Multipoint Genetic Association Studies

- Single-Marker (Single-Point) Analysis

  - Analyzing one mark at a time

  - May lose power to detect gene-disease association when multiple genes cause the disease

  - May not capture the interaction between multiple genes

- Multiple-Marker (Multipoint) Analysis

  - Analyzing multiple markers simultaneously
  - May gain power to detect gene-disease association
  - May capture the gene-gene interactions

- Two types of Mutilpoint Analysis

  - Treating multiple markers as 'longitudinal observations' on the chromosomes and applying statistical methods for longitudinal data such as 'GEE'(Generalized Estimating Equation)
  - Treating multiple markers (when tightly linked) on each of the chromosome as a genetic unit, termed 'haplotype', then analyzing the association between disease and haplotypes

- Generalized T^2 Test

  - coding for J binary markers:

$$X_{ij} = \begin{cases} 1 & A_j A_j \\ 0 & A_j a_j \\ -1 & a_j a_j \end{cases}, \quad j = 1, \ldots, J$$

- $X_i = (X_{i1}, \ldots, X_{iJ})'$, $Y_i = (Y_{i1}, \ldots, Y_{iJ})'$, where X denotes case samples, and Y denotes control samples

- $\bar{X}_j = \sum_{i=1}^{n_X} X_{ij}/n_X$, $\bar{Y}_j = \sum_{i=1}^{n_Y} Y_{ij}/n_Y$,
n_X and n_Y: numbers of cases and controls

- $\bar{X} = (\bar{X}_1, \ldots, \bar{X}_J)'$, $\bar{Y} = (\bar{Y}_1, \ldots, \bar{Y}_J)'$

- pooled-sample variance matrix

$$S = \frac{1}{n_X + n_Y - 2} \left[\sum_{i=1}^{n_X} (X_i - \bar{X})(X_i - \bar{X})' + \sum_{i=1}^{n_Y} (Y_i - \bar{Y})(Y_i - \bar{Y})' \right.$$

- Hotelling's T^2 statistic

$$T^2 = \frac{n_X n_Y}{n_X + n_Y}(\bar{X} - \bar{Y})' S^{-1} (\bar{X} - \bar{Y})$$

-under H_0: no LD exists between any marker being tested and a disease locus, $T^2 \sim \chi^2(df = J)$

- GEE approach for multipoint association analysis: Case-Parent Trio Design

  - M markers at $0 < t_1 < t_2 < t_3 < ... < t_M < T$ cM
  - the transmission statistic $Y(t)$ at location t:

$$Y(t) = Y_1(t) - Y_2(t), \text{ where}$$

$$Y_1(t) = \begin{cases} 1 & \text{the transmitted paternal allele at } t \text{ is } H(t) \\ 0 & \text{the transmitted paternal allele at } t \text{ is } h(t) \end{cases}$$

$$Y_2(t) = \begin{cases} 1 & \text{the non-transmitted paternal allele at } t \text{ is } H(t) \\ 0 & \text{the non-transmitted paternal allele at } t \text{ is } h(t) \end{cases}$$

$H(t)$: the target allele at marker t, $h(t)$: the non-target allele

- Define $X(t) = X_1(t) - X_2(t)$ similarly for the maternal transmission statistic.

Y(X)		Transmitted H	Transmitted h
Nontransmitted	H	0	-1
	h	1	0

*Note: Sum of $Y(t) + X(t)$ across trios $= b - c$, the numerator of TDT

-$E\{Y(t)\} = 0$ where marker t is either unlinked to or in linkage equilibrium with the disease gene.

-Let

Φ be the event that the offspring is affected;

τ be the location of disease gene.

