

# **Recombination Disequilibrium in Ideal and Natural Populations**

Yuan-De Tan

Institute of Personalized Medicine, Department of Biochemistry and Cellular Biology, PENN State College of Medicine, Hershey, PA, 17033

Corresponding author: Yuan-De Tan

Email address: [tanyuande@gmail.com](mailto:tanyuande@gmail.com)

## Abstract

Following Hardy-Weinberg disequilibrium (HWD) occurring at a single locus and linkage disequilibrium (LD) between two loci in generations, we here proposed the third genetic disequilibrium in population: recombination disequilibrium (RD). RD is a measurement of crossover interference among multiple loci in a random mating population. In natural populations besides recombination interference, RD may also be due to selection, mutation, gene conversion, drift and/or migration. Therefore, similarly to LD, RD will also reflect the history of natural selection and mutation. In breeding populations, RD purely results from recombination interference and hence can be used to build or evaluate and correct a linkage map. Several practical examples from  $F_2$ , testcross and human populations indeed demonstrate that RD is useful for measuring recombination interference between two short intervals and evaluating linkage maps. As with LD, RD will be important for studying genetic mapping, association of haplotypes with disease, plant breeding and population history.

In the early days of the last century, after rediscovery of Mendel's genetic laws, one began to look at the genetic behavior of a pair of alleles at a single locus in an ideal population. As a result, Hardy and Weinberg each independently discovered an equilibrium law of frequencies of both alleles and genotypes at a single locus in a very large population in 1908(HARDY 1908; WEINBERG 1908). This equilibrium law, called Hardy and Weinberg equilibrium (HWE), is a foundation of population and evolutionary genetics and an important landmark in population genetics. Now it is a common practice to check whether the observed genotypes conform to HWE expectations in disease gene mapping. These expectations appear to hold for most human populations but at some particular marker sites deviation of allele frequencies from HWE may occur, suggesting that the problems are associated with genotyping or population structure or, in samples of affected individuals, the markers are associated with disease susceptibility (WIGGINTON *et al.* 2005).

When, however, our interest was extended from one-locus system to two-locus system where each locus has a pair of alleles, a very important phenomenon, linkage disequilibrium (LD), was uncovered (GEIRINGER 1944; HILL and ROBERTSON 1968; LEWONTIN and KOJANA 1960; LEWONTIN 1964; ROBBINS 1918). LD plays a fundamental role in gene fine mapping and especially in study of genome-wide association of genetic variants and diseases of interest. Patterns of LD have become a tool for fine mapping of genes for a complex disease of study (HASTINGS 1984; PRITCHARD and PRZEWORSKI 2001) and are a topic of a great current interest due to single nucleotide polymorphisms (SNPs). LD is also of interest for what it can reveal the evolution of population because the patterns of LD are determined, in part, by population history. LD throughout genome reflects the population history and the pattern of geographic subdivision,

whereas in a genomic region LD reflects the history of natural selection, gene conversion, mutation (SLATKIN 2008).

What happens when a three-locus system is considered? Several studies attempted to generalize LD from two-locus system to three-locus system. For example, Bennett(BENNETT 1954) defined three-locus LD as

$$D_{ABC} = p_{ABC} - p_A D_{BC} - p_B D_{AC} - p_C D_{AB} - p_A p_B p_C \quad (1)$$

where  $p_{ABC}$  is the frequency of gamete ABC,  $p_A$ ,  $p_B$ ,  $p_C$  are frequencies of alleles A at locus  $a$ , B at locus  $b$ , and C at locus  $c$ , respectively, and  $D_{ij} = p_{ij} - p_i p_j$  is the standard two-locus LD definition where  $i = A, B, j = B, C$  and  $i \neq j$ . But these attempts have not had substantial progress in the past several decades because LD at three-locus level or higher level becomes very complicated and so poor application has also made in practice due to recombination interference. Hastings (HASTINGS 1984) indicated that commonly used measures of linkage disequilibrium are not appropriate for a multilocus setting. Thomson and Baur (THOMSON and BAUR 1984) also showed by an example that combinations of allele frequencies and pairwise linkage disequilibrium terms, each of which is permissible at the two-locus level, may not be permissible at three-locus level. Actually, in two-locus system the significance of linkage disequilibrium in biology and evolution is quite obvious, while in three-locus or multilocus system, double or multiple crossovers and crossover interference become a major component of linkage. Recall crossover interference occurring in three-locus system. In classical genetics, crossover interference is measured by coefficient of coincidence defined as  $\lambda = \frac{r_{DC}}{E(r_{DC})}$  where  $r_{DC}$  and

$E(r_{DC})$  are the observed and expected frequencies of double crossover types between two neighboring intervals  $a-b$  and  $b-c$ , respectively, where the locus order is assumed to be  $a-b-c$ .  $E(r_{DC}) = r_{ab} r_{bc}$  under condition of no crossover interference where  $r_{ab}$  and  $r_{bc}$  are the observed recombination fractions between loci  $a$  and  $b$  and between loci  $b$  and  $c$ , respectively. Coefficient of coincidence as a measurement of crossover interference is due to the fact that only the positive

