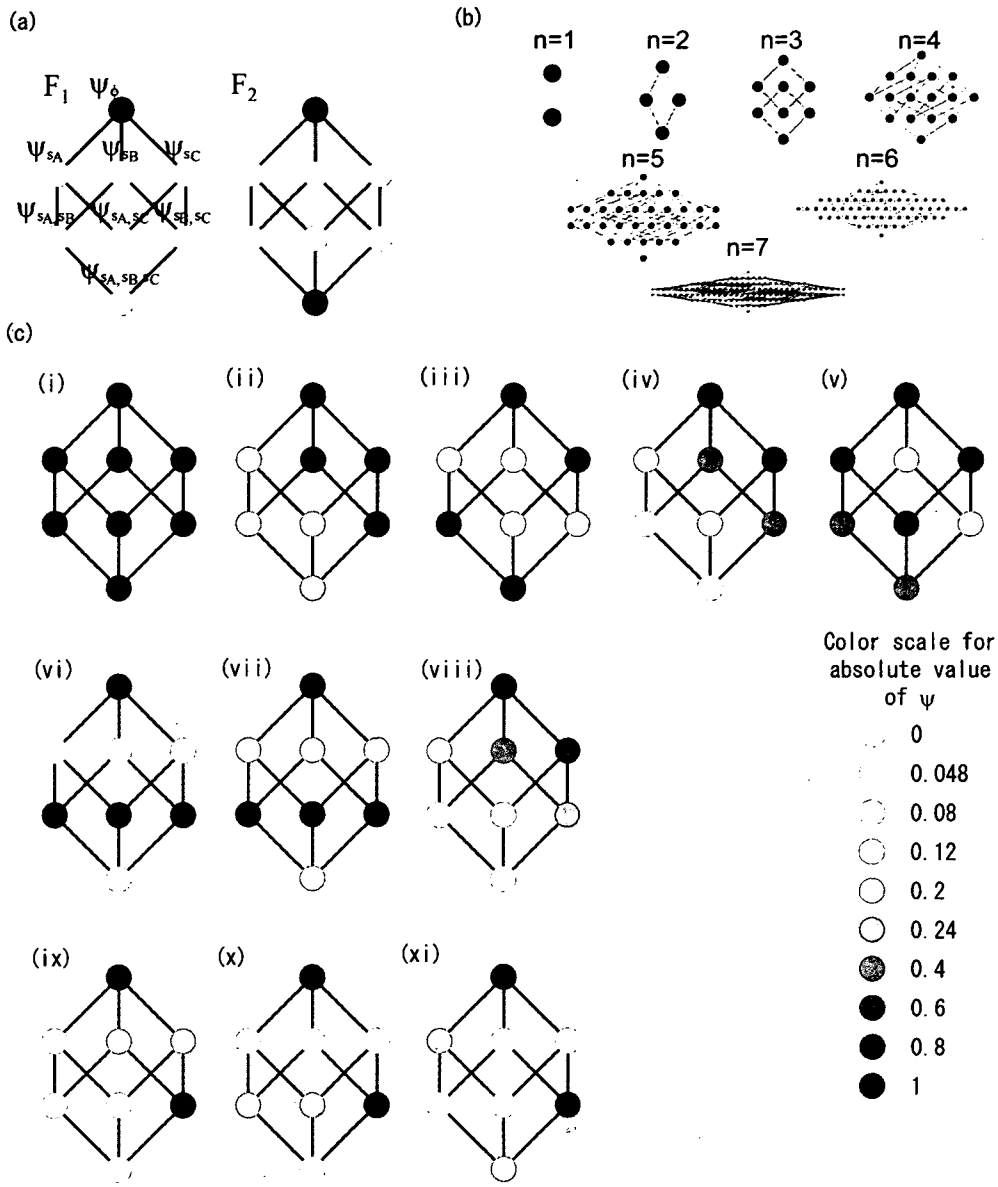


Fig. 1. (a) ψ plot of two haplotype frequency sets, F_1 and F_2 are drawn. ψ plots for three sites are consisted of four rows. The top row is for the empty set, the second top row is for three single sites, the third top row is for three site-pairs and the bottom row is for the site-trio. The circle on the left in the third top row for ψ_{s_A,s_B} is connected to two circles in the second top row, ψ_{s_A} and ψ_{s_B}, representing the relation of $\{s_A\} \subset \{s_A, s_B\}$ and $\{s_B\} \subset \{s_A, s_B\}$. The bottom row has one white (F_1) or black (F_2) circle, corresponding to the site-trio. It is connected to three circles in the third top row, because every site-pair is a subset of the trio. The circles are connected when numbers of elements of two subsets are different by one and the smaller subset is a subset of the larger. Black circles represent ψ value being 1 and white 0. (b) Power sets for 1–7 element set are drawn. They are also ψ plots of clones with $n = 1, \ldots, 7$ are shown. Every ψ plot has one circle at the top as the empty set and one circle at the bottom corresponding to the self subset. (c) Various ψ plots are displayed. The corresponding haplotype frequencies and D_g values are shown in Table I (columns 3 to 14). ψ values were shown in gray scale in (a) and (c).

For the case of F_1, $D_g = \{0,0,0,0,0,0,0\}$ and for F_2, $D_g = \{0,0,0,1,1,1,1\}$. The elements for the divisions of SNP pairs into single SNPs are 0 for both cases, which corresponds to the fact r^2 of three SNP pairs are 0. For F_1 all the other elements of D_g are also 0, indicating that the three sites are truly in linkage equilibrium (LE). On the other hand, the last four elements of D_g for F_2 are different from zero. These four elements represent components of LD in these three sites that can not be described by pairwise LD indices but should be described by taking account of LD for the trio.

Because the number of divisions into subsets becomes very large when the number of sites is

increased, we propose to choose a part of the elements of D_g for visual presentation of D_g and they are plotted into two triangles. One triangle is consisted of d_g's for divisions of all the site-pairs into single sites (pairwise triangle). The other triangle is consisted of d_g's for divisions of all the subsets of sites whose elements are in tandem into single sites (tandem triangle). In case of four sites, the pairwise triangle is consisted of $d_g((s_1,s_2) \to \{(s_1),(s_2)\})$, $d_g((s_1,s_3) \to \{(s_1),(s_3)\})$, $d_g((s_1,s_4) \to \{(s_1),(s_4)\})$, $d_g((s_2,s_3) \to \{(s_2),(s_3)\})$, $d_g((s_2,s_4) \to \{(s_2),(s_4)\})$ and $d_g((s_3,s_4) \to \{(s_3),(s_4)\})$. The tandem triangle is consisted of $d_g((s_1,s_2) \to \{(s_1),(s_2)\})$, $d_g((s_2,s_3) \to \{(s_2),(s_3)\})$, $d_g((s_3,s_4) \to \{(s_3),(s_4)\})$, $d_g((s_1,s_2,s_3) \to \{(s_1),(s_2),(s_3)\})$, $d_g((s_2,s_3,s_4) \to \{(s_2),(s_3),(s_4)\})$ and $d_g((s_1,s_2,s_3,s_4) \to \{(s_1),(s_2),(s_3),(s_4)\})$. Because the pairwise triangle and the tandem triangle share d_g's for divisions of the site-pairs in tandem, the corresponding parts of the two triangles are overlapped and displayed as shown in Figure 2(a). D_g plots for F_1 and F_2 with three sites are shown in Figure 2(b).

MORE EXAMPLES OF Ψ AND D_g FOR THREE SNPS

Now Ψ and D_g are calculated in additional examples with three sites. When all three sites are monomorphic (See column 3 "Clone1" and column 4 "Clone2" in Table I and Fig. 1(c)–(i)), all $|\psi|$'s are 1 and all d_g's are 0. When one of three sites are polymorphic (See column 5 "1 SNP" in Table I and Fig. 1(c)–(ii)), $|\Psi|$'s for the subsets containing the polymorphic site are not 1. All d_g's are 0. Three examples with two SNPs are shown in column 6 "two SNPs in absolute LD", column 7 "two SNPs in LE" and column 8 "two SNPs in partial LD", and Figure 1(c)–(iii),(iv),(v)). When two SNPs are in the absolute LD, a black circle of ψ_{s_B,s_C} in Figure 1(c)–(iii) represents the allelic association. All d_g's are 0 when two SNPs are in LE. When two SNPs in LD, $d_g((s_A,s_B) \to \{(s_A),(s_B)\})$ stands for the strength of LD between the two SNPs. (Columns 6 and 8 of Table I).

Columns 9 and 10, "3 SNPs in absolute LD-1" and "3 SNPs in absolute LD-2" in Table I and Fig. 1(c)–(vi),(vii) present examples where all the three sites are polymorphic and only two haplotypes exist. These two differ in their allele frequencies. The former has two haplotypes with the same frequency and the latter's haplotypes have different frequency. All three SNP pairs are in their absolute LD ($r^2 = 1$). Ψ plots distinguish these two by gray color in the circles for three single SNPs and the trio. The difference between two examples is observed in $d_g((s_A,s_B,s_C) \to \{(s_A),(s_B),(s_C)\})$.

When three SNPs are in LE as shown in column 11 of Table I, all d_g's are 0 (Fig. 1(c)–(viii)). The gradation in gray scale from top to bottom in the Ψ plot is a feature of LE throughout the sites. Columns 12, 13 and 14 are the examples with partial LD in three polymorphic sites. All of them have the same four haplotypes out of eight and their frequencies are 0.2 or 0.3. s_A and s_B are in the absolute LD for all the three examples, as their $d_g((s_B,s_C) \to \{(s_B),(s_C)\}) = 1$. Their difference appears in the distribution of non-zero values in their D_g and their Ψ plots (Fig. 1(c)–(ix),(x),(xi)). Their D_g plots are shown in Figure 2(c).

Estimation of haplotype frequency using Ψ from unphased genotype data. When unphased genotype data of SNP pairs are given, frequencies of four haplotypes have to be estimated, and Ψ transforms this estimation into a monovariate problem.