$$
\begin{aligned}
E\left[Y(t)|\Phi\right] &= E\left[Y_1(t) - Y_2(t)|\Phi\right] \\
&= \Pr\left[Y_1(t) = 1|\Phi\right] - \Pr\left[Y_2(t) = 1|\Phi\right] \\
&= \sum_{g_1,g_2} \{\Pr\left[Y_1(t) = 1|Y_1(\tau) = 1, g_1, g_2, \Phi\right] b\left(g_1, g_2\right) \\
&+ \Pr\left[Y_1(t) = 1|Y_1(\tau) = 0, g_1, g_2, \Phi\right] \left[1 - b\left(g_1, g_2\right)\right] \\
&- \Pr\left[Y_2(t) = 1|Y_1(\tau) = 1, g_1, g_2, \Phi\right] b\left(g_1, g_2\right) \\
&- \Pr\left[Y_2(t) = 1|Y_1(\tau) = 0, g_1, g_2, \Phi\right] \left[1 - b\left(g_1, g_2\right)\right]\} \\
&\times \Pr\left(g_1, g_2|\Phi\right)
\end{aligned}
$$

where $b(g_1, g_2) = \Pr\left[Y_1(\tau) = 1|g_1, g_2, \Phi\right]$,

(g_1, g_2) are haplotypes for the father at loci t and τ.

Except for

$g^* = [H(t)h(\tau), h(t)H(\tau)]$ and $g^{**} = [H(t)H(\tau), h(t)h(\tau)]$,

the terms within the bracket cancel each other out.

Let $\theta_{t,\tau}$ be the recombination rate between loci t and τ.

Then

$$
\begin{aligned}
E[Y(t)|\Phi] = &\{\theta_{t,\tau} b(g^*) + (1-\theta_{t,\tau})[1-b(g^*)] \\
&- (1-\theta_{t,\tau}) b(g^*) - \theta_{t,\tau}[1-b(g^*)]\} \Pr(g^*|\Phi) \\
&+ \{(1-\theta_{t,\tau}) b(g^{**}) + \theta_{t,\tau}[1-b(g^{**})] \\
&- \theta_{t,\tau} b(g^{**}) - (1-\theta_{t,\tau})[(1-b(g^{**}))]\} \Pr(g^{**}|\Phi) \\
&= (1-2\theta_{t,\tau}) \{2 \Pr[Y_1(\tau) = 1 | H(\tau), h(\tau), \Phi] - 1\} \\
&\times [\Pr(g^{**}|\Phi) - \Pr(g^*|\Phi)]
\end{aligned}
$$

where we have used the assumption: There is only one disease locus, so that

$$\begin{aligned} b(g^*) &= \Pr[Y_1(\tau) = 1 | H(t)h(\tau), h(t)H(\tau), \Phi] \\ &= \Pr[Y_1(\tau) = 1 | h(\tau), H(\tau), \Phi] = b(g^{**}) \end{aligned}$$

Using the same assumption we have

$$\Pr(\Phi | g^*) = \Pr(\Phi | g^{**}) = \Pr[\Phi | H(\tau), h(\tau)]$$

and

$$\begin{aligned}\Pr(g^{**}|\Phi) - \Pr(g^{*}|\Phi) = & \frac{\Pr[\Phi|H(\tau), h(\tau)]}{\Pr(\Phi)} \\ & \times \{\Pr\left[H(t)H(\tau), h(t)h(\tau)\right] \\ & - \Pr\left[H(t)h(\tau), h(t)H(\tau)\right]\}\end{aligned}$$

By Hardy-Weinberg Equilibrium,

$$\begin{aligned}\Pr\left[H(t)H(\tau), h(t)h(\tau)\right] - \Pr\left[H(t)h(\tau), h(t)H(\tau)\right] \\ = \Pr\left[H(t)H(\tau)\right] - \Pr\left[H(t)\right]\Pr\left[H(\tau)\right]\end{aligned}$$

and $\Pr(g^{**}|\Phi) - \Pr(g^{*}|\Phi) = \Pr\left[H(\tau), h(\tau)|\Phi\right] d(t)$

where

$$\begin{aligned}d(t) &= \frac{\Pr\left[H(t)H(\tau)\right] - \Pr\left[H(t)\right]\Pr\left[H(\tau)\right]}{\Pr[H(\tau)]\Pr[h(\tau)]} \\ &= \Pr[H(t)|H(\tau)] - \Pr[H(t)|h(\tau)]\end{aligned}$$

Consequently,

$$E\left[Y(t)|\Phi\right] = (1 - 2\theta_{t,\tau})E\left[Y(\tau)|\Phi\right] d(t)$$

-$E[Y(t)|\Phi] = 0$

if $\theta_{t,\tau} = \frac{1}{2}$ (the marker t is unlinked to the disease locus)

or $d(t) = 0$ (the marker t is in linkage equilibrium with the disease locus)