crossover interference has been discovered in early genetic study. With the advance of technologies in molecular genetics, in particular, with the wide application of genotyping at molecular markers such as microsatellites and single nucleotide polymorphisms (SNPs), negative interference has widely been discovered. The coefficient of coincidence is not appropriate to describe negative interference because, for example, for positive interference,  $0 \leq r_{DC} \leq r_{ab}r_{bc}$  so that  $0 \leq \lambda \leq 1$ , but for negative interference,  $r_{DC} \geq r_{ab}r_{bc} \geq 0$ , then  $\lambda \geq 1$ . Thus the coefficient of coincidence  $\lambda$  is asymmetric for interference in the positive and negative directions, which leads to difficulty of testing for the positive or negative interference in statistics. For example, Esch (ESCH 2005) trialed to adopt null simulation way instead of direct way to test coefficient of coincidences. However, simulation has a big defect that the deviation of coefficient of coincidence from expected value  $E(\lambda)=1$  (no interference) would significantly increase with decreasing distance between loci (ESCH 2005). So null simulation also is not a good way. However, both the positive and negative interferences in effect cause recombination disequilibrium (RD). RD is defined as difference ( $D_r$ ) between the observed and expected frequencies of double or multiple crossover types in a population. Obviously RD is symmetric in both negative and positive interference directions. It will be easy to establish statistics for RD test. In this paper, we propose a mathematical definition of RD and a statistical method for hypothesis test of RD in ideal and natural populations and then we give some practical applications.

To estimate and test for RD among loci, it is required to estimate the three-locus gamete frequencies. In a natural population, especially, in a human population, for codominant markers, many existing powerful methods such as PHASE(STEPHENS *et al.* 2001), fastPHASE(SCHEET and STEPHENS 2006), BEAGLE(BROWNING and BROWNING 2007), IMPUTE2(HOWIE *et al.* 2009) and MaCH(LI *et al.* 2010) can be used to estimate haplotype frequencies. In  $F_2$  population, also a lot of expectation maximization (EM) algorithms (ESCH 2005; LI *et al.* 2007; SERGEEV

and ARAPOVA 2002) can be used to estimate triplet gametic frequencies way can be utilized to distinguish dominant homozygous genotypes from heterozygous genotypes (ESCH 2005; LIU 1998 ; SERGEEV and ARAPOVA 2002), the EM algorithms have low power to estimate gametic frequencies in  $F_2$  population. Tan and Fu (TAN and FU 2007) to develop a new approach (TF) to promote estimation accuracy of three-locus dominant gamete frequencies in  $F_2$  population. Different from the current EM methods(ESCH 2005; LI *et al.* 2007; SERGEEV and ARAPOVA 2002), the TF method is based on such an assumption that sister-gametes have the same frequencies in  $F_2$  population, which greatly reduce complexity of mathematical equations so that the power is significantly promoted. This assumption is correct because as indicated above,  $F_2$  population is ideal population without mutation, selection, genetic drift and immigration. In our current work, we use a modified TF method (see Supplementary Notes S1 and S3 for detail) to estimate dominant triplet gametic frequencies and a very fast and highly accurate method to estimate frequencies of triplet codominant gametes in an  $F_2$  population (see Supplementary Notes 2 and S3) because their algorithms are too trivial and too complex and have no open source software package to be used. In this paper we focus on RD instead of estimation of gametic frequencies.

## DEFINITION AND TESTS

Since these four  $q$ 's are estimated separately, sum of them does not always satisfy a constraint of  $q = q_1 + q_2 + q_3 + q_4 = 0.5$ . For this reason, we constrain sum of estimates to 0.5:

$$\begin{cases} p_1 = \frac{\hat{q}_1}{2\hat{q}}, & p_3 = \frac{\hat{q}_3}{2\hat{q}} \\ p_2 = \frac{\hat{q}_2}{2\hat{q}}, & p_4 = \frac{\hat{q}_4}{2\hat{q}} \end{cases} \quad (2)$$

For linked loci, frequencies of the four pairs of nonsister gametes can be used to distinguish coupling phase from repulsion phase between loci, which allows us to find double crossover types. For example, suppose an order of three loci is a-b-c. Then if  $p_4$  is smallest and  $p_1$  is largest, each allele pair is in coupling phase and  $p_4$  is frequency of double crossover types; if  $p_4$  is largest and  $p_1$  is smallest, then allele A at locus  $a$  and allele C at locus  $c$  are in coupling phase but allele B at locus  $b$  is in repulsion phase, so  $p_1$  is frequency of double crossover types. On the other hand, if  $p_2$  is largest and  $p_3$  is smallest, then allele A at locus  $a$  and allele B at locus  $b$  are in coupling phase but allele C at locus  $c$  is in repulsion phase, then  $p_3$  is frequency of double crossover types. Similarly, if  $p_2$  is smallest and  $p_3$  is largest, then  $p_2$  is frequency of double crossover types. Thus the smallest  $p$  must be frequency of double crossover types. When, however, negative interference occurs, the case would be different from the above: the gamete with the smallest frequency is not necessarily double crossover types because negative interference increases frequency of double crossover types in a population. But the law that the gamete with the largest frequency in a population is parental type still holds because negative interference occurs only when one or both two neighboring intervals are very short.

In coupling phase  $p_4$  is frequency of double crossover types. Thus, recombination fractions between loci  $a$  and  $b$ , between loci  $b$  and  $c$ , and between loci  $a$  and  $c$  can be calculated by

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

Recombination fractions between loci in the other orders in coupling phase are also given in a similar way.

In repulsion phase, the order (*a-b-c*) determines  $p_1$  to be frequency of double crossover types, thus the recombination fractions between loci *a* and *b*, between loci *b* and *c*, and between loci *a* and *c* are

$$\begin{cases} r_{ab} = 2(p_3 + p_1) \\ r_{bc} = 2(p_2 + p_1) \\ r_{ac} = 2(p_2 + p_3). \end{cases} \quad (3)$$

The recombination fractions between loci in the other orders in repulsion phase can be estimated in a similar fashion.

