Linkage Disequilibrium

This term reserved for association between pairs of alleles – one at each of two loci.

When gametic data are available, could refer to gametic disequilibrium.

When genotypic data are available, but gametes can be inferred, can make inferences about gametic and non-gametic pairs of alleles.

When genotypic data are available, but gametes cannot be inferred, can work with composite measures of disequilibrium.

Linkage Disequilibrium

For alleles A and B are two loci, the usual measure of linkage disequilibrium is

$$D_{AB} = P_{AB} - p_A p_B$$

Whether or not this is zero does not provide a direct statement about linkage between the two loci. For example, consider marker YFM and disease DTD:

		A	N	Total
YFM	+	1	24	25
	−	0	75	75
Total		1	99	100

$$D_{A+} = \frac{1}{100} - \frac{1}{100}\frac{25}{100} = 0.0075, \text{ (maximum possible value)}$$

Aside: Gametic Linkage Disequilibrium

For loci **A, B** define indicator variables x, y that take the value 1 for allele A, B and 0 for any other alleles. If gametes within individuals are indexed by j, $j = 1, 2$ then for expectations over samples from the same population

$$\begin{aligned}
\mathcal{E}(x_j) = p_A, \;\; j = 1,2 \;\; &, \;\; \mathcal{E}(y_j) = p_B \;\; j = 1,2 \\
\mathcal{E}(x_j^2) = p_A, \;\; j = 1,2 \;\; &, \;\; \mathcal{E}(y_j^2) = p_B \;\; j = 1,2 \\
\mathcal{E}(x_1 x_2) = P_{AA} \;\; &, \;\; \mathcal{E}(y_1 y_2) = P_{BB} \\
\mathcal{E}(x_1 y_1) = P_{AB} \;\; &, \;\; \mathcal{E}(x_2 y_2) = P_{AB}
\end{aligned}$$

The variances of x_j, y_j are $p_A(1-p_A), p_B(1-p_B)$ for $j = 1, 2$ and the covariance and correlation coefficients for x and y are

$$\text{Cov}(x_1, y_1) = \text{Cov}(x_2, y_2) \;\; = \;\; P_{AB} - p_A p_B = D_{AB}$$

$$\text{Corr}(x_1, y_1) = \text{Corr}(x_2, y_2) \;\; = \;\; D_{AB} / \sqrt{[p_A(1-p_A)p_B(1-p_B)]} = \rho_{AB}$$

Estimation of LD

With random sampling of gametes, gamete counts have a multinomial distribution:

$$\Pr(n_{AB}, n_{Ab}, n_{aB}, n_{ab}) = \frac{n!(P_{AB})^{n_{AB}}(P_{Ab})^{n_{Ab}}(P_{aB})^{n_{aB}}(P_{ab})^{n_{ab}}}{n_{AB}!n_{Ab}!n_{aB}!n_{ab}!}$$

The data are the counts of four gamete types, so there are three degrees of freedom. There are three parameters: p_A, p_B, D_{AB} so Bailey's method leads directly to MLE's:

$$\hat{D}_{AB} = \tilde{P}_{AB} - \tilde{p}_A\tilde{p}_B$$

$$\hat{\rho}_{AB} = r_{AB} = \frac{\hat{D}_{AB}}{\sqrt{\tilde{p}_A\tilde{p}_a\tilde{p}_B\tilde{p}_b}}$$

Testing LD

The MLE of D_{AB} is

$$\hat{D}_{AB} = \tilde{P}_{AB} - \tilde{p}_A\tilde{p}_B = \frac{1}{n^2}(n_{AB}n_{ab} - n_{Ab}n_{aB})$$

where n is the number of gametes in the sample. For large n, this estimate is normally distributed about the parametric value D_{AB}, so if $D_{AB} = 0$

$$X^2_{AB} = \frac{\hat{D}^2_{AB}}{\text{Var}(\hat{D}_{AB})} \sim \chi^2_{(1)}$$

When $D_{AB} = 0$, $\text{Var}(\hat{D}_{AB}) = p_A(1-p_A)p_B(1-p_B)/n$ and the test statistic is calculated as

$$X^2_{AB} = \frac{n\hat{D}^2_{AB}}{\tilde{p}_A(1-\tilde{p}_A)\tilde{p}_B(1-\tilde{p}_B)}$$

This can be written as $X^2_{AB} = nr^2_{AB}$, by analogy to the test statistic $X^2 = n\tilde{f}^2$ for Hardy-Weinberg equilibrium.