Example: when genotype counts are observed for two SNPs, allele frequencies of both SNPs are calculated based on their own genotype counts. They give $\psi_{s_A} = f_A - f_a$ and $\psi_{s_B} = f_B - f_b$. In order to estimate frequencies of all four haplotypes, ψ_{s_A,s_B} is the only variable. Therefore estimation of haplotype frequencies of SNP pairs turns to be the same with maximal likelihood estimation of monovariate, ψ_{s_A,s_B}. Once ψ_{s_A,s_B} is estimated, $\{f_{AB}, f_{Ab}, f_{aB}, f_{ab}\}$ can be given by

$$f_{AB} = \tfrac{1}{4}(\psi_{s_A,s_B} + \psi_{s_A} + \psi_{s_B} + \psi_\phi),$$
$$f_{Ab} = \tfrac{1}{4}(-\psi_{s_A,s_B} + \psi_{s_A} - \psi_{s_B} + \psi_\phi),$$
$$f_{aB} = \tfrac{1}{4}(-\psi_{s_A,s_B} - \psi_{s_A} + \psi_{s_B} + \psi_\phi),$$
$$f_{ab} = \tfrac{1}{4}(\psi_{s_A,s_B} - \psi_{s_A} - \psi_{s_B} + \psi_\phi).$$

This topic is discussed in the section "Usage of Ψ for haplotype frequency inference".

NOTATIONS

SITES

- Consider a set of DNA sequences with the same length n. The sites are not necessarily polymorphic.
- Let $S(n)(1_{st})$ denote the first set of n sites. All the sites are potentially diallelic although some of them can be monomorphic;

$$S(n)(1_{st}) = \{s_1, s_2, \ldots, s_n\}.$$

- Let $\mathrm{Pow}(S(n)(1_{st}))$ denote a power set of $S(n)(1_{st})$ (Fig. 1(b)) [Weisstein, 2006a].

$$\begin{aligned}\mathrm{Pow}(S(n)(1_{st})) = \{&S(0)(1_{st}),\\ &S(1)(1_{st}), S(1)(2_{nd}), \ldots, S(1)(n_{th}),\\ &S(2)(1_{st}), S(2)(2_{nd}), \ldots, S(2)({}_nC_{2_{th}}),\\ &S(3)(1_{st}), \ldots, S(3)({}_nC_{3_{th}}), \ldots,\\ &S(n-1)(1_{st}), \ldots, S(n-1)({}_nC_{n-1_{th}}),\\ &S(n)(1_{st})\}.\end{aligned}$$

- $S(i)(j_{th})$ represents the j_{th} subset with i sites in $\mathrm{Pow}(S(n)(1_{st}))$, $(i = 0,1,\ldots,n; j = 1,2,\ldots,{}_nC_i)$. $S(0)(1_{st})$ is an empty set and the last element of $\mathrm{Pow}(S(n)(1_{st}))$ is $S(n)(1_{st})$ itself.

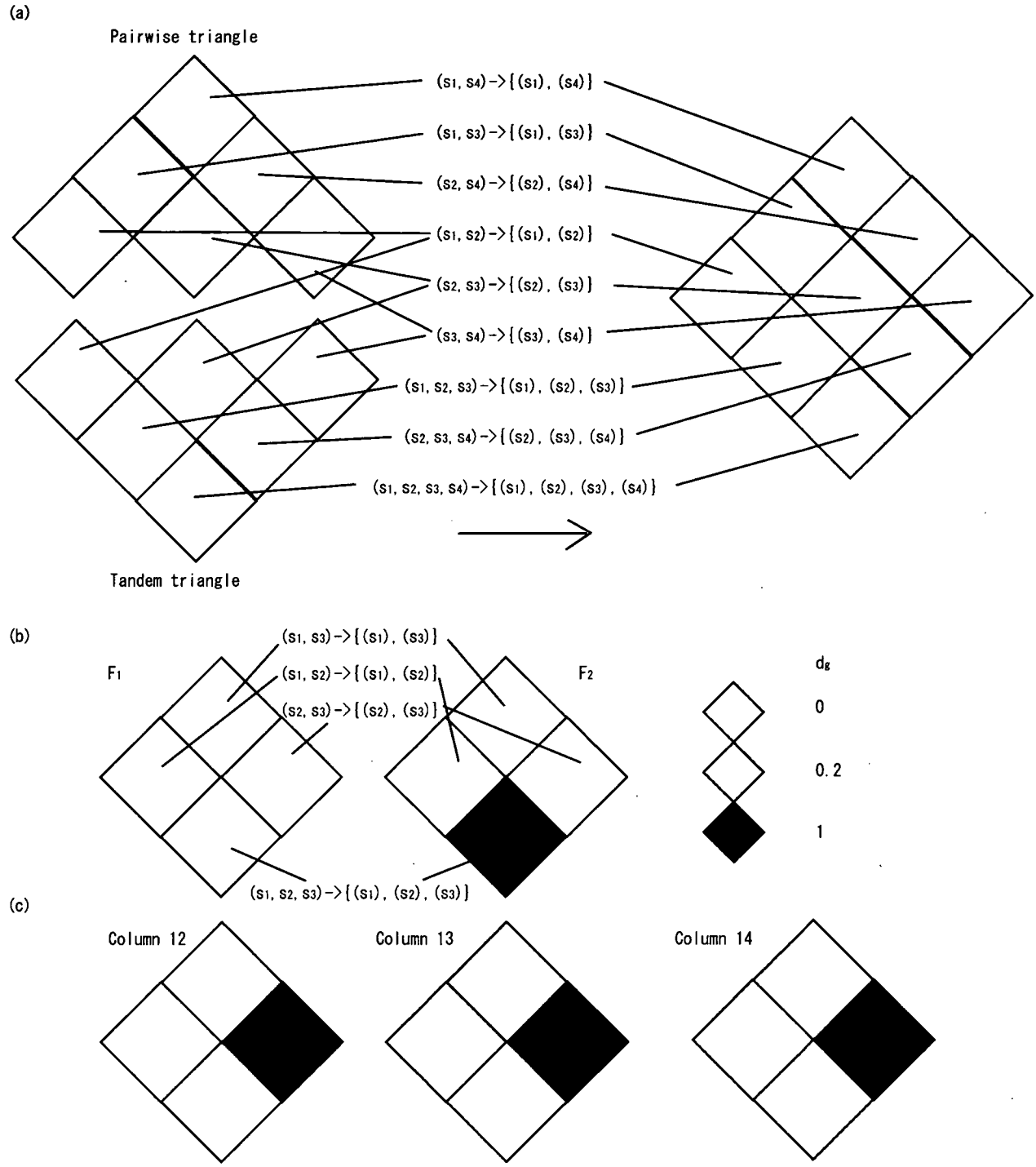


Fig. 2. D_g plots. (a) The pairwise triangle, the upper half of D_g plots, are consisted of squares for the site pairs. The tandem triangle, the lower half of D_g plots, are consisted of squares for the subsets of sites in tandem. The squares for the site pairs in tandem are arranged on the bottom of the pairwise triangle and on the top of the tandem triangle. Therefore they are overlapped in the right drawing. (b) D_g plots for F_1 and F_2. Their difference appears in the lower tandem triangle but not in the upper pairwise triangle. (c) D_g plots for examples of Columns 12, 13 and 14 (Table I).

- Number of subsets which have i elements is ${}_nC_i$. The total number of elements of Pow($S(n)(1)$) is $\sum_{i=0}^{n} {}_nC_i = 2^n$.

HAPLOTYPES

- Let $H(i)(j_{\text{th}})$ and $F(i)(j_{\text{th}})$ denote the 2^i haplotypes and their frequency of $S(i)(j_{\text{th}})$.

$$H(i)(j_{\text{th}}) = \{h_1(i)(j_{\text{th}}), h_2(i)(j_{\text{th}}), \ldots, h_{2^i}(i)(j_{\text{th}})\},$$

$$F(i)(j_{\text{th}}) = \{f_1(i)(j_{\text{th}}), f_2(i)(j_{\text{th}}), \ldots, f_{2^i}(i)(j_{\text{th}})\}.$$

- Let $V(i)(j_{\text{th}})$ denote alternating positive/negative signs for $H(i)(j_{\text{th}})$ as mentioned in the section "INTRODUCTORY EXAMPLES".

$$V(i)(j_{\text{th}}) = \{v_1(i)(j_{\text{th}}), v_2(i)(j_{\text{th}}), \ldots, v_{2^i}(i)(j_{\text{th}})\}.$$

Initially dummy value, 1 or −1, is given to two alleles of individual n sites in $S(n)(1_{\text{st}})$. When a haplotype has even number of sites of value −1, dummy value of the haplotype is 1, and when it has odd number, the value is −1.
Example: Assume $i = 3$ and two alleles of s_1 are A and T, for s_2, G and C and for s_3, C and A where value 1 is assigned to the first allele of each site. Two of three sites (the first and third sites) of haplotype TGA is the allele of (−1), therefore the dummy value of haplotype TGA is 1.

- Partial haplotype. When $S(p_i)(q_{i_{\text{th}}}) \subset S(p_j)(q_{j_{\text{th}}})$ and $h_{k_i}(p_i)(q_{i_{\text{th}}})$ is a part of $h_{k_j}(p_j)(q_{j_{\text{th}}})$, $h_{k_i}(p_i)(q_{i_{\text{th}}})$ is called as a partial haplotype of $h_{k_j}(p_j)(q_{j_{\text{th}}})$in $S(p_i)(q_{i_{\text{th}}})$. Let $u^{p_i,q_{i_{\text{th}}}}_{kj,pj,qjth}$ denote the ordinal number to indicate a partial haplotype in $S(p_i)(q_{i_{\text{th}}})$ for $S(p_j)(q_{j_{\text{th}}})$.

Example: Consider the sample example in the previous bullet. $S(2)(1_{\text{st}}) = \{s_1, s_2\}$ is a subset of $S(3)(1_{\text{st}}) = \{s_1, s_2, s_3\}$, $(S(2)(1_{\text{st}}) \subset S(3)(1_{\text{st}}))$.

$$\begin{aligned} H(3)(1_{\text{st}}) &= \{h_1(3)(1_{\text{st}}), h_2(3)(1_{\text{st}}), h_3(3)(1_{\text{st}}), h_4(3)(1_{\text{st}}), \\ &\quad h_5(3)(1_{\text{st}}), h_6(3)(1_{\text{st}}), h_7(3)(1_{\text{st}}), h_8(3)(1_{\text{st}})\} \\ &= \{\text{"}AGC\text{"}, \text{"}AGA\text{"}, \text{"}ACC\text{"}, \text{"}ACA\text{"}, \text{"}\underline{TG}C\text{"}, \\ &\quad \text{"}\underline{TG}A\text{"}, \text{"}TCC\text{"}, \text{"}TCA\text{"}\}, \\ H(2)(1_{\text{st}}) &= \{h_1(2)(1_{\text{st}}), h_2(2)(1_{\text{st}}), h_3(2)(1_{\text{st}}), h_4(2)(1_{\text{st}})\} \\ &= \{\text{"}AG\text{"}, \text{"}AC\text{"}, \text{"}\underline{TG}\text{"}, \text{"}TC\text{"}\}, \end{aligned}$$

$h_3(2)(1_{\text{st}}) = \text{"}\underline{TG}\text{"}$ is a part of $h_5(3)(1_{\text{st}}) = \text{"}\underline{TG}C\text{"}$ and $h_6(5)(1_{\text{st}}) = \text{"}\underline{TG}A\text{"}$. Then $u^{2,1_{\text{st}}}_{5,3,1_{\text{st}}} = 3$ and $u^{2,1_{\text{st}}}_{6,3,1_{\text{st}}} = 3$.

