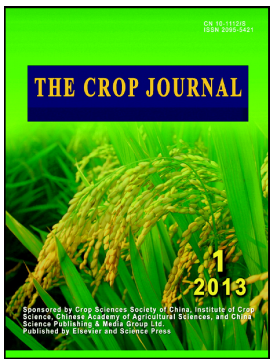

Linkage analysis and integrated software GAPL for pure-line populations derived from four-way and eight-way crosses

Luyan Zhang, Lei Meng, Jiankang Wang

PII: S2214-5141(18)30136-3
DOI: https://doi.org/10.1016/j.cj.2018.10.006
Reference: CJ 334

To appear in: *The Crop Journal*

Received date: 28 August 2018
Revised date: 25 October 2018
Accepted date: 26 November 2018

Please cite this article as: Luyan Zhang, Lei Meng, Jiankang Wang , Linkage analysis and integrated software GAPL for pure-line populations derived from four-way and eight-way crosses. Cj (2018), https://doi.org/10.1016/j.cj.2018.10.006

Linkage analysis and integrated software GAPL for pure-line populations derived from four-way and eight-way crosses

Luyan Zhang, Lei Meng, Jiankang Wang*

The National Key Facility for Crop Gene Resources and Genetic Improvement, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China

Abstract: Pure lines derived from multiple parents provide abundant variation for genetic study. However, efficient genetic analysis methods and user-friendly software are still lacking. In this study, we developed linkage analysis methods and integrated analysis software for pure-line populations derived from four-way and eight-way crosses. First, polymorphic markers are classified into different categories according to the number of identifiable alleles in the inbred parents. Expected genotypic probability is then derived for each pair of complete markers, and based on them a maximum likelihood estimate (MLE) of recombination frequency is calculated. An EM algorithm is proposed for calculating recombination frequencies in scenarios that at least one marker is incomplete. A linkage map can thus be constructed using estimated recombination frequencies. We describe a software package called GAPL for recombination frequency estimation and linkage map construction in multi-parental pure-line populations. Both simulation studies and results from a reported four-way cross recombinant inbred line population demonstrate that the proposed method and software can build more accurate linkage maps in shorter times than other published software packages. The GAPL software is freely available from www.isbreeding.net and can also be used for QTL mapping in multi-parental populations.

Keywords: Pure lines; Four-way cross; Eight-way cross; Recombination frequency estimation; Integrated software

1. Introduction

Multi-parent Advanced Generation InterCross (MAGIC) populations are becoming more and more common in genetic studies. Compared with conventional bi-parental populations, multi-parental populations harbor increased allelic and phenotypic diversity, leading to denser recombination events and higher mapping accuracy [1–3]. Compared with natural populations, kinship in progenies from multi-parental crosses is clear, so that there is no uncertainty of population structure [4]. Pure-line populations can be repeatedly planted in multiple years and locations to increase accuracy of phenotyping and detection power for quantitative trait loci (QTL) and to perform QTL-by-environment interaction analysis [5–8]. These advantages have accelerated the development of multi-parental pure-line populations during the last decade.

Multi-parental pure-line designs were first proposed in mice [9], and have since been applied in plant species. For example, in *Arabidopsis thaliana*, Kover et al. [1] described the first set of MAGIC lines and developed analytical methods to fine-map QTL. Klasen et al. [10] evaluated the statistical powers of QTL detection in multi-parental recombinant inbred line (RIL) populations. In rice, Bandillo et al. [11] developed four

*Corresponding author: Jiankang Wang, E-mail address: wangjiankang@caas.cn, Tel.: +86-10-82105846; Fax: +86-10-82105819

Received: 2018-08-28; Revised: 2018-10-25; Accepted: 2018-11-26.

multi-parental pure-line populations and used genome-wide association mapping for QTL identification. Ponce et al. [12] developed an eight-parent RIL population and used association mapping to detect QTL associated with cooking and eating quality in *indica* rice. In wheat, Huang et al. [13] constructed a linkage map in a four-parent RIL population. Würschum et al. [14] performed association mapping in a six-parent doubled-haploid (DH) triticale population. In soybean, Shivakumar et al. [15] developed an eight-parent MAGIC population by employing two-way, four-way, and eight-way intercross hybridization. In barley, Sannemann et al. [16] incorporated multi-locus QTL analysis and cross validation for flowering time in the first eight-parent DH population.

Linkage-map construction is a crucial step of genetic analysis, providing basic chromosomal information for map-based gene cloning and marker-assisted breeding [17]. Linkage-analysis methodology has been less investigated in multi-parental pure-line populations than in bi-parental populations [18]. The number of alleles and marker types at each locus in multi-parental populations is much larger than that in bi-parental populations. This property complicates methods for recombination frequency estimation and linkage map construction. To date, most studies aimed at gene detection in multi-parental pure-line populations have been based on association mapping, where no linkage map is needed. To fully exploit the potential of these populations in genetic studies, efficient and accurate methods for linkage analysis are needed. Some R packages provide functions for linkage map construction in multi-parental RIL populations, including R/qtl [19], R/happy [20], and R/mpMap [21]. However, these packages lack user-friendly interfaces, and the efficiency of linkage analysis methods in these tools has not been investigated systematically. In addition, these packages are not adapted to multi-parental DH populations.

In this study, we focused on pure-line populations derived from four-way and eight-way crosses. Our objectives were: (1) to estimate recombination frequencies between markers with various numbers of identifiable alleles and then construct a linkage map, (2) to develop an integrated software package for genetic analysis, and (3) to demonstrate the advantages of the proposed method and software by simulation studies and analysis of a reported four-way cross wheat RIL population.

2. Materials and methods

2.1. Classification of markers in genotyping

Four inbred parental lines, parents A, B, C and D, are needed to make a four-way cross. Two kinds of pure lines can be derived from a four-way cross. One, DH lines, is produced by embryo rescue and pollen culture technology, and the other, RILs, is produced by repeated selfing and single-seed descent from the four-way cross F_2 (Fig. 1). Some markers may have four identifiable alleles in four parental lines, but others may have fewer. According to the number of identifiable alleles in four parental lines, 14 marker categories may be defined for any polymorphic marker: ABCD, AACD, ABCC, ABAD, ABCA, ABBD, ABCB, AACC, ABAB, ABBA, ABBB, ABAA, AACA, and AAAD [8]. Markers belonging to category ABCD carry complete information, with the four parents carrying four identifiable alleles, denoted by *A*, *B*, *C*, and *D*. The corresponding genotypes are denoted as *AA*, *BB*, *CC*, and *DD*, following the Mendelian ratio of 1:1:1:1 in DHs or RILs when no distortion occurs. Markers belonging to the other 13 categories carry incomplete information, such that the four alleles in the parents cannot be distinguished

unambiguously. For example, for a marker belonging to category AACD, alleles in parents A and B are the same, but differ from the alleles in parents C and D. In derived DH or RIL populations, only three homozygous genotypes can be observed at a marker locus: *AA*, *CC*, and *DD*, following the Mendelian ratio of 2:1:1 when no distortion has occurred.