Thus, if, in coupling phase,  $p_4$  is frequency of double crossover types, RD is defined as difference between the observed and expected frequencies of double crossover types, that is,

$$\begin{aligned} D_r &= r_{DC} - r_{ab}r_{bc} \\ &= 2p_4 - 4(p_2 + p_4)(p_3 + p_4) \\ &= 2p_4 - 4(p_2p_3 + p_2p_4 + p_3p_4 + p_4^2) \\ &= 2p_4 - 4p_2p_3 - 4p_4(p_2 + p_3 + p_4) \\ &= 2p_4 - 4p_2p_3 - 4p_4(0.5 - p_1) \\ &= 2p_4 - 4p_2p_3 - 2p_4 + 4p_1p_4 \\ &= 4p_1p_4 - 4p_2p_3 \\ &= 4(p_1p_4 - p_2p_3). \end{aligned} \quad (4a)$$

In repulsion phase, we also have

$$D_r = r_{DC} - r_{ab}r_{bc} = 4(p_1p_4 - p_2p_3) \quad (4b)$$

if frequency of double crossover types is  $p_1$ , or

$$D_r = r_{DC} - r_{ab}r_{bc} = 4(p_2p_3 - p_1p_4) \quad (4c)$$



if  $p_2$  or  $p_3$  is frequency of double crossover types. Obviously,  $D_r$  is a symmetric measurement of positive and negative interferences:  $D_r = 0$  means that recombination between two neighboring intervals is in equilibrium, say, recombination events occur independently between these two neighboring intervals. Negative recombination interference occurs if  $D_r > 0$  or positive recombination interference occurs if  $D_r < 0$ . The maximum RD is  $D_r = 4p_1p_4$  and the minimum RD is  $D_r = -4p_2p_3$  in systems AbC/aBc and ABC/abc or the maximum RD is  $D_r = 4p_2p_3$  and the minimum RD is  $D_r = -4p_1p_4$  in systems aBC/Abc and ABc/abC. Both positive and negative recombination interferences occur in two very short intervals of chromosomes. But positive interference may be physical interference due to rigidity of DNA chains while negative interference may be biological interference due to activation of enzymes. One can predict that activation of enzymes on promoting recombination of DNA would become weak or disappear as interval distances increase. Similarly to the LD test, a standard chi-square statistic can be used to test for RD among three loci. In the ABC/abc system, the chi-square statistic is

$$\chi^2_{D_r} = \frac{(O_{DC} - E_{DC})^2}{E_{DC}} \quad (5)$$

where  $O_{DC}$  and  $E_{DC}$  are numbers of observed and expected double crossover types and  $N$  is number of individuals sampled from  $F_2$  population:  $O_{DC}$  can be given by using frequency of double crossover types and sample size  $N$ :  $O_{DC} = Nr_{DC}$  and  $E_{DC}$  can be obtained from expected frequency of double crossover types and  $N$ :  $E_{DC} = NE(r_{DC}) = N(r_{ab}r_{bc})$ . Inputting them into Equation (5), we have

$$\begin{aligned}
 \chi_{D_r}^2 &= \frac{(Nr_{DC} - Nr_{ab}r_{bc})^2}{Nr_{ab}r_{bc}} = \frac{N(r_{DC} - r_{ab}r_{bc})^2}{r_{ab}r_{bc}} \\
 &= \frac{N[4(p_1p_4 - p_2p_3)]^2}{4(P_2 + P_4)(P_3 + P_4)} = \frac{16N(p_1p_4 - p_2p_3)^2}{4(P_2P_3 + P_2P_4 + P_3P_4 + P_4^2)} = \frac{4N(p_1p_4 - p_2p_3)^2}{P_2P_3 + P_4(P_2 + P_3 + P_4)} \\
 &= \frac{4(p_1p_4 - p_2p_3)^2}{P_2P_3 + P_4(\frac{1}{2} - p_1)} N = \frac{4(p_1p_4 - p_2p_3)^2}{P_2P_3 - p_1p_4 + \frac{1}{2}P_4} N = \frac{\frac{4}{16}[4(p_1p_4 - p_2p_3)]^2}{-\frac{4}{4}(p_1p_4 - p_2p_3) + \frac{1}{2}P_4} N \\
 &= \frac{\frac{4}{16}D_r^2N}{-\frac{1}{4}D_r + \frac{1}{2}P_4} = \frac{D_r^2N}{2p_4 - D_r} = \frac{D_r^2N}{r_{DC} - D_r}.
 \end{aligned} \tag{6a}$$

In the AbC/aBc system, the chi-square statistic for RD is

$$\chi_{D_r}^2 = \frac{4(p_1p_4 - p_2p_3)^2 N}{p_2p_3 - p_1p_4 + \frac{1}{2}p_1} = \frac{D_r^2N}{2p_1 - D_r} = \frac{D_r^2N}{r_{DC} - D_r}. \tag{6b}$$

In the aBC/Abc system, the chi-square statistic for RD is

$$\chi_{D_r}^2 = \frac{4(p_2p_3 - p_1p_4)^2 N}{p_1p_4 - p_2p_3 + \frac{1}{2}p_2} = \frac{D_r^2N}{2p_2 - D_r} = \frac{D_r^2N}{r_{DC} - D_r}. \tag{6c}$$

And in the ABC/abC system, the chi-square statistic for RD is

$$\chi_{D_r}^2 = \frac{4(p_1p_4 - p_2p_3)^2 N}{p_1p_4 - p_2p_3 + \frac{1}{2}p_3} = \frac{D_r^2N}{2p_3 - D_r} = \frac{D_r^2N}{r_{DC} - D_r}. \tag{6d}$$

For codominant markers, we will convert AbC/aBc and ABC/abc to 101/010 and 111/000.

Similar way can also be used to the other genotypes. Under the null hypothesis  $H_0$ , these four chi-square statistics have an approximate  $\chi^2$  distribution with 1 degree of freedom.