DIVISION OF A SET OF SITES INTO SUBSETS

- Divisions of a set of sites.

Consider a division of a set of n sites. $S(n)(1_{\text{st}})$ is divided into m non-empty subsets that are mutually exclusive $(S(n)(1_{\text{st}}) = \bigcup_{i=1}^{m} S_i(p_i)(q_{i_{\text{th}}}); i = 1, 2, \ldots, m; p_i \neq 0; n = \sum_{i=1}^{m} n_i; q_i = 1, 2, \ldots, {}_nC_{p_i}; S_i \cap S_j = \{\phi\}$ for any i and j $(i \neq j))$. Let $S(n)(1_{\text{st}}) \rightarrow \{S_1(p_1)(q_{1_{\text{th}}}), \ldots, S_m(p_m)(q_{m_{\text{th}}})\}$ denote this division pattern.
Example: The above mentioned example $S(3)(1_{\text{st}}) = \{s_1, s_2, s_3\}$ is divided into four different division patterns; Into $S(2)(1_{\text{st}}) = \{s_1, s_2\}$ and $S(1)(3_{\text{rd}}) = \{s_3\}$, or into $S(2)(2_{\text{nd}}) = \{s_1, s_3\}$ and $S(1)(2_{\text{nd}}) = \{s_2\}$, or into $S(2)(3_{\text{rd}}) = \{s_2, s_3\}$ and $S(1)(1_{\text{st}}) = \{s_1\}$, or into three single sites $S(1)(1_{\text{st}}) = \{s_1\}$, $S(1)(2_{\text{nd}}) = \{s_2\}$ and $S(1)(3_{\text{rd}}) = \{s_3\}$.

- Division of a haplotype into partial haplotypes.

When $S(n)(1_{\text{st}})$ is divided into m subsets, $H(n)(1_{\text{st}})$ is also divided into their m partial haplotypes, each of which is a haplotype in $S_i(p_i)(q_{i_{\text{th}}})$. $h_k(n)(1_{\text{st}})$, the kth haplotype of length n in $S(n)(1_{\text{st}})$, is expressed as a set,

$$\begin{aligned} h_k(n)(1_{\text{st}}) = (&h_{u^{p_1,q_{1_{\text{th}}}}_{k,n,1_{\text{st}}}}(p_1)(q_{1_{\text{th}}}), h_{u^{p_2,q_{2_{\text{th}}}}_{k,n,1_{\text{st}}}}(p_2)(q_{2_{\text{th}}}), \ldots, h_{u^{p_i,q_{i_{\text{th}}}}_{k,n,1_{\text{st}}}} \\ &(p_i)(q_{i_{\text{th}}}), \ldots, h_{u^{p_m,q_{m_{\text{th}}}}_{k,n,1_{\text{st}}}}(p_m)(q_{m_{\text{th}}})). \end{aligned}$$

where $h_{u^{p_i,q_{i_{\text{th}}}}_{k,n,1_{\text{st}}}}(p_i)(q_{i_{\text{th}}})$ represents the $u^{p_i,q_{i_{\text{th}}}}_{k,n,1_{st}}$ haplotype in $S_i(p_i)(q_{i_{\text{th}}})$, that is a part of $h_k(n)(1_{\text{st}})$.
Example: when $S(3)(1_{\text{st}}) = \{s_1, s_2, s_3\}$ is divided into $S(2)(2_{\text{nd}}) = \{s_1, s_3\}$ and $S(1)(2_{\text{nd}}) = \{s_2\}$, $h_5(3)(1_{\text{st}}) =$ "TGC" is divided into $h_3(2)(2_{\text{nd}}) =$ "TC" and $h_1(1)(2_{\text{nd}}) =$ "G". Therefore $u^{2,2_{\text{nd}}}_{5,3,1_{\text{st}}} = 3$ and $u^{1,2_{\text{nd}}}_{5,3,1_{\text{st}}} = 1$.

$$\begin{aligned} h_5(3)(1_{\text{st}}) &= (h_{u^{2,2_{\text{nd}}}_{5,3,1_{\text{st}}}}(2)(2_{\text{nd}}), h_{u^{1,2_{\text{nd}}}_{5,3,1_{\text{st}}}}(1)(2_{\text{nd}})) \\ &= (h_3(2)(2_{\text{nd}}), h_1(1)(2_{\text{nd}})). \end{aligned}$$

- Dummy values and division and partial haplotypes.

Dummy value of $h_k(n)(1_{\text{st}})$ is also expressed as,

$$v_k(n)(1_{\text{st}}) = \prod_{i=1}^{m} v_{u^{p_i,q_{i_{\text{th}}}}_{k,n,1_{\text{st}}}}(p_i)(q_{i_{\text{th}}})$$

Example: $v_5(3)(1_{\text{st}}) = v_3(2)(2_{\text{nd}}) \times v_1(1)(2_{\text{nd}}) = (-1) \times (1) = -1$.

SNP-BASED HETEROGENEITY TENSOR Ψ

- We define Ψ for $S(n)(1_{\text{st}})$, which is consisted of 2^n elements, each of which is a value for an element of Pow($S(n)(1_{\text{st}})$).

$$\Psi = \{\psi(i)(j_{\text{th}})\},$$

where $\psi(i)(j_{\text{th}})$ represents a value for an element, $S(i)(j_{\text{th}})$, in Pow($S(n)(1_{\text{st}})$), the j_{th} subset with i sites.

Because the elements of Ψ are arranged in the multidimensional structure with indices, i and j, we call Ψ as "SNP-based heterogeneity tensor" (tensor: a multidimensional array). [Rowland and Weissltein, 2006].

Further details of Ψ will be defined in the following sections.

SNP-BASED HETEROGENEITY TENSOR Ψ

DEFINITIONS

Here we give basic rules for $\psi(i)(j_{st})$ so that all of Ψ s are defined by a systematic way, that correspond to subsets of SNPs with various size.

- $\psi(0)(1_{st})$ is defined as 1.
- $\psi(i)(j_{st}); i>0$ is defined below:

$$\psi(i)(j_{st}) = \sum_{k=1}^{2^i} (v_k(i)(j_{st}) \times f_k(i)(j_{st})) \quad (1)$$

With these definitions, Ψ has the following features:

- When DNA sequence population is a clone, absolute value of all the elements of Ψ is 1 or -1.
- When DNA sequence population is in the limit randomness, all the elements of Ψ, except for $\psi(0)(1_{st})$, are 0.
- Otherwise $\psi(i)(j_{th})$ ranges from -1 to 1 according to the heterogeneity condition of the population.

Ψ GIVES A BASE FOR HAPLOTYPE FREQUENCY SPACE

This section gives a note on relation between haplotype frequencies, $F(n)(1_{st})$ and Ψ. Both $F(n)(1_{st})$ and Ψ have 2^n elements. They are in one-to-one correspondence. This bijective relation can be proven by showing that the determinant of the matrix, that transforms $F(n)(1_{st})$ to Ψ, is different from zero, which will recurrently be proven with Laplace expansion of determinant. (proof not shown) [Weisstein, 2006b].

The transformation from $F(n)(1_{st})$ to Ψ is expressed by

$$\psi(i)(j_{th}) = \sum_{k=1}^{2^i} (v_k(i)(j_{th}) \times f_k(i)(j_{th})), i = 0, 1, \ldots, n;$$
$$j = 1, 2, \ldots, {}_nC_i; \quad k = 1, 2, \ldots, 2^i. \quad (2)$$

The reverse transformation from Ψ back to $F(n)(1_{st})$ is expressed by

$$f_k(n)(1_{st}) = \frac{1}{2^n} \times \sum_{p=0}^{n-1} \sum_{q=1}^{{}_nC_p} v_{u^{p,q_{th}}_{k,n,1_{st}}}(p)(q_{th}) \times \psi(p)(q_{th}),$$
$$p = 0, 1, \ldots, n; \quad q = 1, 2, \ldots {}_nC_p; \quad k = 1, 2, \ldots, 2^n. \quad (3)$$

Each element of $F(n)(1_{st})$ represents the frequency of one of the 2^n distinct haplotypes. Because the sum of the 2^n elements of $F(n)(1_{st})$ is 1 and fixed, their degree of freedom is 2^n-1. Because $F(n)(1_{st})$ and Ψ are mutually in one-to-one correspondence, the degree of freedom of Ψ should be also 2^n-1. One of the 2^n elements of Ψ is 1 and constant, therefore all the other 2^n-1 elements are mutually independent. This means that the dimension of the haplotype frequency space is 2^n-1 and 2^n-1 elements of Ψ except for $\psi(0)(1_{st})$ consist a base of the space.

LINKAGE DISEQUILIBRIUM AND Ψ

INTER-SITE RANDOMNESS AND INDEPENDENCY

The inter-site randomness is defined for division patterns of a SNP set as follows. Consider a division of $S(n)(1_{st})$ into m mutually exclusive non-empty subsets $\mathrm{Div}: S(n)(1_{st}) \rightarrow \{S_1(p_1)(q_{1_{th}}), \ldots, S_m(p_m)(q_{m_{th}})\}$. In this situation, when the inter-site randomness is at its maximal conformation, the frequency of all haplotypes in $S_i(p_i)(q_{i_{th}})$ is mutually independent. In this condition, $f_k(n)(1_{st}) = \prod_{i=1}^{m} f_{u^{p_i,q_{i_{th}}}_{k,n,1_{st}}}(p_i)(q_{i_{th}})$, and $v_k(n)(1_{st}) = \prod_{i=1}^{m} v_{u^{p_i,q_{i_{th}}}_{k,n,1_{st}}}(p_i)(q_{i_{th}})$. By a simple transformation, Ψ in the maximized inter-site randomness can be expressed by

$$\begin{aligned}\psi(n)(1_{st}) &= \sum_{i=k}^{2^n} (f_k(n)(1_{st}) \times v_k(n)(1_{st})) \\ &= \sum_{k=1}^{2^n} \left(\prod_{i=1}^{m} f_{u^{p_i,q_{i_{th}}}_{k,n,1_{st}}}(p_i)(q_{i_{th}}) \times \prod_{i=1}^{m} v_{u^{p_i,q_{i_{th}}}_{k,n,1_{st}}}(p_i)(q_{i_{th}}) \right) \\ &= \prod_{i=1}^{m} \left(\sum_{j=1}^{2_i^p} (f_j(p_i)(q_{i_{th}}) \times v_j(p_i)(q_{i_{th}})) \right) \\ &= \prod_{i=1}^{m} \psi(p_i)(q_{i_{th}}). \end{aligned} \quad (4)$$

Figure 1(c)–(viii) shows an example of $n=3$ in the maximized inter-site randomness LE for all division patterns, in which the equation (4) is satisfied.