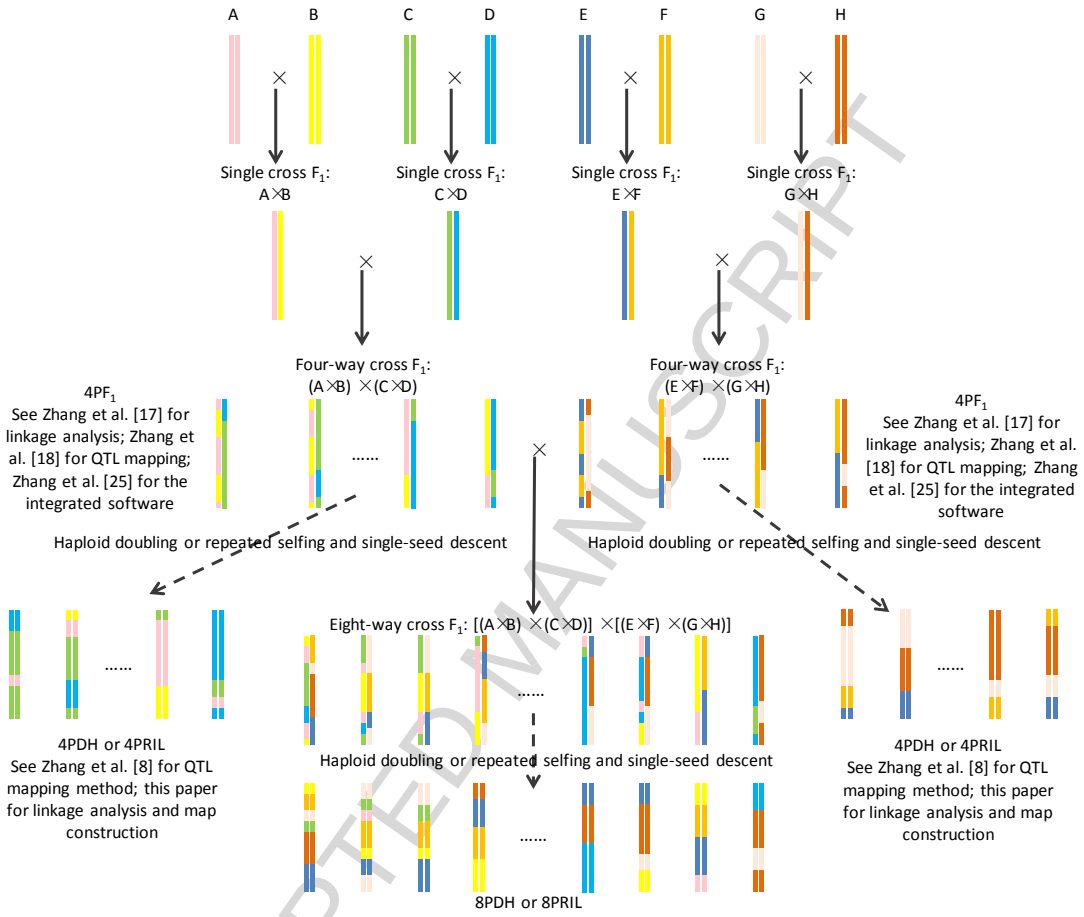

Fig. 1 – Diagram of the development of pure-line populations derived from four-way and eight-way crosses. Starting from eight parental lines, four single crosses are made first, followed by two four-way (or double) crosses (on respectively the left and right sides in the figure). Genetic analysis methods in four-way cross F_1 populations have been previously reported [17, 18, 25]. From each four-way cross, a pure-line population is called 4PDH when generated by haploid doubling and 4PRIL when generated by repeated selfing and single-seed descent. QTL mapping methods in 4PDH and 4PRIL have been reported [8]. One eight-way cross can be made between the two four-way cross populations. A pure-line population is called 8PDH when generated by haploid doubling and 8PRIL when generated by repeated selfing and single-seed descent.

Eight inbred parental lines, parents A, B, C, D, E, F, G, and H, are needed to make an eight-way cross. Similarly, DH lines and RILs can be produced from the eight-way cross (Fig. 1). For pure lines from an eight-way cross, a total of 4139 marker categories may be defined. Markers belonging to category ABCDEFGH represent the ideal situation, in which the parents carry eight identifiable alleles, denoted by *A*, *B*, *C*, *D*, *E*, *F*, *G*, and *H*. Their corresponding genotypes are denoted as *AA*, *BB*, *CC*, *DD*, *EE*, *FF*, *GG*, and *HH*. Markers belonging to the remaining categories are called incomplete loci. For example, marker categories may be AACDEFGH,

ABCDEFGG, AAADEFGH, and so on. Missing genotypes in both populations are coded as *XX*.

For clarity, pure-line populations are denoted as 4PDH and 4PRIL when derived from four-way crosses, and 8PDH and 8PRIL when derived from eight-way crosses.

2.2. Ideal scenario for estimating recombination frequency between two loci

For 4PDH and 4PRIL, assume that marker loci 1 and 2 are linked, falling into one of the 14 categories previously described. Let A_1, B_1, C_1, and D_1 denote the four alleles at locus 1 and A_2, B_2, C_2, and D_2 the four alleles at locus 2. The one-meiosis recombination frequency between the two loci is denoted as r. Based on the 14 marker categories, 105 scenarios may be considered to estimate r. The ideal scenario is represented by the case of two complete markers. Table 1 shows the theoretical probabilities of the 16 identifiable genotypes. The likelihood function (L) and logarithm of the likelihood (lg L) are given in Equation (1) for 4PDH and Equation (2) for 4PRIL, respectively. For 4PDH,

$$L = \frac{n!}{n_1!\cdots n_{16}!} r^{n-(n_1+n_6+n_{11}+n_{16})}(1-r)^{2(n_1+n_6+n_{11}+n_{16})+n_2+n_5+n_{12}+n_{15}},$$

$$\begin{aligned}\lg L = {} & C + \left[n - \left(n_1 + n_6 + n_{11} + n_{16}\right)\right]\lg r \\ & + \left[2\left(n_1 + n_6 + n_{11} + n_{16}\right) + n_2 + n_5 + n_{12} + n_{15}\right]\lg\left(1 - r\right)\end{aligned} \quad (1)$$

and for 4PRIL,

$$L = \frac{n!}{n_1!\cdots n_{16}!}\left(\frac{1-r}{4(1+2r)}\right)^{n_1+n_6+n_{11}+n_{16}}\left(\frac{r}{4(1+2r)}\right)^{n-(n_1+n_6+n_{11}+n_{16})},$$

$$\begin{aligned}\lg L = {} & C + \left(n_1 + n_6 + n_{11} + n_{16}\right)\lg\left(1 - r\right) \\ & + \left[n - \left(n_1 + n_6 + n_{11} + n_{16}\right)\right]\lg r - n\lg\left(1 + 2r\right)\end{aligned} \quad (2)$$

where n_1, n_2, …, and n_{16} are the sample sizes of the 16 genotypes and C is a constant independent of the unknown recombination frequency. For examples, n_1 to n_4 are the sample sizes of genotypes $A_1A_1A_2A_2$, $A_1A_1B_2B_2$, $A_1A_1C_2C_2$, and $A_1A_1D_2D_2$, respectively, and n_{13} to n_{16} are the sample sizes of genotypes $D_1D_1A_2A_2$, $D_1D_1B_2B_2$, $D_1D_1C_2C_2$, and $D_1D_1D_2D_2$, respectively.

Table 1 – Probabilities of 16 pairwise marker types in pure-line populations from four-way crosses when both markers belong to category ABCD. A_1, B_1, C_1, and D_1 are the four alleles at one marker locus. A_2, B_2, C_2, and D_2 are the four alleles at the other locus. r is the one-meiosis recombination frequency.

