Linkage disequilibrium.

We consider first genetic markers of two alleles in diploid case. Then the allelic marker has two levels "+" and "−", and the genotype marker has three levels "++", "+−" and "−−".

1: Allelic linkage disequilibrium (LDE) is based on the alele frequencies p_+ and p_-. For any two fixed locuses denote p_{xy} is the frequency of the variant x at the first locus and variant y at the second one. In these notations we have the following frequency table:

Table 1.

Locus 1\Locus 2	"+"	"−"	Total
"+"	p_{++}	p_{+-}	$p_{+\bullet}$
"−"	p_{-+}	p_{--}	$p_{-\bullet}$
Total	$p_{\bullet+}$	$p_{\bullet-}$	1

Introduce the coefficient $D = p_{++} - p_{\bullet+}p_{+\bullet}$, which is the covariance between the two random variables taking values 0 under + variant and 1 under − variant (or vice versa).

There are two ways to determine the allelic LDE coefficients:

$$r = \frac{D}{\sqrt{p_{+\bullet}(1-p_{+\bullet})p_{\bullet+}(1-p_{\bullet+)}}}$$

and r^2 is commonly used, and

$$D' = |D|/D_{\max},$$

where

$$D_{\max} = \begin{cases} \min(p_{+\bullet}p_{\bullet+}, p_{-\bullet}p_{\bullet-}), \text{ under } D < 0 \\ \min(p_{+\bullet}p_{\bullet-}, p_{-\bullet}p_{\bullet+}), \text{ under } D > 0. \end{cases}$$

We are using the signed version of the D' coefficient

$$D'_s = D/D_{\max}$$

(or the r coefficient) to measure allelic LDE in two locuses.

2: Composite LDE. In the diploid organisms all the genome variants in two locuses (haplotypes) are collected in the following table:

Table 2.

Locus 1\Locus 2	"++"	"+−"	"−+"	"−−"	Total
"++"	p_{++++}	p°_{+++-}	p°_{++-+}	p_{++--}	$p_{++\bullet}$
"+−"	p°_{+-++}	p°_{+-+-}	p°_{+--+}	p°_{+---}	$p_{+-\bullet}$
"−+"	p°_{-+++}	p°_{-++-}	p°_{-+-+}	p°_{-+--}	$p_{-+\bullet}$
"−−"	p_{--++}	p°_{--+-}	p°_{---+}	p_{----}	$p_{--\bullet}$
Total	$p_{\bullet++}$	$p_{\bullet+-}$	$p_{\bullet-+}$	$p_{\bullet--}$	1

After unification cells with identical combination we obtain 10-cells table:

Table 3.

Locus 1\Locus 2	"++"	"+−"	"−+"	"−−"	Total
"++"	p_{++++}	p_{+++-}		p_{++--}	$p_{++\bullet}$
"+−"	p_{+-++}	p_{+-+-}	p_{+--+}	p_{+---}	$p_{+-\bullet}$
"−−"	p_{--++}	p_{--+-}		p_{----}	$p_{--\bullet}$
Total	$p_{\bullet++}$	$p_{\bullet+-}$		$p_{\bullet--}$	1

were $p_{ijks} = 2p^{\circ}_{ijks} = 2p^{\circ}_{ijsk} = 2p^{\circ}_{jiks}$.

The haploid (gametic) part of the LDE: $D_g = p_g - p_{+\bullet}p_{\bullet+}$, where

$$p_g = p_{++++} + p_{+++-}/2 + p_{+-++}/2 + (p_{+-+-} + p_{+--+})/4.$$

The diploid (non gametic) part LDE: $D_d = p_d - p_{+\bullet}p_{\bullet+}$ linkage disequilibrium, where

$$p_d = p_{++++} + p_{+++-}/2 + p_{+-++}/2 + p_{+-+-}/2.$$

The composite LDE:

$$\Delta = D_g + D_d = p_g + p_d - 2p_{+\bullet}p_{\bullet+}.$$

If genotypes are known only the composite LDE cannot be obtained. Then one use the within chromosome equilibrium $p_{+-+-} = p_{+--+}$. Under the within chromosome equilibrium $\Delta = 2D_g$.

The composite linkage disequilibrium correlation is defined as follows:

$$r = \frac{\Delta}{\sqrt{(p_{+\bullet}p_{-\bullet} + D_1)(p_{\bullet+}p_{\bullet-} + D_2)}},$$

where $D_1 = HWD_1 = p_{++\bullet} - p_{+\bullet}^2$, $D_2 = HWD_2 = p_{\bullet++} - p_{\bullet+}^2$.