Theoretically, RD reflects recombination disequilibrium due to interference in an ideal population (such as  $F_2$  or backcross population). In a natural population, RD could result from evolutionary factors such as selection, migration, mutation, drift, and gene conversion because these factors change allele and gamete

frequencies which may make frequencies of sister gametes unequal. For the case of  $p(ABC) \neq p(abc)$ ,  $p(Abc) \neq p(aBC)$ ,  $p(ABc) \neq p(abC)$ , or  $p(AbC) \neq p(aBc)$ , the RD can be given by

$$D_r = P_1P_4 - P_2P_3 \quad (7a)$$

in coupling phase or

$$D_r = P_2P_3 - P_1P_4 \quad (7b)$$

in repulse phase where

$$P_1 = p_1 + p'_1 = p(ABC) + p(abc), \quad P_2 = p_2 + p'_2 = p(Abc) + p(aBC),$$

$$P_3 = p_3 + p'_3 = p(ABc) + p(abC), \quad P_4 = p_4 + p'_4 = p(AbC) + p(aBc).$$

Derivation of Equation (7) is given in Supplementary Note S3. Chi-square test for RD among loci in natural population has the same expression with Equation (6). Supplementary Note S4 gives derivation of Equation (6a) in the case of which sister gametes have unequal frequencies. In an ideal population, Equation (B6) is reduced to Equation (4), underlying that if all three loci follow Hard-Weinberg equilibrium and sister gametes have equal frequencies, their RD derives only from recombination interference, otherwise, the RD results from selection, mutation, gene conversion, migration or drift and/or recombination interference.

## APPLICATIONS

### Test for recombination disequilibrium among dominant and codominant molecular markers in F<sub>2</sub> population

As an example for illustrating construction of linkage map by MAPMAKER/EXP (version 3.0b), LANDER *et al.* (LANDER *et al.* 1987) provided a RFLP dataset of 333 F<sub>2</sub> mice. Since RFLP markers are codominant in the genotype dataset, uppercase letters A, H, and B are used to denote homozygote A (a pair of the same alleles coming from parent A), heterozygote H, and homozygote B (a pair of the same alleles coming from parent B) at a locus, respectively. We applied our moment method (Supplementary Note S2) to the original dataset of the first 6 codominant markers and the modified TF (Supplementary Note S1) to dominant marker data converted by changing B to H to estimate frequencies of four nonsister gametes. Using these gamete frequencies, we calculated and tested for these 20 triplet RDs. The results are summarized in Tables 1~3. Table 1 displays the frequencies of the four nonsister gametes estimated by TF in 333 F<sub>2</sub> mouse individuals. One can find that RDs in 7 triplets (123), (125), (126), (135), (145), (235), and (245) are significant where  $\chi^2_{D_r} > \chi^2_{\alpha=0.05} = 3.84$  at  $p - value < 0.05$ . However, for the same genotype data, Table 2 obtained by performing the modified TF method shows that only triplets (123) and (145) have  $\chi^2_{D_r} > \chi^2_{\alpha=0.05} = 3.84$ . As codominant markers can display the full linkage information for estimation of gamete frequencies in F<sub>2</sub> population and have no phase problem (LIU 1998), the moment estimates of gamete frequencies in the codominant three-locus systems are pretty accurate (Table 3), which make RD estimation more reliable. The result obtained from codominant data shows that only triplet (123) had a significant RD. These indicate that estimate of gametic frequencies in a dominant three-locus system in F<sub>2</sub> population provided by the modified TF method are very close to the estimates of gametic frequencies in codominant three-locus system. The three methods all found that triplet (123) had a significant positive RD, therefore, we can make a conclusion that there was a significant

negative crossover interference among loci 1(T175), 2 (C35), and 3(T24). This conclusion is well validated by mapping of these three markers into two short intervals (T175-C35 in 3.1 cM and C35-T93 in 13.5cM (TAN and FU 2007)).

## Test for recombination disequilibrium

### among three genes in testcross phenotype data

A well-known example for testing for crossover interference among three genes in testcross/backcross phenotype data is the female testcross  $\frac{ts4}{+} \frac{lg2}{+} \frac{N}{T} \frac{al}{+} \times \frac{ts4}{+} \frac{lg2}{+} \frac{N}{+} \frac{al}{+}$  for the 3L.34 breakpoint in maize where  $T$  and  $N$  denote translocation breakpoint and normal point, respectively. The data can be found in Auger and Sheridan (AUGER and SHERIDAN 2001) and Tan and Fornage (TAN and FORNAGE 2008). In this testcross, the single crossover within interval II ( $lg2-T$ ) did not occur but double crossovers between adjacent intervals II ( $lg2-T$ ) and III ( $T-al$ ) did occur. The recombination fractions within intervals II ( $lg2-T$ ) and III ( $T-al$ ) are  $r_{lg2T} = 0+0.0047 = 0.0047$  and  $r_{Tal} = 0.2025+ 0.0047= 0.2072$  and frequency of the observed double crossover types is  $r_{DC} = 0.0047$ . RD among the three loci ( $lg2-T-al$ ) is  $D_r = r_{DC} - r_{lg2T}r_{Tal} = 0.00372$ , which has

$$\chi^2_{D_r} = \frac{630(r_{DC} - r_{lg2T}r_{Tal})^2}{r_{lg2T}r_{Tal}} = \frac{630(0.0047 - 0.0047 \times 0.2072)^2}{0.0047 \times 0.2072} = 8.98206 > \chi^2_{\alpha=0.05} = 3.84 \text{ at p-value}$$

<0.05 level. These two examples have a common point that at least one of two adjacent intervals is very short (TAN and FORNAGE 2008; TAN and FU 2007).