Marker type	4PDH				4PRIL			
	A_2A_2	B_2B_2	C_2C_2	D_2D_2	A_2A_2	B_2B_2	C_2C_2	D_2D_2
A_1A_1	$\frac{1}{4}(1-r)^2$	$\frac{1}{4}r(1-r)$	$\frac{1}{8}r$	$\frac{1}{8}r$	$\frac{1-r}{4(1+2r)}$	$\frac{r}{4(1+2r)}$	$\frac{r}{4(1+2r)}$	$\frac{r}{4(1+2r)}$
B_1B_1	$\frac{1}{4}r(1-r)$	$\frac{1}{4}(1-r)^2$	$\frac{1}{8}r$	$\frac{1}{8}r$	$\frac{r}{4(1+2r)}$	$\frac{1-r}{4(1+2r)}$	$\frac{r}{4(1+2r)}$	$\frac{r}{4(1+2r)}$
C_1C_1	$\frac{1}{8}r$	$\frac{1}{8}r$	$\frac{1}{4}(1-r)^2$	$\frac{1}{4}r(1-r)$	$\frac{r}{4(1+2r)}$	$\frac{r}{4(1+2r)}$	$\frac{1-r}{4(1+2r)}$	$\frac{r}{4(1+2r)}$
D_1D_1	$\frac{1}{8}r$	$\frac{1}{8}r$	$\frac{1}{4}r(1-r)$	$\frac{1}{4}(1-r)^2$	$\frac{r}{4(1+2r)}$	$\frac{r}{4(1+2r)}$	$\frac{r}{4(1+2r)}$	$\frac{1-r}{4(1+2r)}$

Solving the likelihood equation by setting $\frac{d \lg L}{dr} = 0$ yields the maximum likelihood estimate (MLE) of recombination frequency given in Equations (3) and (4) for the two respective populations. For 4PDH,

$$\hat{r} = \frac{n - (n_1 + n_6 + n_{11} + n_{16})}{2(n_1 + n_6 + n_{11} + n_{16}) + n_2 + n_5 + n_{12} + n_{15}}, \quad (3)$$

and for 4PRIL,

$$\hat{r} = \frac{n - (n_1 + n_6 + n_{11} + n_{16})}{n + 2(n_1 + n_6 + n_{11} + n_{16})} \quad (4)$$

where n is the total sample size, i.e. $n = n_1 + \ldots + n_{16}$.

Table 2 shows the theoretical probabilities of the 64 identifiable genotypes in the ideal scenario for 8PDH and 8PRIL. Let n_1, n_2, …, and n_{64} be the sample sizes of the 64 genotypes. For example, n_1 to n_8 are the sample sizes of genotypes $A_1A_1A_2A_2$, $A_1A_1B_2B_2$, …, $A_1A_1H_2H_2$; n_{57} to n_{64} are the sample sizes of genotypes $H_1H_1A_2A_2$, $H_1H_1B_2B_2$, …, $H_1H_1H_2H_2$. The MLE of recombination frequency is shown in Equations 5 and 6 for the two respective populations. For 8PDH,

$$\hat{r} = \frac{M_1}{M_1 + M_2}. \quad (5)$$

and for 8PRIL, $\hat{r}$ is the solution of Equation (6),

$$(-2M_1 - 2M_3 + 2n)\hat{r}^2 + (M_1 - M_3 - 2n)\hat{r} + M_1 = 0. \quad (6)$$

where

$$M_1 = n - (n_1 + n_{10} + n_{19} + n_{28} + n_{37} + n_{46} + n_{55} + n_{64}),$$

$$\begin{aligned} M_2 = {} & 3(n_1 + n_{10} + n_{19} + n_{28} + n_{37} + n_{46} + n_{55} + n_{64}) \\ & + 2(n_2 + n_9 + n_{20} + n_{27} + n_{38} + n_{45} + n_{56} + n_{63}) \\ & + n_3 + n_4 + n_{11} + n_{12} + n_{17} + n_{18} + n_{25} + n_{26} + n_{39} + n_{40} \\ & + n_{47} + n_{48} + n_{53} + n_{54} + n_{61} + n_{62}, \text{ and} \end{aligned}$$

$$\begin{aligned} M_3 = {} & 2(n_1 + n_{10} + n_{19} + n_{28} + n_{37} + n_{46} + n_{55} + n_{64}) \\ & + (n_2 + n_9 + n_{20} + n_{27} + n_{38} + n_{45} + n_{56} + n_{63}) \end{aligned}$$

Table 2 – Probabilities of 64 pairwise marker types in pure-line populations from eight-way crosses when both markers belong to category ABCDEFGH. A_1, B_1, C_1, D_1, E_1, F_1, G_1, and H_1 denote the eight alleles at one marker locus. A_2, B_2, C_2, D_2, E_2, F_2, G_2, and H_2 denote the eight alleles at the other locus. r is the one-meiosis recombination frequency.

Pop.	Type	A_2A_2	B_2B_2	C_2C_2	D_2D_2	E_2E_2	F_2F_2	G_2G_2	H_2H_2
8PDH	A_1A_1	$\frac{1}{8}(1-r)^3$	$\frac{1}{8}r(1-r)^2$	$\frac{1}{16}r(1-r)$	$\frac{1}{16}r(1-r)$	$\frac{1}{32}r$	$\frac{1}{32}r$	$\frac{1}{32}r$	$\frac{1}{32}r$
	B_1B_1	$\frac{1}{8}r(1-r)^2$	$\frac{1}{8}(1-r)^3$	$\frac{1}{16}r(1-r)$	$\frac{1}{16}r(1-r)$	$\frac{1}{32}r$	$\frac{1}{32}r$	$\frac{1}{32}r$	$\frac{1}{32}r$
	C_1C_1	$\frac{1}{16}r(1-r)$	$\frac{1}{16}r(1-r)$	$\frac{1}{8}(1-r)^3$	$\frac{1}{8}r(1-r)^2$	$\frac{1}{32}r$	$\frac{1}{32}r$	$\frac{1}{32}r$	$\frac{1}{32}r$
	D_1D_1	$\frac{1}{16}r(1-r)$	$\frac{1}{16}r(1-r)$	$\frac{1}{8}r(1-r)^2$	$\frac{1}{8}(1-r)^3$	$\frac{1}{32}r$	$\frac{1}{32}r$	$\frac{1}{32}r$	$\frac{1}{32}r$
	E_1E_1	$\frac{1}{32}r$	$\frac{1}{32}r$	$\frac{1}{32}r$	$\frac{1}{32}r$	$\frac{1}{8}(1-r)^3$	$\frac{1}{8}r(1-r)^2$	$\frac{1}{16}r(1-r)$	$\frac{1}{16}r(1-r)$
	F_1F_1	$\frac{1}{32}r$	$\frac{1}{32}r$	$\frac{1}{32}r$	$\frac{1}{32}r$	$\frac{1}{8}r(1-r)^2$	$\frac{1}{8}(1-r)^3$	$\frac{1}{16}r(1-r)$	$\frac{1}{16}r(1-r)$
	G_1G_1	$\frac{1}{32}r$	$\frac{1}{32}r$	$\frac{1}{32}r$	$\frac{1}{32}r$	$\frac{1}{16}r(1-r)$	$\frac{1}{16}r(1-r)$	$\frac{1}{8}(1-r)^3$	$\frac{1}{8}r(1-r)^2$
	H_1H_1	$\frac{1}{32}r$	$\frac{1}{32}r$	$\frac{1}{32}r$	$\frac{1}{32}r$	$\frac{1}{16}r(1-r)$	$\frac{1}{16}r(1-r)$	$\frac{1}{8}r(1-r)^2$	$\frac{1}{8}(1-r)^3$
8PRIL	A_1A_1	$\frac{(1-r)^2}{8(1+2r)}$	$\frac{r(1-r)}{8(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$
	B_1B_1	$\frac{r(1-r)}{8(1+2r)}$	$\frac{(1-r)^2}{8(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$
	C_1C_1	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{(1-r)^2}{8(1+2r)}$	$\frac{r(1-r)}{8(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$
	D_1D_1	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r(1-r)}{8(1+2r)}$	$\frac{(1-r)^2}{8(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$
	E_1E_1	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{(1-r)^2}{8(1+2r)}$	$\frac{r(1-r)}{8(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$
	F_1F_1	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r(1-r)}{8(1+2r)}$	$\frac{(1-r)^2}{8(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$
	G_1G_1	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{(1-r)^2}{8(1+2r)}$	$\frac{r(1-r)}{8(1+2r)}$
	H_1H_1	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r}{16(1+2r)}$	$\frac{r(1-r)}{8(1+2r)}$	$\frac{(1-r)^2}{8(1+2r)}$