-$E[Y(t)|\Phi] \uparrow E[Y(\tau)|\Phi]$ when $t \longrightarrow \tau$

-Under initial complete LD, random mating, and constant Pr$[H(\tau)]$ over time,

$d(t) = (1-\theta_{t,\tau})^N$ Pr$[h(t)|h(\tau)]$, so

$$E[Y(t)|\Phi] = (1-2\theta_{t,\tau})E[Y(\tau)|\Phi](1-\theta_{t,\tau})^N \text{ Pr}[h(t)|h(\tau)]$$

$$
\begin{aligned}
E[Y_i(t_j)|\Phi] &= E[X_i(t_j)|\Phi] \\
&= (1-2\theta_{t_j,\tau})C(1-\theta_{t_j,\tau})^N\pi_j \\
&= \mu(t_j;\tau,C,N,\pi_j)
\end{aligned}
$$

$i = 1,...,n(trios),\ j = 1,...,M(markers)$

where

$$C = E[Y(\tau)|\Phi] = E[X(\tau)|\Phi]$$

$$\pi_j = \mathsf{Pr}[h(t_j)|h(\tau)]$$

$\theta_{t_j,\tau}$ is a function of $\left|t_j - \tau\right|$ using Haldane map function.

Note that π_j can be estimated by $\frac{\sum_{i=1}^{n}[(1-Y_{i2}(t_j))+(1-X_{i2}(t_j))]}{2n}$ assuming dominant mode of inheritence and the disease is rare.

-the parameters $\delta = (C, \tau, N)$ are estimated by solving

$$\begin{aligned} S(\delta) = \sum_{i=1}^{n} [&\frac{\partial \mu(\delta,\pi)}{\partial \delta} Cov^{-1}(Y_i) \{Y_i - \mu(\delta,\pi)\} \\ &+ \frac{\partial \mu(\delta,\pi)}{\partial \delta} Cov^{-1}(X_i) \{X_i - \mu(\delta,\pi)\}] \\ &= 0 \end{aligned}$$

where

$$Y_i = \begin{pmatrix} Y_i(t_1) \\ Y_i(t_2) \\ \cdots \\ Y_i(t_M) \end{pmatrix}, X_i = \begin{pmatrix} X_i(t_1) \\ X_i(t_2) \\ \cdots \\ X_i(t_M) \end{pmatrix}, \mu(\delta, \widehat{\pi}) = \begin{pmatrix} \mu(t_1; \delta, \widehat{\pi}_1) \\ \mu(t_2; \delta, \widehat{\pi}_2) \\ \cdots \\ \mu(t_M; \delta, \widehat{\pi}_M) \end{pmatrix},$$

-$Cov(Y_i)$ is the covariance matrix of Y_i, which can be naively set as a diagonal matrix (as if the components of Y_i are independent)

-This approach is essentially an application of generalized estimating equation (GEE).

For details, see Liang and Zeger (1986 Biometrika); Liang et al. (2001 Am J Hum Genet).

- Haplotype Analysis

  - Haplotype:
    the combination of closely linked alleles on a single chromosome

  - Haplotype Association Analysis:
    using haplotypes as a basic genetic unit for dissecting the genetic basis of the disease

    * haplotype composed of closely linked markers can have more of a biological role

* haplotypes can sometimes provide greater power than single-marker analysis for genetic disease association, because haplotypes can capture ancestral structure and gene-gene interactions. Besides, using haplotypes instead of multiple markers usually reduce the number of variables since the number of haplotypes within candidate genes is much smaller than the number of all possible haplotypes.

– Drawback of Haplotypes: Ambiguity

Haplotype information can usually be obtained indirectly from unphased genotype data; that is, at each

locus, we can only observe the two alleles appear on the two chromosomes but cannot observe which allele appear at each of the two chromosomes.

ex.

unphased genotype: AaBb

phased genotype (diplotype) can be (AB,ab) or (Ab,aB) where 'AB','ab','Ab','aB' are haplotypes respectively.

– Statistical Methods are required to reconstruct haplotypes from observed unphased genotype data.

