the cross (C.Cavanagh, personal communication), and we have limited the study to two chromosomes for computational expediency. Two QTL with heritability 0.043 are positioned on chromosome 2 at 20 and 80 cM. As described in Section 2, the QTL are calculated by summing QTL effects of size 0.5 with a random deviate drawn from a normal distribution with mean 0 and variance 1.

We analyze a single replicate of the data with 2 cM inter-marker spacing as follows. First, we build the marker map using the approach described previously. We begin by estimating recombination fractions between all marker loci and ordering these loci using R/seriation. We then refine the marker map using multipoint ordering on windows of three markers. We found this approach computationally tractable while maintaining good performance. The final map order on both chromosomes matched the simulated marker map exactly, with chromosome lengths inflated slightly to 117 and 135 cM. Second, we investigate the haplotype structure of individuals in the experimental population using multipoint haplotype probabilities. After estimating parental origin of alleles at each locus at a threshold of 70%, 6% of each chromosome could not be classified; otherwise, alleles were evenly distributed between all four founders as expected. There were 1.8 and 1.7 recombinations per 100 cM for chromosomes 1 and 2. The 50 most recombinant lines are depicted in Figure 1a for Chromosome 1 to illustrate the mosaic of genomes inherited from the four founders. Third, we perform composite interval mapping. Both QTL are detected on Chromosome 2; none are detected on Chromosome 1 (Fig. 1b). The detected QTL are located within 10 cM of the correct position, and both have (empirical) genome-wide $P < 10^{-7}$.

While extensive simulations are beyond the scope of this article, we can also compare some of the approaches described previously through simulated data. We construct marker maps for the first chromosome in 1000 replicates of data with 2 cM marker spacing.

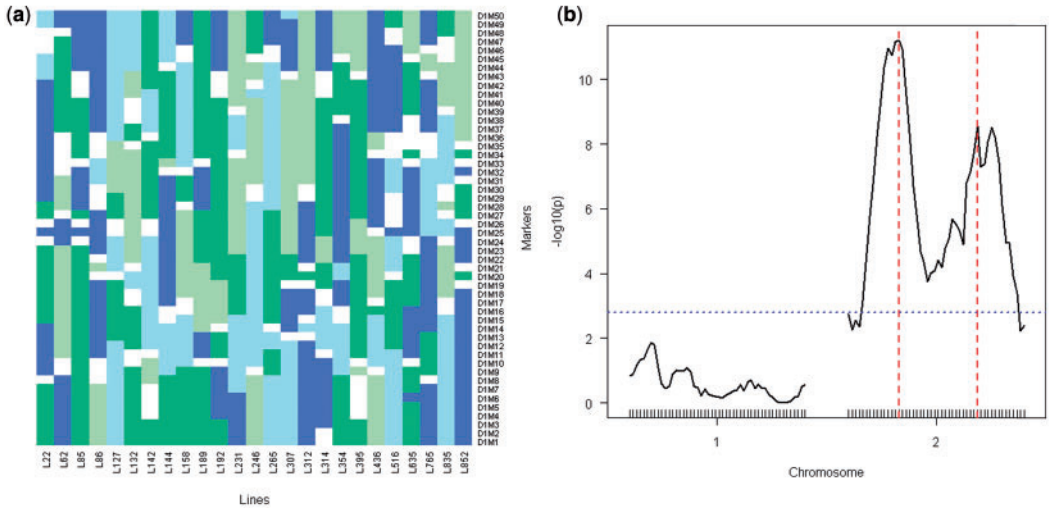


**Fig. 1.** Analysis of simulated data replicate. (**a**) Haplotype mosaics on Chromosome 1 for the 25 most recombinant lines. The columns are indexed by lines; the rows indexed by markers. Light blue indicates regions inherited from the first founder; dark blue from the second; light green from the third and dark green from the fourth. White regions indicate loci where alleles cannot be attributed to any founders at the threshold of 70% probability. (**b**) Transformed *P*-value [−log10(p)] profiles for Chromosomes 1 and 2, with detected QTL positions denoted by red dashed lines. The genome-wide significance threshold of 2.79, computed from 10 000 empirical null simulations, is shown by the horizontal blue dotted line.

In 42% of these replicates, the true map order was found; in the remaining replicates, the median number of markers incorrectly ordered was 2. For haplotype reconstruction, we compare the three described approaches that rely on functions in R/mpMap, R/qtl and R/happy, respectively. We found that the dynamic programming approach to computing multipoint probabilities in R/happy had lower probabilities of imputing a parental origin for loci than did the other two methods. Even at imputation thresholds as low as 60%, <20% of values were imputed. As expected, the multipoint computation implemented in R/qtl performed better than calculating three-point probabilities in terms of number of genotypes classified. However, when parental origin could be assigned, both approaches had similar accuracy around 90%. Supplementary Figure S1 shows the effect of changing marker density on these two approaches for varying thresholds.

## 4 DISCUSSION

In this article, we have described an exciting new computer package for the analysis of marker and trait data from complex multiparent RIL crosses. With mpMap, users are now in a position to simulate data, construct marker maps, examine the genomic structure of individuals and perform QTL mapping on these cutting-edge experimental designs. Our package differs from existing *R* packages mainly in the extensive facility for map construction. In addition, mpMap unifies some of the functionality currently available in other packages and adds new approaches and visualization functions. It thus equips breeders and geneticists with a wide variety of practical tools to dissect the genetics of complex traits.

We do recognize that for large $2^n$-crosses, computation will be demanding. Estimating recombination fractions scales linearly with lines and founders, but quadratically with markers. We are currently working to parallelize the analysis process through the use of machines with Graphics Processing Units. Preliminary results indicate that we can reduce computation time by two orders of magnitude.

*Conflict of Interest*: none declared.

## REFERENCES

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

Broman,K. (2005) The genomes of recombinant inbred lines. *Genetics,* **169**, 1133–1146.

Cavanagh,C. *et al.* (2008) From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. *Curr. Opin. Plant Biol.*, **11**, 215–221.

Complex Trait Consortium (2004) The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nat. Genet.*, **36**, 1133–1137.

Gilmour,A.R. *et al*. (2009) ASReml User Guide Release 3.0. VSN International Ltd., Hemel Hampstead, UK.

Hahsler,M. *et al*. (2009) Seriation: Infrastructure for Seriation. R package version 1.0–1.

Kover,P.X. *et al*. (2009) A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana. PLoS Genet.*, **5**, e1000551.

Mott,R. *et al*. (2000) A new method for fine-mapping quantitative trait loci in outbred animal stocks. *Proc. Natl Acad. Sci. USA*, **97**, 12649–12654.