The Author(s) *BMC Bioinformatics* 2017, **18**(Suppl 11):404
DOI 10.1186/s12859-017-1804-8

BMC Bioinformatics

RESEARCH Open Access

New statistical methods for estimation of recombination fractions in F_2 population

Yuan-De Tan[1], Xiang H. F. Zhang[1,2,3,4*] and Qianxing Mo[1,5*]

From The International Conference on Intelligent Biology and Medicine (ICIBM) 2016
Houston, TX, USA. 08-10 December 2016

Abstract

Background: Dominant markers in an F_2 population or a hybrid population have much less linkage information in repulsion phase than in coupling phase. Linkage analysis produces two separate complementary marker linkage maps that have little use in disease association analysis and breeding. There is a need to develop efficient statistical methods and computational algorithms to construct or merge a complete linkage dominant marker maps. The key for doing so is to efficiently estimate recombination fractions between dominant markers in repulsion phases.

Result: We proposed an expectation least square (ELS) algorithm and binomial analysis of three-point gametes (BAT) for estimating gamete frequencies from F_2 dominant and codominant marker data, respectively. The results obtained from simulated and real genotype datasets showed that the ELS algorithm was able to accurately estimate frequencies of gametes and outperformed the EM algorithm in estimating recombination fractions between dominant loci and recovering true linkage maps of 6 dominant loci in coupling and unknown linkage phases. Our BAT method also had smaller variances in estimation of two-point recombination fractions than the EM algorithm.

Conclusion: ELS is a powerful method for accurate estimation of gamete frequencies in dominant three-locus system in an F_2 population and BAT is a computationally efficient and fast method for estimating frequencies of three-point codominant gametes.

Keywords: Dominant marker, Codominant marker, Gamete frequency, EM algorithm, ELS algorithm

Background

A great advance has been made in building genetic maps of various species due to the development of large-scale molecular marker technologies [1–7] and statistical methods [4, 8–18]. However, mapping of numerous molecular markers has been complicated by linkage phases of dominance [14–16, 19]. In two-point analysis, markers in repulsion phase provide quite less linkage information than in coupling phase [14, 15, 20, 21]. This is especially true for dominant markers in F_2 population [14]. In practical mapping experiments, although the linkage phase for each dominant marker is random, a half of markers are derived from one of two coupling phases. The phase between couplings is repulsion [14, 15]. This situation results in two separate partner linkage maps for dominant markers: high linkage information content of markers in the coupling phase and low linkage information content of markers in the repulsion phase. Thus one has to build two complementary linkage maps [14, 15, 21, 22]. To date, there has not yet been an effective way to integrate both into a complete map. Mester et al. [15] attempted to use pairs of codominant and dominant (CD) markers to merge such two complementary maps because pairs of the CD markers in repulsion phase have much higher linkage information content than pairs of dominant-only markers in repulsion phase. However, this strategy demands that all dominant markers be paired with codominant markers, which is not a general case in mapping practice, otherwise, local and global disturbance will then violently affect the reliability of the integrated map.

The two-point analysis implemented by the expectation maximization (EM) algorithm [11–13, 23–25] is a

* Correspondence: xiangz@bcm.edu; qmo@bcm.edu
[1]Dan L. Ducan Cancer Center, Baylor College of Medicine, Houston, TX, USA
Full list of author information is available at the end of the article

highly powerful approach to estimate recombination fractions between codominant loci and between dominant loci in coupling phase, but the EM algorithm has very low power in estimation of recombination fractions between dominant loci in repulsion phase. This is because it is difficult for the EM algorithm to distinguish genotypes in coupling phase from those in repulsion phase for dominant markers.

Therefore, the key of developing a powerful method for mapping dominant loci in an intersection population is to overcome the difficulty of distinguishing coupling phase from repulsion phase. Since two-point analysis, as pointed out above, performs very poorly in the estimation of recombination fractions between dominant loci, three-point analysis is alternatively taken into account. However, few three-point EM algorithms can be applied to dominant markers because dominant markers are less informative for maximum likelihood estimation [26]. One effective way to carry out three-point analysis is to dissect three-point genotypes into various gamete components that are informative for distinction between coupling and repulsion phases, and then, to estimate their frequencies. With these estimated gamete frequencies, one can immediately estimate recombination fractions between dominant loci in couple and repulsion phases. A key to this strategy is to obtain estimate of gamete frequencies. On the basis of dissection of genotypes, Tan and Fu proposed a binomial analysis of three-point (BAT) to estimate frequencies of dominant gametes [19]. However, this binomial approach is limited to the frequency of the three-point recessive gamete abc. The accuracy of estimation is completely dependent on the observed frequency of its phenotype (aabbcc). We have developed a new method called "expectation least square" (ELS) to address this problem. ELS estimation, similarly to expectation maximum algorithm, is realized on the basis of Tan and Fu's BAT method [19]. That is, the expectation of phenotype frequencies can be given by using Eqs. (1-9) in the BAT of Tan and Fu [19], and the difference between estimated and expected values of phenotype frequencies is given using least square. The expectation and least square steps are iterated so that the difference between estimated and expected values is less than tolerant value. In addition, we have also developed a fast binomial approach to estimate frequencies of codominant gametes.

Methods

Real data collection

Mouse genotype data: A RFLP dataset of 333 F_2 mice was obtained from MAPMAKER/EXP (version 3.0b) [13].

Simulation

For dominant loci, we just took unknown phase into account in simulation and followed a point process model [27] and scheme of Tan and Fu [19] to perform simulations. In N F_1 meioses, recombination events occurred at random between two adjacent loci. Here for the simplicity, we allowed for only independent crossovers during procedure of recombination occurrence between nonsister chromatides. We generated N F_2 individuals with ratio = phenotype A: phenotype a = 3:1 at each dominant locus or A(homozygote): H(heterozygote): B(homozygote) = 1:2:1 at each codominant locus. We set three levels for sample size: N = 100, 200, and 300 F_2 individuals and 100 iterations and used variance (equivalent to mean square error, MSE) that quantifies deviation of estimated recombination fraction between two adjacent loci from its true value to evaluate these estimators. Since the ELS and BAT estimators work in three-point system, three-point recombination fractions were incorporated to two-point recombination fractions by using Tan and Fu [19] method. Simulation of codominant and dominant F_2 populations and the ELS and BAT estimations of gamete frequencies in F_2population were implemented by our R functions (Additional file 1, source code).

Results

Estimation of the frequencies of three-locus gametes in an F_2 population

Since our ELS method for accurate estimation of the frequencies of three-locus gametes in a population with random union of gametes is based on dissection of phenotypes, for convenience, we start by presenting the BAT method of Tan and Fu [19].