– Testing for association between traits and haplotypes

* Generalized Linear Models (GLMs)

· y: trait

· X_g: a vector of numerical codes for genotype g

ex.

$$X_g = \begin{cases} 1 & \text{if } g = AA \\ 1 & \text{if } g = Aa \\ 0 & \text{if } g = aa \end{cases}$$

$\Rightarrow$ dominant model

ex.

$$X_g = \begin{cases} 1 & \text{if } g = AA \\ 0 & \text{if } g = Aa \\ 0 & \text{if } g = aa \end{cases}$$

$\Rightarrow$ recessive model

ex.

$$X_g = \begin{cases} 2 & \text{if } g = AA \\ 1 & \text{if } g = Aa \\ 0 & \text{if } g = aa \end{cases}$$

$\Rightarrow$ additive (codominant) model

ex. $X_g = (X_{g_1}, X_{g_2})$ where

$$X_{g_1} = \begin{cases} 1 & \text{if } Aa \\ 0 & \text{o.w.} \end{cases}$$

$$X_{g_2} = \begin{cases} 1 & \text{if } AA \\ 0 & \text{o.w.} \end{cases}$$

$\Rightarrow$ general model

- X_e: environmental variables (age, gender, race, ...), including intercept.

- linear predictor

$$\eta = X_e'\alpha + X_g'\beta$$

α: regression coefficients for the intercept and environmental variables

β: regression coefficients for the genotype (the effect of genotype on the trait)

- testing for gene-trait association, adjusting for environmental factors:

$$H_0 : \beta = 0$$

- GLM for exponential family data:

  likelihood:

$$L(y|X_e, X_g) = \exp\left[\frac{y\eta - b(\eta)}{a(\phi)} + C(y, \phi)\right]$$

$$E(y) = f^{-1}(\eta) = b'(\eta)$$

$$Var(y) = b''(\eta)a(\phi)$$

$$\eta = X_e'\alpha + X_g'\beta$$

Distribution	$E(y)$	$a(\phi)$	$b''(\eta)$
Normal	η	σ^2	1
Binomial	$\frac{e^\eta}{1+e^\eta}$	1	$E(y)[1-E(y)]$
Poisson	e^η	1	$E(y)$

* Score Tests for gene-trait association

  · Score function:

$$U = \sum_{i=1}^{N} \frac{\partial \ln L(y_i|X_{ei}, X_{gi})}{\partial(\alpha, \beta)}$$
$$= \sum_{i=1}^{N} \begin{pmatrix} X_{ei} \\ X_{gi} \end{pmatrix} \frac{y_i - E(y_i)}{a(\phi)}$$

N: number of subjects

- Score statistic for testing $H_0 : \beta = 0$

$$U_\beta = \sum_{i=1}^{N} X_{gi} \frac{y_i - E(y_i)}{a(\phi)}|_{\beta=0,\alpha=\hat{\alpha}}$$

where $\widehat{\alpha}$ is the solution to

$$U_\alpha = \sum_{i=1}^{N} X_{ei} \frac{y_i - E(y_i)}{a(\phi)}|_{\beta=0} = 0$$

$Var(U_\beta)$ under H_0:

$$V_\beta = V_{\beta\beta} - V_{\beta\alpha} V_{\alpha\alpha}^{-1} V_{\alpha\beta}|_{\beta=0,\alpha=\widehat{\alpha}}$$

where V_{ij} are corresponding submatrices of

$$Var(V) = \sum_{i=1}^{N} \frac{b''(\eta)}{a(\phi)} Z_i Z_i',$$

where $Z_i = \begin{pmatrix} X_{ei} \\ X_{gi} \end{pmatrix}$

Under H_0: $\beta = 0$,

$$S = U_\beta V_\beta^{-1} U_\beta \sim \chi^2(p)$$

where $p = dim(\beta)$ if V_β is full rank; when V_β is not full rank. we use

$$S = U_\beta' V_\beta^{-} U_\beta$$

where V_β^- is the generalized inverse of V_β and $S \sim \chi^2(p')$, where $p' = \text{rank}(V_\beta)$

* Score test for Ambiguous Haplotypes

  When haplotype information is derived from unphased genotype data, X_g is incompletely observed. We can apply the EM algorithm to obtain the score statistic.

