Family-based Association Studies

- Why Family-based Association Study

  Although genetic linkage studies, which collects pedigrees with affected individuals to find disease genes, have been used successfully to map simple Mendelian diseases, they have not yielded consistent evidence for mapping complex disease genes.

  Theoretical studies have shown that linkage methods may be inferior in power compared to studies that directly utilize allelic association. Practical studies also confirm this. For example, evidence from conventional linkage studies for the involvement of insulin gene region in insulin-dependent diabetes mellitus lagged behind that from association studies.

Population-based case-control association studies have been a popular alternative to the linkage studies. However, population stratification and admixture may lead to spurious association, and hence biased finding on the gene-disease association.

Family-based association designs offer a compromise between traditional linkage studies and case-control association studies. In these studies, association is assessed within family, and hence the confounding due to population heterogeneity can be eliminated.
Moreover, compared to the linkage analysis, family-based association analysis can be more powerful in detecting moderate size of genetic effects (as low as relative risk of 1.5).

– Transmission Disequilibrium Test (TDT) :
the most popular family-based association analysis.

– Case-Parent (Trio) Design (nuclear family):
each family contains 3 memernbers: one affected child and his/her two parents, all genotyped at one or more genetic markers.

– TDT for the simplest case: one biallelic marker.
Given the parental genotype data, we can determine the alleles transmitted to the offspring and the alleles not transmitted from the parents. Viewing the transmitted pair of alleles as 'case' and the non-transmitted pair of alleles as 'control', and recognizing that each

trio (family) contributes a ‘matched’ pair of ‘case’ and ‘control’, we can summarize the data as the following 2X2 table:

		Non-transmitted allele	
		M	m
transmitted	M	a	b
allele	m	c	d

and apply the McNemar test (originally developed for the matched case-control study) to test for linkage or association between the disease gene and the marker:

$$T = \frac{(b-c)^2}{b+c} \overset{H_o}{\sim} \chi^2(df = 1)$$

where H_o: no linkage or association. (Note that the homozygous parents do not contribute to the test statistic). This test is named the Transmission/Disequilibrium Test (TDT).

– Limitation of TDT :

- ∗ For families containing two or more affected children, TDT can use only one of these children since TDT is invalid (i.e. has incorrect type I error rate) when using multiple offsprings in a family.

- ∗ Using only affected offspring.

- Some Generalizations of TDT

  - model-based TDT

    g_c, g_f, g_m: the marker genotypes of the affected offspring and the parents (father and mother).

D: the event that the offspring is affected.

– conditional on parental genotype likelihood :

$$L = \prod_{i=1}^{n} Pr(g_{ci}|g_{mi}, g_{fi}, D)$$

where

$$Pr(g_c|g_m, g_f, D) = \frac{P(D|g_c, g_m, g_f)P(g_c|g_m, g_f)}{\sum_{g^* \in G} P(D|g_*, g_m, g_f)P(g_*|g_m, g_f)}$$

where g_* is one of the four possible genotypes of the offspring conditional on parental genotypes.

Assuming

$$P(D|g_c, g_m, g_f) = P(D|g_c)$$

$$Pr(g_c|g_m, g_f, D) = \frac{r(g_c)}{\sum_{g_* \in G} r(g_*)}$$

where $r(g) = \frac{Pr(D|g)}{Pr(D|g_o)}$ is the genotype relative risk, with g_o representing a chosen reference genotype.

– in general, for H distinct alleles, we require $H(H+1)/2$ distinct relative risk parameters.

– simplified models

we can model

$$\mathsf{log}[r(g)] = X'\beta$$

where X is the coded vector of genotype g. The null hypothesis of no association can be performed by testing $H_o : \beta = 0$ using the score statistic

$$S = U'V^{-1}U$$

where $U = \frac{\partial \ln L}{\partial \beta}|_{\beta=0}$ and $V_{ij} = -E[\frac{\partial^2 \ln L}{\partial \beta_i \partial \beta_j}]_{\beta=0}$

$$S \overset{H_o}{\sim} \chi^2 \text{ (df=\# of components in } \beta)$$

- Association Analysis Incorporating Unaffected Offsprings

  – General phenotype Y, coded genotype X

  Assuming generalized linear model:

  $Y_{ij}|X_{ij} \sim$ exponential family

  Where $i = 1, ..., n$ indexes families, $j = 1, ..., n_i$ indexes the offsprings in family i .

  i.e.

$$f(y_{ij}|x_{ij} = \exp[\frac{y_{ij}\eta_{ij}}{a(\phi)} - b(\eta_{ij}) + C(y_{ij})],$$

where $\eta_{ij} = \beta_0 + \beta_1 X_{ij}$ is a canonical link function, and

$$E(y_{ij}) \equiv \mu_{ij} = b'(\eta_{ij})$$

$$Var(y_{ij}) = b''(\eta_{ij})a(\phi)$$

ex: Some common distributions

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

– log-likelihood

$$\log L(\beta_0, \beta_1) \propto \sum_{ij} [Y_{ij}\eta_{ij} - b(\eta_{ij})]$$

– score function with respect to β_1:

$$U(\beta_0, \beta_1) = \frac{\partial}{\partial \beta} \log L(\beta_0, \beta_1) = \sum_{ij} X_{ij}(Y_{ij} - \mu_{ij})$$

– score statistic for testing $H_o : \beta_1 = 0$

$$S = U(\beta_0, \beta_1 = 0) = \sum X_{ij}(Y_{ij} - \mu_0)$$

where μ_0 is the population mean of Y under H_o,

$[S - E(S)]' Var(S) [S - E(S)] \overset{H_o}{\sim} \chi^2$ (df= dimension of S)

$E(S)$ and $Var(S)$ are computed using permutation distribution with known parental genotype.

- Special case for binary outcome:

  When $Y_{ij} = 1$ if affected and 0 if unaffected, $X_{ij} = \#$ of M alleles in the jth offspring of the ith family

$$S = (1 - \mu_0)S_a - \mu_0 S_u$$

S_a : # of M alleles transmitted to the affected offspring

S_u : # of M alleles transmitted to the unaffected offspring

Note: when $\mu_0 \approx 0$ (rare disease) $S \approx S_a$ (in this case most information is contained in the affected individuals).

HW: Under H_o: no association , show that

$$S_a - E(S_a) = \frac{b_a - c_a}{2}$$

$$Var(S_a) = \frac{b_a + c_a}{4}$$

where b_a and c_a are respectively # of times that M is transmitted and not transmitted from heterozygous parents to the affected offspring.

This implies that when $S = S_a$ the resulting test is just TDT. In general when $0 < \mu_0 < 1$, under H_o and given parental genotypes we have

$$\begin{aligned} E(S) &= (1-\mu_0)E(S_a) - \mu_0 E(S_u) \\ &= (1-\mu_0)\frac{b_a - c_a}{2} - \mu_0 \frac{b_u - c_u}{2} \end{aligned}$$

Where b_u and c_u are respectively # of times that M is transmitted and not transmitted from heterozygous parents to the unaffected offspring

$$Var(S) = (1-\mu_0)^2 \frac{b_a + c_a}{4} + \mu_0^2 \frac{b_u + c_u}{4}$$

$$\frac{[S - E(S)]^2}{Var(S)} = \frac{[(1-\mu_0)(b_a - c_a) - \mu_0(b_u - c_u)]^2}{(1-\mu_0)^2(b_a + c_a) + \mu_0^2(b_u + c_u)}$$

$$\overset{H_o}{\sim} \chi^2(df = 1)$$

* Note : $\mu_0 = 0$ leads to TDT.

* Note : $\mu_0 = E(y)$ leads to most powerful test under multiplicative relative-risk model.

- TDT Without Parental Genotypes

  When parental data are unavailable, information on the parental genotypes may be contained in the genotypes of the proband and his/her siblings.

  Sib-TDT:

  – there are at least one affected sib and one unaffected sib in each sibship

  – the sibs must not have the same genotype

- Idea : whether the marker allele frequencies among affected offspring differ significantly from the frequencies among their unaffected sibs.

- a valid test of linkage

- a valid test of association if each family consists only one affected.