ELS estimation of frequencies of dominant marker gametes

Our study here is restricted to three biallelic dominant markers. We use *A* and *a*, *B* and *b*, *C* and *c* to represent two alleles at three loci where upper letters (A, B and C) stand for dominant alleles and lower letters (a, b and c) for recessive alleles. A triple-heterozygote individual via meiosis produces eight types of gametes at the three loci: *ABC, ABc, Abc, AbC, aBC, abC, aB*c and *abc*. Gametes *ABC* and *abc* are a pair of sister gametes on which two alleles at the all three loci are different and come from two different parents. Similarly, *Abc* and *aBC*, *abC* and *ABc*, *AbC* and *aBc* are also pairs of sister gametes. Two sister gametes theoretically have equal frequency in an F_2population because no mutation, no migration, no gene conversion and no selection occur in such a random mating population. From the expectation that sister-gametes have equal frequencies, we have in an F_2 population $f(ABC) = f(abc) = q_1$, $f(ABC) = f(aBC) = q_2$, $f(ABc) = f(aBC) = q_3$, $f(AbC) = f(aBc) = q_4$. These gamete frequencies are constrained by $2q_1 + 2q_2 + 2q_3 + 2q_4 = 1$. The individuals in the population can be classified into four categories: category 0 in which all individuals possess

0 dominant locus, that is, all individuals have three recessive loci; categories 1, 2 and 3 in which all individuals have respectively only one, two and three homozygous or heterozygous dominant loci. To accurately estimate gamete frequencies, we dissect a phenotype into different zygote types (genotypes) in each category using sister gametes. In category 1, for example, *aabbC_* has only locus c with one or two dominant alleles. Therefore it can be dissected into three zygote types:

$$aabbC\_\rightarrow\begin{cases} aabbCC\rightarrow(abC)^2 \quad : \quad (f(abC))^2=q_3^2 \\ aabbCc\rightarrow(abC)(abc):f(abC)f(abc)=q_3q_1 \\ aabbcC\rightarrow(abc)(abC):f(abc)f(abC)=q_1q_3 \end{cases}. \tag{1a}$$

Phenotypes *aaB_cc* and *A_bbcc* are dissected in a similar fashion. Category 2 also has three phenotypes and each of them can be dissected into four zygote types that are comprised of five pairs of sister gametes. For instance, phenotype type *A_B_cc* can be dissected into

$$A\_B\_cc\rightarrow\begin{cases} AABBcc\rightarrow(ABc)(ABc):f(ABc)f(ABc)=q_3^2 \\ AaBbcc\rightarrow(ABc)(abc):f(ABc)f(abc)=2q_3q_1 \\ AABbcc\rightarrow(ABc)(Abc):f(ABc)f(Abc)=2q_3q_2 \\ AaBBcc\rightarrow(ABc)(aBc):f(ABc)f(aBc)=2q_3q_4 \\ AaBbcc\rightarrow(Abc)(aBc):f(Abc)f(aBc)=2q_2q_4 \end{cases}. \tag{1b}$$

Category 3 has only one phenotype. The phenotype is comprised of 8 zygote types (genotypes) and therefore it is not useful for estimate of gamete frequencies. We use Q_1, Q_2, Q_3, Q_4, Q_5, Q_6, and Q_7 to respectively represent the frequency expectations of phenotypes *aabbcc*, *aabbC_*, *aaB_cc*, *A_bbcc*, *A_B_cc*, *A_bbC_*, and *aaB_C_* in a population. The frequency of phenotype *aabbcc* is

$$f(aabbcc)=Q_1=q_1^2 \tag{2}$$

The other 6 phenotypes have their frequencies:

$$\begin{cases} f(aabbC\_)=Q_2=q_3^2+2q_1q_3 \\ f(aaB\_cc)=Q_3=q_4^2+2q_1q_4 \\ f(A\_bbcc)=Q_4=q_2^2+2q_1q_2 \end{cases}. \tag{3}$$

$$\begin{cases} f(A\_B\_cc)=Q_5=q_3^2+2q_1q_3+2(q_3q_2+q_3q_4+q_2q_4) \\ f(A\_bbC\_)=Q_6=q_4^2+2q_1q_4+2(q_3q_2+q_3q_4+q_2q_4) \\ f(aaB\_C\_)=Q_7=q_2^2+2q_1q_2+2(q_3q_2+q_3q_4+q_2q_4) \end{cases}. \tag{4}$$

Using $Q = 2\ (q_2q_3 + q_2q_4 + q_3q_4)$, Eq. (4) is simplified as

$$\begin{cases} Q_5=Q_2+Q \\ Q_6=Q_3+Q \\ Q_7=Q_4+Q \end{cases}. \tag{5}$$

Estimates of $q_1, \ldots, q_4$ can be obtained from the above sets of equations by replacing Q_k with their observed frequencies where k = 1, 2,...,7 for 7 phenotypes. Theoretically, eqs. (1) and (3) are sufficient to make solutions for the frequencies of four types of gametes. However, Eq. (5) can be used to further minimize noise in the observed frequencies. That is, Q_2, Q_3,and Q_4 can be alternatively estimated as

$$\begin{cases} \hat{Q}_2^{\#}=\hat{Q}_5-\hat{Q}=0.25-\left(\hat{Q}_1+\hat{Q}_6+\hat{Q}_7\right) \\ \hat{Q}_3^{\#}=\hat{Q}_6-\hat{Q}=0.25-\left(\hat{Q}_1+\hat{Q}_5+\hat{Q}_7\right) \\ \hat{Q}_4^{\#}=\hat{Q}_7-\hat{Q}=0.25-\left(\hat{Q}_1+\hat{Q}_5+\hat{Q}_6\right) \end{cases} \tag{6}$$

where $Q = Q_5 + Q_6 + Q_7 + Q_1 - 0.25$ [19]. It implicates that Q_2, Q_3, and Q_4 can also be estimated from the estimated frequencies of Q_1, Q_5, Q_6, and Q_7. Thus, we can combine the two sets of estimates of Q_2, Q_3, and Q_4 into one set:

$$\begin{cases} \hat{Q}_2^{*}=\frac{1}{a_2+b_2}\left(a_2\hat{Q}_2+b_2\hat{Q}_2^{\#}\right) \\ \hat{Q}_3^{*}=\frac{1}{a_3+b_3}\left(a_3\hat{Q}_3+b_3\hat{Q}_3^{\#}\right) \\ \hat{Q}_4^{*}=\frac{1}{a_4+b_4}\left(a_4\hat{Q}_4+b_4\hat{Q}_4^{\#}\right) \end{cases} \tag{7}$$

where a_k and b_k are weights of $\hat{Q}_k$ and $\hat{Q}_k^{\#}$, respectively, where k = 2, 3, and 4. $\hat{Q}_k$ and $\hat{Q}_k^{\#}$ are respectively estimates of Q_k and $Q_k^{\#}$. In general case, $a_k = b_k$ (see Additional file 3: Appendix B). An alternative method for weighting is $a_k=\hat{Q}_k/\left(\hat{Q}_k+\hat{Q}_k^{\#}\right)$ and $b_k = 1 - a_k$. When the sample is small, it is likely that $\hat{Q}_k^{\#}\leq 0$ or $\hat{Q}_k = 0$. In such a case, one can set a_k = 1 and b_k = 0 for $\hat{Q}_k^{\#}\leq 0$, or a_k = 0 and b_k = 1 for $\hat{Q}_k^{\#} > 0$ and $\hat{Q}_k$ = 0. Since $Q_2 = q_3^2 + 2q_1q_3 + q_1^2 - q_1^2 = (q_3 + q_1)^2 - q_1^2$, q_3 can be given by

$$q_3=\sqrt{Q_2+Q_1}-\sqrt{Q_1}. \tag{8a}$$

Similarly,

$$q_2=\sqrt{Q_4+Q_1}-\sqrt{Q_1}, \tag{8b}$$

$$q_4=\sqrt{Q_3+Q_1}-\sqrt{Q_1}. \tag{8c}$$

Q_1, Q_2, Q_3, and Q_4 are respectively estimated by $\hat{Q}_1$, $\hat{Q}_2^{*}$, $\hat{Q}_3^{*}$, $\hat{Q}_4^{*}$, therefore q_3, q_2, q_4, and q_1 are respectively estimated by

$$\hat{q_3}=\sqrt{\hat{Q}_2+\hat{Q}_1}-\sqrt{\hat{Q}_1}, \tag{9a}$$

$$\hat{q}_2 = \sqrt{\hat{Q}_4 + \hat{Q}_1} - \sqrt{\hat{Q}_1}, \tag{9b}$$

$$\hat{q}_4 = \sqrt{\hat{Q}_3^* + \hat{Q}_1} - \sqrt{\hat{Q}_1}, \tag{9c}$$

$$\hat{q}_1 = \sqrt{\hat{Q}_1}. \tag{9d}$$

In Eq. (9), accurate estimation of q_1 is a key contribution to accurate estimations of q_2, q_3, and q_4. Equations (3) and (4) show that $Q_2 \sim Q_7$ can also provide information of solution to q_1. But it is impossible to directly obtain a solution for q_1 from $Q_2 \sim Q_7$. To estimate q_1 from $Q_1 \sim Q_7$, we here proposed a seeking method, named "expectation least square" (ELS) method.