GENERALIZED LINKAGE DISEQUILIBRIUM INDEX, D_g

When alleles at the sites are associated on the same chromosome, they are called to be in LD. In LE no allelic association is present. Therefore when the equation (4), $\psi(n)(1_{st}) = \prod_{i=1}^{m} \psi(p_i)(q_{i_{th}})$ for $\mathrm{Div}, S(n)(1_{st}) \rightarrow \{S_1(p_1)\ (q_{1_{th}}), \ldots, S_m(p_m)(q_{m_{th}})\}$, is satisfied, it can be said that $S(n)(1_{st})$ is in LE for the particular division pattern Div. The deviation from the equation (4) represents the degree of LD for $S(n)(1_{st})$ with

respect to the particular Div. Therefore LE and LD are defined for ways to divide a set of sites.

We introduce generalized linkage disequilibrium index, d_g(Div) for Div as:

$$d_g(\mathrm{Div}) = \max\left(\left(1 - \frac{\psi(n)(1_{st}) + 1}{\prod_{i=1}^{m} \psi_i(p_i)(q_{i_{st}}) + 1}\right), \times \left(1 - \frac{\psi(n)(1_{st}) - 1}{\prod_{i=1}^{m} \psi_i(p_i)(q_{i_{st}}) - 1}\right)\right). \quad (5)$$

When denominator of either expression in the parenthesis of "max" is zero, the other value should be selected for d_g(Div).

By this definition, d_g(Div) satisfies:

- d_g(Div) takes a value in the range 0–1.
- D_g(Div) takes zero in LE.

For a pair of two sites and its division into two single SNPs, this is expressed by

$$D_g(\mathrm{Pair}) = \max\left(\left(1 - \frac{\psi(2)(1_{st}) + 1}{\psi(1)(1_{st})\psi(1)(2_{nd}) + 1}\right), \times \left(1 - \frac{\psi(2)(1_{st}) - 1}{\psi(1)(1_{st})\psi(1)(2_{nd}) - 1}\right)\right). \quad (6)$$

For a SNP pair, $F(2)(1_{st}) = \{f_1(2)(1_{st}), f_2(2)(1_{st}), f_3(2)(1_{st}), f_4(2)(1_{st})\}$. For simplicity, $F = \{f_1, f_2, f_3, f_4\}$ will be used hereafter. The numerator of equation (6) is expressed as $\psi(1)(1_{st})\psi(1)(2_{nd}) - \psi(2)(1_{st}) = -((f_1 - f_2 - f_3 + f_4) - ((f_1 + f_2) - (f_3 + f_4)) \times ((f_1 + f_3) - (f_2 + f_4)))$. The right-hand side of the equation is transformed into $-4 \times (f_1 \times f_4 - f_2 \times f_3)$ (Appendix 1). This expression of numerator is proportional to the numerator of other conventional LD indices including D′ and r^2 [Devlin and Risch, 1995], which indicates D_g(Pair) is an appropriate index for SNP pairs. D_g(Pair) has a standardized value of the numerator that takes 1 when $\psi(2)(1_{st}) = \pm 1$, and takes 0 in case of LE. Values of D_g(Pair) are plotted for comparison with other conventional LD indices, D′, r^2 and r in Figure 3. Let A/a and B/b denote alleles of two SNPs. In Figure 3(a), major alleles of two SNPs, A and B, are fixed at 0.8. Frequency of haplotype AB (f_1), is parameterized from 0.6 to 0.8 under the condition where frequency of haplotype Ab (f_2) equals the one of aB (f_3). In Figure 3(b), the frequency of haplotype AB (f_1) is parameterized from 0 to 0.8 under the condition where frequency of haplotype aB is fixed at zero. D_g(Pair) takes the same value with D' and r, when $f_1 \times f_4 - f_2 \times f_3$ is positive and when $f_2 = f_3$. However, when the symmetry of $f_2 = f_3$ is lost, the values of D_g(Pair), D' and r diverge. Both D_g(Pair) and r^2 are 1 when $f_1 + f_4 = 1$, and both converge to zero when $f_1 + f_3 = 1$.

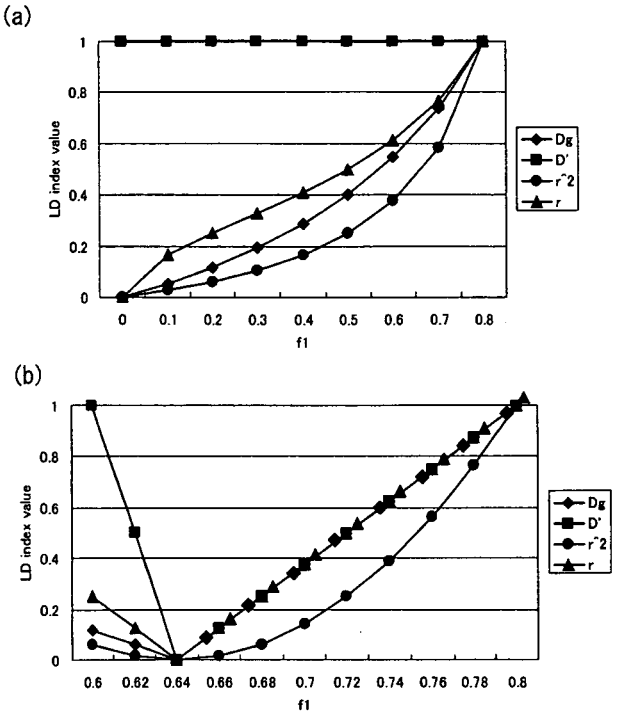


Fig. 3. Plots of D_g, D', r^2 and r. (a) SNP A and SNP B with respectively alleles A, a and B, b. Allele frequencies $P(A)$ and $P(B)$ are fixed at 0.8, and $P(AB)$ is parameterized from 0.6 to 0.8 under the condition of $P(Ab) = P(aB)$. (b) $P(AB)$ is parameterized from 0 to 0.8 under the condition of $P(aB) = 0$.

ADDITIONAL EXAMPLES

In addition to the examples presented in the section INTRODUCTORY EXAMPLES, a few more examples will be helpful.

Ψ FOR TWO SITES

Assume there are two sites, $S(2)(1_{st}) = \{s_1, s_2\}$.
The power set $\mathrm{Pow}(S(2)(1_{st})) = \{\{\phi\}, \{s_1\}, \{s_2\}, \{s_1, s_2\}\}$.

$$H(2)(1_{st}) = \{h_1(2)(1_{st}), h_2(2)(1_{st}), h_3(2)(1_{st}), h_4(2)(1_{st})\} = \{\text{"00"}, \text{"01"}, \text{"10"}, \text{"11"}\}.$$

For a subset $S(2)(1_{st}) = \{s_1, s_2\}$,

$$F(2)(1_{st}) = \{f_1(2)(1_{st}), f_2(2)(1_{st}), f_3(2)(1_{st}), f_4(2)(1_{st})\},$$
$$V(2)(1_{st}) = \{1, -1, -1, 1\},$$
$$\psi(2)(1_{st}) = f_1(2)(1_{st}) - f_2(2)(1_{st}) - f_3(2)(1_{st}) + f_4(2)(1_{st}).$$

The single site frequencies are derived by summing full haplotype frequencies across alleles at the other sites.

For a subset $S(1)(1_{st}) = \{s_1\}$,

$$F(1)(1_{st}) = \{(f_1(2)(1_{st}) + f_2(2)(1_{st})), (f_3(2)(1_{st}) + f_4(2)(1_{st}))\},$$
$$V(1)(1_{st}) = \{1, -1\},$$
$$\psi(1)(1_{st}) = (f_1(2)(1_{st}) + f_2(2)(1_{st})) - (f_3(2)(1_{st}) + f_4(2)(1_{st})).$$

For a subset $S(1)(2_{nd})=\{s_2\}$,

$$F(1)(2_{nd})=\{(f_1(2)(1_{st})+f_3(2)(1_{st})),(f_2(2)(1_{st})+f_4(2)(1_{st}))\},$$
$$V(1)(2_{nd})=\{1,-1\},$$
$$\psi(1)(2_{nd})=(f_1(2)(1_{st})+f_3(2)(1_{st}))-(f_2(2)(1_{st})+f_4(2)(1_{st})).$$

For a subset $S(0)(1_{st})$,

$$\psi(0)(1_{st})=1.$$

$$\Psi=\{\psi(0)(1_{st}),\psi(1)(1_{st}),\psi(1)(2_{nd}),\psi(2)(1_{st})\},$$

Ψ plots of these cases are shown and explained in Figure 4.