  · likelihood:

$$\begin{aligned} L &= \Pr(y, m|X_e) \\ &= \sum_{g \in G} \Pr(y|X_e, X_g) \Pr(g) \end{aligned}$$

  where m is the unphased genotype data, $\Pr(g)$ is the marginal haplotype distribution, G is the set

of haplotype pairs that are consistent with the observed genotype data m.

eg. When $m = AaBb$

then $G = \{(AB, ab), (Ab, aB), (ab, AB), (aB, Ab)\}$

*Note: we usually assume Hardy-Weinberg Equilibrium for $\Pr(g)$, i.e. $\Pr(g) = \Pr(g_1, g_2) = \Pr(g_1)\Pr(g_2)$, so that the parameters for $\Pr(g)$ can be reduced from $L^2 - 1$ to $L - 1$, where $L = \#$ of haplotypes.

- the EM algorithm:

  the incomplete-data score function U_β can be obtained by the conditional expectation of the complete-data score function, conditional on the observed data (y, X_e, m).

  Since the score statistic is derived under H_0, the EM algorithm is taken under H_0: $\beta = 0$ and $\alpha = \hat{\alpha}$, and the incomplete-data score function (under H_0) is

$$\tilde{U}_\beta = \sum_{i=1}^{N} \frac{y_i - E(y_i)}{a\phi} E(X_g|m)|_{\beta=0,\alpha=\hat{\alpha}}$$

where

$$E(X_g|m) = \frac{\sum_{g \in G} X_g \Pr(g)}{\sum_{g \in G} \Pr(g)}$$

is the conditional expectation of X_g given the observed genotype m. (Note that this conditional expectation does not depend on y because it is evaluated under H_0)

The estimate of the genotype distribution is given by $\hat{\Pr}(g) = \hat{\Pr}(g_1, g_2) = \hat{\Pr}(g_1)\hat{\Pr}(g_2)$ (assuming HWE), where $\hat{\Pr}(g_i)$, $i = 1, 2$, is the haplotype frequencies that

can be estimated by the EM algorithm:

$$\widehat{\Pr}(g_1 = h) = \frac{1}{2N} \sum_{i=1}^{N} E\{I(g_{1i} = h) + I(g_{2i} = h)|m_i\}$$

$$= \frac{1}{2N} \sum_{i=1}^{N} \frac{\sum_{g \in G_i} \{I(g_1 = h) + I(g_2 = h)\} \widehat{\Pr}(g_1) \widehat{\Pr}(g_2)}{\sum_{g \in G_i} \widehat{\Pr}(g_1) \widehat{\Pr}(g_2)},$$

$$h = 1, ..., L$$

where G_i is the set of haplotype pairs that are consistent with the observed genotype m_i

*Note: The two sides of the above equation both involve the unknown $\widehat{\Pr}(g_1)$ hence it must be solved in an iterative manner: staring from an initial set of values for $\left\{\Pr^0(g_1), g_1 = 1, ..., L\right\}$, at the $(i+1)$th iteration, we obtain the updated values for $\Pr(g_1)$ by

$$\widehat{\Pr}^{(i+1)}(g_1 = h) =$$

$$\frac{1}{2N} \sum_{i=1}^{N} \frac{\sum_{g \in G_i} \{I(g_1 = h) + I(g_2 = h)\} \widehat{\Pr}^{(i)}(g_1) \widehat{\Pr}^{(i)}(g_2)}{\sum_{g \in G_i} \widehat{\Pr}^{(i)}(g_1) \hat{P}r^{(i)}(g_2)}$$

where $h = 1, ...L$, $i = 0, 1, 2, ...$

- Variance of the incomplete-data score from the EM algorithm:

  Let $S_\beta = \frac{\partial}{\partial \beta} \ln L(C)$, $\tilde{S}_\beta = \frac{\partial}{\partial \beta} \ln L(O)$,where C and O denotes the complete and incomplete observation, $L(\cdot)$ is the likelihood function so that

$$L(O) = \int_O L(C) dC$$

  By the EM algorithm, we have $\tilde{S}_\beta = E(S_\beta | O)$

HW : Show that

$$\begin{aligned} Var(\tilde{S}_\beta) &= -\frac{\partial^2}{\partial\beta^2} \ln L(O) \\ &= E[-\frac{\partial^2}{\partial\beta^2} \ln L(C)|O] - [E(S_\beta S'_\beta|O) - \tilde{S}_\beta \tilde{S}'_\beta] \end{aligned}$$