2.3. Recombination frequency estimation in other scenarios

The EM algorithm [22] is used for estimating recombination frequency in the other scenarios. The initial value of r is set at 0.25. In the E step, the sample size n_i is calculated or updated by probabilities of genotypes in the ideal scenario. Consider, for example, 4PDH and 4PRIL, and the scenario in which one marker category is ABCD and the other AACD. Locus 1 has four identifiable genotypes: A_1A_1, B_1B_1, C_1C_1, and D_1D_1, and locus 2 has three identifiable genotypes: $A_2A_2 + B_2B_2$ (i.e. A_2A_2 or B_2B_2), C_2C_2, and D_2D_2. Table 3 shows the theoretical probabilities of the 12 identifiable genotypes, with sample sizes were represented by N_1, N_2, …, N_{12}, respectively.

Table 3 – Probabilities of 12 pairwise marker classes in pure-line populations from four-way crosses when one locus belongs to category ABCD and the other belongs to category AACD. A_1, B_1, C_1, and D_1 are the four alleles at one locus. A_2, B_2, C_2 and D_2 are the four alleles at the other locus, but A_2 and B_2 cannot be distinguished. r is the one-meiosis recombination frequency. The last column represents the observed sample size of each genotype.

Marker class	Locus 1	Locus 2	Probability in 4PDH	Probability in 4PRIL	Sample size
1	A_1A_1	A_2A_2+B_2B_2	$\frac{1}{4}(1-r)$	$\frac{1}{4(1+2r)}$	N_1
2	A_1A_1	C_2C_2	$\frac{1}{8}r$	$\frac{r}{4(1+2r)}$	N_2
3	A_1A_1	D_2D_2	$\frac{1}{8}r$	$\frac{r}{4(1+2r)}$	N_3
4	B_1B_1	A_2A_2+B_2B_2	$\frac{1}{4}(1-r)$	$\frac{1}{4(1+2r)}$	N_4
5	B_1B_1	C_2C_2	$\frac{1}{8}r$	$\frac{r}{4(1+2r)}$	N_5
6	B_1B_1	D_2D_2	$\frac{1}{8}r$	$\frac{r}{4(1+2r)}$	N_6
7	C_1C_1	A_2A_2+B_2B_2	$\frac{1}{4}r$	$\frac{r}{2(1+2r)}$	N_7
8	C_1C_1	C_2C_2	$\frac{1}{4}(1-r)^2$	$\frac{1-r}{4(1+2r)}$	N_8
9	C_1C_1	D_2D_2	$\frac{1}{4}r(1-r)$	$\frac{r}{4(1+2r)}$	N_9
10	D_1D_1	A_2A_2+B_2B_2	$\frac{1}{4}r$	$\frac{r}{2(1+2r)}$	N_{10}
11	D_1D_1	C_2C_2	$\frac{1}{4}r(1-r)$	$\frac{r}{4(1+2r)}$	N_{11}
12	D_1D_1	D_2D_2	$\frac{1}{4}(1-r)^2$	$\frac{1-r}{4(1+2r)}$	N_{12}

Eight of the 12 genotypes in Table 3 are exactly the same as in the ideal scenario in Table 1. The other four genotypes are combinations of two genotypes in Table 1. For example, N_1 is the sample size of the genotype having A_1A_1 at locus 1 and A_2A_2+B_2B_2 at locus 2. The ratio of probabilities from genotypes A_2A_2 and B_2B_2 is $\frac{1}{4}(1-r)^2 : \frac{1}{4}r(1-r) = (1-r):r$ in both 4PDH and 4PRIL. So N_1 is divided into n_1 (corresponding to genotype A_2A_2 at locus 2) and n_2 (corresponding to genotype B_2B_2 at locus 2) in the ratio $(1-r):r$. Similarly, N_4 is divided into n_5 (corresponding to genotype A_2A_2 at locus 2) and n_6 (corresponding to genotype B_2B_2 at locus 2) in the ratio $r:(1-r)$; N_7 is divided into n_9 (corresponding to genotype A_2A_2 at locus 2) and n_{10} (corresponding to genotype B_2B_2 at locus 2) in the ratio 1:1; and N_{10} is divided into n_{13} (corresponding to genotype A_2A_2 at locus 2) and n_{14} (corresponding to genotype B_2B_2 at locus 2) in the ratio of 1:1. Thus,

$$n_1 = (1-r)N_1,\ \ n_2 = rN_1,\ \ n_5 = rN_4,\ \ n_6 = (1-r)N_4,$$

$$n_9 = n_{10} = 0.5N_7,\ n_{13} = n_{14} = 0.5N_{10}$$

In the M step, r is updated using Equations (3) and (4) and used to recalculate the probabilities of the 16 genotypes in Table 1. The EM iterations continue until the difference in r between two consecutive iterations reaches a predefined precision criterion of, by default, 1×10^{-6}. The MLE of recombination frequency is thus obtained and then used for linkage map construction. A combination of a nearest-neighbor algorithm and a two-opt algorithm for the Traveling Salesman Problem (TSP, [23]) is used for marker ordering, as in our previous studies [17, 24, 25]. The nearest-neighbor algorithm is used to determine an initial solution and the two-opt algorithm is then used for improving the solution.

2.4. Development of integrated software GAPL for genetic analysis

The proposed methods for estimating recombination frequencies and building linkage maps have been implemented in a software package named "Genetic Analysis of Multi-parental Pure-line Populations", or GAPL, which is freely available from www.isbreeding.net. GAPL is an integrated software package combining linkage analysis, map construction, and QTL mapping for pure-line populations from four-way and eight-way crosses. Core modules for recombination frequency estimation, linkage-map construction, and QTL mapping algorithms were written in FORTRAN 90/95. The user interface for the software was written in C#. GAPL runs on Windows XP/Vista/7/8/10, with Microsoft.NET Framework 2.0 (×86) or higher versions. GAPL is project-based software: all operations and files can be stored in projects, as with package QTL IciMapping for bi-parental populations [24] and package GACD for clonal F_1 and four-way crosses [25].

2.5. Simulation study

To investigate the efficiency of our methods, two chromosomes were simulated. Twenty markers were unevenly distributed on chromosome I. The minimum and maximum recombination frequency between two neighboring markers was 0.005 and 0.101, equivalent to 0.5025 cM and 11.2823 cM in mapping distance under the Haldane mapping function. One thousand each of 4PDH, 4PRIL, 8PDH, and 8PRIL populations consisting of 200 pure lines were simulated with the genetics and breeding simulation tool of QuLine [26, 27]. No missing marker data points were simulated. For 4PDH and 4PRIL populations, seven markers were randomly chosen and assigned to category ABCD and the other 13 were randomly assigned to the other 13 categories. In the end, markers 1, 5, 8, 12, 17, 18, and 20 belonged to category ABCD and the other markers belonged to incomplete categories. In the 8PDH and 8PRIL populations, all markers belonged to category ABCDEFGH for simplicity.

