Published March 20, 2015

SimpleMap: A Pipeline to Streamline High-Density Linkage Map Construction

Abdulqader Jighly,* Reem Joukhadar, and Manickavelu Alagu

Abstract

The recent development in high-throughput genotyping techniques requires new statistical methods to analyze large datasets. The current available linkage mapping software are time consuming and limited in terms of the maximum number of markers that can be mapped on a single linkage group. In this paper, we propose the Perl pipeline, SimpleMap. This tool can significantly improve the speed of currently available linkage mapping software with minimal impact on marker order and map length by limiting the consideration of duplicated and tightly linked molecular markers during linkage group development. SimpleMap works with the following three main steps: (i) generating a subset of markers for which each pair has a number of recombinants higher than a threshold determined by the user (the repulsion threshold), (ii) mapping this subset with any external mapping tool, and (iii) intersecting the remaining unmapped markers to the constructed map. The script was tested on 15 wheat (*Triticum aestivum* L.) linkage groups derived from two different crosses. In 13 genetic groups, the computational time was reduced from ~8 h to ~8 min, while it was impossible to map the remaining two linkage groups without applying SimpleMap first. SimpleMap is a very time-efficient tool, and considering a repulsion threshold equivalent to 1 cM results in a number of markers similar to map lengths that can be analyzed on a simple personal computer. SimpleMap can be downloaded from http://simplemap-aj.sourceforge.net/.

THE RAPID DEVELOPMENT of high-throughput, low-cost molecular markers and the availability of single nucleotide polymorphism (SNP) detection and genotyping-by-sequencing (GBS) technologies makes it possible for breeders to screen populations with a large number of markers at a very reasonable time and cost (Davey et al., 2011). However, the computational time to analyze these large datasets is still challenging and requires high-performance computing systems, especially when constructing high-density linkage maps.

Genetic mapping is the determination of the linear arrangement of the genes or markers alongside chromosomes. The distances among those loci are measured as the percentage of recombination between them in a biparental population and the map unit is the centimorgan (cM). Two genetic loci are said to be 1 cM apart if they exhibit 1% recombinant lines between them (Collard et al., 2005). For this reason, genetic distance differs from physical distance, which is usually measured by base pairs and defines the actual position of the genes on a reference assembly. If no recombination occurs between two or more linked genetic markers in a population, they will be overlapped on the linkage group and will have an identical genetic position even if their physical position is different. From a computational prospective, even

Published in The Plant Genome 8
doi: 10.3835/plantgenome2014.09.0056

5585 Guilford Rd., Madison, WI 53711 USA
An open-access publication

A. Jighly, The International Center for Agricultural Research in the Dry Areas (ICARDA), P.O. Box 5466, Aleppo, Syria; Dep. of Environment and Primary Industries, AgriBio, 5 Ring Rd., Bundoora, Vic. 3083, Australia; and School of Applied Systems Biology, La Trobe Univ., Bundoora, Vic. 3083, Australia. R. Joukhadar, Dep. of Botany, La Trobe Univ., Bundoora, Vic. 3083, Australia. M. Alagu, Kihara Institute for Biological Research, Yokohama City Univ., Maioka 641-12, Yokohama 2440813, Japan. Received 26 Sept. 2014. Accepted 18 Dec. 2014. *Corresponding author (a.jighly@cgiar.org).

Abbreviations: DArT, diversity arrays technology; DH, double haploid; GBS, genotyping-by-sequencing; MSD, mean standard deviation; RIL, recombinant inbred lines; SNP, single nucleotide polymorphism.

though those markers have an identical haplotype, their presence together during the map construction analysis will cost more computing time.

This paper suggests a three-stage mapping approach using a tool that can reduce the number of markers during the long mapping process by selecting a subset of representative markers based on their haplotype similarities with other markers. The markers excluded in the first step flank or overlap with the mapped markers that cosegregate with them.