Also recall that when phase is known,

$$Var(U_\beta) \equiv V_\beta = V_{\beta\beta} - V_{\beta\alpha} V_{\alpha\alpha}^{-1} V_{\alpha\beta}$$

with V_{ij} the appropriate submatrices of the information matrix $E(-\frac{\partial^2}{\partial\beta^2} \ln L)$

When phase is unknown (ambiguous haplotype information), we can use the similar result for $Var(\widetilde{U}_\beta)$ except that V_{ij} are replaced by the appropriate submatrices of the incomplete-data information matrix $E[-\frac{\partial^2}{\partial\beta^2}\ln L(O)]$, which in turn can be obtained by the formula given above for the EM algorithm. In fact, now

$$\widetilde{V}_{\alpha\alpha} = \sum_{i=1}^{N} \frac{b''(\eta_i)}{a(\phi)} X_{ei} X_{ei}'$$

$$\widetilde{V}_{\alpha\beta} = \sum_{i=1}^{N} \frac{b''(\eta_i)}{a(\phi)} X_{ei} E(X_{gi}'|m_i)$$

$$\widetilde{V}_{\beta\beta} = \sum_{i=1}^{N} \{\frac{b''(\eta_i)}{a(\phi)} - \frac{[y_i - E(y_i)]^2}{a(\phi)^2}\} E(X_{gi} X'_{gi} | m_i)$$

$$+ \frac{[y_i - E(y_i)]^2}{a(\phi)^2} E(X_{gi}|m_i) E(X'_{gi}|m_i)$$

where $\eta_i = X'_{ei}\widehat{\alpha}$, and $Var(\widetilde{U}_\beta) = \widetilde{V}_\beta \equiv \widetilde{V}_{\beta\beta} - \widetilde{V}_{\beta\alpha}\widetilde{V}_{\alpha\alpha}^{-1}\widetilde{V}_{\alpha\beta}$

- the score test for ambiguous haplotype data is then

$$S = \widetilde{U_\beta}' \widetilde{V_\beta}^{-1} \widetilde{U_\beta},$$

under H_0, $S \sim \chi^2(df = p = dim(\beta))$

$\star$ Note: when $\widetilde{V}_\beta$ is not full rank, $df = p' = rank(\widetilde{V}_\beta)$.

$\star$ Note: $\widetilde{V}_\beta$ is not affected by the estimation of $\widehat{Pr}(g)$. This can be seen from

$$\frac{\partial \widetilde{U}_\beta}{\partial \gamma} = \frac{y - E(y)}{a(\phi)} Cov[X_g, \frac{\partial}{\partial \gamma} \ln \Pr(g)|m]$$

and hence $E(\frac{\partial \widetilde{U}_\beta}{\partial \gamma}) = 0$ where γ is the parameters determining $\Pr(g)$.

$\star$ Note: We have employed the HWE assumption to estimate $\Pr(g)$. Even if this assumption is violated, the score test is still valid for testing $H_0 : \beta = 0$; namely, the score test statistic S still follows a $\chi^2(p)$ distribution under $H_0 : \beta = 0$, even if the HWE does not hold.

HW: Show that the score test statistic S is valid even if the distribution $\Pr(g)$ is misspecified.

- Empirical P values

  When haplotype data is sparse (exist some rare haplotypes), the χ^2 distribution may not be accurate and we may need to compute empirical P values by simulations. Under H_0, none of the haplotypes are associated with the traits, so the empirical P values can be computed by repeatedly first permuting the trait values among the subjects and then computing the score statistics. The empirical distribution of the score statistics from the repetitions can then be used to find the P value of the observed score statistic.

- Contrasting LD patterns between cases and controls

  It has been noted that the extent of LD can be different between cases and controls in a region of genetic association, and the case-control LD comparison can aid the association analysis.

- LD coefficient

$$D_{AB} = P_{AB} - P_A P_B$$

  needs the use of please information

- composite LD coefficient: requires no phase information

- gametic LD (intragametic) coefficient

$$D_{AB} = P_{AB} - P_A P_B$$

where P_{AB} is the haplotype frequency of AB, and $P_{A|B} = P_{AB,AB} + 1/2(P_{AB,Ab} + P_{AB,aB} + P_{AB,ab})$.