Another famous example also comes from *Drosophila*. There we have

$v\_cv^+\_ct^+$  (parental type): 580;  $v^+\_cv\_ct$  (parental type): 592;  $v^+\_cv\_ct^+$  (single crossover between genes  $ct$  and  $cv$ ): 45;  $v\_cv^+\_ct$  (single-crossover between genes  $ct$  and  $cv$ ): 40;  $v\_cv\_ct$  (single crossover between genes  $v$  and  $ct$ ): 89;  $v^+ - cv^+ - ct^+$  (single-crossover between genes  $v$  and  $ct$ ): 94;  $v^+\_cv^+\_ct$  (double-crossover): 3; and  $v\_cv\_ct^+$  (double crossover): 5. Total number of gametes is 1448 (ANTHONY *et al.* 2008). From this data, the frequencies of the four types of nonsister gametes are estimated as

$$p_2 = p(v\_cv^+\_ct^+) = (592+580)/(2 \times 1448) = 0.404696,$$

$$p_3 = p(v^+\_cv^+\_ct) = (3+5)/(2 \times 1448) = 0.002762431,$$

$$p_1 = p(v\_cv\_ct) = (89+94)/(2 \times 1448) = 0.06319061,$$

$$p_4 = p(v\_cv^+\_ct) = (45+40)/(2 \times 1448) = 0.02935083.$$

Thus, RD among the three genes  $v$ ,  $cv$ , and  $ct$  is  $D_r = 4(p_2p_3 - p_1p_4) = -0.0029470$  and

$$\chi_{D_r}^2 = \frac{ND_r^2}{r_{DC} - D_r} = 2.9688 < 3.84. \text{ Therefore, this positive interference among the three genes is}$$

due to experimental error.

## Evaluation of genetic map

An important application of RD is to evaluate genetic map of a chromosome or a chromosome fragment. If a chromosome map is correct and the three-locus RD is significant, then the three loci must be located within an adjacent region. Otherwise, some markers of this map would be falsely linked because a significant RD implicates a strong recombination interference that occurs between two adjacent shorter intervals. To elucidate this point, we here chose a genetic

map of rice chromosome 12 made by Harushima et al (HARUSHIMA *et al.* 1998). This chromosome map (here is called Harushima rice map 12 for convenience) consisting of 93 loci was built with 126 markers in an  $F_2$  population, among which 90 are codominant and 36 are dominant. For convenience, we here just considered the 90 codominant markers and used the raw data (downloaded from the website: <http://rgp.dna.affrc.go.jp/public> ) to construct 117481 triplets. 4 non-sister gamete frequencies  $p_1$ ,  $p_2$ ,  $p_3$ , and  $p_4$  in each triplet were estimated by applying our method (Supplementary Note S2). The results are summarized in Supplementary Table 1 (excel sheet1). By testing for RDs using Equations (6) (since there are strong linkage relationships among loci and each marker locus repeatedly occurs in many triplets, Bonferroni or Benjamini-Hochberg multiple tests are not available to test the 117481 RDs), we found 137 triplets with significant RDs at significant level of 0.05. Supplementary Table 2S (excel sheet2) shows that these 137 triplets form at least 12 larger linkage triplet groups: G24B, V57B, R3266, F8, S790B, C362A, C1116A, R769B, S2572, G193, R2253B, and C732B. Linkage triplet group G24B consists of 8 triplets (G24B, R2253B, C930), (G24B, S10637A, S10363), (G24B, V110, S1830), (G24B, R1709, R2672B), (G24B, C87, R496), (G24B, S861, R2672B), (G24B, L714, C930), (G24B, R1759, S13561). Among them, markers R2253B, V110, L714, and S10637A have their own linkage triplet groups, markers S10363, S1830, R1709, R2672B, C87, R496, S861, L714, C930, R1759, and S13561 are separately scattered in triplet groups V57B, R3266, V9A, R2253B, R642B, S894, R769B, S2572, V9A, S790B, R1957, F8, V110, C1116A. However, these markers were mapped into an end (72.9cM ~ 109.3cM) of Harushima rice map 12, while G24B was mapped into another end (5.5 cM) (see Fig.1A). On the other hand, the markers V57B, R3266, F8, G193, R2253B, C732B, S790B, R3025S, C362A, C104A, G1112, R328A, W120B, R1957, and C1116A that are tightly linked to G24B on the rice map12 are not members of G24B

triplet group. Again, markers R496, R1759, S1830, G1106, R2292, C930, S13561, and L405B were located at the same position (108.8 cM) on the Harushima rice map 12, but except for C930, S1830, R496, and S13561 that are members of G24B and V57B linkage triplet groups, all the others are scattered in different linkage triplet groups, indicating that it is impossible that these 8 markers were located at the same linkage position. Markers V124, S11679, M10C, and T5 were also located at the same linkage position of 70.7cM, but they are not found in these 137 linkage triplets with significant RD. To globally display RD chi-square tests of 137 triplets, we constructed a 4D plot (Fig.1B). To easily see incorrect of the Harushima rice map 12(Fig1.A), we gave the RDs expected to occur between two adjacent intervals among three adjacent loci mapped in the Harushima rice map 12. The expected RDs follow on red diagonal dots. Theoretically, if the Harushima rice map 12 is correct, then RDs should distribute along the red diagonal dots between original point (0,0,0) and end point (90,90,90) and in its neighboring regions. However, Fig.1B shows that almost all of these 137 RD points are far away from this diagonal line. These indicate that Harushima et al's map 12 is not correct.