The first simulated population from each population type was selected for the demonstration of linkage-map construction. These populations are denoted as Pop1 to Pop4 for 4PDH, 4PRIL, 8PDH, and 8PRIL, respectively. To investigate the effect of distortion on map construction, populations with distorted markers were generated. Pop1 to Pop4 were used as a start, with no markers showing distortion. The steps for generating distorted populations were as follows. Considering types at marker 12, fifty *AA* or *BB* individuals were randomly deleted in Pop1 and Pop2 and fifty *AA*, *BB*, *CC*, or *DD* individuals were randomly deleted in Pop3 and Pop4. The distorted populations had only 150 individuals, denoted as Pop5 to Pop8, respectively. For example, in Pop5, the percentages of genotypes *AA*, *BB*, *CC*, and *DD* at marker 12 were 16.00%, 12.67%, 36.00%, and 35.33%, where the *P*-value of the χ^2 test for distortion was equal to 4.29×10^{-6}. Markers 5, 6, 8, 9, 11, 13, 14, and 15 also showed segregation distortion at the significance level of 0.05, owing to their linkage with marker 12.

Two hundred markers were evenly distributed on chromosome II. The marker distance between any two adjacent markers was set at 1 cM and the Haldane mapping function was used to convert mapping distance to recombination frequency. One bi-parental RIL population with 200 lines was simulated with QTL IciMapping [24] and denoted as Pop9. Pop9 can be regarded as a special case of 4PRIL considering the third parent to be the same as the first one and the fourth parent to be the same as the second one. Categories of all markers were ABAB. Before linkage map construction, the marker order was shuffled.

Recombination frequency estimation and linkage map construction in Pop1 to Pop9 were performed with

GAPL. The Haldane mapping function was used to convert recombination frequency (*r*) to map distance (*d*) in cM. For comparison, the R/qtl and R/mpMap packages were also used for linkage map construction in Pop 9. The best order in R/qtl was determined by function "orderMarker" where the initial order was established by a greedy algorithm and then refined by rippling. A window size of 3 was used for both ordering and rippling. Countxo (comparing orders by counting the number of obligate crossovers) and likelihood methods were used in rippling. The best order in R/mpMap was determined by function "mporder" with two-point ordering selected. Multi-point ordering in R/mpMap was not used for comparison, as its algorithm was based on R/qtl. The other parameters were set at their default values.

2.6. Real data of a 4PRIL population in wheat

The real dataset used in this study was derived from four Australian wheat cultivars (Yitpi, Baxter, Chara and Westonia [13]). A total of 1063 pure lines were generated by single-seed decent, and genotyped with SNPs, DArT markers, and microsatellites. Verbyla et al. [28] used the R/mpMap package as well as manual intervention to build the linkage map. The full genome was 5787.73 cM in length with 3230 markers distributed across the 21 wheat chromosomes plus three additional linkage groups. Marker intervals were predominantly shorter than 5 cM and the average marker interval was 1.79 cM. For comparison, the linkage map was rebuilt with GAPL. The numbers of groups and of markers in each group were the same as those in Verbyla et al..

3. Results

3.1. Functionalities in integrated software GAPL

Four functionalities have been implemented in the integrated software package GAPL version 1.2, i.e. (1) SNP, SNP genotypic data conversion; (2) BIN, binning of redundant markers; (3) PLM, map construction in multi-parental pure-line populations; and (4) PLQ, QTL detection in multi-parental pure-line populations. The four functionalities can act as a pipeline. The input file for one functionality can be found in the outputs of the previous functionality. Several examples are provided in an example folder in the software in three formats, i.e. pure text, Microsoft Excel 2003 and Excel 2007. Missing genotypes are allowed in any functionality, and are not used in recombination frequency estimation but may be imputed for QTL mapping using the linkage information.

The SNP functionality helps convert SNP data of DNA bases (i.e. A, T, G, or C) into a format that can be recognized in GAPL (A, B, C, D, E, F, G, or H) (Fig. 2-A). SNPs showing non-polymorphism in parents or progenies or missing in one or more parents are deleted in this functionality. The output of the functionality can be directly used as input to the next functionality of redundant-marker removal or linkage-map construction.

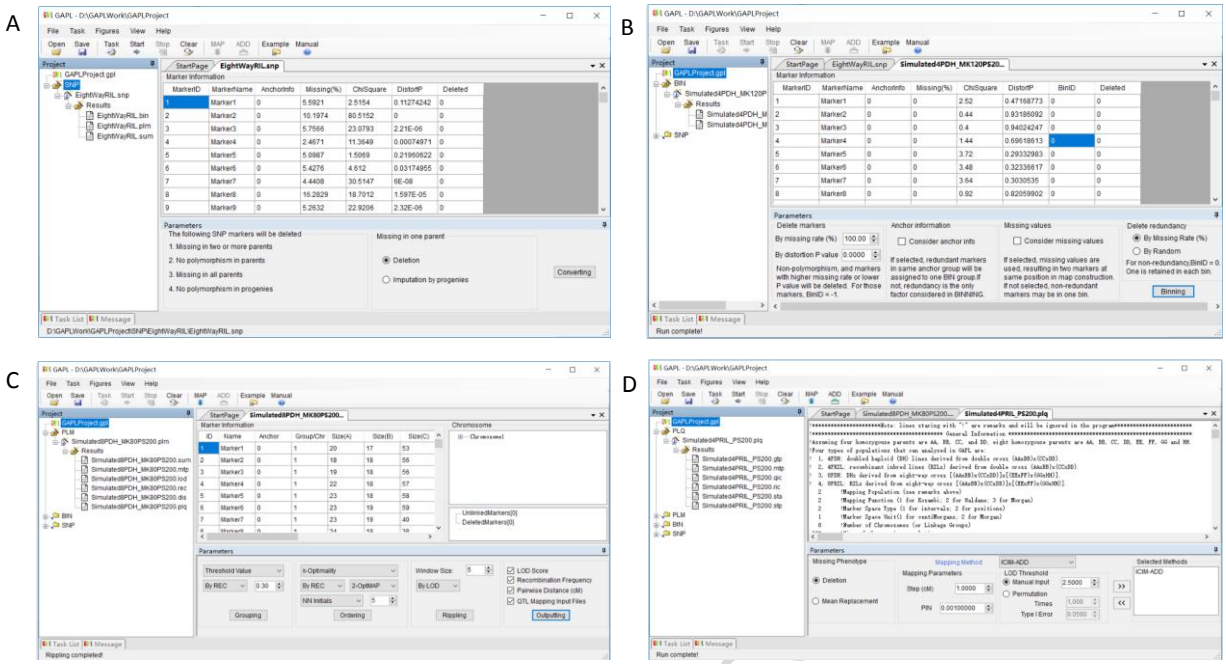

Fig. 2 – User interface of the integrated software package GAPL. A, interface of functionality SNP; B, interface of functionality BIN; C, interface of functionality PLM; D, interface of functionality PLQ.

The BIN functionality in GAPL (Fig. 2-B) is similar to those of QTL IciMapping [24] and GACD [25]. Users can use BIN to remove redundancy and perform quality control of markers, for example by deleting markers with high missing rates or severe segregation distortion. The output of the functionality can be used as input to the next functionality of linkage map construction.

The PLM functionality was designed for linkage analysis and map construction in multi-parental pure-line populations (Fig. 2-C). Three steps are involved in map construction: grouping, ordering, and rippling. Algorithms used in the three steps are the same as those in QTL IciMapping [24] and GACD [25]. Users can build linkage maps by clicking buttons for grouping, ordering and rippling in turn. They can also modify the constructed map at any step using the interface. PLM generates several files, including summary information of linkage maps, LOD scores, recombination frequencies and genetic distances between markers, and an input file for the next functionality, QTL mapping.

The PLQ functionality was developed for QTL mapping in multi-parental pure-line populations (Fig. 2-D). Three mapping methods are available in PLQ: single-marker analysis (SMA, [29]), interval mapping (IM, [30]), and inclusive composite interval mapping (ICIM, [8]). For each method, several parameters should be determined before mapping, for example LOD threshold and scan step size. A LOD threshold can also be determined by permutation testing. Both plain text files and figures are available to display mapping results including QTL positions, LOD scores, and effects at all scan positions.