Aside: Testing LD

Writing the MLE of D_{AB} as

$$\hat{D}_{AB} = \frac{1}{n^2}(n_{AB}n_{ab} - n_{Ab}n_{aB})$$

where n is the number of gametes in the sample, allows the use of the "Delta method" to find

$$\begin{aligned}\text{Var}(\hat{D}_{AB}) &\approx \frac{1}{n}[p_A(1-p_A)p_B(1-p_B) \\ &\quad + (1-2p_A)(1-2p_B)D_{AB} - D_{AB}^2]\end{aligned}$$

When $D_{AB} = 0$, $\text{Var}(\hat{D}_{AB}) = p_A(1-p_A)p_B(1-p_B)/n$.

If $\hat{D}_{AB}$ is assumed to be normally distributed then

$$X_{AB}^2 = \frac{\hat{D}_{AB}^2}{\text{Var}(\hat{D}_{AB})} = n\hat{\rho}_{AB}^2 = nr_{AB}^2$$

is appropriate for testing $H_0 : D_{AB} = 0$. When H_0 is true, $X_{AB}^2 \sim \chi_{(1)}^2$. Note the analogy to the test statistic for Hardy-Weinberg equilibrium: $X^2 = n\hat{f}^2$.

Goodness-of-fit Test

The test statistic for the 2×2 table

n_{AB}	n_{Ab}	n_A
n_{aB}	n_{ab}	n_a
n_B	n_b	n

has the value

$$X^2 = \frac{n(n_{AB}n_{ab} - n_{Ab}n_{aB})^2}{n_A n_a n_B n_b}$$

$$= \frac{n\hat{D}_{AB}^2}{\tilde{p}_A \tilde{p}_a \tilde{p}_B \tilde{p}_b}$$

For DTD/YFM example, $X^2 = 3.03$. This is not statistically significant, even though disequilibrium was maximal.

Composite Disequilibrium

When genotypes are scored, it is often not possible to distinguish between the two double heterozygotes AB/ab and Ab/aB, so that gametic frequencies cannot be inferred.

Under the assumption of random mating, in which genotypic frequencies are assumed to be the products of gametic frequencies, it is possible to estimate gametic frequencies with the EM algorithm. To avoid making the random-mating assumption, however, it is possible to work with a set of composite disequilibrium coefficients.

Composite Disequilibrium

Although the separate digenic frequencies p_{AB} (one gamete) and $p_{A,B}$ (two gametes) cannot be observed, their sum can be since

$$\begin{aligned} p_{AB} &= P_{AB}^{AB} + \frac{1}{2}P_{Ab}^{AB} + \frac{1}{2}P_{aB}^{AB} + \frac{1}{2}P_{ab}^{AB} \\ p_{A,B} &= P_{AB}^{AB} + \frac{1}{2}P_{Ab}^{AB} + \frac{1}{2}P_{aB}^{AB} + \frac{1}{2}P_{aB}^{Ab} \\ p_{AB} + p_{A,B} &= 2P_{AB}^{AB} + P_{Ab}^{AB} + P_{aB}^{AB} + \frac{P_{ab}^{AB} + P_{aB}^{Ab}}{2} \end{aligned}$$

Digenic disequilibrium is measured with a composite measure Δ_{AB} defined as

$$\begin{aligned} \Delta_{AB} &= p_{AB} + p_{A,B} - 2p_A p_B \\ &= D_{AB} + D_{A,B} \end{aligned}$$

which is the sum of the gametic ($D_{AB} = p_{AB} - p_A p_B$) and nongametic ($D_{A,B} = p_{A,B} - p_A p_B$) coefficients.

Composite Disequilibrium

If the counts of the nine genotypic classes are

	BB	Bb	bb
AA	n_1	n_2	n_3
Aa	n_4	n_5	n_6
aa	n_7	n_8	n_9

the count for pairs of alleles in an individual being A and B, whether received from the same or different parents, is

$$n_{AB} = 2n_1 + n_2 + n_4 + \frac{1}{2}n_5$$

and the MLE for Δ is

$$\hat{\Delta}_{AB} = \frac{1}{n}n_{AB} - 2\tilde{p}_A\tilde{p}_B$$

Composite LD and Allele Dosage

The allele dosage for a SNP is the number of copies of the (say) the reference allele carried by an individual. If A is the reference allele for SNP **A**, then genotypes AA, Aa, aa have dosages X_A of 2,1,0.