## Haplotypes

Most of SNP haplotype data were derived from human populations that are not ideal population. To demonstrate that the RD test can be also applied to natural populations, we here used a haplotype dataset published by Fallin et al(FALLIN *et al.* 2001) as an example of RD analysis in natural populations. The haplotypes consist of 4 SNPs scattered in a 205kbp region on human chromosome 19. The map of the four SNP loci is M1M3M4M6. In the results of Fallin et al, no LD was found between M1 and M4 and between M1 and M6 but LD existed between M4 and M6, between M3 and M4, and between M3 and M6 with p-value<0.001, suggesting that M3, M4



and M6 are in short region. To perform triplet RD tests, we constructed four combinations of three-SNP haplotypes: M1M3M4, M1M4M6, M3M4M6 and M1M3M6 each having 8 haplotypes. The results of our RD analysis of four combinations of three-haplotypes were summarized in Table 4. Table 4 shows that there are no RDs among M1, M4 and M6, among M1, M3 and M6, while M3, M4 and M6 had very significant RD, which is very consistent with significant LDs between M3 and M4, between M4 and M6 and between M3 and M6.

## DISCUSSION

Estimation of gamete frequencies is the first step to test for RD among three or multiple loci. Therefore a good method for accurately estimating gamete frequencies in a population is required. In  $F_2$  population, for dominant loci, the current existing EM methods have low power (LIU 1998 ) because these EM methods still inefficiently distinguish dominant homozygous genotypes from dominant heterozygous genotypes and have to use a lot of trivial and complex algebraic equations to overcome this difficulty, while the TF method efficiently utilizes sister gamete genotypes with equal frequencies to reduce complexity of estimation of gametic frequencies. The modified TF method tries to find a desirable value of  $q_1$  such that difference between the observed and expected frequencies of gametes is relatively small. Therefore, it can give more accurate estimation of gamete frequencies than the TF method. This point was supported by the above results of testing for RD among loci obtained from the real genotype data (Tables 1-3). For codominant loci, the EM algorithms have really high power to estimate frequencies of three-locus gametes (Li e tal 2005 and Esch 2005), however, the computational burden would be extremely huge when number (n) of loci is very large because there would be

$n^3$  triplet gametes (Esch 2005). For example, in general GWAS data, number of SNPs is more than 2000000, so there would be  $2000000^3 = 8 \times 10^{18}$  three-locus gametes. To significantly reduce computational burden, estimation of gamete frequencies of genome-wide loci requires both fast and powerful method. This is why here we want to propose this new method. Since all genotypes can be recognized and are informative for estimation of gamete frequencies, there is not phase problem to be solved. In addition, likewise, the assumption that sister gametes also have the same frequencies reduces complexity of estimation equations. For example, we just use 10 simple binomial equations to estimate frequencies of four non-sister gametes, while Li et al (2005)'s EM method requires 18 complex algebraic equations and Hospital et al (HOSPITAL *et al.* 1996)' method requires 20 equations. In natural or outbreeding populations, frequencies of gametes or haplotypes can be estimated by using the existing powerful methods such as PHASE(STEPHENS *et al.* 2001), fastPHASE(SCHEET and STEPHENS 2006) , BEAGLE (BROWNING and BROWNING 2007) , IMPUTE2 (HOWIE *et al.* 2009) and MaCH(LI *et al.* 2010).

One application of RD is the construction of genetic maps for fine mapping. As seen in Tables 1-3, some triplets have significant RD, but most of triplets have no significant RD. For the triplets without significant RD, their coincident coefficients are set to be 1; for those with significant negative or positive RD, the coincident coefficients are given by Tan and Fornage definition (TAN and FORNAGE 2008). Thus, one can use Tan and Fornage mapping functions (TAN and FORNAGE 2008) to accurately calculate the genetic distances between these markers.

Since RD provides information of recombination interference among three loci, it is useful for the evaluation of the linkage map. As seen above, Harushima rice map12 is very

inconsistent with results of our triplet analysis. One possible reason is that this map was based on two-point analysis and the two-point approach does not utilize information of recombination interference occurring between two adjacent intervals, while recombination interference disturbs linkage information between loci. As a result, two-point analysis would generate incorrect linkage between loci. Using three-locus EM methods or our methods (Supplementary Note S2), one can obtain recombination fractions among three adjacent loci, and then use the method provided by Tan and Fu (TAN and FU 2007) to convert three-locus recombination fractions into two-locus recombination fractions. A correct genetic map can be made by using UG mapping method (TAN and FU 2006) or other mapping methods, the chi-square test for the three-locus RD and Tan-Fornage map functions (TAN and FORNAGE 2008).

In ideal population such as  $F_2$  or backcross population, RD is purely due to recombination interference between short intervals but in natural populations, in addition to recombination interference, RD may also result from selection, mutation, drift, and gene conversion because these evolutionary factors also change frequencies of gametes. So as with LD, evolutionary history of multiple genes can be revealed by testing for RD among genes. In addition, RD can also be applied to genome-wide association analysis of haplotypes with diseases. For this application, we have furthermore developed a new method for studying association of haplotypes with disease of study and in somewhere we will publish it.