D_g FOR SIX SITES

In the section "INTRODUCTORY EXAMPLES", F_2 was shown to have LD components that are not detected by pairwise LD measures but detected by D_g. In this section, we deal with six site examples, that are consisted of two sets of three sites; $S = \{S_p, S_q\} = \{s_A, s_B, s_C, s_{A'}, s_{B'}, s_{C'}\}$. Haplotype frequencies for the former and the latter three sites are identical with F_2; $F_p = \{f_{ABC}, f_{ABc}, f_{AbC}, f_{Abc}, f_{aBC}, f_{aBc}, f_{abC}, f_{abc}\} = \{0.25,0,0,0.25,0,0.25,0.25,0\}$ and $F_q=\{f_{A'B'C'}, f_{A'B'c'}, f_{A'b'C'}, f_{A'b'c'}, f_{a'B'C'}, f_{a'B'c'}, f_{a'b'C'}, f_{a'b'c'}\} = \{0.25,0,0,0.25, 0,0.25,0.25,0\}$. In the first case of six sites, case1, the haplotypes of the former three sites and the haplotypes of the latter are in one-to-one correspondence;

$$F(\text{case1})=\{f_{ABCA'B'C'}, f_{AbcA'b'c'}, f_{aBca'B'c'}, f_{abCa'b'C'}\}$$
$$=\{0.25,0.25,0.25,0.25\}.$$

The D_g plot of case1 is shown in Figure 5(a). The black square on the bottom representing the division of all the six sites into six single sites, explains LD in the region as a whole. Four black squares in the third row from the bottom, representing site-trios, which are in LD themselves. Three black squares in the third row from the top, representing site-pairs intervened by two sites, are also in LD. In the second case, case2, each haplotype in the former site-set are evenly connected to every haplotype in the latter site-set.

$$F(\text{case2})=\{f_{ABCA'B'C'}, f_{ABCA'b'c'}, f_{ABCa'B'c'}, f_{ABCa'b'C'}, f_{AbcA'B'C'},$$
$$f_{AbcA'b'c'}, f_{Abca'B'c'}, f_{Abca'b'C'},$$
$$f_{aBcA'B'C'}, f_{aBcA'b'c'}, f_{aBca'B'c'}, f_{aBca'b'C'}, f_{abCA'B'C'},$$
$$f_{abCA'b'c'}, f_{abCa'B'c'}, f_{abCa'b'C'}\}$$
$$=\{0.0625,0.0625,0.0625,0.0625,0.0625,0.0625,$$
$$0.0625,0.0625,$$
$$0.0625,0.0625,0.0625,0.0625,0.0625,0.0625,$$
$$0.0625,0.0625\}.$$

D_g plot of this case is shown in Figure 5(b), All the pairwise d_g's are 0. The black square on the bottom representing the division of all the six sites into six single sites, explains LD in the region as a whole. Two

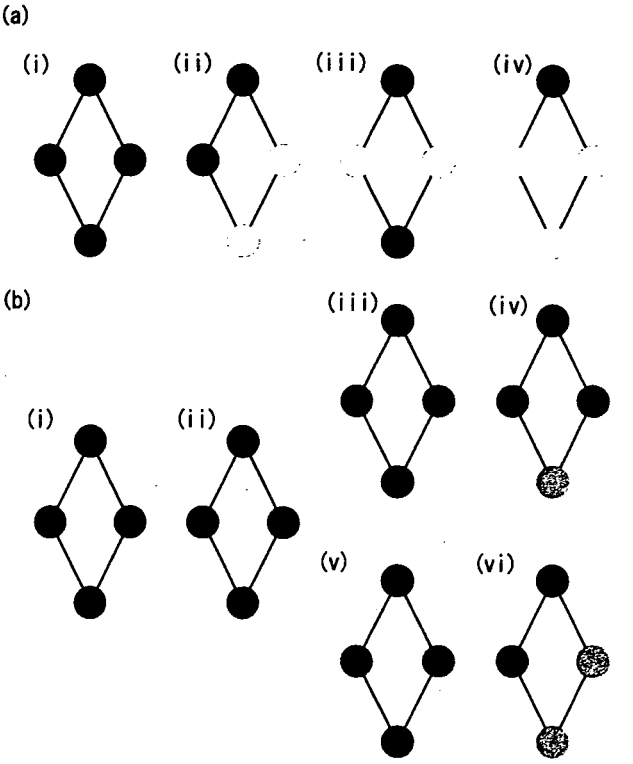


Fig. 4. ψ plots for two sites. (a) Four patterns of ψ where two sites are monomorphic or their allele frequency is 0.5. (a)–(i) $F(2)(1_{st}) = \{1,0,0,0\}$ (a clone). $\Psi = \{1,1,1,1\}$. (a)–(ii) One site is monomorphic and the other site is polymorphic and its allele frequency is 0.5. $F(2)(1_{st}) = \{0.5,0.5,0,0\}$. $\Psi = \{1,1,0,0\}$. (a)–(iii) and (a)–(iv) Allele frequencies of both sites are 0.5. (a)–(iii) Absolute LD. $F(2)(1_{st}) = \{0.5,0,0,0.5\}$ and $\Psi = \{1,0,0,1\}$. (a)–(iv) LE. $F(2)(1_{st}) = \{0.25,0.25,0.25,0.25\}$ and $\Psi = \{1,0,0,0\}$. The black-and-white circles distinguish the four patterns. The absolute LD was indicated by $\psi(2)(1_{st}) = 1$ and LE was by $\psi(2)(1_{st}) = 0$. (b) shows patterns of Ψ where allele frequency of two sites are not necessarily 0.5. (b)–(i) is an example of a clone. When one site is polymorphic and its allele frequency is not 0.5, ψ plot appears like (b)–(ii) $(F(2)(1_{st}) = \{0.6,0.4,0,0\}$. $\Psi = \{1,1,0.2,0.2\}$. When both sites are polymorphic and their allele frequencies are the same but not 0.5, they are in the absolute LD and its ψ plot is (b)–(iii) $(F(2)(1_{st}) = \{0.6,0,0,0.4\}$. $\Psi = \{1,0.2,0.2,1\}$. (b)–(iv) is a plot when two sites are in LE and allele frequencies of both sites are the same $(F(2)(1_{st} = \{0.36,0.24,0.24,0.16\}$, $\Psi = \{1,0.2,0.2,0\})$. (b)–(v) represents two sites having different allele frequencies and being in LD with $D' = 1$ but their $r^2 \neq 1$. $(F(2)(1_{st}) = \{0.6,0.2,0,0.2\}$, $\Psi = \{1,0.6,0.2,0.6\}$. (b)–(vi) is a plot when two sites have different allele frequencies and they are in LE. $F(2)(1_{st}) = \{0.48,0.32,0.12,0.08\}$, $\Psi = \{1,0.6,0.2,0.12\}$.

black squares for $(s_A,s_B,s_C) \rightarrow \{(s_A),(s_B),(s_C)\}$ and $(s_{A'},s_{B'}, s_{C'}) \rightarrow \{(s_{A'}),(s_{B'}),(s_{C'})\}$ stands for LD at the trio level.

When the latter three sites are monomorphic for three of four haplotypes in the former set (case3),

$$F(\text{case3})=\{f_{ABCA'B'C'}, f_{ABCA'b'c'}, f_{ABCa'B'c'}, f_{ABCa'b'C'}, f_{AbcA'b'c'},$$
$$f_{aBcA'B'C'}, f_{abCA'B'C'}, f_{abCA'b'c'}\}$$
$$=\{0.0625,0.0625,0.0625,0.0625,0.25,0.25,0.25\}.$$

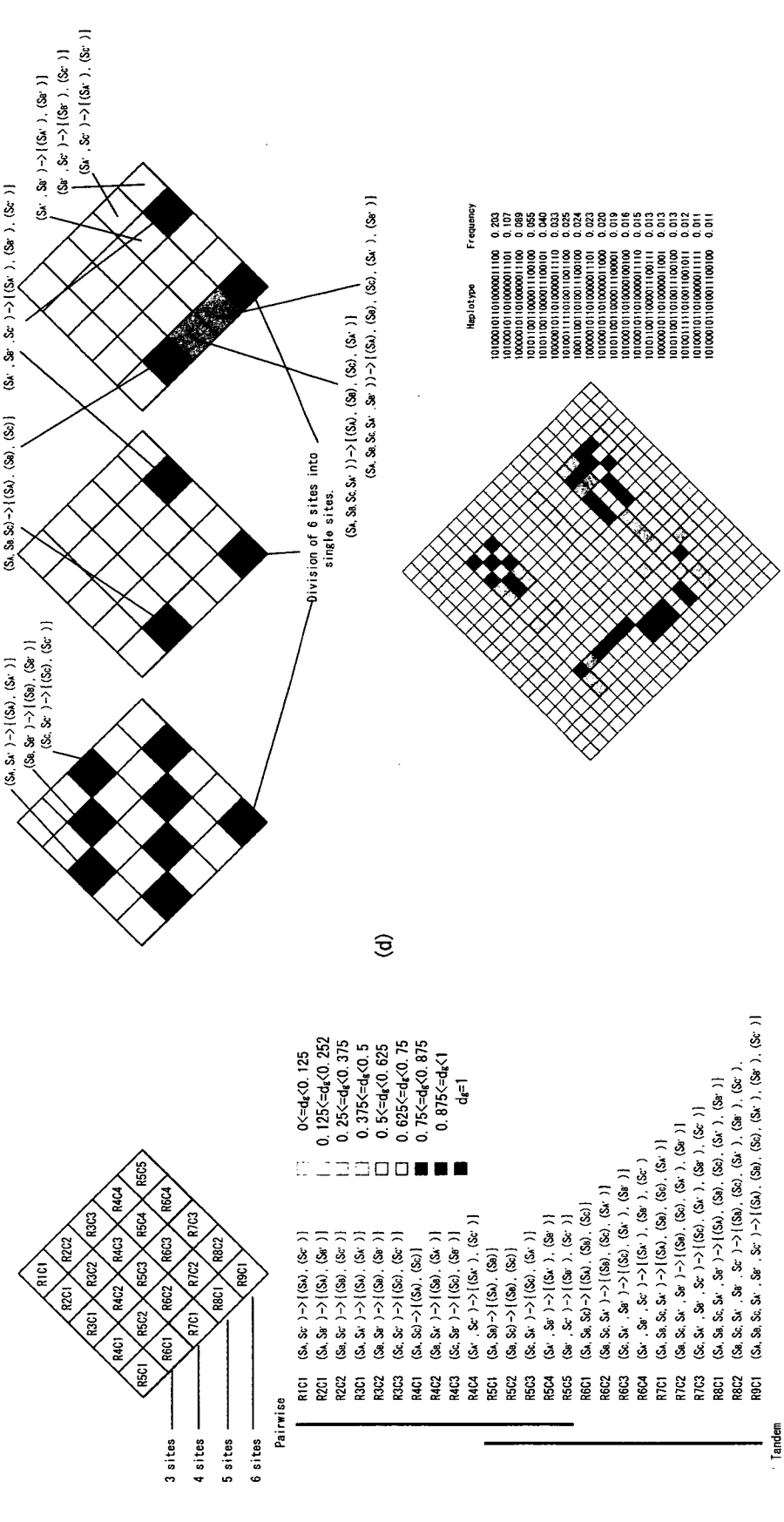


Fig. 5. D_g plots for examples of six sites. The upper halves are d_g's for site-pairs and the lower halves are for sites in tandem. All the three cases are in strong LD, but the components of LD are different each other. (a) Case1. d_g's for three site-pairs separated by two sites, four site-trios in tandem and the whole six site-set indicate strong LD. (b) Case2. All the site-pairs are in LE. d_g's for the two trio-sites on the left and right indicate LD. d_g for division of six sites into single sites also indicates LD. (c) Case3. This plot is similar to (b) with additional colored squares for site-pairs in the right-most three-site segment and for divisions of four tandem sites or five tandem sites containing all the three sites in the left-most. See text for haplotype frequencies of the three examples. (d) D_g plot for 22 site region. The upper pairwise triangle displayed extension of LD in the region, and the lower tandem triangle indicated LD components that were not captured bythe pairwise d_g's.

