Bioinformatics, 38(22), 2022, 5134–5136
https://doi.org/10.1093/bioinformatics/btac662
Advance Access Publication Date: 4 October 2022
Applications Note

Genetics and population analysis

statgenMPP: an R package implementing an IBD-based mixed model approach for QTL mapping in a wide range of multi-parent populations

Wenhao Li [1], Martin P. Boer[1,*], Bart-Jan van Rossum[1], Chaozhi Zheng [1], Ronny V.L. Joosen[2] and Fred A. van Eeuwijk[1]

[1]Biometris, Wageningen University and Research Center, Wageningen, 6700 AC, The Netherlands and [2]Rijk Zwaan Breeding B.V., De Lier 2678 ZG, The Netherlands

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on July 22, 2022; revised on August 23, 2022; editorial decision on September 28, 2022; accepted on October 3, 2022

Abstract

Motivation: Multi-parent populations (MPPs) are popular for QTL mapping because they combine wide genetic diversity in parents with easy control of population structure, but a limited number of software tools for QTL mapping are specifically developed for general MPP designs.

Results: We developed an R package called *statgenMPP*, adopting a unified identity-by-descent (IBD)-based mixed model approach for QTL analysis in MPPs. The package offers easy-to-use functionalities of IBD calculations, mixed model solutions and visualizations for QTL mapping in a wide range of MPP designs, including diallele, nested-association mapping populations, multi-parent advanced genetic inter-cross populations and other complicated MPPs with known crossing schemes.

Availability and implementation: The R package *statgenMPP* is open-source and freely available on CRAN at https://CRAN.R-project.org/package=statgenMPP

Contact: martin.boer@wur.nl

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

1 Introduction

Multi-parent population (MPP) designs capture wide genetic diversity and overcome the drawbacks of low minor-allele frequencies and population structures (Arrones *et al.*, 2020; Scott *et al.*, 2020). General MPPs, e.g. diallele (Coles *et al.*, 2010; Giraud *et al.*, 2017), NAM (Yu *et al.*, 2008) and MAGIC (Gardner *et al.*, 2016; Huang *et al.*, 2015) are nowadays widely used in genetic studies and plant breeding programs. Statistical models that can be used for MPP designs are either family-based (linkage mapping) or population-based (linkage disequilibrium) methods (Giraud *et al.*, 2014; Würschum *et al.*, 2012; Xu *et al.*, 2017), but a limited number of software tools is specifically developed for MPP designs.

In this Application Note, we present an easy-to-use R package, *statgenMPP*, adopting an identity-by-descent (IBD)-based mixed model approach for QTL mapping. Compared to other tools, *statgenMPP* integrates a framework of IBD calculation (Boer and van Rossum, 2021b; Zheng *et al.*, 2014, 2015) with linear mixed models (Boer and van Rossum, 2021a). The IBD-based mixed model approach estimates random QTL effects in relation to IBD probabilities of parental origins across the offspring genome (Li *et al.*, 2021; Wei and Xu, 2016) while accounting for polygenic and family background genetic variation, which has been proven to increase the mapping power and resolution of QTLs for simulated and empirical MPPs (Li *et al.*, 2021). Single crosses can also be analyzed with *statgenMPP*, for a full list of available population types, see Boer and van Rossum (2021b).

2 Materials and methods

Figure 1a demonstrates the workflow of IBD-based QTL mapping using *statgenMPP* that comprises two main steps: IBD calculation and QTL mapping.

2.1 IBD calculation

The framework of hidden Markov models (HMM) and inheritance vectors (Zheng *et al.*, 2014, 2015; Boer and van Rossum, 2021b) is employed for the IBD calculation. For a wide range of MPP designs,