  When a family consists of more than one affected offspring, the transmission of the marker from parents to the offsprings are not independent when there exists linkage.

HW: Explain why sib-TDT is not valid as a test for association when there is more than one affected child in a family, and propose a possible way to correct the bias.

- construction of Sib-TDT

For a family:

	Aff.	Unaff.	
M_1M_1	x		r
M_1M_2	y		s
M_2M_2			
	a	u	t

Under H_o (affected status is independent of the genotypes), for fixed marginals a,u,r,s, x and y are hypergeometric and we can show that

$$E(x) = ra/t, \qquad Var(x) = r(t-r)au/[t^2(t-1)]$$
$$E(y) = sa/t, \qquad Var(y) = s(t-s)au/[t^2(t-1)]$$

$$Cov(x, y) = -rsau/[t^2(t-1)]$$

Accordingly, the number $N = 2x + y$ of the M allele among the affected sib in a family has mean $A = (2r + s)a/t$ and variance $V = au[4r(t - r - s) + s(t - s)]/[t^2(t - 1)]$

HW: Derive all the above identities.

The Sib-TDT used the total number of M_1 alleles across all sibships (families) as the test statistic:

$$\frac{\sum N_i - \sum A_i}{\sqrt{\sum V_i}} \overset{H_o}{\sim} N(0,1)$$

– A permutation test equivalent to Sib-TDT:
Within each family, we choose randomly affected and unaffected sibs regardless of the genotypes, with the number of affected and unaffected sibs equal to the actual numbers. Then calculate the total number of M_1 alleles in affected sibs across families. Replicate the sampling a large number of times, we can then get a simulation distribution for the number of M_1 alleles in the affected sibs under H_o by which we can obtain an empirical P value for the observed data.

- General Likelihood Principle for Family-based Association Study

  - $g_i = (g_{M_i}, g_{F_i})$: parental genotypes
    $h_i = (h_{i1}, ..., h_{in_i})$: offspring genotypes
    $\Phi_i = (\Phi_{iF}, \phi_{iNF})$: observed phenotypes
    for family members.

  - The ith family's likelihood:

$$L_i(\theta; g_i, h_i) = P(g_i, h_i | \Phi_i; \theta)$$

$$\begin{aligned} &= \frac{P(g_i;\gamma)P(\Phi_{iF}|g_i;\beta)P(h_i|g_i)P(\Phi_{iNF}|h_i;\beta)}{\sum_{g'} P(g'\gamma)P(\Phi_{iF}|g';\beta)\sum_{h''} P(h''|g')P(\Phi_{iNF}|h'';\beta)} \\ &\equiv L_i^F(\theta;g_i)\times L_i^{NF}(\beta;h_i|g_i) \end{aligned}$$

where

$$L_i^F(\theta;g_i) = \frac{P(g_i;\gamma)P(\Phi_{iF}|g_i;\beta)\sum_{h'} P(h'|g)P(\Phi_{iNF}|h';\beta)}{\sum_{g'} P(g';\gamma)P(\Phi_{iF}|g';\beta)\sum_{h''} P(h''|g')P(\Phi_{iNF}|h'';\beta}$$

⇒ “Founder-Likelihood”;

$$L_i^{NF}(\beta;h_i|g_i) = \frac{P(h_i|g_i)P(\Phi_{iNF}|h_i;\beta)}{\sum_{h'} P(h'_i|g)P(\Phi_{iNF}|h'_i;\beta)}$$

⇒ “Non-Founder Likelihood”

– Score function:

$$u_i^F(\theta; g) = \left(\begin{array}{c} \frac{\partial}{\partial\beta} \log L_i^F(\theta; g_i) \\ \frac{\partial}{\partial\gamma} \log L_i^F(\theta; g_i) \end{array} \right)_{\beta=0} = \left(\begin{array}{c} u_{i1}^F(\theta; g_i) \\ u_{i2}^F(\theta; g_i) \end{array} \right)$$