SimpleMap Features

The Perl programming language (v. 5.8.8; http://www.perl.org/) was used to design the software. SimpleMap consist of two Perl scripts: Before_Mapping.pl and After_Mapping.pl. The Before_Mapping.pl script checks the number of recombination events for each pair of markers to detect pairs with no or few recombination. If the script detects such pair of markers, it will remove one of them from the mapping step and save the number of recombination between both markers. This script requires the user to select the genotype file, to enter the number of markers and the number of lines, and to specify a threshold of the maximum accepted number of recombinant lines between two markers, that is, the repulsion threshold. This script will generates two files: For_Mapping.txt, which contains the haplotype of the representative markers selected for mapping; and the Repulsion.txt file, which involves a list of the markers that were missed in the mapping file, the closest markers to them in the mapping file, and the number of recombinant lines between the mapped and the unmapped markers. The user then can use any mapping software to map the markers in the file For_Mapping.txt; in this paper, we used JoinMap 4 (Van Ooijen 2006). The final step is to use the script After_Mapping.pl. This script anchors unmapped markers to the exported map using the guide file Repulsion.txt and requires the user to name the input and the output maps. However, the markers reported as unlinked with JoinMap, or any other tool, will be unlinked using SimpleMap as well.

Applying SimpleMap does not affect the map quality since an external tool generates the backbone of the map, and SimpleMap adds only markers that are tightly linked to other markers on this backbone. For this reason, SimpleMap uses the recombinant fraction θ (percentage of recombinant lines between two markers) to estimate the genetic distance for the unmapped markers from the mapped ones. The value of θ will be very similar to the distance estimated using multilocus methods and using mapping functions like Kosambi and Haldane for tiny distances (Mihovilovich et al., 2008). However, it is highly recommended to not select a repulsion threshold over 2 cM for Haldane or 3 cM for Kosambi functions.

Implementation of SimpleMap

SimpleMap was tested on the wheat (*Triticum aestivum* L.) reference double haploid (DH) population Synthetic-W7984 × Opata-M85 (Sorrells et al., 2011), genotyped with a diversity arrays technology (DArT) chip and microsatellite markers with repulsion threshold of 1. Out of the 21 linkage groups in the DH population, only groups with total number of markers over 60 were used in this study, which were 1A, 1B, 2B, 3B, 4A, 5B, 6B, 7A, and 7B (Table 1; Supplemental Figure S1). The DH population involved 163 lines and was genotyped with 917 DArT and 44 SSR markers for the adapted groups. SimpleMap was also tested on six linkage groups adapted from the map of the wheat recombinant inbreed lines (RIL) population derived from the cross CS × Syn (Alagu et al., unpublished data, 2015). The RIL population has 104 lines and the six groups were genotyped with the GBS 1.0 V array containing 236 SNP and 798 DArT markers for the six groups (Table 1; Supplemental Figure S1). We tested the CS × Syn population due to the presence of heterozygote lines in RIL populations and to avoid the clustering nature of the markers in the wheat DArT chip (Semagn et al., 2006). All analyses were performed on a personal computer with Core-i7-3632QM 2.2 GHz processor and 8 GB of RAM.

The Before_Mapping.pl script reduced the number of markers in the DH population to 239 (24.9%) and in the RIL population to 425 (41.1%) (Table 1) and took about 5 min to finish the analysis for both populations. This process reduced the computational time for the mapping step in the DH population, except for the 3B group, from approximately 4.5 h to 3 min; and in the RIL population for the first five groups from approximately 3.5 h to 5 min. For the sixth linkage group in the RIL population (Table 1), it took ~10 h to be mapped with 230 markers using JoinMap after applying SimpleMap with a repulsion threshold of 1. On the other hand, it was impossible to map this group using JoinMap with a total of 571 markers. Similarly, JoinMap fails to map the group 3B in the DH population with 312 markers. The remaining markers then intersected to the map using the After_Mapping.pl script and the analysis took only a few seconds for all linkage groups.

Comparing the positions of the markers when using JoinMap to construct a linkage group with the whole dataset and when using SimpleMap to reduce the number of markers for mapping revealed consistent positions. The mean standard deviation (MSD) for the positions of the marker mapped using JoinMap in both maps ranged between ±0.2 and ±5.76 (Table 1) with an overall MSD of ±2.18 for the DH population and ±0.77 for the RIL population, while the MSD for the position of the markers intersected using SimpleMap ranged from ±0.3 to ±3.05 (Table 1) with an overall MSD of ±1.77 for the DH population and ±0.76 for the RIL population. These results show that the positions from SimpleMap have similar or lower error rates than those mapped only with JoinMap. Moreover, the order of the markers along the

Table 1. Summary of the tested linkage groups, including the total number of markers, the number of markers used for mapping after applying the script Before_Mapping.pl, the length of the linkage group when mapping all markers in JoinMap, the length of the linkage group when applying SimpleMap, the mean standard deviation (MSD) in cM of the map positions between both maps for the markers mapped using JoinMap and for the markers intersected to the map using the script After_Mapping.pl, and the MSD in cM of the marker positions for ten mapping replicates using JoinMap and SimpleMap. JoinMap fails to construct the group 3B in the double haploid (DH) population with 312 and sixth linkage group in the recombinant inbred lines (RIL) population with 571 markers.