D_g plot of this case is shown in Figure 5(c). Pairwise d_g's are weakly present for the site-pairs in the latter three sites. The square for the division of all the six sites into six single sites indicates strong LD in this region overall and the square for the division of the latter three sites into single sites also indicates the presence of LD.

D_g PLOTS FOR REAL DATA

A region with 22 sites was chosen from HapMap project data set [The International HapMap Consortium, 2005] and haplotype frequency was estimated with fastPhase, one of the popular haplotype inferrence applications for large scale data [Scheet and Stephens, 2006]. Nineteen haplotypes were inferred and their D_g plot was shown in Figure 5(d), that displayed that the pairwise triangle and the tandem triangle captured LD components of the region differently.

USAGE OF Ψ FOR HAPLOTYPE FREQUENCY INFERENCE

LIKELIHOOD FUNCTION OF GENOTYPE DATA FOR SNP PAIRS IS EXPRESSED AS A MONOVARIATE FUNCTION OF Ψ AND THE HAPLOTYPE FREQUENCY IS OBTAINED BY SOLVING THE DERIVATIVES OF UNIVARIATE FUNCTION.

Although Ψ is calculable when frequency of all haplotypes are given, the majority of LD mapping studies are based on unphased genotype data of SNPs, where the haplotype frequency has to be inferred. As described in the section "Ψ Gives a Base for Haplotype Frequency Space", $F(n)(1_{st})$ and Ψ are in one-to-one correspondence. Therefore the inference of $F(n)(1_{st})$ is equivalent to the inference of Ψ.

Consider haplotype frequency inference from unphased genotype data of a SNP pair. For two SNPs, the four haplotype frequencies are expressed with Ψ as:

$$\begin{aligned} f_1 &= \tfrac{1}{4}(\psi(2)(1_{st}) + \psi(1)(1_{st}) + \psi(1)(2_{nd}) + \psi(0)(1_{st})), \\ f_2 &= \tfrac{1}{4}(-\psi(2)(1_{st}) + \psi(1)(1_{st}) - \psi(1)(2_{nd}) + \psi(0)(1_{st})), \\ f_3 &= \tfrac{1}{4}(-\psi(2)(1_{st}) - \psi(1)(1_{st}) + \psi(1)(2_{nd}) + \psi(0)(1_{st})), \\ f_4 &= \tfrac{1}{4}(\psi(2)(1_{st}) - \psi(1)(1_{st}) - \psi(1)(2_{nd}) + \psi(0)(1_{st})). \end{aligned} \quad (7)$$

$\ln(L)$, logarithm of likelihood function to obtain a unphased genotype data is expressed as a function of f_1;

$$\ln(L) = G_1 \log(f_1) + G_2 \log(f_2) + G_3 \log(f_3) + G_4 \log(f_4) + G_5 \log(f_1 f_4 + f_2 f_3) + C,$$

where $G_i (i = 1,\ldots,4)$ represents the number of chromosomes that are deterministically known from unphased genotype data, and G_5 is the number of double heterozygotes, and C is a constant.

The EM algorithm attempts to maximize L by handling f_1, f_2, f_3 and f_4 as variables where $f_1 + f_2$ and $f_1 + f_3$ are fixed at the value given by method of moments. Because f_i is expressed with Ψ, $\ln(L)$ is also a function of Ψ. Although Ψ for SNP pairs has four elements, $\psi(0)(1_{st})$ is always constant and value of $\psi(1)(1_{st})$ and $\psi(1)(2_{nd})$ are known under the condition where $f_1 + f_2$ and $f_1 + f_3$ are given by the method of moments ($\psi(1)(1) = (f_1 + f_2) - (f_3 + f_4)$ and $\psi(1)(2_{nd}) = (f_1 + f_3) - (f_2 + f_4)$). Therefore the equations (6) are transformed to:

$$\begin{aligned} f_1 &= \tfrac{1}{4}(\psi(2)(1_{st}) + c_1), \\ f_2 &= \tfrac{1}{4}(-\psi(2)(1_{st}) - c_2), \\ f_3 &= \tfrac{1}{4}(-\psi(2)(1_{st}) - c_3), \\ f_4 &= \tfrac{1}{4}(\psi(2)(1_{st}) + c_4). \end{aligned} \quad (8)$$

where c_i denotes constant terms of frequency with appropriate signs.

It is shown that $\ln(L)$ is expressed as a monovariate function of $\psi(2)(1_{st})$. $\ln(L)$ is defined for the finite range of $\psi(2)(1_{st})$, where $0 \le f_i \le 1$, and the function is continuous and differentiable in the range. Therefore the global maximum can be obtained by solving its derivatives with conventional searching methods.

Equation transformations and its Newton-Raphson estimation of the derivatives are described in Appendix 2.

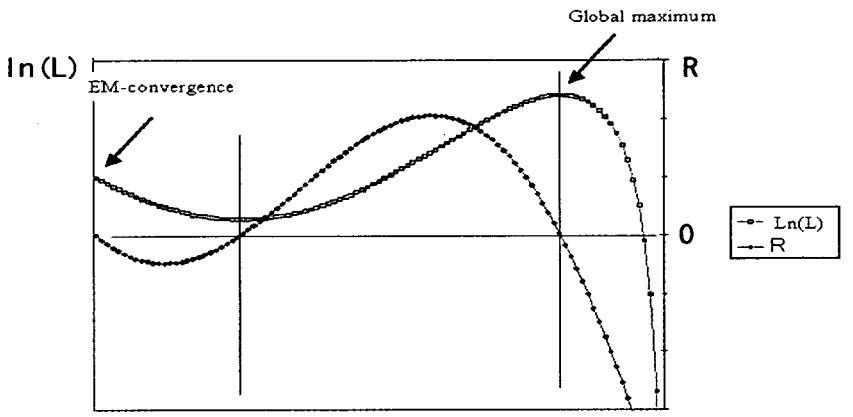
In the case of $n = 2$, maximum likelihood estimates of Ψ was obtained by solving a univariate likelihood function, as above. Similarly, when Ψ is solved for all subsets of $S(n)(1_{st})$ except for $S(n)(1_{st})$ itself where all the elements of Ψ for $S(n)(1_{st})$ but $\psi(n)(1_{st})$ are given, the likelihood function can be expressed as a univariate function of $\psi(n)(1_{st})$. Appendix 3 gives this generalization of likelihood function expressed as a univariate functionof $\psi(n)(1_{st})$ $(n = 1,2,\ldots)$.

COMPARISON WITH THE EM ALGORITHM

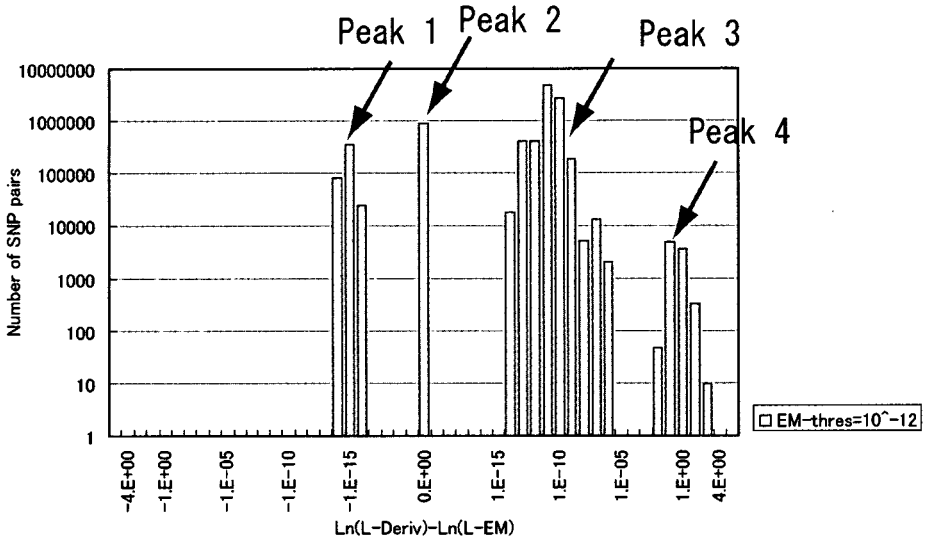
The EM algorithm is known to give reliable estimates of haplotype frequencies of SNP pairs in the majority of cases, but is susceptible to convergence to a local maximum [Nin, 2004]. Figure 6a shows an example of convergence to a local maximum of $\ln(L)$ for a SNP pair from the HapMap Project. We evaluated how frequently the standard EM algorithm converges to a local minimum but not to the global maximum using HapMap Project data [The International HapMap Consortum, 2005].

Ψ-based method and the EM algorithm were applied to 10 million SNP pairs of chromosome 10 within a 250 kb window with 45 unrelated Japanese of the HapMap project. The average number of iterations of EM method was 22.2, and the average number of iterations to solve five derivatives in Ψ-based method was 126.1. The results of the Ψ-based method indicated that 39.8% of the pairs did not have

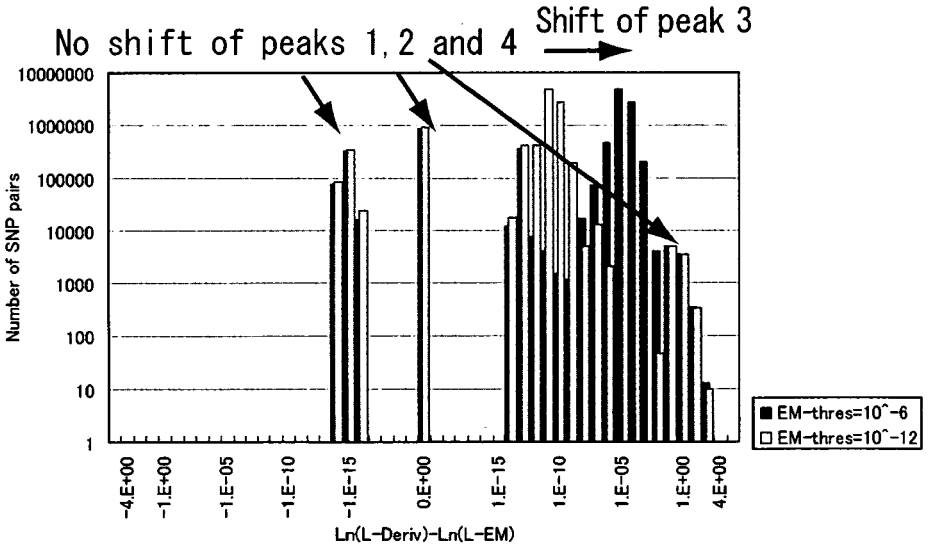
(a)