3.2. Recombination frequency estimation and linkage map construction for simulated chromosome I

Table 4 shows the average recombination frequency between neighboring markers estimated from 1000 simulated populations. The recombination frequency estimate was unbiased irrespective of population type. For example,

the true recombination frequency between markers 1 and 2 was 0.090, and the estimated values were 0.090, 0.091, 0.091, and 0.090 in 4PDH, 4PRIL, 8PDH, and 8PRIL, respectively. The corresponding standard errors were 0.017, 0.014, 0.014, and 0.014, indicating the high estimation accuracy.

Table 4 – Means and standard errors of estimated recombination frequencies between neighboring markers on simulated chromosome I in simulated DH and RIL populations derived from four-way and eight-way crosses.

Marker	True value	Category in 4PDH and 4PRIL	Estimated recombination frequency with the next marker			
	value		4PDH	4PRIL	8PDH	8PRIL
1	0.090	ABCD	0.090 ± 0.017	0.091 ± 0.014	0.091 ± 0.014	0.090 ± 0.014
2	0.005	AACD	0.005 ± 0.010	0.005 ± 0.010	0.005 ± 0.000	0.005 ± 0.000
3	0.077	ABAA	0.076 ± 0.053	0.081 ± 0.048	0.077 ± 0.014	0.077 ± 0.010
4	0.053	AACA	0.054 ± 0.017	0.053 ± 0.014	0.053 ± 0.010	0.054 ± 0.010
5	0.065	ABCD	0.065 ± 0.014	0.065 ± 0.014	0.065 ± 0.010	0.065 ± 0.010
6	0.010	ABCC	0.010 ± 0.010	0.010 ± 0.010	0.010 ± 0.000	0.010 ± 0.000
7	0.055	ABBA	0.056 ± 0.014	0.055 ± 0.014	0.055 ± 0.010	0.055 ± 0.010
8	0.063	ABCD	0.063 ± 0.014	0.063 ± 0.014	0.064 ± 0.010	0.063 ± 0.010
9	0.037	ABAD	0.037 ± 0.014	0.038 ± 0.010	0.037 ± 0.010	0.037 ± 0.010
10	0.063	ABCA	0.063 ± 0.014	0.064 ± 0.014	0.064 ± 0.010	0.064 ± 0.010
11	0.043	ABBD	0.043 ± 0.010	0.043 ± 0.010	0.043 ± 0.010	0.044 ± 0.010
12	0.006	ABCD	0.006 ± 0.000	0.006 ± 0.000	0.006 ± 0.000	0.006 ± 0.000
13	0.026	AACC	0.027 ± 0.022	0.027 ± 0.017	0.026 ± 0.010	0.026 ± 0.000
14	0.061	AAAD	0.062 ± 0.030	0.063 ± 0.026	0.061 ± 0.010	0.061 ± 0.010
15	0.043	ABCB	0.043 ± 0.010	0.044 ± 0.010	0.043 ± 0.010	0.043 ± 0.010
16	0.046	ABAB	0.046 ± 0.014	0.046 ± 0.010	0.046 ± 0.010	0.047 ± 0.010
17	0.075	ABCD	0.075 ± 0.014	0.075 ± 0.014	0.075 ± 0.010	0.075 ± 0.010
18	0.101	ABCD	0.102 ± 0.022	0.102 ± 0.022	0.101 ± 0.014	0.101 ± 0.014
19	0.015	ABBB	0.015 ± 0.010	0.015 ± 0.010	0.015 ± 0.000	0.015 ± 0.000
20	-	ABCD	-	-	-	-

In Pop1 to Pop4, no marker was distorted, and marker orders were the same as predefined, i.e. from marker 1 to marker 20. The estimated map lengths were 99.53, 99.34, 100.39, and 99.79 cM, respectively, close to the true value 100.03 cM. In Pop5 to Pop8, some markers were distorted, but the correct orders were still achieved. The estimated map lengths were 97.84, 99.97, 101.65, and 100.30 cM, also close to the true value. It can be concluded that segregation distortion had little effect on recombination frequency estimation and linkage map construction.

3.3. Linkage map construction for simulated chromosome II

Figure 3-A shows the linkage map constructed by GAPL in Pop 9. The marker order was the same as predefined, i.e. from marker 1 to marker 200. The length of the map was estimated at 200.58 cM, close to the true length of 199.0 cM. All markers were approximately evenly distributed. There were five intervals with length 0 cM: between markers 26 and 27, 82 and 83, 88 and 89, 135 and 136, and 195 and 196, as no crossovers were observed in those intervals in Pop9. The maximal length of marker interval was 2.70 cM, between markers 103 and 104 and markers 108 and 109. Differences in lengths of marker intervals were caused by random recombination events in a limited-size population. Grouping, ordering, and rippling required 34 s on a personal computer, Lenovo W510

(Windows 10, Intel Core i7 Q720 CPU @ 1.60GHz).

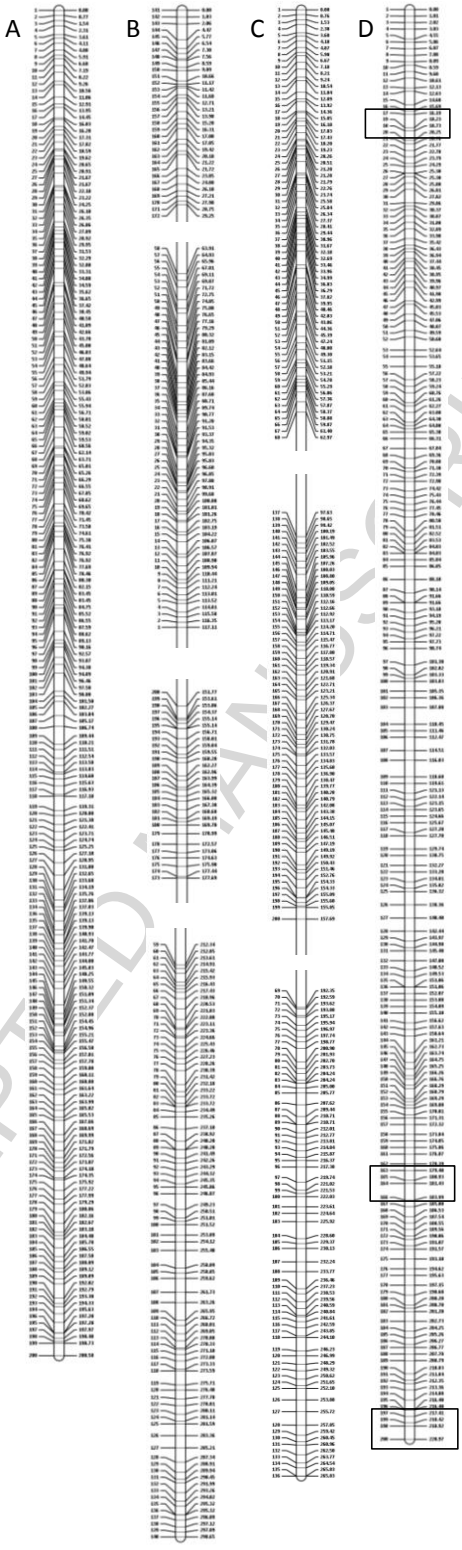

Fig. 3 – Linkage maps in Pop9 constructed with different software packages. A, GAPL: the correct order was achieved; B, R/qtl with the countxo method: the constructed map was broken by the three largest intervals; C, R/qtl with likelihood method: the constructed map was broken by the two largest intervals; D, R/mpMap: markers are in incorrect order, as indicated in rectangular boxes.