Similar to the EM method [11, 25, 28, 29], the ELS method also consists of two steps. The first step is the expectation step, denoted by E-step, and the second step is the least-square step, denoted by LS-step. q_1 is initialized to be $\hat{q}_1^0 = \sqrt{\hat{Q}_1}$. We use $\hat{q}_1^0$ to estimate q_2, q_3, and q_4 and get $\hat{q}_2^0$, $\hat{q}_3^0$, and $\hat{q}_4^0$ from Eqs. (9). Then, we calculate the expected values of $Q_2 \sim Q_7$ from Eqs. (3) ~ (4) with $\hat{q}_2^0$, $\hat{q}_3^0$, and $\hat{q}_4^0$. At iteration j, we realize E-step and LS-step to get $\hat{q}_2^j$, $\hat{q}_3^j$, and $\hat{q}_4^j$:

E-step:

Calculate the expected values $E\left(Q_2^j\right) \sim E\left(Q_7^j\right)$ of $Q_2 \sim Q_7$ by replacing $\hat{q}_1^j$, $\hat{q}_2^j$, $\hat{q}_3^j$, and $\hat{q}_4^j$ into Eqs. (3) ~ (4) where $\hat{q}_2^j$, $\hat{q}_3^j$, and $\hat{q}_4^j$ are obtained by

$$\hat{q}_2^j = \sqrt{Q_4^{*j} + \left(\hat{q}_1^j\right)^2} - \hat{q}_1^j,$$

$$\hat{q}_3^j = \sqrt{Q_2^{*j} + \left(\hat{q}_1^j\right)^2} - \hat{q}_1^j,$$

$$\hat{q}_4^j = \sqrt{Q_3^{*j} + \left(\hat{q}_1^j\right)^2} - \hat{q}_1^j$$

where

$$Q_{i^{*j}} = \frac{1}{a+b}\left(a\hat{Q}_i + bQ_i^{\#j}\right)$$

where $i = 2, \ldots, 4$ and $Q_i^{\#j} = \hat{Q}_{i+3} - E\left(Q^{j-1}\right)$ where $E\left(Q^{j-1}\right) = 2\left(\hat{q}_2^{j-1}\hat{q}_3^{j-1} + \hat{q}_2^{j-1}\hat{q}_4^{j-1} + \hat{q}_3^{j-1}\hat{q}_4^{j-1}\right)$.

LS-step:

Calculate square value using

$$S_j^2 = \sum_{i=2}^{7}\left(\hat{Q}_i - E\left(Q_i^j\right)\right)^2. \tag{10}$$

Note that $\hat{q}_1^j$ is a value we want to seek for, therefore, Eq. (10) does not contain $\left(\hat{Q}_1 - EQ_1^j\right)^2$. As it is very difficult to directly get solutions for these four q-values from the derivative approach, we use an iteration approach to minimize square value:

$$\hat{q}_1^{\,j-1} = \arg\min(S_{j-1}^2, S_j^2). \tag{11}$$

Use $\hat{q}_1^j = \hat{q}_{1^{j-1}} \pm \Delta$ to calculate $\hat{q}_2^j$, $\hat{q}_3^j$, and $\hat{q}_4^j$ where j is the jth iteration, $j = 1, \ldots$, and Δ is specified with a very small value. Here our algorithm to realize LS-step is

If $S_j^2 > S_{j-1}^2$, then

if $\hat{q}_1^j > \hat{q}_{1^{j-1}}$, then $\hat{q}_1^j = \hat{q}_1^{\,j-1} - \Delta$,

otherwise, $\hat{q}_1^j = \hat{q}_{1^{j-1}} + \Delta$

else if $S_j^2 < S_{j-1}^2$, then

if $\hat{q}_1^j > \hat{q}_{1^{j-1}}$, then $\hat{q}_1^j = \hat{q}_{1^{j-1}} + \Delta$,

otherwise, $\hat{q}_1^j = \hat{q}_{1^{j-1}} - \Delta$.

Note that there are not $S_j^2 = S_{j-1}^2$ and $\hat{q}_1^j = \hat{q}_{1^{j-1}}$ in this algorithm. The iteration will stop at $S_j^2 \le t$ where t is a given tolerant value. Once the final estimate ($\hat{q}_1^f$) of q_1 is found at a given tolerant value where $j = f$, the final estimates of q_2, q_3, and q_4 are obtained. Then we let $\hat{q}_1 = \hat{q}_1^f$, $\hat{q}_2 = \hat{q}_2^f$, $\hat{q}_3 = \hat{q}_3^f$, and $\hat{q}_4 = \hat{q}_4^f$.

BAT for estimation of the frequencies of codominant marker gametes in F_2 population

To avoid confusing notations in codominant loci with those in dominant loci, we let 0 and 1 code for homozygote from two parents, respectively, and 2 code for heterozygote at a locus. Since homozygote and heterozygote at three loci can be recognized, most of zygotes are informative for estimation of the frequencies of four pairs of sister gametes. We still assume that the sister-gametes have equal frequencies, that is, $q_1 = f(111) = f(000)$, $q_2 = f(100) = f(011)$, $q_3 = f(110) = f(001)$, $q_4 = f(101) = f(010)$ in F_2 population. Here these complementary zygote type pairs are listed as follows:

Zygote gamete frequency expected	Zygote gamete frequency expected
$111,000 \rightarrow \begin{cases}(111)(111): q_1^2\\(000)(000): q_1^2\end{cases}$,	$100,011 \rightarrow \begin{cases}(100)(100): q_2^2\\(011)(011): q_2^2\end{cases}$,
$110,001 \rightarrow \begin{cases}(110)(110): q_3^2\\(001)(001): q_3^2\end{cases}$,	$101,010 \rightarrow \begin{cases}(101)(101): q_4^2\\(010)(010): q_4^2\end{cases}$,

$$200,211\rightarrow\begin{cases}(000)(100):2q_1q_2\\(111)(011):2q_1q_2\end{cases},\quad 112,002\rightarrow\begin{cases}(000)(001):2q_1q_3\\(111)(110):2q_1q_3\end{cases},$$

$$121,020\rightarrow\begin{cases}(000)(010):2q_1q_4\\(111)(101):2q_1q_4\end{cases},\quad 021,120\rightarrow\begin{cases}(011)(001):2q_2q_3\\(110)(100):2q_2q_3\end{cases},$$

$$102,012\rightarrow\begin{cases}(100)(101):2q_2q_4\\(011)(010):2q_2q_4\end{cases},\quad 201,210\rightarrow\begin{cases}(001)(101):2q_3q_4\\(110)(010):2q_3q_4\end{cases},$$

$$122\rightarrow\begin{cases}(111)(100):2q_1q_2\\(110)(101):2q_3q_4\end{cases},\quad 221\rightarrow\begin{cases}(111)(001):2q_1q_3\\(011)(101):2q_2q_4\end{cases},$$

$$022\rightarrow\begin{cases}(000)(011):2q_1q_2\\(001)(010):2q_3q_4\end{cases},\quad 220\rightarrow\begin{cases}(000)(110):2q_1q_3\\(100)(010):2q_2q_4\end{cases},$$

$$212\rightarrow\begin{cases}(111)(010):2q_1q_4\\(110)(011):2q_2q_3\end{cases},\quad 202\rightarrow\begin{cases}(000)(101):2q_1q_4\\(100)(001):2q_2q_3\end{cases}.$$