Global maximum
ln(L)
EM-convergence
R
0
Ln(L)
R

(b)


Peak 1
Peak 2
Peak 3
Peak 4
Number of SNP pairs
10000000
1000000
100000
10000
1000
100
10
1
-4.E+00
-1.E+00
-1.E-05
-1.E-10
-1.E-15
0.E+00
1.E-15
1.E-10
1.E-05
1.E+00
4.E+00
Ln(L-Deriv)-Ln(L-EM)
EM-thres=10^-12

(c)


No shift of peaks 1, 2 and 4
Shift of peak 3
Number of SNP pairs
10000000
1000000
100000
10000
1000
100
10
1
-4.E+00
-1.E+00
-1.E-05
-1.E-10
-1.E-15
0.E+00
1.E-15
1.E-10
1.E-05
1.E+00
4.E+00
Ln(L-Deriv)-Ln(L-EM)
EM-thres=10^-6
EM-thres=10^-12

local extrema, while 61.1% of pairs had a single local extreme in the search range and 0.025% had multiple local extrema. Among the pairs with one local extreme, 81.5% of them was a local maximum, and the remainder was a local minimum. Difference of ln(L) between inferences of the two methods was shown in Fig. 6(b). Peak 1 (Fig. 6(b)) represented 4.5% of SNP pairs for which the EM algorithm gave slightly higher likelihood. The EM algorithm gave better inference due to luck to start at the best value for the majority of SNP pairs in the peak 1. Peak 2 (Fig. 6(b)) represented 9.3% of pairs and two methods gave almost identical results. Peaks 3 and 4 (Fig. 6(b)) represented 86.1% of pairs for which the Ψ-based method gave slightly better result. When we allowed the EM algorithm method to stop earlier with looser convergence threshold, peaks 1, 2 and 4 did not change but a part of peak 3 shifted to right (Fig. 6(c)). This change indicated that the EM algorithm method could give better estimate for a part of SNP pairs in peak 3 by modifying its parameters but that the EM algorithm method converged to a local maximum for SNP pairs in peak 4. However the peak 4 represented only 0.09% of total SNP pairs. More detailed characterization of SNP pairs for which the EM algorithm method converged to a local maximum were described in the Appendix 3. Conditions of inference of the standard EM algorithm and the Ψ-based algorithm are available in the Appendix 5.

DISCUSSION

In this paper, a novel tensor Ψ was introduced to quantitate genetic heterogeneity with SNPs in populations. The Ψ was consisted of 2^n elements for a sequence with n sites that were mutually transformable with 2^n values of haplotype frequency. Actually 2^n-1 non-constant variables in Ψ were the base of the haplotype frequency space with 2^n-1 dimensions. Each element of Ψ represented one of subsets of n sites and they were arranged in a structure of tensor and gave information on two types of randomness of the population, the allele frequency randomness and the inter-site randomness. As an example of utility of Ψ, we proposed a generalized LD index, D_g(Pair), between two SNPs was formulated using the elements of Ψ, and its basic feature was compared with D' and r^2. Moreover LD index for a set of multiple sites more than two, D_g(Div) was also defined as a natural extension of D_g(Pair). The components of D_g for SNP pairs were drawn in the pairwise triangle and the representative components of D_g for multiple sites were drawn in the tandem triangle. For another practical purpose, Ψ offered the absolute maximum haplotype frequency inference for SNP pairs with tolerable increase of computational burden and it overcomes the problem to converge to a local maximum by the EM algorithm method. Application of the Ψ-based haplotype inference algorithm to larger SNP sets seemed possible but modifications to limit computational burdens would be necessary.

Because populational DNA sequence heterogeneity is a product of many genetic events over years and Ψ carry complete information on heterogeneity of individual sites and inter-site dependency for any combinations of sites in the region, it is necessarily complex. In order to describe the complexity, Ψ has almost fully simplified formula. (i) It uses minimum number of variables (2^n for sequence of length n). (ii) All the variables are recurrently defined so that each element represents a subset of the set of n sites. (iii) The variables are arranged in a structure based on their mutual relations (tensor structure). Although it seems still difficult to use all the information included in Ψ in order to untangle genetic heterogeneity of species, Ψ would contribute to formulate and understand interspecies genetic heterogeneity.

ACKNOWLEDGMENTS

The authors thank all the contributors to the HapMap Project, and the members, Particularly Dr. Alexandre Vasilescu, of the Center for Genomic Medicine, Graduate School of Medicine, Kyoto University and the SNP Research Center, RIKEN for valuable discussion. This work was supported in part by the CREST program (JST), Research on Measures for Intractable Diseases and Research on Human Genome and Tissue Engineering (Ministry of Health, Labour and Welfare, Japan) and BioBankJapan project.

ELECTRONIC DATABASE INFORMATION

See also HapMap: http://www.hapmap.org/index.html; Program sources and tools to calculate D_g are available, http://www.genome.med.Kyoto-u.ac.jp/ra/statgenet/index_en.html

Fig. 6. Comparison of Ψ based-haplotype inference method and EM methods. (a) Plots of ln(L) and R for an SNP pair from the HapMap Project (See Appendix 2 for the definition of R. The pair has a genotype distribution of 10, 11, 6, 1, 12, 0, 0, 0, 0, for AABB, AABb,..., aabb, and the estimated global maximum of haplotype frequency is $h(AA) = 0.547$, $h(AB) = 0.291$, $h(aB) = 0.053$, $h(ab) = 0.109$, $D' = 0.45$. The standard EM method converges to $h(AB) = 0.438$, $h(Ab) = 0.400$, $h(aB) = 0.162$, $h(ab) = 0.00$, $D' = 1.00$. Vertical lines denote $R = 0$, at which ln(L) takes a local minimum and local maximum. (b) Distribution of difference in ln(L) between the two methods for 10^6 SNP pairs from the HapMap Project. The convergence threshold for the EM method is 10^{-12}. (c) Comparison of EM convergence thresholds (10^{-6} and 10^{-12}).

REFERENCES

Aquadro CG, Begun DJ, Kndahl EC. 1994. Selection, recombination and DNA polymorphism in Drosophila. In: Golding B, editors. Non-Neutral Evolution: Theories and Molecular Data. New York: Chapman & Hall.

Collins FS, Green ED, Guttmacher AE, Guyer MS. 2003. A vision for the future of genomics research. Nature 422:835–847.

Devlin B, Risch N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311–322.

Hartl DL, Clark AG. 1997a. Principles of Population Genetics, 3rd edition. MA: Sinauer Associates, Inc. p 294–296.

Hartl DL, Clark AG. 1997b. Principles of Population Genetics, 3rd edition. MA: Sinauer Associates, Inc. p 95–106.

Hartl DL, Clark AG. 1997c. Principles of Population Genetics, 3rd edition. MA: Sinauer Associates, Inc. p 57–61.

Kidd KK, Pakstis AJ, Speed WC, Kill JR. 2004. Understanding human DNA sequence variation. J Hered 95:406–420.

Morton NE. 2005. Linkage disequilibrium maps and association mapping. J Clin Invest 115:1425–1430.

Navarro A, Barton NH. 2003. Accumulating postzygotic isolation genes in parapartry: a new twist on chromosomal speciation. Evolution 57:447–459.

Niu T. 2004. Algorithms for inferring haplotypes. Genet Epidemiol 27:334–347.

Noor MA, Grams KL, Bertucci LA, Reiland J. 2001. Chromosomal inversions and the reproductive isolation of species. Proc Natl Acad Sci 98:12084–12088.

Nothnagel R, Furst R, Rohde K. 2002. Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. Hum Hered 54:186–198.

Rieseberg LH. 2001. Chromosomal rearrangements and speciation. Trends Ecol Evol 98:12084–12088.

Rieseberg LH, Livingstone K. 2003. Evolution. Chromosomal speciation in primates. Science 300:267–268.

Rowland T, Weisstein EW. 2006. Tensor. From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/Tensor.html.

Scheet P, Stephens M. 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.

The International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437:1299–1320.

Weisstein EW. 2006a. Power Set. From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/PowerSet.html.

Weisstein EW. 2006b. Determinant Expansion by Minors. From MathWorld—A Wolfram Web Resource. http://mathworld.wolfram.com/DeterminantExpansionbyMinors.html.

Zapata C. 2000. The D′ measure of overall gametic disequilibrium between pairs of multiallelic lici. Evolution 54:1809–1812.

APPENDIX

APPENDIX 1

Equivalence of D_g^{Pair} to conventional pair-wise linkage disequilibrium.

$$
\begin{aligned}
&(f_1-f_2-f_3+f_4)-((f_1+f_2)-(f_3+f_4))\times((f_1+f_3)-(f_2+f_4))\\
&=(f_1-f_2-f_3+(1-f_1-f_2-f_3))-(f_1+f_2-f_3\\
&\quad-(1-f_1-f_2-f_3))\times(f_1-f_2+f_3-(1-f_1-f_2-f_3))\\
&=(1-2(f_2+f_3))-(2(f_1+f_2)-1)\times(2(f_1+f_3)-1)\\
&=(1-2(f_2+f_3))-4(f_1+f_2)\times(f_1+f_3)+2((f_1+f_2)\\
&\quad+(f_1+f_3))-1\\
&=4(f_1-(f_1+f_2)\times(f_1+f_3))\\
&=4(f_1-f_1\times(f_1+f_2+f_3)-f_2\times f_3)\\
&=4(f_1\times(1-f_1-f_2-f_3)-f_2\times f_3)\\
&=4(f_1\times f_4-f_2\times f_3).
\end{aligned}
$$

APPENDIX 2

Monovariate likelihood function expressed as a function of $\psi(2)(1_{st})$ and its maximal likelihood estimation.

$$
\begin{aligned}
f_1&=\tfrac{1}{4}(\psi(2)(1_{\mathrm{st}})+c_1),\\
f_2&=\tfrac{1}{4}(-\psi(2)(1_{\mathrm{st}})-c_2),\\
f_3&=\tfrac{1}{4}(-\psi(2)(1_{\mathrm{st}})-c_3),\\
f_4&=\tfrac{1}{4}(\psi(2)(1_{\mathrm{st}})+c_4),
\end{aligned}
\qquad (A7)
$$

where c_i denote constant terms of frequency with appropriate signs.