## Acknowledgments

## References

- ANTHONY, J., F. GRIFFITHS, S. R. WESSLER and R. C. LEWONTIN, 2008 Introduction to Genetic Analysis, pp. 141-143, edited by J. CORREA. Sara Tenney, New York.
- AUGER, D. L., and W. F. SHERIDAN, 2001 Negative crossover interference in maize translocation heterozygotes. *Genetics* **159**: 1717-1726.
- BENNETT, J. H., 1954 On the theory of random mating. *Ann Eugen* **18**: 311-317.
- BROWNING, S. R., and B. L. BROWNING, 2007 Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**: 1084-1097.
- ESCH, E., 2005 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* **111**: 100-109.
- FALLIN, D., A. COHEN, L. ESSIOUX, I. CHUMAKOV, M. BLUMENFELD *et al.*, 2001 Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. *Genome Res* **11**: 143-151.
- GEIRINGER, H., 1944 On the probability theory of linkage in Mendelian heredity. *Annals of Math. Stat* **15**: 25-57.
- HARDY, G. H., 1908 Mendelian Proportions in a Mixed Population. *Science* **28**: 49-50.
- HARUSHIMA, Y., M. YANO, A. SHOMURA, M. SATO, T. SHIMANO *et al.*, 1998 A high-density rice genetic linkage map with 2275 markers using a single F2 population. *Genetics* **148**: 479-494.
- HASTINGS, A., 1984 Linkage disequilibrium, selection and recombination at three Loci. *Genetics* **106**: 153-164.
- HILL, W. G., and A. ROBERTSON, 1968 The effects of inbreeding at loci with heterozygote advantage. *Genetics* **60**: 615-628.
- HOSPITAL, F., C. DILLMANN and A. E. MELCHINGER, 1996 A general algorithm to compute multilocus genotype frequencies under various mating systems. *Comput Appl Biosci* **12**: 455-462.
- HOWIE, B. N., P. DONNELLY and J. MARCHINI, 2009 A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**: e1000529.
- LANDER, E. S., P. GREEN, J. ABRAHAMSON, A. BARLOW, M. J. DALY *et al.*, 1987 MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1**: 174-181.
- LEWONTIN, R., and K. KOJANA, 1960 The evolutionary dynamics of complex polymorphisms. *Evolution* **14**: 458-472.
- LEWONTIN, R. C., 1964 The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics* **49**: 49-67.
- LI, Y., C. J. WILLER, J. DING, P. SCHEET and G. R. ABECASIS, 2010 MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* **34**: 816-834.
- LI, Y., S. WU, K. HAN, Z. WANG, W. HOU *et al.*, 2007 Estimation of multilocus linkage disequilibria in diploid populations with dominant markers. *Genetics* **176**: 1811-1821.
- LIU, B. H., 1998 *Statistical genomics: Linkage, mapping, and QTL analysis*. CRC Press LLC,.
- LIU, B. H., (Editor), 1998 *Statistical genomics: Linkage, mapping, and QTL analysis*. CRC Press, Florida.
- PRITCHARD, J. K., and M. PRZEWORSKI, 2001 Linkage disequilibrium in humans: models and data. *Am J Hum Genet* **69**: 1-14.
- ROBBINS, R. B., 1918 Applications of Mathematics to Breeding Problems II. *Genetics* **3**: 73-92.
- SCHEET, P., and M. STEPHENS, 2006 A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78**: 629-644.
- SERGEEV, A. S., and R. K. ARAPOVA, 2002 [The use of the expectation-maximization (EM) algorithm for maximum likelihood estimation of gametic frequencies of multilocus polymorphic codominant systems based on sampled population data]. *Genetika* **38**: 407-418.
- SLATKIN, M., 2008 Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* **9**: 477-485.
- STEPHENS, M., N. J. SMITH and P. DONNELLY, 2001 A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68**: 978-989.
- TAN, Y. D., and M. FORNAGE, 2008 Mapping functions. *Genetica* **133**: 235-246.
- TAN, Y. D., and Y. X. FU, 2006 A novel method for estimating linkage maps. *Genetics* **173**: 2383-2390.
- TAN, Y. D., and Y. X. FU, 2007 A new strategy for estimating recombination fractions between dominant markers from an F2 population. *Genetics* **175**: 923-931.
- THOMSON, G., and M. P. BAUR, 1984 Third order linkage disequilibrium. *Tissue Antigens* **24**: 250-255.

- WEINBERG, W., 1908 Über den Nachweis der Vererbung beim Menschen. Jahresh. Wuerth. Ver. vaterl. Natkd **64**: 369-382.
- WIGGINTON, J. E., D. J. CUTLER and G. R. ABECASIS, 2005 A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet **76**: 887-893.

Table 1. The moment estimated frequencies of nonsister gametes in triplets  
of 6 dominant loci in 333 F2 mice<sup>a</sup> and recombination disequilibrium tests

Locus			$p_1 =$	$p_2 =$	$p_3 =$	$p_4 =$	RD	
a	b	c	$p(abc)$	$p(Abc)$	$p(abC)$	$p(aBc)$	$D_r$	$\chi^2$
1	2	3	0.200547	0.089443	0.097526	0.112484	0.055341	5.99423*
1	2	4	0.156654	0.081497	0.162688	0.099161	-0.0091	0.159813
1	2	5	0.197112	0.082379	0.109589	0.11092	0.051343	5.13317*
1	2	6	0.215739	0.072776	0.123004	0.088481	0.040548	4.01347*
1	3	4	0.124449	0.091552	0.176875	0.107124	0.011447	0.251911
1	3	5	0.236392	0.049819	0.118733	0.095055	0.066221	11.7868*
1	3	6	0.176321	0.05693	0.167597	0.099152	0.031765	2.017614
1	4	5	0.186545	0.095081	0.071434	0.14694	0.082476	10.5217*
1	4	6	0.180297	0.096093	0.109022	0.114587	0.040734	2.888049
1	5	6	0.172502	0.079552	0.149065	0.098881	0.020795	0.808814
2	3	4	0.143189	0.087697	0.173774	0.09534	0.006351	0.079209
2	3	5	0.239605	0.069337	0.103666	0.087392	0.055007	8.41204*
2	3	6	0.141078	0.084521	0.169385	0.105016	-0.002	0.007681
2	4	5	0.201763	0.110214	0.069613	0.11841	0.064874	8.07710*
2	4	6	0.074161	0.142371	0.142371	0.141097	0	0
2	5	6	0.172724	0.088155	0.142527	0.096595	0.016479	0.511758
3	4	5	0.157883	0.115863	0.089558	0.136697	0.044822	2.926938
3	4	6	0.098566	0.141188	0.131539	0.128707	0.023542	0.76588
3	5	6	0.15671	0.10181	0.153616	0.087864	-0.00748	0.101766
4	5	6	0.137416	0.101387	0.156667	0.10453	0.00608	0.062021