Let P_1, P_2, P_3 and P_4 represent the frequencies of complementary homozygote types (111/000), (100/011), (110/001), and (101/010) in each of which all three loci are homozygous; let P_{12}, P_{13}, P_{14}, P_{23}, P_{24}, and P_{34} be the frequencies of complementary two-locus homozygote types (200/211), (002/112), (121/020), (021/120), (102/012), and (201/210) in each of which only one locus are heterozygous and let P_{1234}, P_{1324}, P_{1423} be the frequencies of complementary one-locus homozygote types (122/022), (221/220) and (212/202) in each of which two loci are heterozygous. Then, $P_1 = 2q_1^2$, $P_2 = 2q_2^2$, $P_3 = 2q_3^2$, $P_4 = 2q_4^2$, $P_{12} = 4q_1q_2$, $P_{13} = 4q_1q_3$, $P_{14} = 4q_1q_4$, $P_{23} = 4q_2q_3$, $P_{24} = 4q_2q_4$, $P_{34} = 4q_3q_4$, $P_{1234} = 4q_1q_2 + 4q_3q_4$, $P_{1324} = 4q_1q_3 + 4q_2q_4$, $P_{1423} = 4q_1q_4 + 4q_2q_3$. From the zygote type pair list above, we find that the frequencies of these 12 pairs of zygote types can constitute two sets of 6 binomial equations:

$$Q_{12}^1 = \frac{1}{2}(P_1 + P_{12} + P_2) = q_1^2 + 2q_1q_2 + q_2^2 = (q_1 + q_2)^2, \tag{12a}$$

$$Q_{13}^1 = \frac{1}{2}(P_1 + P_{13} + P_3) = q_1^2 + 2q_1q_3 + q_3^2 = (q_1 + q_3)^2, \tag{12b}$$

$$Q_{14}^1 = \frac{1}{2}(P_1 + P_{14} + P_4) = q_1^2 + 2q_1q_4 + q_4^2 = (q_1 + q_4)^2, \tag{12c}$$

$$Q_{23}^1 = \frac{1}{2}(P_2 + P_{23} + P_3) = q_2^2 + 2q_2q_3 + q_3^2 = (q_2 + q_3)^2, \tag{12d}$$

$$Q_{24}^1 = \frac{1}{2}(P_2 + P_{24} + P_4) = q_2^2 + 2q_2q_4 + q_4^2 = (q_2 + q_4)^2, \tag{12e}$$

$$Q_{34}^1 = \frac{1}{2}(P_3 + P_{34} + P_4) = q_3^2 + 2q_3q_4 + q_4^2 = (q_3 + q_4)^2 \tag{12f}$$

$$Q_{12}^2 = \frac{1}{2}(P_1 + P_{1234} - P_{34} + P_2) = q_1^2 + 2q_1q_2 + q_2^2 = (q_1 + q_2)^2, \tag{13a}$$

$$Q_{13}^2 = \frac{1}{2}(P_1 + P_{1324} - P_{24} + P_3) = q_1^2 + 2q_1q_3 + q_3^2 = (q_1 + q_3)^2, \tag{13b}$$

$$Q_{14}^2 = \frac{1}{2}(P_1 + P_{1423} - P_{23} + P_4) = q_1^2 + 2q_1q_4 + q_4^2 = (q_1 + q_4)^2, \tag{13c}$$

$$Q_{23}^2 = \frac{1}{2}(P_2 + P_{1423} - P_{14} + P_3) = q_2^2 + 2q_2q_3 + q_3^2 = (q_2 + q_3)^2, \tag{13d}$$

$$Q_{24}^2 = \frac{1}{2}(P_2 + P_{1324} - P_{13} + P_4) = q_2^2 + 2q_2q_4 + q_4^2 = (q_2 + q_4)^2, \tag{13e}$$

$$Q_{34}^2 = \frac{1}{2}(P_3 + P_{1234} - P_{12} + P_4) = q_3^2 + 2q_3q_4 + q_4^2 = (q_3 + q_4)^2. \tag{13f}$$

We use arithmetic mean to get frequencies of these zygote types in F_2 population:

$$Q_{ij} = \left(a_{ij}Q_{ij}^1 + b_{ij}Q_{ij}^2\right) = \left(q_i + q_j\right)^2, \tag{14}$$

where $a_{ij} = \hat{Q}^1{}_{ij}/\left(\hat{Q}_{ij}^1 + \hat{Q}_{ij}^2\right)$ and $b_{ij} = 1 - a_{ij}$. $\left(a_{ij}Q_{ij}^1 + b_{ij}Q_{ij}^2\right) = a_{ij}\left(q_i + q_j\right)2 \quad + b_{ij}(q_i + q_j)^2 = \ (a_{ij} + b_{ij})(q_i + q_j)^2 = (q_i + q_j)^2$ where i and j are gamete types i and j (i = 1, 2, 3 and j = 2, 3, 4 and $i \neq j$). Thus, the frequencies of four types of non-sister gametes in a codominant three-locus system in an F_2 population are easily and fast estimated by

$$\hat{q}_1 = \frac{1}{2}\left(\frac{\sqrt{\hat{Q}_{12}} + \sqrt{\hat{Q}_{13}} + \sqrt{\hat{Q}_{14}} - \left(\sqrt{\frac{1}{2}\hat{P}_2} + \sqrt{\frac{1}{2}\hat{P}_3} + \sqrt{\frac{1}{2}\hat{P}_4}\right)}{3} + \sqrt{\frac{\hat{P}_1}{2}}\right), \tag{15a}$$

$$\hat{q}_2 = \frac{1}{2}\left(\frac{\sqrt{\hat{Q}_{12}} + \sqrt{\hat{Q}_{23}} + \sqrt{\hat{Q}_{24}} - \left(\sqrt{\frac{1}{2}\hat{P}_1} + \sqrt{\frac{1}{2}\hat{P}_3} + \sqrt{\frac{1}{2}\hat{P}_4}\right)}{3} + \sqrt{\frac{\hat{P}_2}{2}}\right), \tag{15b}$$

$$\hat{q}_3 = \frac{1}{2}\left(\frac{\sqrt{\hat{Q}_{13}} + \sqrt{\hat{Q}_{23}} + \sqrt{\hat{Q}_{34}} - \left(\sqrt{\frac{1}{2}\hat{P}_1} + \sqrt{\frac{1}{2}\hat{P}_2} + \sqrt{\frac{1}{2}\hat{P}_4}\right)}{3} + \sqrt{\frac{\hat{P}_3}{2}}\right), \tag{15c}$$

$$\hat{q}_4 = \frac{1}{2}\left(\frac{\sqrt{\hat{Q}_{14}} + \sqrt{\hat{Q}_{24}} + \sqrt{\hat{Q}_{34}} - \left(\sqrt{\frac{1}{2}\hat{P}_1} + \sqrt{\frac{1}{2}\hat{P}_2} + \sqrt{\frac{1}{2}\hat{P}_3}\right)}{3} + \sqrt{\frac{\hat{P}_4}{2}}\right), \tag{15d}$$

where $\hat{Q}_{ij}$ and $\hat{P}_k$ are respective estimates of Q_{ij} and P_k in F_2 population where k = 1,...,4 denote gamete types 1, ..., 4.

