Crosslink: A fast, scriptable genetic mapper for outcrossing species

Robert J. Vickerstaff1,* and Richard J. Harrison1

1Department of Genetics, Genomics and Breeding, NIAB EMR, New Road, East Malling, UK, ME19 6BJ

*To whom correspondence should be addressed.

Associate Editor: XXXXXX
Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Crosslink is genetic mapping software for outcrossing species designed to run efficiently on large datasets by combining the best from existing tools with novel approaches. Tests show it runs much faster than several comparable programs whilst retaining a similar accuracy.

Availability and implementation: Available under the GNU General Public License version 2 from https://github.com/eastmallingresearch/crosslink

Contact: robert.vickerstaff@emr.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online and from https://github.com/eastmallingresearch/crosslink/releases/tag/v0.5.

1 Introduction

Genetic maps are valuable for such key tasks as quantitative trait loci discovery (Hancock et al., 2016), map-based cloning (Huang et al., 2003) and genome assembly scaffolding (Fierst, 2015). Modern high-throughput genotyping technologies routinely provide more markers than can be conveniently mapped using older mapping programs designed around smaller datasets. Even with the ongoing advances in long read sequencing technology the final stage of many genome assemblies is likely to rely on a genetic map for some time to come. This motivates continuing efforts to improve mapping algorithms to facilitate analyses of larger datasets and to increase the level of automation and reliability.

Here we focus specifically on the case where the map is constructed using an outcross between two heterozygous parents, which requires a more complex mapping approach than traditional inbred-based F2 or backcross mapping (van Ooijen & Jansen, 2013).

We present Crosslink, an outcross mapper which adapts the minimum spanning tree (MST) method from the inbred mapper MSTmap to provide a rapid initial approximate marker ordering. It also modifies the Gibbs sampler method from outcross mapper JoinMap 4.1 (van Ooijen, 2011) to propagate information unidirectionally leading to shorter convergence times for imputing the missing information inherent in outcross markers heterozygous in both parents. Time complexity of the genetic algorithm used to finalise the marker order is improved by adopting differential recombination counting, such that only the markers adjacent to the break points created by the reordering mutations need to be examined to decide whether to accept or reject the mutation instead of rescanning the entire ordering. Calculation of recombination counts between marker pairs is expedited by converting genotype calls to bit strings with masks representing missing information, counts are then stored in a cache to prevent redundant recalculation for previously examined marker pairs.

2 Methods

Crosslink consists of two main programs: crosslink_group which performs marker phasing, missing genotype call imputation using the k nearest-neighbour method (Troyanskaya et al., 2001), formation of linkage groups and approximate marker ordering; and crosslink_map which performs final marker ordering and Gibbs sampler imputation of missing information to allow multi-point recombination fractions to be calculated. Further details are given in the supplement section 1. Input files are encoded using the genotype code conventions of JoinMap (see page 2 of the manual https://github/ eastmallingresearch/ crosslink/ blob/ docs/ crosslink_manual.pdf and the sample_data directory of the github repository).

3 Results

Multiple simulated diploid outcross data sets were generated, each for a single 100 centimorgan chromosome and 200 progeny, varying either error/missing data rates or marker density (see supplement and Fig.1 for details) and used to compare Crosslink to LepMap version Lep-MAP2 v0.2 (Rastas, Paulin, Hanski, Lehtonen, & Auvinen, 2013), OneMap version onemap_2.0 with R version 3.1.2 stable version (Margarido, Souza, & Garcia, 2007), Tmap version 1.1 (Cartwright, Troggio, Velasco, & Gutin, 2007) and MSTmap (all tests ran using automated scripts on a high performance Linux computer cluster). LepMap and Tmap used their default settings, OneMap used unidirectional growth mode. Crosslink was tested using either approximate ordering only (running crosslink_group but not crosslink_map) or full ordering (crosslink_group then crosslink_map). MSTmap was tested by prephasing and recoding the data into separate maternal and paternal backcross-type inputs, and then combining the two output maps into a consensus. These extra steps were not counted as part of the execution time. n=8 matched replicates were used for all treatments except for LepMap, Tmap and OneMap at 1000 markers (n=3) or above (n=0) due to excessive run times. JoinMap 4.1 maximum likelihood mapping was also tested, running manually under Windows XP on a desktop computer (therefore running times are not strictly comparable) testing only the first replicate (n=1) for all treatments up to 5000 markers (above which memory was insufficient for JoinMap).
this package

We thank Eric van de Weg for helpful comments throughout the development of this package.

Conflicts of interest: none declared.

References

Cartwright, D. A., Troggio, M., Velasco, R., & Gutin, A. (2007) Genetic Mapping in the Presence of Genotyping Errors. *Genetics*, 176, 2521–2527.

Fierst, J. L. (2015) Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. *Frontiers in Genetics*, 6, 220.

Hancock, J. F., Sooriyapathirana, S. S., Bassil, N. V, Stegmeir, T., Cai, L., Finn, C. E., … Weebadde, C. K. (2016) Public Availability of a Genotyped Segregating Population May Foster Marker Assisted Breeding (MAB) and Quantitative Trait Loci (QTL) Discovery: An Example Using Strawberry. *Frontiers in Plant Science*, 7, 619.

Huang, L., Brooks, S. A., Li, W., Fellers, J. P., Trick, H. N., & Gill, B. S. (2003) Map-Based Cloning of Leaf Rust Resistance Gene La21 From the Large and Polyploid Genome of Bread Wheat. *Genetics*, 164, 655–664.

Margarido, G., Souza, A., & Garcia, A. (2007) OneMap: software for genetic mapping in outcrossing species. *Hereditas*, 144, 78–79.

Rastas, P., Paulin, L., Hanski, I., Lehtonen, R., & Auvinen, P. (2013) Lep-MAP: fast and accurate linkage map construction for large SNP datasets. *Bioinformatics*, 29, 3128–3134.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., … Altman, R. B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17, 520–525.

van Ooijen, J. (2011) Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. *Genetics Research*, 93, 343–349.

van Ooijen, J., & Jansen, J. (2013). *Genetic mapping in experimental populations*. Cambridge University Press.

Acknowledgements

We thank Eric van de Weg for helpful comments throughout the development of this package.