By R/qtl, if the countxo method was used in rippling, the marker order was 141 – ... – 172 – 58 –…– 1 – 200– …– 173 –59 – …– 140 (Fig. 3-B). Here "x – ... – y" represents continuous markers between two markers x and y. For example, "141 – ... – 172" is order 141, 142 to 171, 172. "Marker 58 – ... –1" is order 58, 57 to 2, 1. The map length was estimated at 298.65 cM, 99.65 cM longer than the true order. Markers were not evenly distributed, and three long gaps were observed on the chromosome (Fig. 3-B). The maximum length of a marker interval was 34.66 cM, between markers 172 and 58, 1 and 200, and 173 and 59. Linkage map construction required 131 s on the same personal computer, much slower than by GAPL.

By R/qtl, if the likelihood method was used in rippling, the marker order was 1 – ...–68 –137–…– 200 – 69 – …– 136 (Fig. 3-C). The map length was estimated at 265.83 cM, 66.83 cM longer than the true length. Two long gaps were observed on the constructed map, each with length 34.66 cM, between markers 68 and 137 and markers 200 and 69. Map construction cost 5 h 21 min and 36 s on the same computer, much slower than by GAPL and the countxo method in R/qtl.

By R/mpMap as well, the marker order was not the same as the predefined order. Markers 18 and 19, 164 and 165, and 198 and 199 were reversed (Fig. 3-D). The map length was estimated at 220.97 cM, 21.97 cM longer than the true length. The maximum length of a marker interval was 2.56 cM, between markers 96 and 97, 103 and 104, and 164 and 166. Linkage map construction required 48 s on the same computer, slower than in GAPL but faster than in R/qtl.

3.4. Linkage maps in real wheat 4PRIL population

The newly built map by the GAPL software was 3807.10 cM in length: 1457.21 cM for the A genome, 1277.86 cM for the B genome, 1019.90 cM for the D genome, and 52.13 cM for the three additional groups (Fig. 4) This length is 1980.63 cM less than that reported by Verbyla et al. [28], and closer to those of other published linkage maps of wheat. The average marker interval was 1.01 cM. Each chromosome or group produced by GAPL was shorter than that produced by Verbyla et al.

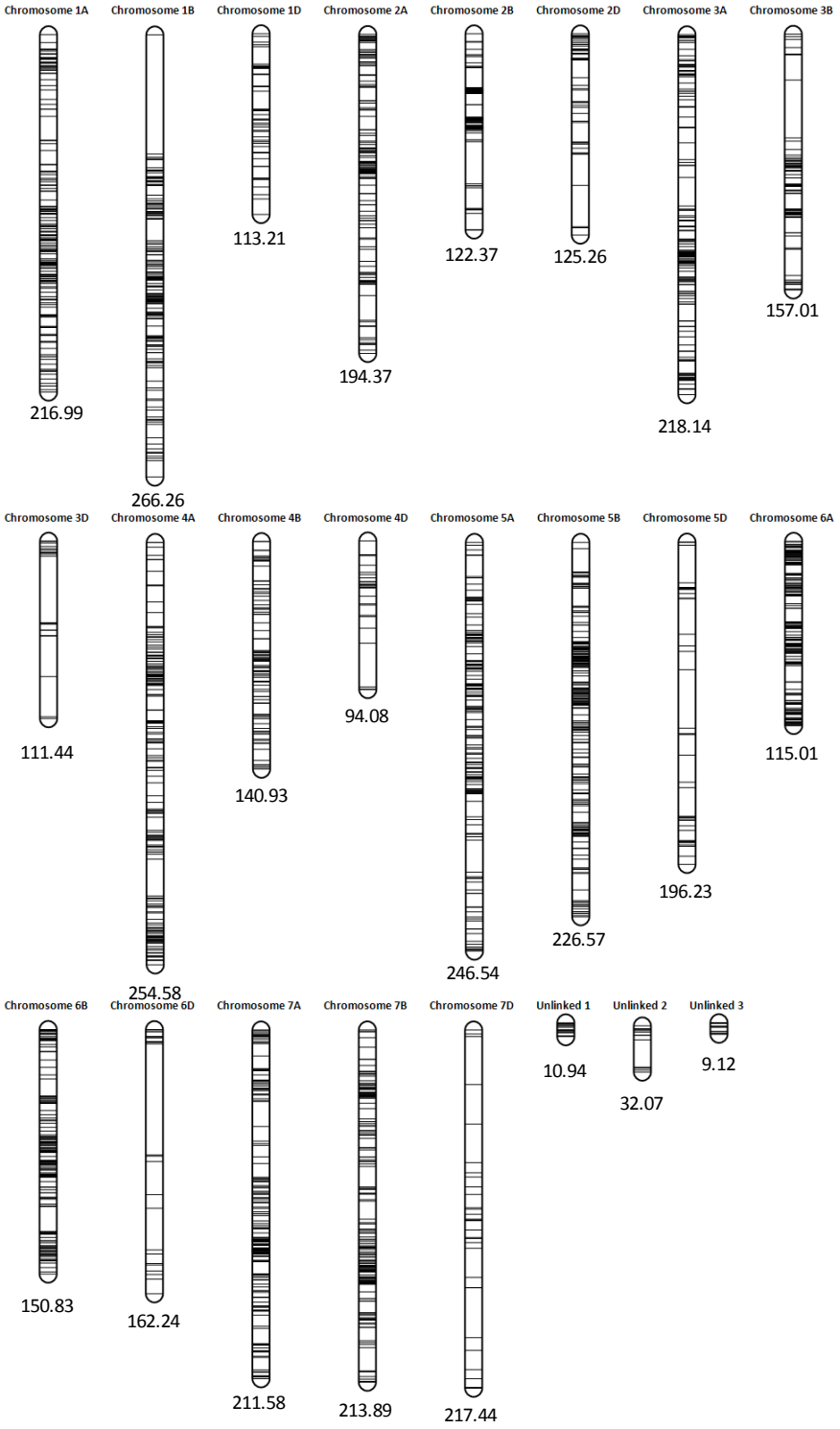

Fig. 4 – Linkage map constructed by GAPL using real data from a four-way cross RIL wheat population. Marker grouping is the same as reported in Verbyla et al. [28].

4. Discussion

In this study, we developed linkage-analysis methods for multi-parental pure-line populations. The unbiasedness and efficiency of our methods in recombination frequency estimation was confirmed by simulation study and a reported four-way cross RIL population. Our methods have been implemented in software package GAPL,

available for linkage map construction and QTL mapping in multi-parental pure-line populations. GAPL built better maps in much shorter time than R/qtl and R/mpMap. It is the first software package that is freely available for DH populations derived from four-way and eight-way crosses.

Missing markers contribute no information in linkage analysis, but their genotypes can be imputed from the constructed linkage map and then used for the next step of QTL mapping. An accurate linkage map leads to better imputation and consequently more reliable QTL identification and genomic analysis. Some algorithms have been proposed (for example, [31, 32]) for detecting and removing genotyping errors to improve the accuracy of linkage analysis. However, these algorithms were developed only for bi-parental populations and may not be suitable for multi-parental populations. Efficiency of algorithms for missing-genotype imputation and genotyping error correction in multi-parental populations needs further study. Once tested, suitable algorithms will be implemented in the next version of software GAPL.