Because

$$\frac{\mathrm{d}f_i}{\mathrm{d}\psi(2)(1_{\mathrm{st}})}=\frac{1}{4},\quad \text{for } i=1,4,$$

$$\frac{\mathrm{d}f_i}{\mathrm{d}\psi(2)(1_{\mathrm{st}})}=-\frac{1}{4},\quad \text{for } i=2,3,$$

and

$$\frac{\mathrm{d}}{\mathrm{d}\psi(2)(1_{\mathrm{st}})}(f_1f_4+f_2f_3)=\psi(2)(1_{\mathrm{st}})$$

from the equations (7), we have

$$
\begin{aligned}
&\frac{\mathrm{d}\ln(L)}{\mathrm{d}(\psi(2)(1_{\mathrm{st}}))}(\psi(2)(1_{\mathrm{st}}))\\
&=\frac{1}{4}\left(\frac{G_1}{f_1}-\frac{G_2}{f_2}-\frac{G_3}{f_3}+\frac{G_4}{f_4}+\frac{G_5}{f_1f_4+f_2f_3}(\psi(2)(1_{\mathrm{st}}))\right).
\end{aligned}
$$

The global maximum of $\ln(L)$ is given by $\psi(2)(1_{\mathrm{st}})$ among the solutions of $[\mathrm{d}\ln(L)/\mathrm{d}(\psi(2)(1_{\mathrm{st}}))](\psi(2)(1_{\mathrm{st}}))=0$ in the defined range of $\psi(2)(1_{\mathrm{st}})$ or the two endpoints of the range. Because $\ln(L(\psi(2)(1_{\mathrm{st}}))$ and $[\mathrm{d}\ln(L)/\mathrm{d}(\psi(2)(1_{\mathrm{st}}))](\psi(2)(1_{\mathrm{st}}))$ are both continuous in the defined range where $0\le f_i\le 1$, a conventional searching algorithm gives the estimate of $\psi(2)(1_{\mathrm{st}})$ corresponding to the global maximum of $\ln(L(\psi(2)(1_{\mathrm{st}}))$. The followings are the steps to solve the derivative.

Let $\ln(L(\psi(2)(1_{\mathrm{st}}))=0$ take the form of $\ln(L(\psi(2)(1_{\mathrm{st}})))=\frac{R(\psi(2)(1_{\mathrm{st}}))}{T(\psi(2)(1_{\mathrm{st}}))}=0$, so that all the solutions

of $\ln(L(\psi(2)(1_{st}))) = 0$ are included in the solutions of $R(\psi(2)(1_{st})) = 0$.

$$R(\psi(2)(1_{st})) = (G_1f_2f_3f_4 - G_2f_1f_3f_4 - G_3f_1f_2f_4 + G_4f_1f_2f_3)(f_1f_4 + f_2f_3) + (\psi(2)(1_{st}))G_5f_1f_2f_3f_4 = 0.$$

Now solutions of $R(\psi(2)(1_{st}))$ cover all candidate values of $\psi(2)(1_{st})$ as the global maximum of $\ln(L(\psi(2)(1_{st}))$. Then, R can be re-expressed as:

$$\begin{aligned}
&R(\psi(2)(1_{st}))\\
&= \left(\tfrac{1}{4}\right)^5((G_1(\psi(2)(1_{st}) + c_2)(\psi(2)(1_{st}) + c_3)(\psi(2)(1_{st}) + c_4)\\
&+ G_2(\psi(2)(1_{st}) + c_1)(\psi(2)(1_{st}) + c_3)(\psi(2)(1_{st})c_4)\\
&+ G_3(\psi(2)(1_{st}) + c_1)(\psi(2)(1_{st}) + c_2)(\psi(2)(1_{st}) + c_4)\\
&+ G_4(\psi(2)(1_{st}) + c_1)(\psi(2)(1_{st}) + c_2)(\psi(2)(1_{st}) + c_3))\\
&\times ((\psi(2)(1_{st}) + c_1)(\psi(2)(1_{st}) + c_4) + (\psi(2)(1_{st}) + c_2)\\
&\times (\psi(2)(1_{st}) + c_3))\\
&+ 4G_5\psi(2)(1_{st})(\psi(2)(1_{st}) + c_1)(\psi(2)(1_{st}) + c_2)(\psi(2)(1_{st})\\
&+ c_3)(\psi(2)(1_{st}) + c_4)) = 0
\end{aligned}$$

$R(\Psi(2)(1_{st}))$ is a fifth-order polynomial equation, and its first through fifth derivative equations are obtained by regular transformation. Actually the fifth derivative is given as

$$\begin{aligned}
&\frac{d^5}{(d\psi(2)(1_{st}))^5}R(\psi(2)(1_{st}))\\
&= \left(\frac{1}{4}\right)^5 \times 5 \times 4 \times 3 \times 2 \times (2G_1 + 2G_2 + 2G_3 + 2G_4 + 4G_5)\\
&= \left(\frac{1}{4}\right)^5 \times 240 \times N_{\text{chromosomes}},
\end{aligned}$$

where $N_{\text{chromosomes}}$ stands for number of chromosomes in the genotype data. $[d^4/(d\psi(2)(1_{st}))^4]$ $R(\psi(2)(1_{st})) = 0$ is a first-order function and it is solved arithmetically. Thereafter solutions of $[d^3/(d\psi(2)(1_{st}))^3]R(\psi(2)(1_{st})) = 0$, $[d^2/(d\psi(2)(1_{st}))^2]R(\psi(2)(1_{st})) = 0$, $[d/(d\psi(2)(1_{st}))]R(\psi(2)(1_{st})) = 0$ and $R(\psi(2)(1_{st})) = 0$ are obtained using the Newton-Raphson method. The value of $\ln(L)\psi(2)(1_{st})$ for all local maxima and the two endpoints are then calculated and the absolute maximum is determined.

APPENDIX 3

Generalization of likelihood function expressed as a function of $\psi(n)(1)$.

Assume n SNPs that construct $\Gamma = \{\gamma_i\}$ composite genotypes. α_i individuals are observed to have a genotype γ_i. Further, assume γ_i has n_i heterozygous sites, and let $\Theta(\gamma_i) = \{(\theta_1, \hat{\theta}_1), (\theta_2, \hat{\theta}_2), \cdots, (\theta_{nc_i}, \hat{\theta}_{nc_i})\}$ denote the set of potential haplotype pairs for γ_i, where nc_i is the number of haplotype pairs for γ_i: ($nc_i = 1$ when $n_i = 0$, and $nc_i = 2^{(n_i-1)}$) otherwise. The $\ln(L)$ for the observed genotype data is expressed as

$$\ln(L) = \sum_{\gamma_i \in \Gamma} \alpha_i \times \ln\left(\sum_{j=1}^{nc_i} (f(\theta_j)f(\hat{\theta}_j))\right) + C \quad (*)$$

where $f(\theta_i)$ denotes frequency of θ_i. When all Ψ's except for $\psi(n)(1)$ are solved, $\psi(n)(1)$ is the only unsolved variable in Ψ. Therefore all $f(\theta_j)$ and $f(\hat{\theta}_j)$ are expressed as a univariate function of $\psi^{(n)}$ and equation (*) is also a univariate function of $\psi(n)(1)$ and differentiable as follows:

$$\frac{d}{d(\psi(n)(1))}\ln(L) = \sum_{\gamma_i \in \Gamma} \alpha_i \times \frac{\frac{d}{d(\psi(n)(1))}\left(\sum_{j=1}^{nc_i} (f(\theta_j)f(\hat{\theta}_j))\right)}{\sum_{j=1}^{nc_i} (f(\theta_j)f(\hat{\theta}_j))}.$$

Denote the subset of n_i SNPs that are heterozygous in genotype γ_i by $S^{(n_i)}_{\text{hetero}}(\gamma_i)$, and let $P(S^{(n_i)}_{\text{hetero}}(\gamma_i))$ be its power set and let $S(p_i)(q_i)(S^{(n_i)}_{\text{hetero}}(\gamma_i))$ be an element of $P(S^{(n_i)}_{\text{hetero}}(\gamma_i))$. Because $[d/d(\psi(n)(1))]f(\theta_j) = \pm(1/2^n)$, numerator of an element in (*), $[d/d(\psi(n)(1))]\left(\sum_{j=1}^{nc_i} (f(\theta_j)f(\hat{\theta}_j))\right)$, can be expressed as

$$\begin{aligned}
&d/d(\psi(n)(1))\left(\sum_{j=1}^{nc_i} (f(\theta_j)f(\hat{\theta}_j))\right) = \frac{1}{2^n} \times 2^{(nc_i+1)}\\
&\sum_{S(p_i)(q_i)(S^{(n_i)}_{\text{hetero}}(\gamma_i)) \in P(S^{(n_i)}_{\text{hetero}}(g_i)), u \neq S^{(n_i)}_{\text{hetero}}(g_i)} v(p_i)(q_i) \times \psi(p_i)(q_i),
\end{aligned}$$

where $\psi(u)$ denotes Ψ for a subset u and $v(u)$ is the value of corresponding haplotype.

APPENDIX 4

Classification of SNP pairs for which the EM algorithm did not direct toward the global maximum.

The SNP pairs that were not affected by the tightening of the threshold can be grouped into four categories (Patterns 1–4). The SNP pairs of Pattern 1 (85.0% of unaffected pairs) had a symmetric distribution of deterministic chromosomes for only two haplotypes with double heterozygotes. Such pairs exhibited two global maximum estimates at the two ends of the range of $\psi(2)(1_{st})$. As the EM algorithm started from the symmetric haplotype frequency in LE, the solution did not move from the LE condition due to this symmetry. For pairs in Pattern 2 (10.7%), the EM algorithm converged to $D' = 1$, whereas the $D' \neq 1$ condition gave the global maximum. Pairs in Pattern 3 (4.1%) were the opposite case, where the EM method converged to $D' \neq 1$ and the Ψ-based method converged to $D' = 1$. Pairs in Pattern 4 (0.28%) had multiple local maxima and the EM converged to a local maximum that was not the global maximum.

APPENDIX 5

Settings of programs to perform the standard EM algorithm and the Ψ-based algorithm.

For the standard EM, the maximum number of iterations was set at 10^9, and the calculation was stopped when the difference in $\log_{10}L$ between iterations became less than 10^{-12}. Without limitation on the maximum number of iterations, calculation did not end due to the slowness of convergence for some cases. For the Ψ-based method, no limitation was applied on the maximum number of iterations, and the iteration was stopped only when the difference in estimated x between iterations became less than 10^{-6}. Convergence of the Newton-Raphson method was fast in this case and it was unnecessary to set a limitation on the maximum number of iterations for the Ψ-based method.