A modified BAT method (BAT II) for estimating the frequencies of eight gamete types without assumption that the sister gametes have equal frequencies in any generation population is given in Additional file 2, Appendix A.

Estimation of recombination fractions

Since these four qs are estimated separately, sum of them does not always satisfy a constraint of $\hat{q}_1 + \hat{q}_2 + \hat{q}_3 + \hat{q}_4 = 0.5$. For this reason, we normalize our estimates as

$$\begin{cases} p_1 = \dfrac{\hat{q}_1}{2\hat{q}}, & p_3 = \dfrac{\hat{q}_3}{2\hat{q}} \\ p_2 = \dfrac{\hat{q}_2}{2\hat{q}}, & p_4 = \dfrac{\hat{q}_4}{2\hat{q}} \end{cases}. \tag{16}$$

For three linked loci, the frequencies of the four gamete pairs can be used to find the double crossover types by distinguishing coupling phase from repulsion phase between loci. For example, for an order a-b-c of the three loci a, b and c, p_4 is determined to be the frequency of double crossover types if its value is the smallest and/or p_1 is the largest, which are produced at three coupling loci or p_1 is found to be the frequency of double crossover types if its value is the smallest and/or p_4 is the largest, which are formed at loci a and c in coupling phase and locus b in repulsion phase. In a similar way, we can also define p_3 or p_2 as the frequency of double crossover types.

If p_4 is frequency of double crossover types, then the recombination fractions between loci a and b, between loci b and c, and between loci a and c can be estimated by

$$\begin{cases} r_{ab} = 2(p_3 + p_4) \\ r_{bc} = 2(p_2 + p_4) \\ r_{ac} = 2(p_2 + p_3) \end{cases}. \tag{17}$$

For the linkage orders a-c-b and b-a-c, the recombination fractions between loci are also estimated in a similar way.

In the repulsion phase, the linkage a-b-c order of three loci determines p_1 to be the frequency of double crossover types, so estimates of recombination fractions between loci a and b, between loci b and c, and between loci a and c are

$$\begin{cases} r_{ab} = 2(p_2 + p_1) \\ r_{bc} = 2(p_3 + p_1) \\ r_{ac} = 2(p_2 + p_3) \end{cases}. \tag{18}$$

For the linkage orders b-a-c and a-c-b, the recombination fractions between three loci in the repulsion phase can be estimated in this way.

r_{ab}, r_{bc}, and r_{ac} are simple notations of three recombination fractions in a triple. However, when n markers on a chromosome or a fragment are genotyped, it is difficult to use these notations of three recombination frequencies to denote recombination fractions in $n(n-1)(n-2)/6$ triples. To notate recombination fractions in multiple triples, we let $r_{ab} = r_{abc}$ where c is referred to as a reference marker for recombination fraction between markers a and b, $r_{ac} = r_{acb}$ where b as reference marker for that between loci a and c, and $r_{bc} = r_{bca}$ where a as reference locus for that between markers b and c, in a three-locus system consisting of markers a, b, and c [19]. In more general fashion, we denote i for the first marker, j for the second maker, and k for the last marker. Thus, the rest $n-2$ markers are combined with loci i and j into $n-2$ three-points, therefore, there are $n-2$ estimates of the recombination fraction between markers i and j. Hence estimate of recombination fraction between loci i and j is given by Tan and Fu's method [19]:

$$\theta_{ij} = \frac{1}{n-2}\sum_{k=1}^{n-2} r_{ijk}. \tag{19}$$

Practical examples

Here we used RFLP (restriction fragment length polymorphisms) data of 333 F_2 mice from MAPMAKER/EXP (version 3.0b), LANDER et al. [13] to illustrate performances of our ELS and BAT methods to estimate recombination fractions between dominant and codominant loci. RFLP markers are codominant markers. In genotype data of 333 F2 mice, "A" stands for homozygote A (two alleles from parent A), "H" for heterozygote H (an allele from parent A and the other from parent B), and "B" for homozygote B (two alleles from parent B). We arbitrarily selected 6 codominant markers from the original genotype data. To evaluate our ELS algorithm, we converted the codominant genotype data into dominant genotype data by changing B to H. For convenience, we used arabic digits (1, 2,...,6) to label these six markers: marker 1, marker 2, ..., marker 6. Sometime we also used locus 1, locus 2, ..., locus 6 to mark these six marker loci. The frequencies of 20 non-sister gametes were estimated by respectively performing ELS on the dominant data and BAT on the codominant genotype data, normalized by using Eq. (16) and the results are summarized in Tables 1 and 2. For the ELS estimation, three non-sister gametes containing loci 4

Table 1 The ELS estimated frequencies of four nonsister gametes in 20 triplets of 6 dominant loci in 333 F2 mice[a]

locus			frequency of non-sister gamete				Chi-square test	
a	b	c	*p1 = f(abc)*	*p2 = f(Abc)*	*p4 = f(aBc)*	*p3 = f(abC)*	*p*-value	ratio
1	2	3	0.208668	0.086162	0.094698	0.110472	0.000339	
1	2	4	0.14751	0.087958	0.160866	0.103665	0.028	
1	2	5	0.200976	0.080676	0.108494	0.109854	0.00092	
1	2	6	0.192229	0.084408	0.126033	0.09733	0.0023	
1	3	4	0.140566	0.079252	0.181795	0.098387	0.0038	
1	3	5	0.209093	0.065783	0.12237	0.102753	0.00012	
1	3	6	0.16895	0.063323	0.165648	0.102079	0.0011	
1	4	5	0.173539	0.100396	0.079665	0.1464	0.0069	
1	4	6	0.16482	0.102771	0.113819	0.118591	0.0837	1:1:1:1
1	5	6	0.173447	0.07932	0.148645	0.098588	0.0059	
2	3	4	0.141958	0.088919	0.172784	0.096339	0.012	
2	3	5	0.202566	0.085775	0.112098	0.099561	0.00084	
2	3	6	0.139672	0.086032	0.16843	0.105866	0.0212	
2	4	5	0.173634	0.117152	0.085607	0.123607	0.0212	
2	4	6	0.10113	0.133668	0.133668	0.131535	0.3134	1:1:1:1
2	5	6	0.156859	0.096648	0.142758	0.103735	0.0634	1:1:1:1
3	4	5	0.143582	0.120278	0.098137	0.138002	0.1954	1:1:1:1
3	4	6	0.105025	0.139707	0.129181	0.126086	0.3581	1:1:1:1
3	5	6	0.140841	0.109125	0.152351	0.097682	0.1028	1:1:1:1
4	5	6	0.146847	0.096392	0.156905	0.099856	0.0476	

a: The data came from MAPMAKER/EXP(3.0b) [27]