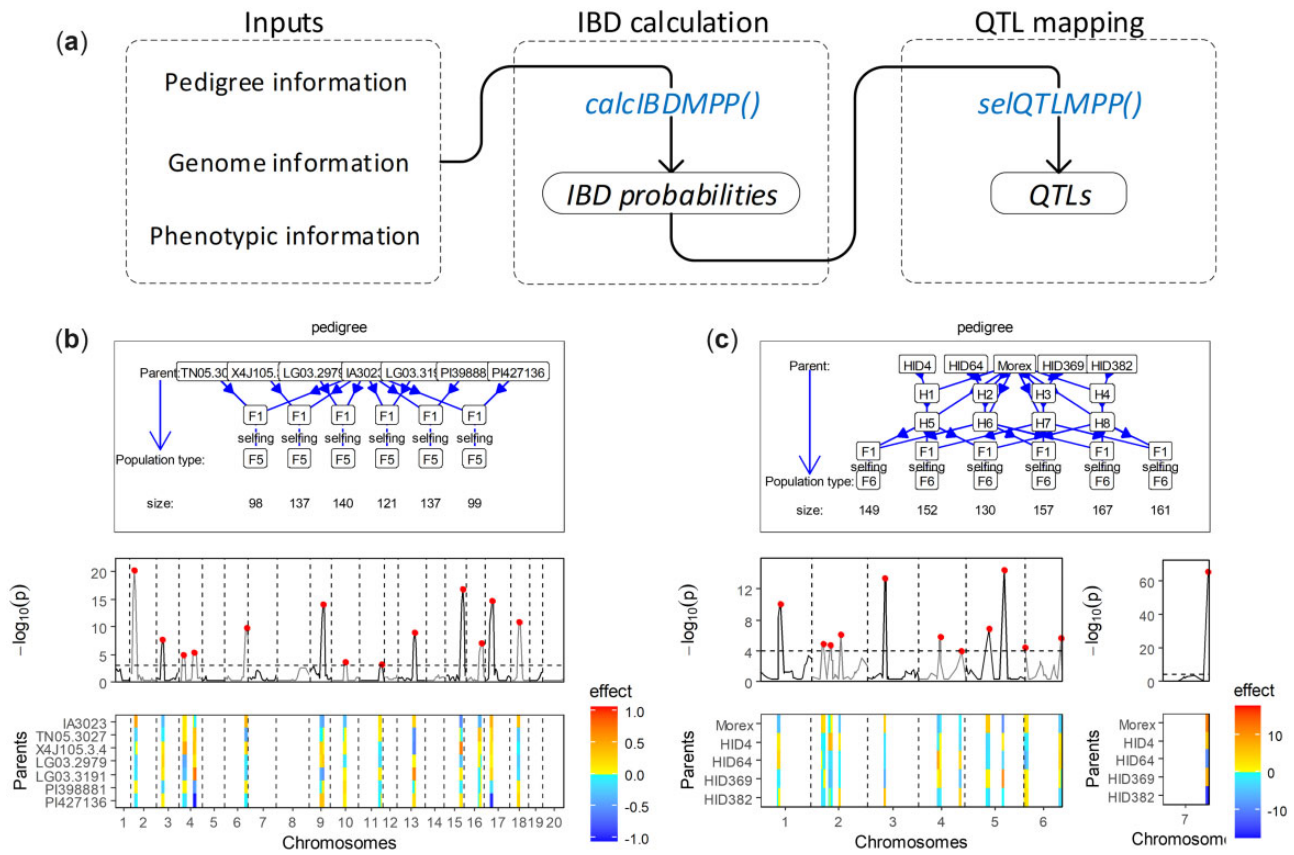

Fig. 1. (**a**) The workflow of IBD-based QTL mapping implemented in statgenMPP. QTL mapping in the examples of (**b**) soybean NAM design for seed weight and (**c**) barley complex design for awn length. Upper panel pedigree plots of MPP designs. Middle panel QTL mapping profiles expressed at −log10(p) scale. Lower panel parental QTL effects

IBD probabilities are calculated by the function *calcIBDMPP()* at a customizable grid (cM) for the specified population type. For complex MAGIC designs and designs with complicated pedigree structures, the IBD probabilities can be first calculated using RABBIT (Zheng, 2019) and then imported by the function *readRABBITMPP()*.