\*P-value <0.05. a: The data came from MAPMAKER/EXP(3.0b)[27]

Table 2. The ELS estimated frequencies of nonsister gametes in triplets of 6 dominant loci in 333 F2 mice<sup>a</sup> and recombination disequilibrium tests

Locus			$p_1 =$	$p_2 =$	$p_3 =$	$p_4 =$	RD	
a	b	c	$p(abc)$	$p(Abc)$	$p(abC)$	$p(aBc)$	$D_r$	$\chi^2$
1	2	3	0.208668	0.086162	0.094698	0.110472	0.059571	4.19998*
1	2	4	0.14751	0.087958	0.160866	0.103665	0.004568	0.038389
1	2	5	0.200976	0.080676	0.108494	0.109854	0.053301	3.454827
1	2	6	0.192229	0.084408	0.126033	0.09733	0.032286	1.529476
1	3	4	0.140566	0.079252	0.181795	0.098387	-0.00231	0.011286
1	3	5	0.209093	0.065783	0.12237	0.102753	0.05374	3.709605
1	3	6	0.16895	0.063323	0.165648	0.102079	0.027028	1.052238
1	4	5	0.173539	0.100396	0.079665	0.1464	0.069631	4.37455*
1	4	6	0.16482	0.102771	0.113819	0.118591	0.031395	1.203758
1	5	6	0.173447	0.07932	0.148645	0.098588	0.021237	0.683502
2	3	4	0.141958	0.088919	0.172784	0.096339	-0.00675	0.088422
2	3	5	0.202566	0.085775	0.112098	0.099561	0.04221	2.458404
2	3	6	0.139672	0.086032	0.16843	0.105866	0.001185	0.002673
2	4	5	0.173634	0.117152	0.085607	0.123607	0.045733	2.356028
2	4	6	0.10113	0.133668	0.133668	0.131535	0	0
2	5	6	0.156859	0.096648	0.142758	0.103735	0.009898	0.150093
3	4	5	0.143582	0.120278	0.098137	0.138002	0.032044	1.109958
3	4	6	0.105025	0.139707	0.129181	0.126086	0.019221	0.440547
3	5	6	0.140841	0.109125	0.152351	0.097682	-0.01147	0.211908
4	5	6	0.146847	0.096392	0.156905	0.099856	-0.00184	0.005869

\*P-value <0.05. a: The data came from MAPMAKER/EXP(3.0b)[27]

Table 3. The moment estimated frequencies of nonsister gametes in triplets of 6 codominant loci in 333 F2 mice<sup>a</sup> and recombination disequilibrium tests

Locus			$p_1 =$	$p_2 =$	$p_3 =$	$p_4 =$	RD	
1	2	3	$p(000)$	$p(110)$	$p(011)$	$p(010)$	$D_r$	$\chi^2$
1	2	3	0.242929	0.066568	0.080146	0.110358	0.085896	18.1688*
1	2	4	0.145838	0.08977	0.162134	0.102258	-0.00143	0.003769
1	2	5	0.196094	0.091387	0.121051	0.091467	0.027495	1.614612
1	2	6	0.17224	0.104184	0.143308	0.080268	-0.00442	0.039436
1	3	4	0.165099	0.068697	0.191983	0.074221	0.00374	0.034531
1	3	5	0.222931	0.079297	0.147828	0.049944	-0.00235	0.018027
1	3	6	0.177615	0.065713	0.187929	0.068743	0.000558	0.000793
1	4	5	0.158699	0.103969	0.114759	0.122573	0.030083	1.37606
1	4	6	0.165089	0.113628	0.139128	0.082155	-0.00898	0.152777
1	5	6	0.155874	0.091722	0.16263	0.089774	0.003693	0.025114
2	3	4	0.142565	0.093943	0.179432	0.08406	0.01949	0.748921
2	3	5	0.216411	0.069533	0.134853	0.079203	0.031055	2.521765
2	3	6	0.160337	0.092614	0.172787	0.074262	0.016382	0.524515
2	4	5	0.172459	0.100044	0.105018	0.12248	0.042465	2.938794
2	4	6	0.154284	0.140079	0.121173	0.084464	-0.01577	0.442962
2	5	6	0.167782	0.072156	0.153072	0.10699	0.027624	1.363535
3	4	5	0.154314	0.118649	0.10895	0.118086	0.021182	0.694969
3	4	6	0.144358	0.108635	0.131647	0.115359	0.009406	0.132324
3	5	6	0.16738	0.053399	0.176627	0.102594	-0.03096	2.317289
4	5	6	0.153613	0.092081	0.124467	0.129838	0.033935	1.68348

\*P-value <0.05. a: The data came from MAPMAKER/EXP(3.0b)[27]



Table 4. RD and chi-square testing RD among three SNPs in four haplotypes (M1M3M4\*M6) where M4\* is C19M4 that is part of ApoE-ε4

	M1M3M4*	M1M4*M6	M3M4*M6	M1M3M6
<b>P1</b>	0.475	0.305	0.589	0.32
<b>P2</b>	0.032	0.199	0.358	0.175
<b>P3</b>	0.472	0.292	0.008	0.316
<b>P4</b>	0.023	0.206	0.047	0.191
<b>RD</b>	-0.0042	0.0047	0.0248	0.0058
<b>X<sup>2</sup></b>	0.237	0.041	10.247	0.067
<b>p-value</b>	0.626	0.839	0.0014	0.795