and 6 (146, 246 and 346) fitted well the ratio of 1:1:1:1 (Chi-square test *p*-value >0.084, Table 1), indicating that loci 4 and 6 are unlinked to loci 1, 2 and 3. In addition, the frequencies of gametes 256, 356, and 345 also fitted the ratio of 1:1:1:1 with *p*-value ≥ 0.063 (Chi-square test, Table 1), but gametes 156, 245 and 145 had the ratios significantly deviating against 1:1:1:1 (Chi-square test *p*-value <0.0212, Table 1), we could infer that locus 5 was linked to loci 1 but independent of locus 3 and unascertained at locus 2. Thus, we definitely excluded loci 4 and 6 in the linkage. By using eqs. (17) – (19), the recombination fractions in four triples (123), (125), (135), and (235) were calculated by following the five given steps: the first step is to determine the linkage order of three loci in triple. For example, in triple (123), p_1 = *f(abc)* = 0.208668 is the largest value while p_2 = *f(Abc)* = 0.086162 is the smallest one, that is to say, gamete *Abc* is double crossover type and *abc* is parental type, so their order is 2(b)-1(a)-3(c). Step2 is to determine linkage phase: since gamete *bac* is parental type and *bAc* is double crossover type, gamete *BAC or bac* is couple phase. At step 3, we abstracted frequencies of gametes 123, 125, 135, 235 (Table 3) from Table 1. At step 4, recombination fractions between loci in a triple were estimated as

$$r_{bac(213)} = 2[f(Abc) + f(aBc)] = 2(0.086162 + 0.11047) = 0.39327$$

$$r_{acb(132)} = 2[f(Abc) + f(abC)] = 2(0.086162 + 0.09469) = 0.36172$$

$$r_{bca(231)} = 2[f(aBc) + f(abC)] = 2(0.086162 + 0.09469) = 0.41034$$

Similarly, we also estimated the recombination fractions in triples (125), (135), and (235) (Table 3). Finally, the three-point estimates of the recombination fractions were incorporated into two-point estimates by applying Eq. (19) to the data in Table 4:

$$\theta_{12} = \frac{r_{213} + r_{215}}{2} = \frac{0.393268 + 0.38106}{2} = 0.387164,$$

$$\theta_{13} = \frac{r_{135} + r_{132}}{2} = \frac{0.337072 + 0.36172}{2} = 0.349396,$$

$$\theta_{15} = \frac{r_{152} + r_{153}}{2} = \frac{0.37834 + 0.376306}{2} = 0.377323,$$

$$\theta_{23} = \frac{r_{231} + r_{235}}{2} = \frac{0.41034 + 0.370672}{2} = 0.390506,$$

Table 2 The BAT estimated frequencies of nonsister gametes in 20 triplets of 6 codominant loci in 333 F2 mice[a]

locus			*frequency of non-sister gamete*				*Chi-square test*	
a	*b*	*c*	*p1 = f*(000)	*p2 = f*(100)	*p3 = f*(001)	*p4 = f*(010)	*p-value*	*ratio*
1	2	3	0.242929	0.066568	0.080146	0.110358	3.348e-07	
1	2	4	0.145838	0.08977	0.162134	0.102258	0.0291	
1	2	5	0.196094	0.091387	0.121051	0.091467	0.0017	
1	2	6	0.17224	0.104184	0.143308	0.080268	0.0098	
1	3	4	0.165099	0.068697	0.191983	0.074221	7.1334e-05	
1	3	5	0.222931	0.079297	0.147828	0.049944	1.0943e-06	
1	3	6	0.177615	0.065713	0.187929	0.068743	2.3780e-05	
1	4	5	0.158699	0.103969	0.114759	0.122573	0.1336	1:1:1:1
1	4	6	0.165089	0.113628	0.139128	0.082155	0.0249	
1	5	6	0.155874	0.091722	0.16263	0.089774	0.0121	
2	3	4	0.142565	0.093943	0.179432	0.08406	0.0050	
2	3	5	0.216411	0.069533	0.134853	0.079203	2.0404e-05	
2	3	6	0.160337	0.092614	0.172787	0.074262	0.0020	
2	4	5	0.172459	0.100044	0.105018	0.12248	0.0365	
2	4	6	0.154284	0.140079	0.121173	0.084464	0.0559	1:1:1:1
2	5	6	0.167782	0.072156	0.153072	0.10699	0.0057	
3	4	5	0.154314	0.118649	0.10895	0.118086	0.2051	1:1:1:1
3	4	6	0.144358	0.108635	0.131647	0.115359	0.3080	1:1:1:1
3	5	6	0.16738	0.053399	0.176627	0.102594	0.0002	
4	5	6	0.153613	0.092081	0.124467	0.129838	0.1092	1:1:1:1

a: The data came from MAPMAKER/EXP(3.0b) [27]

$$\theta_{25} = \frac{r_{251} + r_{253}}{2} = \frac{0.436696 + 0.395746}{2} = 0.416221,$$

$$\theta_{35} = \frac{r_{351} + r_{352}}{2} = \frac{0.450246 + 0.423318}{2} = 0.436782,$$

Table 2 displays frequencies of codominant gametes estimated by our BAT method. It is clear to see that frequencies of gametes 145, 246, 345, 346, and 456 fitted well ratio of 1:1:1:1 with p-value ≥ 0.0559 (Chi-square test), however, the frequencies of gametes 156, 256 and 356 did not fit the ratio of 1:1:1:1 with p-value < 0.0121 (Chi-square test, Table 2), inferring that loci 4 and 6 are unlinked to loci 1, 2 and 3 but locus 5 could not be inferred to linked to them. Again, in codominant genotype data, locus 5 was still unascertained. Following the steps above, we obtained estimates of recombination fractions between these four loci (Table 5). Both ELS estimates of recombination fractions between dominant loci and BAT estimates between codominant loci show that locus 5 could not be tightly linked to any one of loci 1, 2 and 3. Loci 1, 2 and 3 could be determined to have linkage order of 2-1-3. Simulation data also showed that the codominant estimator had higher precision than the dominant estimator (see Simulation data section), suggesting that codominant markers indeed contain higher linkage information than dominant ones.

Table 3 The ELS estimated frequencies of nonsister gametes in triplets of dominant loci 1, 2, 3 and 5 in 333 F2 mice

locus			frequency of gamete			
a	*b*	*c*	*p1 = f(abc)*	*p2 = f(Abc)*	*p3 = f(abC)*	*p4 = f(aBc)*
1	2	3	0.208668	0.086162	0.094698	0.110472
1	2	5	0.200976	0.080676	0.108494	0.109854
1	3	5	0.209093	0.065783	0.12237	0.102753
2	3	5	0.202566	0.085775	0.112098	0.099561

Table 4 The estimated recombination fractions between dominant loci in four triples

triple			Recombination fraction between loci		
a	*b*	*c*	*b-a*	*a-c*	*b-c*
1	2	3	0.39326	0.36172	0.41034
1	2	5	0.38106	0.37834	0.43669
1	3	5	0.33707	0.37631	0.45024
2	3	5	0.37067	0.39575	0.42332

Table 5 Comparison between two estimators of recombination fractions between markers

two loci		the ELS estimate in dominant genotype data	the BAT estimate in codominant genotype data
1	2	0.387164	0.271317
1	3	0.349396	0.275955
1	5	0.377323	0.439563
2	3	0.390506	0.339240
2	5	0.416221	0.426574
3	5	0.436782	0.402158