The covariance of allele dosages X_A, X_B for loci **A, B** is

$$\text{Cov}(X_A, X_B) \;=\; 2\Delta_{AB}$$

By analogy to the tests for within-population inbreeding and for gametic linkage disequilibrium, a test statistic for composite LD is

$$X^2_{AB_c} \;=\; nr^2_{AB_c}$$

where r_{AB_c} is the sample correlation coefficient for allele dosages at the two loci over the n individuals in a sample.

Example

A sample of size 15 has these two-locus genotypes and allele dosages:

		X_A	X_A^2	X_B	X_B^2	X_AX_B
1	$AAbb$	2	4	0	0	0
2	$AAbb$	2	4	0	0	0
3	$AaBB$	1	1	2	4	2
4	$AaBb$	1	1	1	1	1
5	$AaBb$	1	1	1	1	1
6	$AaBb$	1	1	1	1	1
7	$Aabb$	1	1	0	0	0
8	$Aabb$	1	1	0	0	0
9	$Aabb$	1	1	0	0	0
10	$Aabb$	1	1	0	0	0
11	$aaBb$	0	0	1	1	0
12	$aabb$	0	0	0	0	0
13	$aabb$	0	0	0	0	0
14	$aabb$	0	0	0	0	0
15	$aabb$	0	0	0	0	0
Sum		$S_A = 12$	$S_{AA} = 16$	$S_B = 6$	$S_{BB} = 8$	$S_{AB} = 5$

Example (contd.)

The sample means, variances, covariance and correlation of dosages X_A, X_B are:

means: $\bar{X}_A = S_A/n = 12/15; \bar{X}_B = S_B/n = 6/15$

variances: $s_A^2 = (S_{AA} - S_A^2/n)/(n-1) = (16 - 144/15)/14$;
$s_B^2 = (S_{BB} - S_B^2/n)/(n-1) = (8 - 36/15)/14$

covariance: $s_{AB} = (S_{AB} - S_A S_B/n)/(n-1) = (5 - 72/15)/14$

correlation: $r_{AB_c}^2 = s_{AB}^2/s_A^2 s_B^2 = 1/(32 * 28)$

test statistic: $X_{AB_c}^2 = n r_{AB_c}^2 = 0.0168$

The hypothesis of no *composite* LD is not rejected. If there is HWE is this the same as testing for LD.

Aside: Composite Linkage Disequilibrium

For loci **A, B** define indicator variables x, y that take the value 1 for allele A, B and 0 for any other alleles. If gametes within individuals are indexed by j, $j = 1, 2$ then for expectations over samples from the same population

$$\begin{aligned}
\mathcal{E}(x_j) = p_A, \;\; j = 1,2 \;\; &, \;\; \mathcal{E}(y_j) = p_B \;\; j = 1,2 \\
\mathcal{E}(x_j^2) = p_A, \;\; j = 1,2 \;\; &, \;\; \mathcal{E}(y_j) = p_B \;\; j = 1,2 \\
\mathcal{E}(x_1 x_2) = P_{AA} \;\; &, \;\; \mathcal{E}(y_1 y_2) = P_{BB} \\
\mathcal{E}(x_1 y_1) = P_{AB} \;\; &, \;\; \mathcal{E}(x_2 y_2) = P_{AB} \\
\mathcal{E}(x_1 y_2) = P_{A,B} \;\; &, \;\; \mathcal{E}(x_2 y_1) = P_{A,B}
\end{aligned}$$

Write

$$\begin{aligned}
D_A = P_{AA} - p_A^2 \;\; &, \;\; D_B = P_{BB} - p_B^2 \\
D_{AB} = P_{AB} - p_A p_B \;\; &, \;\; D_{A,B} = P_{A,B} - p_A p_B \\
\Delta_{AB} = D_{AB} + D_{A,B} &
\end{aligned}$$

Aside:Composite LD and Allele Dosage

Now set $X = x_1 + x_2, Y = y_1 + y_2$, the allelic dosages at each locus, to get

$$\begin{aligned}
\mathcal{E}(X) = 2p_A \quad &, \quad \mathcal{E}(Y) = 2p_B \\
\mathcal{E}(X^2) = 2(p_A + P_{AA}) \quad &, \quad \mathcal{E}(Y^2) = 2(p_B + P_{BB}) \\
\text{Var}(X) = 2p_A(1-p_A)(1+f_A) \quad &, \quad \text{Var}(Y) = 2p_B(1-p_B)(1+f_B)
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{E}(XY) &= 2(P_{AB} + P_{A,B}) \\
\text{Cov}(X,Y) &= 2(P_{AB} - p_A p_B) + 2(P_{A,B} - p_A p_B) \\
&= 2(D_{AB} + D_{A,B}) = 2\Delta_{AB} \\
\text{Corr}(X,Y) &= \frac{\Delta_{AB}}{\sqrt{p_A(1-p_A)(1+f_A)p_B(1-p_B)(1+f_B)}}
\end{aligned}$$