ID	Marker count	Before mapping	JoinMap length	SimpleMap length	JoinMap MSD	SimpleMap MSD	JoinMap MSD 10 reps	SimpleMap MSD 10 reps
DH population								
1A	74	21	117.3	113.6	±1.20	±1.97	$\pm 5.6 \times 10^{-5}$	0
1B	89	27	126.5	122.8	±1.15	±1.23	0	0
2B	104	26	154.4	142.8	±1.01	±1.08	$\pm 7.3 \times 10^{-4}$	0
3B	312	41	–	161.4	–	–	–	–
4A	77	21	104.5	128.0	±5.76	±1.72	0	0
5B	91	32	194.6	188.9	±3.16	±3.05	±0.013	$\pm 2.9 \times 10^{-5}$
6B	62	25	138.5	132.8	±2.06	±1.90	0	0
7A	67	20	200.8	195.7	±2.13	±2.57	0	0
7B	85	26	173.8	170.5	±1.56	±1.13	±0.009	$\pm 1.4 \times 10^{-4}$
RIL population								
1	89	43	87.0	87.6	±0.20	±0.30	0	0
2	84	35	99.6	101.3	±1.37	±1.15	0	0
3	87	37	90.4	94.1	±1.25	±1.42	0	0
4	62	25	71.3	71.6	±0.62	±0.44	0	0
5	141	55	88.6	88.0	±0.50	±0.53	0	0
6	571	230	–	189.1	–	–	–	–

linkage groups was very similar. Supplemental Figure S1 shows all of the linkage groups mapped with and without SimpleMap adjustment.

To test the accuracy of SimpleMap, we ran the analysis ten times using JoinMap and SimpleMap and the MSD of the ten replicates for all markers were estimated. The results showed high accuracy in both cases with neglected deviations (Table 1). However, SimpleMap accuracy is dependent on the accuracy of the external mapping tool that is used to generate the backbone of the map.

Conclusions

SimpleMap can effectively reduce the computational time for linkage group construction and supports, as well integrates with, any currently available mapping software. It can significantly decrease the number of markers that should be used for mapping by omitting the redundant markers and realigning them to the map after linkage group construction.

Acknowledgments

The authors would like to thank Dr. Hans Daetwyler (Department of Environment and Primary Industry) for his valuable comments to improve the content of the manuscript and his assistance in the reviewer comments and Dr. Anthony Gendall (La Trobe University) for editing the final copy of the manuscript. This project is funded by ICARDA (The International Center for Agricultural Research in the Dry Areas).

References

Collard, B.C.Y., M.Z.Z. Jahufer, J.B. Brouwer, and E.C.K. Pang. 2005. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica 142:169–196. doi:10.1007/s10681-005-1681-5

Davey, J.W., P.A. Hohenlohe, P.D. Etter, J.Q. Boone, J.M. Catchen, and M.L. Blaxter. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12:499–510. doi:10.1038/nrg3012

Mihovilovich, E., E. Simon, and R. Bonierbale. 2008. Construction of genetic linkage maps. In: C. Kole and A.G. Abbott, editors, Principles and practices of plant genomics Vol. 1. Genome mapping. Science Publishers, Enfield, NH. p. 93–139.

Semagn, K., A. Bjornstad, H. Skinnes, A.G. Maroy, Y. Tarkegne, and M. William. 2006. Distribution of DArT, AFLP and SSR markers in a genetic linkage map of a double haploid hexaploid wheat population. Genome 49:545–555. doi:10.1139/G06-002

Sorrells, M.E., J.P. Gustafson, D. Somers, S. Chao, D. Benscher, G. Guedira-Brown, E. Huttner, A. Kilian, P.E. McGuire, K. Ross, J. Tanaka, P. Wenzl, K. Williams, and C.O. Qualset. 2011. Reconstruction of the synthetic W7984 × Opata M85 wheat reference population. Genome 54:875–882. doi:10.1139/g11-054

Van Ooijen, J.W. 2006. JoinMap 4, software for the calculation of genetic linkage maps in experimental populations. Kyazma, B.V., Wageningen, the Netherlands.