Under within chromosome equilibrium assumption if to set

$$\xi_1 = \begin{cases} -1, \text{ if genotype is "++"}, \\ 0, \text{ if genotype is "+−"}, \\ 1, \text{ if genotype is "−−"}, \end{cases} \quad \text{and} \quad \xi_2 = \begin{cases} -1, \text{ if genotype is "++"}, \\ 0, \text{ if genotype is "+−"}, \\ 1, \text{ if genotype is "−−"}, \end{cases}$$

then $\Delta = \mathbf{cov}(\xi_A, \xi_B)$ and $r = \mathbf{corr}(\xi_A, \xi_B)$.

3: Allelic LDE and maximum likelihood method. If we have haplotype data from diploid organism we can easily specify frequencies of allelelic variants in Table 1 from the count table

Table 5.

Locus 1\Locus 2	"++"	"+−"	"−+"	"−−"	Total
"++"	n_{++++}	n_{+++-}		n_{++--}	$n_{++\bullet}$
"+−"	n_{+-++}	n_{+-+-}	n_{+--+}	n_{+---}	$n_{+-\bullet}$
"−−"	n_{--++}	n_{--+-}		n_{----}	$n_{--\bullet}$
Total	$n_{\bullet++}$	$n_{\bullet+-}$		$n_{\bullet--}$	n

Then the corresponding allelic counts are given in the following table

Table 6.

Locus 1\Locus 2	"+"	"−"
"+"	$2n_{++++} + n_{+++-} + n_{+-++} + n_{+-+-}$	$2n_{++--} + n_{+++-} + n_{+--+} + n_{+---}$
"−"	$2n_{--++} + n_{+-++} + n_{+--+} + n_{--+-}$	$2n_{----} + n_{--+-} + n_{+-+-} + n_{+---}$

and, taking account of the $2n$ total number of gametes, the corresponding frequencies are following:

Table 7.

Locus 1\Locus 2	"+"	"−"
"+"	$p_{++++} + p_{+++-}/2 + p_{+-++}/2 + p_{+-+-}/2$	$p_{++--} + p_{+++-}/2 + p_{+--+}/2 + p_{+---}/2$
"−"	$p_{--++} + p_{+-++}/2 + p_{+--+}/2 + p_{--+-}/2$	$p_{----} + p_{--+-}/2 + p_{+-+-}/2 + p_{+---}/2$

In the case of genotypes we have only $n_{+-+-} + n_{+--+}$ instead of n_{+-+-} and n_{+--+}. One can simplify the parametization

Table 8.

Locus 1\Locus 2	"+"	"−"
"+"	$p_{++} = p^*_{++} + p_{+-+-}/2$	$p_{+-} = p^*_{+-} + p_{+--+}/2$
"−"	$p_{-+} = p^*_{-+} + p_{+--+}/2$	$p^*_{--} = p_{--} + p_{+-+-}/2$

and join the observed counts to $n^*_{++}, n^*_{+-}, n^*_{-+}, n^*_{--}, n^*_{+-+-}$, where

$$\begin{aligned}
n^*_{++} &= 2n_{++++} + n_{+++-} + n_{+-++};\\
n^*_{+-} &= 2n_{++--} + n_{+++-} + n_{+---};\\
n^*_{-+} &= 2n_{--++} + n_{+-++} + n_{--+-};\\
n^*_{--} &= 2n_{----} + n_{--+-} + n_{+---};\\
n^*_{+-+-} &= 2(n_{+-+-} + n_{+--+})
\end{aligned}$$

and $n^*_{++} + n^*_{+-} + n^*_{-+} + n^*_{--} + n^*_{+-+-} = 2n$. The obtaind statistical model is overparametrized; one can use the follofing relations $p_{+-+-} = p_{++}p_{--}$ and $p_{+--+} = p_{+-}p_{-+}$ (local Hardy–Weinberg equilibrium in some sense). Then, the log liklihood function in the model can be written as follows:

$$\begin{aligned}
LL(\vec{p}^*|\vec{n}^*) = n^*_{++}\log p^*_{++} + n^*_{+-}\log p^*_{+-} &+ n^*_{-+}\log p^*_{-+} + n^*_{--}\log p^*_{--}\\
&+ n^*_{+-+-}\log(p^*_{++}p^*_{--} + p^*_{+-}p^*_{+-}).
\end{aligned} \qquad (A)$$

The maximum likelihood can be obtained using the EM-algorithm based on the full likelihood, which is unobserved:

$$\begin{aligned}
LL^\circ(\vec{p}|\vec{n}) &= n^*_{++}\log p^*_{++} + n^*_{+-}\log p^*_{+-} + n^*_{-+}\log p^*_{-+} + n^*_{--}\log p^*_{--}\\
&+ 2n_{+-+-}(\log p_{++} + \log p_{--}) + 2n_{+-+-}(\log p_{+-} + \log p_{+-}).
\end{aligned}$$