$$\begin{aligned} u_{i1}^F(\theta; g_i) &= \frac{\partial}{\partial\beta} \log P(\Phi_{iF}|g_i; \beta)|_{\beta=0} + \\ &\frac{\partial}{\partial\beta} \log \frac{\sum_{h'} P(h'|g)P(\Phi_{iNF}|h'; \beta)}{\sum_{g'} P(g'; \gamma)P(\Phi_{iF}|g'; \beta) \sum_{h''} P(h''|g')P(\Phi_{iNF}|h''; \beta)}|_{\beta=0} \\ &= a(\Phi_{iF})Z(g_i) + a(\phi_{iNF})E[Z(h)|g] - \\ &\quad \{a(\Phi_{iF})E[Z(g_i)] + a(\phi_{iNF})E[Z(h)]\} \\ &= a(\Phi_{iF})\{Z(g_i) - E[Z(g_i)]\} + \\ &\quad a(\phi_{iNF})\{E[Z(h)|g] - E[Z(h)]\} \end{aligned}$$

where $\frac{\partial}{\partial\beta} \log P(\phi|x; \beta) = a(\phi)Z(x)$; ϕ : phenotype,

$Z(x)$: covariate for gene factors

$$u_i^{NF}(\theta; g) = \left(\begin{array}{c} \frac{\partial}{\partial\beta} \log L_i^{NF}(\theta; h_i|g_i) \\ \frac{\partial}{\partial\gamma} \log L_i^{NF}(\theta; h_i|g_i) \end{array} \right) = \left(\begin{array}{c} u_{i1}^{NF}(\theta; g_i) \\ 0 \end{array} \right)$$

$$\begin{aligned} u_{i1}^{NF}(\theta; g) &= \frac{\partial}{\partial\beta} \log P(\Phi_{iNF}|h_i; \beta) \\ &- \frac{\sum_{h'} \frac{\partial}{\partial\beta} log P(\Phi_{iNF}|h_i'; \beta) P(\Phi_{iNF}|h_i'; \beta) p(h_i|g_i)}{\sum_{h''} P(\Phi_{iNF}|h_i''; \beta) P(h_i''|g_i)}|_{\beta=0} \\ &= a(\Phi_{iNF}) Z(h_i) - a(\phi_{iNF}) E[Z(h)|g_i] \\ &= a(\Phi_{iNF}) \{Z(h_i) - E[Z(h)|g_i]\} \end{aligned}$$

$\star$ Note:

$E[Z(h)|g]$ is evaluated under H_o: $g(h)$ is unrelated to $\phi_F(\phi_{NF})$, hence is determined only by Mendel's Law.

– Examples:

(1) qualitative trait ($\phi = 0/1$)

Logistic model:

$$P(\phi = 1|Z(x)) = \frac{e^{[\beta_0 + \beta Z(x)]}}{1 + e^{[\beta_0 + \beta Z(x)]}}$$

Hence $a(\phi) = \phi - \frac{e^{\beta_0}}{1+e^{\beta_0}}$

(2) quantitative trait (ϕ ~Normal)

Linear model:

$$P(\phi|Z(x)) \propto \exp\{-\frac{1}{2}(\phi - \beta_0 - \beta Z(x))^2\};$$

$$E(\phi|Z(x)) = \beta_0 + \beta Z(x)$$

Hence $a(\phi) = \phi - \beta_0$

(3) censored time (e.g. age at onset)

$\phi = (t, \delta)$,t=censored time, δ=I(t is the actual failure time)

Proportional hazards model:

$$\lambda(t|Z(x)) = \exp[\beta_0(t) + \beta Z(x)]$$

where $\lambda(\cdot)$ denotes hazard function;

$$p(\phi|Z(x)) \propto \exp\{[\beta_0(t) + \beta Z(x)]\delta - e^{\beta Z(x)}\Lambda_0(t)\}$$

where $\Lambda_0(t) = \int_0^t e^{\beta_0(s)} ds$.

Here, $a(\phi) = \delta - \Lambda_0(t)$.

HW: Derive the $a(\phi)'s$ in (1)-(3).

- Remark:

  The validity of the score test based on the founder likelihood depends on the assumed genotype distribution $P(g;\gamma)$, or on the assumed model for the phenotype, hence is less robust than the non-founder statistic. But, if the assumed model is correct, the founder statistic is more powerful than the non-founder statistic.