GAPL can be directly applied to pure-line populations from less than eight parents. For example, a top-cross represented by (A × B) × C is equivalent to the four-way cross (A × B) × (C × D), where parent C is same as D. A cross between two top-crosses, [(A × B) × C] × [(E × F) × G] is equivalent to the eight-way cross [(A × B) × (C × D)] × [(E × F) × (G × H)], where parent C is the same as D and G is the same as H. Linkage analysis methods and the GAPL software described in this study cannot be directly used for populations derived from more than eight parents. Extension to such populations may not be easy, owing to the difficulty in tracing the allelic origins of progeny to the parents, and the derivation of theoretical probabilities of marker types. In future when such populations become common in genetic studies, the algorithms we have described may need to be expanded to more complex multi-parental populations. The current version of GAPL is 32-bit and the upper limit of marker numbers is around 20,000. If more markers are involved, stack or memory overflow may occur. Another version of GAPL, for example a 64-bit version, a command-line version, or a Linux version might allow increasing the upper limit of marker number.

Availability

GAPL is freely available at http://www.isbreeding.net.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2016YFD0101804), the National Natural Science Foundation of China (31671280), and HarvestPlus (part of the CGIAR Research Program on Agriculture for Nutrition and Health, http://www.harvestplus.org/).

Conflicts of interest: none.

References

[1] P.X. Kover, W. Valdar, J. Trakalo, N. Scarcelli, I.M. Ehrenreich, M.D. Purugganan, C. Durrant, R. Mott, A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*, PLoS Genet. 5 (2009) e1000551.

[2] M. Dell'Acqua, D.M. Gatti, G. Pea, F. Cattnaro, F. Coppens, G. Magris, A.L. Hlaing, H.H. Aung, H. Nelissen, J. Baute, E. Frascaroli,G. A. Churchill, D. Inzé, M. Morgante, M.E. Pè, Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in *Zea mays*, Genome Biol. 16 (2015) 167.

[3] B.E. Huang, K.L. Verbyla, A.P. Verbyla, C. Raghavan, V.K. Singh, P. Gaur, H. Leung, R.K. Varshney, C.R. Cavanagh, MAGIC populations in crops: current status and future prospects, Theor. Appl. Genet. 128 (2015) 999–1017.

[4] R.H. Higgins, C.S. Thurber, I. Assaranurak, P.J. Brown, Multiparental mapping of plant height and flowering time QTL in partially isogenic sorghum families, G3-Genes Genomes Genet. 4 (2014) 1593–1602.

[5] S. Li, J. Wang, L. Zhang, Inclusive composite interval mapping of QTL by environment interactions in biparental populations, PLoS One 10 (2015) e0132414.

[6] C. Yin, H. Li, S. Li, L. Xu, Z. Zhao, J. Wang, Genetic dissection on rice grain shape by the two-dimensional image analysis in one *japonica*×*indica* population consisting of recombinant inbred lines, Theor. Appl. Genet. 128 (2015) 1969–1986.

[7] C. Yin, H. Li, Z. Zhao, Z. Wang, S. Liu, L. Chen, X. Liu, Y. Tian, J. Ma, L. Xu, D. Zhang, S. Zhu, D. Li, J. Wan, J. Wang, Genetic dissection of top three leaf traits in rice using progenies from one *japonica* × *indica* cross, J. Integr. Plant Biol. 59 (2017) 866–880.

[8] S. Zhang, L. Meng, J. Wang, L. Zhang, Background controlled QTL mapping in pure-line genetic populations derived from four-way crosses, Heredity 119 (2017) 256–264.

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

[10] J.R. Klasen, H.P. Piepho, B. Stich, QTL detection power of multi-parental RIL populations in *Arabidopsis thaliana*, Heredity 108 (2012) 626–632.

[11] N. Bandillo, C. Raghavan, P.A. Muyco, M.A.L Sevilla, I.T. Lobina, C.H. Dilla-Ermita, C.W. Tung, S. McCouch, M. Thomson, R. Mauleon, R.K. Singh, G. Gregorio, E. Redoña, H. Leung, Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding, Rice 6 (2013) 11.

[12] K.S. Ponce, G. Ye, X. Zhao, QTL identification for cooking and eating quality in *indica* rice using multi-parent advanced generation intercross (MAGIC) population, Front. Plant Sci. 9 (2018) 868.

[13] B.E. Huang, A.W. George, K.L. Forrest, A. Kilian, M.J. Hayden, M.K. Morell, C.R. Cavanagh, A multiparent advanced generation inter-cross population for genetic analysis in wheat, Plant Biotechnol. J. 10 (2012) 826–839.

[14] T. Würschum, W. Liu, K.V. Alheit, M.R. Tucker, M. Gowda, E.A. Weissmann, V. Hahn, H.P. Maurer, Adult plant development in triticale (× *Triticosecale* Wittmack) is controlled by dynamic genetic patterns of regulation, G3-Genes Genomes Genet. 4 (2014) 1585–1591.

[15] M. Shivakumar, G. Kumawat, C. Gireesh, S.V. Ramesh, S.M. Husain, Soybean MAGIC population: a novel resource for genetic and plant breeding, Curr. Sci. 11 (2018) 906–908.

[16] W. Sannemann, B.E. Huang, B. Mathew, J. Léon, Multi-parent advanced generation inter-cross in barley: high-resolution quantitative trait locus mapping for flowering time as a proof of concept, Mol. Breed. 35 (2015) 86.

[17] L. Zhang, H. Li, J. Wang, Linkage analysis and map construction in genetic populations of clonal F_1 and double cross, G3-Genes Genomes Genet. 5 (2015) 181–189.

[18] L. Zhang, H. Li, J. Ding, J. Wu, J. Wang, Quantitative trait locus mapping with background control in genetic populations of clonal F_1 and double cross, J. Integr. Plant Biol. 57 (2015) 1046–1062.

[19] K.W. Broman, H. Wu, S. Sen, G.A. Churchill, R/qtl: QTL mapping in experimental crosses, Bioinformatics 19 (2003) 889–890.

[20] R. Mott, C.J. Talbot, M.G. Turri, A.C. Collins, J. Flint, A method for fine mapping quantitative trait loci in outbred animal stocks, Proc. Natl. Acad. Sci. U.S.A. 97 (2000) 12649–12654.

[21] B.E. Huang, A.W. George, R/mpMap: a computational platform for the genetic analysis of multi-parent recombinant inbred lines, Bioinformatics 27 (2011) 727–729.

[22] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. Roy. Stat. Soc. B 39 (1977) 1–38.

[23] S. Lin, B.W. Kernighan, An effective heuristic algorithm for the traveling-salesman problem, Oper. Res. 21 (1973) 498–516.

[24] L. Meng, H. Li, L. Zhang, J. Wang, QTL IciMapping: integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations, Crop J. 3 (2015) 269–283.

[25] L. Zhang, L. Meng, W. Wu, J. Wang, GACD: Integrated software for genetic analysis in clonal F_1 and double cross populations, J. Hered. 106 (2015) 741–744.

[26] J. Wang, M. van Ginkel, D. Podlich, G. Ye, R. Trethowan, W. Pfeiffer, I.H. DeLacy, M. Cooper, S. Rajaram, Comparison of two breeding strategies by computer simulation, Crop Sci. 43 (2003) 1764–1773.

[27] J. Wang, M. van Ginkel, R. Trethowan, G. Ye, I. Delacy, D. Podlich, M. Copper, Simulating the effects of dominance and epistasis on selection response in the CIMMYT Wheat Breeding Program using QuCim, Crop Sci. 44 (2004) 2006–2018.

[28] A.P. Verbyla, A.W. George, C.R. Cavanagh, K.L. Verbyla, Whole-genome QTL analysis for MAGIC, Theor. Appl. Genet. 127 (2014) 1753–1770.

[29] J. Wang, H. Li, L. Zhang, Genetic Mapping and Breeding Design, Science Press, Beijing, 2014 (in Chinese).

[30] E.S. Lander, D. Botstein, Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps, Genetics 121 (1989) 185–199.

[31] Y. Wu, P.R. Bhat, T.J. Close, S. Lonardi, Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph, PLoS Genet. 4 (2008) e1000212.

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