Simulation data

We performed simulation study to compare the two estimators of recombination fractions. We followed the simulation scheme of Tan and Fu [19]. Briefly, we set two linkage maps comprised of 6 dominant loci and 6 codominant loci, respectively. Five possible map distances 10, 15, 20, 25, and 30 cM (1 cM = 1%) were randomly assigned to the five adjacent intervals on these two linkage models with equal probability (see Methods for detail). The point process model [27] was used to generate F_2 population. We did not consider recombination interference and linkage disequilibrium. Recombination fractions between adjacent loci in an unknown linkage phase (or say random phase) were estimated by the two-point EM [14, 23] and ELS estimators in 100 repeated samples of 100, 200, and 300 individuals drawn from the simulated F_2 population. These two estimators were rated by the variance that quantifies deviation of estimated recombination fraction between two adjacent loci from its true value and is equivalent to mean squared error (MSE). For dominant markers, simulation shows that the ELS algorithm had much smaller variances in estimation of true recombination fractions between adjacent loci in samples of 100, 200 and 300 F_2 individuals than two-point EM algorithm (Fig. 1). In Table 6, one can find that ELS had slightly higher probability of recovering true linkage maps of 6 loci than EM [14, 23] and BAT in the case of coupling phase and samples of 100 and 200 F_2 individuals. When sample size reached 300 individuals, both ELS and EM recovered true coupling linkage maps with 100% probability and BAT also had 97.9% recovery rate. However, in unknown phase, ELS recovered true linkage maps of 6 loci with 23.4% probability in sample of 100 F_2 individuals and reached 85% recovery rate in sample of 300 F_2 individuals. By contrast, EM had very low recovery rate (23.4%) even when sample size was 300. Therefore, ELS performed much better than two-point EM algorithm in all given scenarios. An inexact comparison can be done between ELS and three-point EM algorithm of Lu et al. [30], Table 4 in Lu et al. showed that their three-point EM algorithms had 98.5% probability of finding the correct linkage

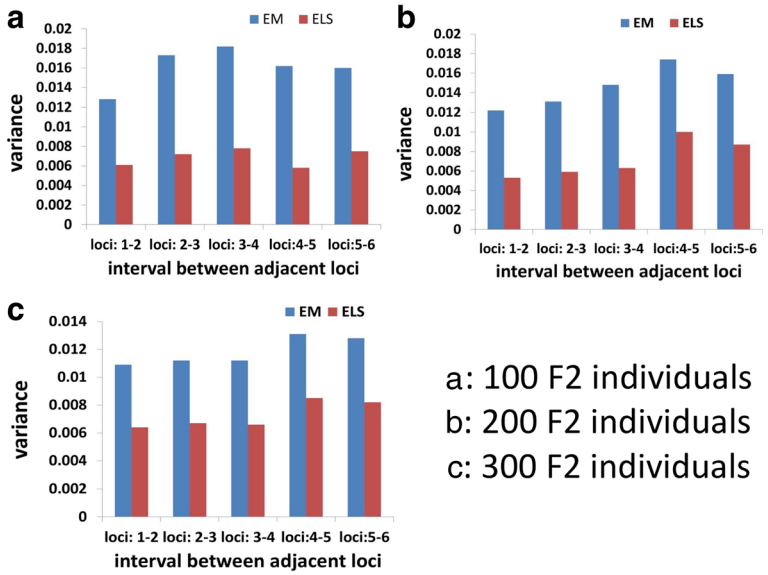

Fig. 1 Variances of estimated recombination fractions between adjacent dominant loci in unknown linkage phase deviated from their respective true values Variance of estimated recombination fraction between adjacent dominant loci is given by simulating 100 estimates around true recombination fraction between adjacent loci. The variance here is equivalent to mean square error (MSE)

Table 6 Efficiencies of estimators of recombination fractions in recovering the true linkage maps of 6 dominant loci in the case of random distance

Estimator	Linkage phase	Sample size		
		100	200	300
Two-point EM	CP	92.3	97.8	100.0
	UP	15.7	22.9	23.4
ELS	CP	96.7	100.0	100.0
	UP	50.5	77.0	85.1
BAT	CP	82.1	95.9	97.9
	UP	26.0	40.9	42.3

CP: Coupling phase and UP: unknown phase

map of three dominant markers in coupling phase from a sample of full-sib 100 individuals (corresponding to 100 F_2 individuals), our ELS had 96.7% probability of recovering true linkage map of 6 dominant markers in coupling phase in 100 F_2 individuals (Table 6). The probability to find a given linkage map will remarkably decrease as number of markers increases. So we can predict that the three-point EM algorithm would not have over 96.7% of the probability to find a given linkage map of 6 dominant markers. For the repulsion phase (or *trans* × *trans*), Lu et al.'s three-point EM algorithm had 99.5% probability of finding a correct linkage map of three markers in 100 full-sib individuals, which is higher than 98.6% in coupling phase. In theory, any EM algorithm should have much lower probability to find a given linkage order in repulsion phase than in coupling phase because the repulsion phase has much less linkage information content than the coupling phase [14, 26]. So, this result may be required to be confirmed in more simulations. Since Lu et al. did not implement simulation of random phase case and the repulsion phase is not random phase, the comparison cannot be made between the three-point EM and ELS algorithms in the random phase. For codominant markers, the BAT method performed with smaller variances than the two-point EM algorithm in the most cases. The results provided strong evidence for the conclusion that a method or algorithm based on three-point gametes can mitigate effect of low linkage information of repulsion phase on estimation of recombination fractions. Compared to the simulated results in Table 3 in [19], one can find that the ELS algorithm is better than the Tan and Fu's BAT method. Table 3 in [19] showed that in case of unknown phase, the BAT method outperformed two-point EM.

Discussion

Accurate estimation of recombination fractions is a key for mapping multiple markers. Therefore, powerful method for estimating recombination fractions is required. For dominant loci, the EM and ML methods have been verified to have low power to estimate frequencies of recombination between loci in repulsion phase [14, 19]. This is because the EM method cannot distinguish dominant homozygous genotypes from dominant heterozygous genotypes.

Compared to the EM algorithm, the ELS algorithm based on Tan and Fu's method [19] has small bias for estimating recombination fractions between dominant loci on a chromosome in a larger F_2 population due to the following reasons: (a) gamete analysis can effectively distinguish marker linkage phases; (b) accurately estimate q_1, and (c) average of estimates of recombination fraction between two loci over all reference loci [Eq. (19)] effectively balances sampling error. Estimation of q_1 is restriction of the Tan and Fu's method. We here proposed iteration expectation-least square algorithm (ELS) to seek for accurate q_1 estimation. This new algorithm is similar to expectation maximum algorithm and its statistical properties will be given by more simulation comparisons in elsewhere. In addition, importance for high efficiency of recombination fraction estimation is $\hat{Q}_k^*$. ELS had much higher recovery rate by using $\hat{Q}_k^*$ than by using $\hat{Q}_k$ in both coupling and unknown phase (Table 6). Correlation analysis also indicated that $\hat{Q}_k^*$ indeed has the linkage behavior similar to $\hat{Q}_k$ (Additional file 3, Appendix B). Furthermore, we found that $\hat{Q}_k^*$ obtained from a data set of 100 simulated samples of 100 F_2 individuals has remarkably smaller variance than $\hat{Q}_k$ (data not shown). To fully confirm that $\hat{Q}_k^*$ is the optimal choice in our ELS method, $\hat{Q}_k^*$ was taken into account where $\hat{Q}_k^* = \left(\hat{Q}_k + \hat{Q}_k^o\right)/2$ if $\hat{Q}_k^o > 0$, otherwise, $\hat{Q}_k^* = \hat{Q}_k$. The simulated result showed that ~31% of linkage maps recovered true order of 6 dominant loci in samples of 100 F_2 individuals, which is apparently lower than that by using $\hat{Q}_k^* = \frac{1}{2}\left(\hat{Q}_k + \hat{Q}_k^o\right)$. For this reason, we chose $\hat{Q}_k^* = \frac{1}{2}\left(\hat{Q}_k + \hat{Q}_k^o\right)$ in our ELS algorithm. Besides the ELS algorithm, average of recombination fraction between two loci over all reference loci greatly reduces noise of recombination fractions.