The expectation step is based on

$$\begin{aligned}
E_{\vec{p}^*_0}(LL(\vec{p}|\vec{n})|\vec{n}^*) = n^*_{++}\log p^*_{++} &+ n^*_{+-}\log p^*_{+-} + n^*_{-+}\log p^*_{-+} + n^*_{--}\log p^*_{--}\\
&+ n^*_{+-+-}\frac{p^*_{0++}p^*_{0--}}{p^*_{0++}p^*_{0--} + p^*_{0+-}p^*_{0-+}}(\log p^*_{++} + \log p^*_{--})\\
&+ n^*_{+-+-}\frac{p^*_{0+-}p^*_{0-+}}{p^*_{0++}p^*_{0--} + p^*_{0+-}p^*_{0-+}}(\log p^*_{+-} + \log p^*_{+-}).
\end{aligned}$$

The algorithm:
(i). One should start from some $\vec{p}^{*(0)}$ such that $LL(\vec{p}|\vec{n}^*)$; set $p^{()}_{++}$ and set $n = 0$

(ii) Loop while $\delta \geq \delta_0$; for k-th iteration
E-step: Calculate the expected counts

$$\begin{aligned}
n^{(k)}_{++} &= n^*_{++} + n^*_{+-+-}\frac{p^{(k)}_{++}p^{(k)}_{--}}{p^{(k)}_{++}p^{(k)}_{--} + p^{(k)}_{+-}p^{(k)}_{-+}} \quad n^{(k)}_{--} = n^*_{--} + n^*_{+-+-}\frac{p^{(k)}_{++}p^{(k)}_{--}}{p^{(k)}_{++}p^{(k)}_{--} + p^{(k)}_{+-}p^{(k)}_{-+}}\\
n^{(k)}_{+-} &= n^*_{+-} + n^*_{+-+-}\frac{p^{(k)}_{+-}p^{(k)}_{-+}}{p^{(k)}_{++}p^{(k)}_{--} + p^{(k)}_{+-}p^{(k)}_{-+}} \quad n^{(k)}_{-+} = n^*_{-+} + n^*_{+-+-}\frac{p^{(k)}_{+-}p^{(k)}_{-+}}{p^{(k)}_{++}p^{(k)}_{--} + p^{(k)}_{+-}p^{(k)}_{-+}}.
\end{aligned}$$

M-step: The expected likelihood

$$E_{\vec{p}^{\,(k)}}(LL(\vec{p}|\vec{n})|\vec{n}^*) = n^{(k)}_{++}\log p^*_{++} + n^{(k)}_{+-}\log p^*_{+-} + n^{(k)}_{-+}\log p^*_{-+} + n^{(k)}_{--}\log p^*_{--}$$

is maximized by

$$p_{++}^{(k+1)} = n_{++}^{(k)}/n_{\bullet}^{(k)}, \quad p_{+-}^{(k+1)} = n_{+-}^{(k)}/n_{\bullet}^{(k)}, \quad p_{-+}^{(k+1)} = n_{-+}^{(k)}/n_{\bullet}^{(k)}, \quad p_{--}^{(k+1)} = n_{--}^{(k)}/n_{\bullet}^{(k)},$$

where $n_{\bullet}^{(k)} = n_{++}^{*} + n_{+-}^{*} + n_{-+}^{*} + n_{--}^{*} + 2n_{+-+-}^{*}$.
Calculate δ(normally, $\delta > 0$):

$$\delta = LL(\vec{p}^{\,(k+1)}|\vec{n}^{*}) - LL(\vec{p}^{\,(k)}|\vec{n}^{*}),$$

where LL is calculated by formula (A).
(iii). If the loop was stopped at K-th iteration one use the following table instead of Table 1 to obtain D.

Table 9.

Locus 1\Locus 2	"+"	"−"	Total
"+"	$p_{++}^{(K)} + p_{++}^{(K)} p_{--}^{(K)}/2$	$p_{+-}^{(K)} + p_{+-}^{(K)} p_{-+}^{(K)}/2$	$p_{+\bullet}^{(K)} + p^{(K)}$
"−"	$p_{-+}^{(K)} + p_{+-}^{(K)} p_{-+}^{(K)}/2$	$p_{--}^{(K)} + p_{++}^{(K)} p_{--}^{(K)}/2$	$p_{-\bullet} + p^{(K)}$
Total	$p_{\bullet+} + p^{(K)}$	$p_{\bullet-} + p^{(K)}$	1

where $P^{(K)} = (p_{++}^{(K)} p_{--}^{(K)} + p_{++}^{(K)} p_{--}^{(K)})/2$.