2.2 QTL mapping

IBD probabilities between parents and offspring, as design vectors, or genetic predictors, are fitted in a mixed model for QTL mapping using the function *selQTLMPP()*. The IBD-based mixed model approach is described by Li *et al.* (2021). To test each position on a 1D grid along the genome, a single locus QTL model is fitted whose effects are modeled as random in a mixed effects model:

$$Y = X\beta + M_q \underline{a_q} + g + \underline{\varepsilon}$$
$$\underline{a_q} \sim MVN(0, I_P\sigma_q^2),\ \underline{g} \sim MVN(0, K\sigma_g^2), \underline{\varepsilon} \sim MVN(0,\ \oplus_{k=1}^{F} I_{n_k}\sigma_{\varepsilon_k}^2)$$

Y is a vector with phenotypes; X is the design matrix indicating to which family each individual belongs; β is a vector of fixed family intercepts; M_q is the design matrix containing the expected number of parental alleles as a function of IBD probabilities; a_q is the vector of random parental effects with the variance–covariance (VCOV) structure $I_P\ \sigma_q^2$, in which I_P is the identity matrix for P parents and σ_q^2 is the genetic variance of the QTL effects; it is optional to include a polygenic term g whose VCOV is described by the kinship matrix K; the residual term ε has a family-specific VCOV structure $\oplus_{k=1}^{F} I_{n_k}\sigma_{\varepsilon_k}^2$ in which $\sigma_{\varepsilon_k}^2$ is the residual variance of individuals in the kth family ($k = 1, 2, ..., F$)with family size n_k.

The linear mixed model is fitted and variance components are estimated based on restricted maximum likelihood. Variance components corresponding to putative QTLs ($\sigma_q^2 = 0$ versus $\sigma_q^2 \neq 0$) are evaluated by likelihood ratio tests (LRT) that approximate a mixture of χ^2 distributions (Self and Liang, 1987). Multiple rounds of genome QTL scans can be performed until either (i) no new QTL outside of a certain window size is found with a −log10(p) value below a predefined threshold or (ii) a predefined maximum number of QTLs is reached.

3 Applications

We demonstrate the main functionalities of *statgenMPP* using two publicly available MPP designs—a soybean NAM design and a barley complex design (datasets are provided in the supplementary material with R codes). Other examples with full details are available in the vignette of *statgenMPP*.

All computations were performed in (64-bit) R 4.2.1 (R Core Team, 2022) and a 3.10 GHz Intel Core i5 processor computer with 16GB of RAM and Windows 10 operating system. We used the parallel option of *statgenMPP* using four cores, and using the default values, not including a kinship matrix.

We selected six families (Fig. 1b, upper panel) with a total population size of 732 genotypes from the soybean NAM project (https://soybase.org/SoybeanNAM/index.php) (Xavier *et al.*, 2018) for analysis to demonstrate the functionalities of *statgenMPP*. The consensus map containing 4289 markers (Song *et al.*, 2017) was used to calculate IBD probabilities on a regular 5 cM grid of evaluation points. We map QTLs for the trait of seed weight ('mean_seedWT') as an example. The total computation time was 1.03 min. The second example is a complex barley MPP design (Liller *et al.*, 2017), with a total population size of 916 genotypes, for QTL mapping for awn length ('Awn_length') (Fig. 1c, upper panel). The total computation time was 1.43 min. IBD calculation for the complex design

was performed by RABBIT, and then the output was imported by the *readRABBITMPP()* function in *statgenMPP*.

Results of QTL mapping for soybean NAM and barley complex MPP designs are shown in Figure 1b and c. For example in the barley complex design, all QTLs for awn length in the barley complex design can be confirmed from the previous study where the strong QTL on chromosome 7 was successfully fine mapped (Liller *et al.*, 2017). Further details on how to use the package for visualization can be found in the *statgenMPP* vignette.

4 Conclusion

An increasing range and number of MPP designs become available for QTL mapping in breeding programs and genetic studies. We introduce the R package *statgenMPP* that covers the demands for versatile QTL analysis in MPP designs. *statgenMPP* contains a unified IBD-based mixed model framework for mapping multi-allelic QTLs. We analyzed multiple MPP data sets with *statgenMPP*, whose data are available within the package, and demonstrated the theoretical and practical advantages of our approach. Extensions of *statgenMPP* will deal with epistasis, pleiotropy and QTL-by-environment analysis, to allow the user to explore the full potential of MPP designs.