ASIDE: Composite Linkage Disequilibrium Test

$$\hat{\Delta}_{AB} = n_{AB}/n - 2\tilde{p}_A\tilde{p}_B$$

where

$$n_{AB} = 2n_{AABB} + n_{AABb} + n_{AaBB} + \frac{1}{2}n_{AaBb}$$

This does not require phased data.

By analogy to the gametic linkage disequilibrium result, a test statistic for $\Delta_{AB} = 0$ is

$$X^2_{AB} = \frac{n\hat{\Delta}^2_{AB}}{\tilde{p}_A(1-\tilde{p}_A)(1+\hat{f}_A)\tilde{p}_B(1-\tilde{p}_B)(1+\hat{f}_B)}$$

This is assumed to be approximately $\chi^2_{(1)}$ under the null hypothesis. The approximation rests on ignoring disequilibria between three and four alleles of the two **A** and two **B** alleles.

Aside: Example

For the data shown on Slide 12:

	BB	Bb	bb	Total
AA	$n_{AABB} = 0$	$n_{AABb} = 0$	$n_{AAbb} = 2$	$n_{AA} = 2$
Aa	$n_{AaBB} = 1$	$n_{AaBb} = 3$	$n_{Aabb} = 4$	$n_{Aa} = 8$
aa	$n_{aaBB} = 0$	$n_{aaBb} = 1$	$n_{aabb} = 4$	$n_{aa} = 5$
Total	$n_{BB} = 1$	$n_{Bb} = 4$	$n_{bb} = 10$	$n = 15$

$$
\begin{aligned}
n_{AB} &= 2 \times 0 + 0 + 1 + \frac{1}{2}(3) = 2.5 \\
n_A &= 12, \;\; \tilde{p}_A = 0.4 \\
n_B &= 6, \;\; \tilde{p}_B = 0.2 \\
\widehat{f}_A &= 1 - \frac{8/15}{0.48} = -0.11 \\
\widehat{f}_B &= 1 - \frac{4/15}{0.32} = 0.17
\end{aligned}
$$