- non-gametic (intergametic) coefficient

$$D_{A|B} = P_{A|B} - P_A P_B$$

where $P_{A|B}$ is non-gametic frequency of AB, and $P_{A|B} = P_{AB,AB} + 1/2(P_{AB,Ab} + P_{AB,aB} + P_{Ab,aB})$, i.e. $P_{A|B}$ is the probability that A, B are on different haplotypes.

- the composite LD is the sum of the gametic and non-gametic LD:

$$\begin{aligned}\Delta_{AB} &= D_{AB} + D_{A|B} \\ &= P_{AB} + P_{A|B} - 2P_AP_B \\ &= 2P_{AB,AB} + P_{AB,Ab} + P_{AB,aB} + 1/2(P_{AB,ab} + P_{Ab,aB}) \\ &\quad -2P_AP_B.\end{aligned}$$

Note that Δ_{AB} depends on double heterozygous (where the haplotype phase can not be uniquely determined) only through their total of $P_{AB,ab} + P_{Ab,aB}$. Also note that $\Delta_{AB} = D_{AB}$ where HWE is assumed.

- MLE of Δ_{AB}:

$$\hat{\Delta}_{AB} = n^{-1}(2n_{AABB} + n_{AABb} + n_{AaBB} + \frac{1}{2}n_{AaBb}) - 2\hat{P}_A\hat{P}_B$$

where $\hat{P}_A, \hat{P}_B$ are estimates of allele frequencies, n is the total number of subjects.

- Variance of $\widehat{\Delta}_{AB}$

$$\begin{aligned} nVar(\widehat{\Delta}_{AB}) \;=\; & (\pi_A + D_A)(\pi_B + D_B) + 1/2\tau_A\tau_B\Delta_{AB} \\ & + \tau_A D_{ABB} + \tau_B D_{AAB} + \Delta_{AABB} \end{aligned}$$

$\pi_A = P_A(1 - P_A)$

$D_A = P_{AA} - P_A^2$

$\tau_A = 1 - 2P_A, \tau_B, D_B, \tau_B$ defined similarly.

$D_{AAB} = P_{AAB} - P_A\Delta_{AB} - P_B D_A - P_A^2 P_B$

$D_{ABB} = P_{ABB} - P_B\Delta_{AB} - P_A D_B - P_A P_B^2$

$\Delta_{AABB} = P_{AABB} - 2P_A D_{ABB} - 2P_B D_{AAB}$

$-2P_A P_B D_{AB} - \Delta_{AB}^2 - 2P_A^2 D_B - P_B^2 D_A - D_A D_B - P_A^2 P_B^2.$

- Testing for $H_0 : \Delta_{AB} = 0$

$$\frac{n\Delta_{AB}^2}{(\hat{\pi}_A + \hat{D}_A)(\hat{\pi}_B + \hat{D}_B)} \sim \chi^2(1)$$

under H_0 , where $\hat{\pi}_A, \hat{\pi}_B, \hat{D}_A, \hat{D}_B$ are estimates of π_A, π_B, D_A, D_B.

- composite correlation

$$\gamma_{AB} = \frac{\Delta_{AB}}{\sqrt{(\pi_A + D_A)(\pi_B + D_B)}}$$

which can be estimated by

$$\widehat{\gamma}_{AB} = \frac{\widehat{\Delta}_{AB}}{\sqrt{(\widehat{\pi}_A + \hat{D}_A)(\widehat{\pi}_B + \hat{D}_B)}}$$

HW: Let

$$X_1 = \begin{cases} 2 & AA \\ 1 & Aa \\ 0 & aa \end{cases}$$

$$X_2 = \begin{cases} 2 & BB \\ 1 & Bb \\ 0 & bb \end{cases}$$

and r =Pearson correlation coefficient between X_1 and X_2. Show that $\widehat{\gamma}_{AB} = r$.

- Sum-of-squared-differences statistic that measures the overall difference in pairwise LD:

$$Z = \text{Trace}[(R_Y - R_N)'(R_Y - R_N)]$$

where R_Y=the matrix of the composite LD correlation for the case group, and R_N is that for the control group.

The significant level (P value) of the statistic Z can be assessed via permutation procedure by permuting the case-control status among the subjects.