Acknowledgements

We thank two reviewers and the Associate Editor for their helpful comments.

Funding

This work was supported by Rijk Zwaan Breeding B.V.

Conflict of Interest: The authors declare that they do not have any conflict of interest.

Data availability

All data are incorporated into the article and its online supplementary material.

References

Arrones,A. *et al.* (2020) The dawn of the age of multi-parent magic populations in plant breeding: novel powerful next-generation resources for genetic analysis and selection of recombinant elite material. *Biology (Basel)*, **9**, 229.

Boer,M.P. and van Rossum,B. (2021a) *LMMsolver: Linear Mixed Model Solver*. https://CRAN.R-project.org/package=LMMsolver

Boer,M.P. and van Rossum,B. (2021b) *statgenIBD: Calculation of IBD Probabilities*. https://CRAN.R-project.org/package=statgenIBD

Coles,N.D. *et al.* (2010) Genetic control of photoperiod sensitivity in maize revealed by joint multiple population analysis. *Genetics*, **184**, 799–812.

Gardner,K.A. *et al.* (2016) A highly recombined, high-density, eight--founder wheat MAGIC map reveals extensive segregation distortion and genomic locations of introgression segments. *Plant Biotechnol. J.*, **14**, 1406–1417.

Giraud,H. *et al.* (2014) Linkage disequilibrium with linkage analysis of multiline crosses reveals different multiallelic QTL for hybrid performance in the flint and dent heterotic groups of maize. *Genetics*, **198**, 1717–1734.

Giraud,H. *et al.* (2017) Reciprocal genetics: identifying QTL for general and specific combining abilities in hybrids between multiparental populations from two maize (*Zea mays* L.) heterotic groups. *Genetics*, **207**, 1167–1180.

Huang,B.E. *et al.* (2015) MAGIC populations in crops: current status and future prospects. *Theor. Appl. Genet.*, **128**, 999–1017.

Li,W. *et al.* (2021) An IBD-based mixed model approach for QTL mapping in multiparental populations. *Theor. Appl. Genet*, **1**, 1–18.

Liller,C.B. *et al.* (2017) Fine mapping of a major QTL for awn length in barley using a multiparent mapping population. *Theor. Appl. Genet.*, **130**, 269–281.

R Core Team (2022) *R: A Language and Environment for Statistical Computing*.

Scott,M.F. *et al.* (2020) Multi-parent populations in crops: a toolbox integrating genomics and genetic mapping with breeding. *Heredity (Edinb)*, **125**, 396–416.

Self,S.G. and Liang,K.Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.*, **82**, 605–610.

Song,Q. *et al.* (2017) Genetic characterization of the soybean nested association mapping population. *Plant Genome*, **10**(2).

Wei,J. and Xu,S. (2016) A random-model approach to QTL mapping in multiparent advanced generation intercross (MAGIC) populations. *Genetics*, **202**, 471–486.

Würschum,T. *et al.* (2012) Comparison of biometrical models for joint linkage association mapping. *Heredity (Edinb.)*, **108**, 332–340.

Xavier,A. *et al.* (2018) Genome-wide analysis of grain yield stability and environmental interactions in a multiparental soybean population. *G3 Genes Genomes, Genet.*, **8**, 519–529.

Xu,Y. *et al.* (2017) Genetic mapping of quantitative trait loci in crops. *Crop J*, **5**, 175–184.

Yu,J. *et al.* (2008) Genetic design and statistical power of nested association mapping in maize. *Genetics*, **178**, 539–551.

Zheng,C. *et al.* (2014) A general modeling framework for genome ancestral origins in multiparental populations. *Genetics*, **198**, 87–101.

Zheng,C. (2019) *RABBIT (v3.2) Manual Book*. https://github.com/chaozhi/RABBIT

Zheng,C. *et al.* (2015) Reconstruction of genome ancestry blocks in multiparental populations. *Genetics*, **200**, 1073–1087.