Aside: Example

The estimated composite disequilibrium coefficient is

$$\hat{\Delta}_{AB} = \frac{2.5}{15} - 2(0.4)(0.2) = 0.0067$$

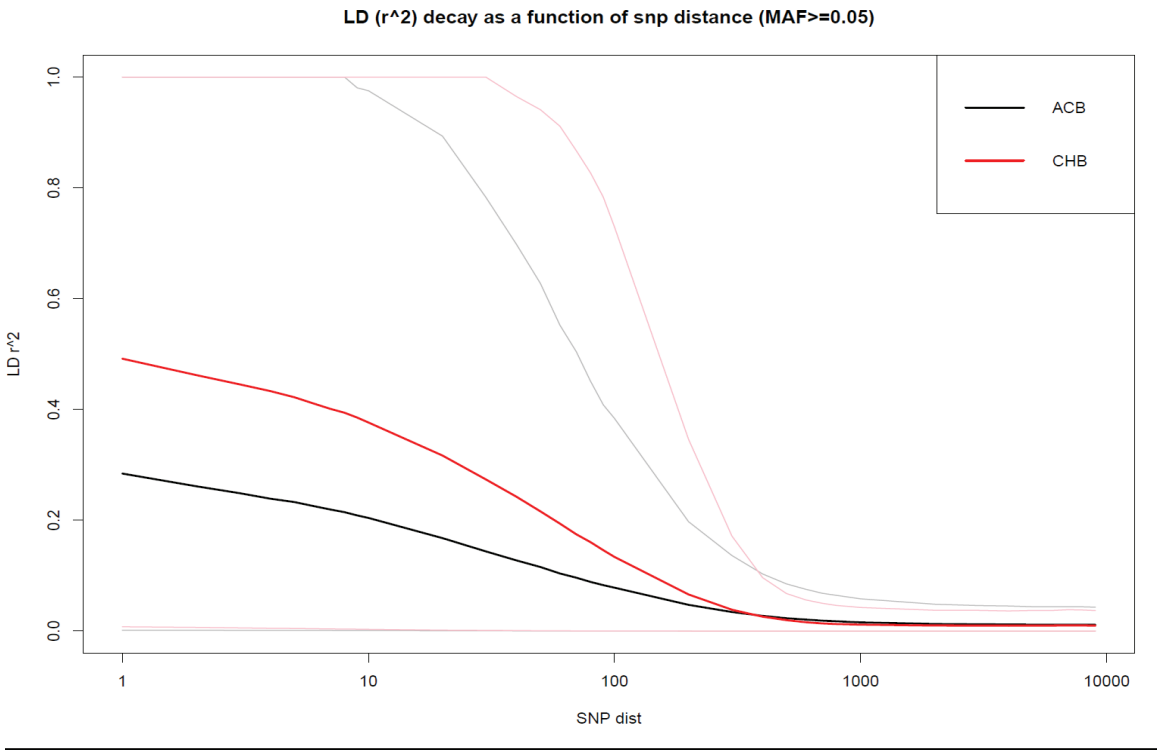
The test statistic is

$$X^2 = \frac{15 \times (0.0067)^2}{0.24 \times 0.89 \times 0.16 \times 1.17} = 0.02$$

Previous work on EM algorithm, assuming HWE, estimated p_{AB} as 0.0893 so

$$\begin{aligned} \hat{D}_{AB} &= 0.0893 - 0.4 \times 0.2 = 0.0093 \\ X^2 &= \frac{30 \times (0.0093)^2}{0.4 \times 0.6 \times 0.2 \times 0.8} = 0.07 \end{aligned}$$

1000 Genomes Example



Allele dosage squared correlations for pairs of SNPs on chromosomes 21 and 22 of the 1000 Genomes ACB and populations. Heavy lines: means. Light lines: 5th and 95th percentiles.

Aside: Multi-locus Entropy

It is difficult to describe associations among alleles at several loci. One approach is based on information theory.

For a locus with sample frequencies $\tilde{p}_u$ for alleles A_u the entropy is

$$H_A = -\sum_u \tilde{p}_u \ln(\tilde{p}_u)$$

For two loci with alleles A_u, B_v, the entropy is

$$H_{AB} = -\sum_u \sum_v \tilde{P}_{uv} \ln(\tilde{P}_{uv})$$

In the absence of linkage disequilibrium $\tilde{P}_{uv} = \tilde{p}_u \tilde{p}_v$ so

$$\begin{aligned} H_{AB} &= -\sum_u \sum_v \tilde{p}_u \tilde{p}_v [\ln(\tilde{p}_u) + \ln(\tilde{p}_v)] \\ &= H_A + H_B \end{aligned}$$

so if $H_{AB} \neq H_A + H_B$ there is evidence of dependence. This extends to multiple loci.

Aside: Conditional Entropy

If the entropy for a multi-locus profile A is H_A then the conditional probability of another locus B, given A, is $H_{B|A} = H_{AB} - H_A$.

In performing meaningful calculations for Y-STR profiles, this suggests choosing a set of loci by an iterative procedure. First choose locus L_1 with the highest entropy. Then choose locus L_2 with the largest conditional entropy $H(L_2|L_1)$. Then choose L_3 with the highest conditional entropy with the haplotype L_1L_2, and so on.

Aside: Conditional Entropy for Y-STR Data

Added Marker	Entropy		
	Single	Multi	Cond.
DYS385ab	4.750	4.750	4.750
DYS481	2.962	6.972	2.222
DYS570	2.554	8.447	1.474
DYS576	2.493	9.318	0.871
DYS458	2.220	9.741	0.423
DYS389II	2.329	9.906	0.165
DYS549	1.719	9.999	0.093
DYS635	2.136	10.05	0.053
DYS19	2.112	10.08	0.028
DYS439	1.637	10.10	0.024
DYS533	1.433	10.11	0.010
DYS456	1.691	10.12	0.006
GATAH4	1.512	10.12	0.005
DYS393	1.654	10.13	0.003
DYS448	1.858	10.13	0.002
DYS643	2.456	10.13	0.002
DYS390	1.844	10.13	0.002
DYS391	1.058	10.13	0.002

Most-discriminating loci may not contribute to the most-discriminating haplotypes. No additional discriminating power beyond 10 loci.