BATII given in Additional file 2, Appendix A, can be used to estimate frequencies of 8 codominant gamete types in any nature population because it does not require the assumption that the sister gametes have equal frequencies in a population. However, its estimation accuracy is not higher than the first BAT method in F_2 population because sister-gametes really have equal frequencies and two-locus heterozygote types are not useful in the BATII. In a natural population, for example, human population, the frequencies of these gametes are not purely derived from recombination events but may be due to selection, genetic drift, migration and mutation. If, however, sister gametes are found to be equal in statistics, then these frequencies can still be used to inference recombination fractions between loci and recombination inference.

Conclusions

Accurate estimation of recombination fractions between loci is given by methodologies developed for accurate estimation of gamete frequencies in a population. Analyses of simulated and real dominant and codominant data show that the ELS method proposed here is a powerful algorithm for accurate estimation of frequencies of gametes with unknown phase in dominant three-locus system in F_2 population and BAT is a computationally efficient and powerful method for estimating frequencies of non-sister three-point codominant gametes.

Additional files

Additional file 1: Source code: three R functions: BAT.R, ELS.R, simulatF2.R. (ZIP 4 kb)

Additional file 2: Appendix A. Binomial analysis of three-point method (BATII) is described in detail. BATII is used to estimate frequencies of sister gametes at codominant loci in natural populations. (DOCX 184 kb)

Additional file 3: Appendix B. A proof of a proposition that equal weights of two datasets combined into a dataset have maximum linkage information and minimum error for linkage analysis is given. (DOCX 41 kb)

Abbreviations

BAT: Binomial analysis of three-point gametes; ELS: Expectation least square; EM: Expectation maximization; RFLP: Restriction fragment length polymorphism

Acknowledgements

Not applicable.

Funding

This research and article's publication was supported in part by grant R01 CA183878. The funding agency played no role in the design or conclusion of the study.

Availability of data and materials

Since all simulation data were temporary and dynamic data and discarded when the programs were ended, we did not have simulation data to be deposited in a repository. R source code for ELS and BAT as well as the program for generating simulated data are available in Additional file 3.

About this supplement

This article has been published as part of BMC Bioinformatics Volume 18 Supplement 11, 2017: Selected articles from the International Conference on Intelligent Biology and Medicine (ICIBM) 2016: bioinformatics. The full contents of the supplement are available online at <https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-18-supplement-11>.

Authors' contributions

Study design: TYD, XHZ, QM. Method development and data analysis: TYD, QM. Manuscript writing: TYD, XHZ, QM. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Dan L. Ducan Cancer Center, Baylor College of Medicine, Houston, TX, USA. [2]Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, TX, USA. [3]Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX, USA. [4]McNair Medical Institute, Baylor College of Medicine, Houston, TX, USA. [5]Department of Medicine, Baylor College of Medicine, Houston, TX, USA.

Published: 3 October 2017

References

1. Bowers JE, Bachlava E, Brunick RL, Rieseberg LH, Knapp SJ, Burke JM. Development of a 10,000 locus genetic map of the sunflower genome based on multiple crosses. G3 (Bethesda). 2012;2(7):721–9.
2. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet. 2011;12(7):499–510.
3. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011;6(5):e19379.
4. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34(8):816–34.
5. Moriguchi Y, Ujino-Ihara T, Uchiyama K, Futamura N, Saito M, Ueno S, Matsumoto A, Tani N, Taira H, Shinohara K, et al. The construction of a high-density linkage map for identifying SNP markers that are tightly linked to a nuclear-recessive major gene for male sterility in Cryptomeria Japonica D. Don. BMC Genomics. 2012;13:95.
6. Sonah H, Bastien M, Iquira E, Tardivel A, Legare G, Boyle B, Normandeau E, Laroche J, Larose S, Jean M, et al. An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS One. 2013;8(1):e54603.
7. Verma S, Gupta S, Bandhiwal N, Kumar T, Bharadwaj C, Bhatia S. High-density linkage map construction and mapping of seed trait QTLs in chickpea (Cicer Arietinum L.) using genotyping-by-sequencing (GBS). Sci Rep. 2015;5:17512.
8. Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, Bowden DW, Smith DR, Lander ES, et al. A genetic linkage map of the human genome. Cell. 1987;51(2):319–37.
9. Ellis THN. Neighbor mapping as method for ordering genetic markers. Genet Res. 1997;69:35–43.
10. Lander ES, Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science. 1987;236(4808):1567–70.
11. Lander ES, Green P. Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci U S A. 1987;84(8):2363–7.
12. Lander ES, Green P. Counting algorithms for linkage: correction to Morton and Collins. Ann Hum Genet. 1991;55(Pt 1):33–8.
13. Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newberg LA. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics. 1987;1(2):174–81.
14. Liu BH. Statistical genomics: linkage, mapping, and QTL analysis. Florida: CRC Press; 1998.
15. Mester DI, Ronin YI, Hu Y, Peng J, Nevo E, Korol AB. Efficient multipoint mapping: making use of dominant repulsion-phase markers. Theor Appl Genet. 2003;107(6):1102–12.
16. Ronin Y, Mester D, Minkov D, Belotserkovski R, Jackson BN, Schnable PS, Aluru S, Korol A. Two-phase analysis in consensus genetic mapping. G3 (Bethesda). 2012;2(5):537–49.
17. Tan YD, Fu YX. A novel method for estimating linkage maps. Genetics. 2006;173(4):2383–90.
18. Zhang L, Li H, Wang J. Linkage analysis and map construction in genetic populations of clonal F1 and double cross. G3 (Bethesda). 2015;5(3):427–39.
19. Tan YD, Fu YX. A new strategy for estimating recombination fractions between dominant markers from an F2 population. Genetics. 2007;175(2):923–31.

20. Allard RW. Formulas and tables to facilitate the calculation of recombination values in heredity. California: University of California; 1956.
21. Knapp SJ, Holloway JL, Bridges WC, Liu BH. Mapping dominant markers using F_2 matings. Theor Appl Genet. 1995;91(1):74–81.
22. Peng J, Korol AB, Fahima T, Roder MS, Ronin YI, Li YC, Nevo E. Molecular genetic maps in wild emmer wheat, Triticum Dicoccoides: genome-wide coverage, massive negative interference, and putative quasi-linkage. Genome Res. 2000;10(10):1509–31.
23. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B Methodol. 1977;39(1):1–38.
24. Morton NE, Collins A. Counting algorithms for linkage. Ann Hum Genet. 1990;54(Pt 2):103–6.
25. Ott G. Analysis of human genetic linkage. Baltimore/London: John Hopkins University Press; 1991.
26. Esch E. Estimation of gametic frequencies from F2 populations using the EM algorithm and its application in the analysis of crossover interference in rice. Theor Appl Genet. 2005;111(1):100–9.
27. Foss E, Lande R, Stahl FW, Steinberg CM. Chiasma interference as a function of genetic distance. Genetics. 1993;133(3):681–91.
28. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. JR Statist Soc. 1977;39B:1–38.
29. Niu T, Ding AA, Kreutz R, Lindpaintner K. An expectation-maximization-likelihood-ratio test for handling missing data: application in experimental crosses. Genetics. 2005;169(2):1021–31.
30. Lu Q, Cui Y, Wu R. A multilocus likelihood approach to joint modeling of linkage, parental diplotype and gene order in a full-sib family. BMC Genet. 2004;5:20.