行政院國家科學委員會專題研究計畫　成果報告

三元體家庭資料連鎖不平衡檢定方法群之發展理論研究

(3/3)

計畫類別：個別型計畫
計畫編號：NSC92-2118-M-002-011-
執行期間：92 年 08 月 01 日至 93 年 07 月 31 日
執行單位：國立臺灣大學公共衛生學院流行病學研究所

計畫主持人：戴政

報告類型：完整報告

處理方式：本計畫可公開查詢

中　華　民　國 93 年 8 月 18 日

摘 要

利用三元體(父、母、一個染病孩子) 資料之標識基因是否由親代傳至染病子代之傳遞-不傳遞訊息來建立檢定連鎖方法，已有許多被發展出來。推廣這些方法至其它方面，諸如：多對偶基因、缺失資料、異質性、數量性狀、多標識基因等問題，亦已有許多研究可見。雖然如此，此方面仍存有二個主要問題值得探究：(1) 在一般化之遺傳模式下 (如顯、隱、累加、相乘模式)，影響連鎖之干擾參數為何？ (2) 在一般化模式下，除已提出之連鎖檢定方法外，是否還有其它？在本論文中，將先導出三元體資料於一般化遺傳模式下之標識基因傳遞-不傳遞資料結構，由此資料結構推得於無連鎖情形下的邊際同質與對稱性，利用這些特性建立九種檢定連鎖統計量 (其中三種為新的方法)，並推導出於一般化遺傳模式下對應於各統計量之干擾參數組成結構。最後，以模擬方法探討干擾參數對檢定連鎖的影響。

關鍵字：遺傳模式；邊際同質性；對稱性；傳遞

Abstract

Using the information of transmission-nontransmission of a parental marker allele to the affected child in a case-parent triad family, currently there have been many transmission linkage tests proposed for the inference of linkage between a marker and a disease-susceptibility gene. Extensions of these methods to a variety of problems such as multiple alleles, parent-missing, genetic heterogeneity and quantitative trait loci were also widely studied. However, there still remain a couple of problems that are worth studying: (1) There is not a comprehensive study that investigates all the possible transmission linkage tests under the general disease models. Therefore, it is not clear whether there still are, other than the existing ones, some statistics that are valid for the test of linkage and how they perform. (2) It is known that there are several nuisance parameters (linkage disequilibrium, marker and disease allele frequencies, penetrances of disease genes) that can interfere with the test of linkage, but it is not clear that how they jointly interfere with the test of linkage under the general disease models. Here, we distinguish three types of transmission-nontransmission data structure generated from the case-parent triad data under the general disease models. For the three types of data structure we discover twelve properties of symmetry or marginal homogeneity under the null hypothesis of no linkage. Based on these properties, a systematic class of nine transmission linkage tests under general disease models is developed. Three in it are new. The joint interference structures of the studied nuisance parameters for the respective tests are formulated. Simulations are conducted to examine the joint effect of the nuisance parameters on the test of linkage under the general situations.

Key words: inheritance mode; marginal homogeneity; symmetry; transmission

Purpose of the Project

It is known that association could interfere with the test of linkage in case-parent triad (CPT) data under the recessive mode of inheritance (MOI) of a disease. Yet, it is not very clear that what other nuisance parameters may also interfere with the test of linkage if the recessive disease model is extended to the general models; and what other tests exist beyond the known ones. Here, we look into the possible nuisance parameters that may interfere with the test of linkage and investigate the valid linkage test statistics for CPT data under general disease models. We will first identify three types of transmission-nontransmission data structure for CPT data by relaxing the recessive assumption. For each type of data structure we formulate the relevant properties under the null hypothesis of no linkage between two loci. Utilizing these properties we systematically construct a class of valid linkage test statistics for CPT data. Last, we will investigate the power performance of all the methods.

Method

P-Based, PM-Based and CPM–Based Transmission-Nontransmission Probabilities

Consider a disease locus and a marker locus in a study. There are two alleles A and a at the disease locus and two codominant alleles B_1 and B_2 at the marker locus. The three genotypes AA, Aa and aa of the disease locus may all lead to the occurrence of the disease. Let p and q be the frequencies of A and a and r and s be the frequencies of B_1 and B_2, respectively. Assume that the population is in Hardy-Weinberg equilibrium. The frequencies of the four haplotypes at the two loci are $P(AB_1) = x_1 = pr + D$, $P(AB_2) = x_2 = ps - D$, $P(aB_1) = x_3 = qr - D$, $P(aB_2) = x_4 = qs + D$, where D is the linkage disequilibrium or association parameter. Denote the recombination fraction between two loci by θ, the penetrances of the genotypes AA, Aa and aa by f_2, f_1 and f_0, and the prevalence rate in the population by ψ, $\psi = p^2f_2 + 2pqf_1 + q^2f_0$. The four parental marker alleles in a CPT family are classified into two groups: the transmission group (or the case group) and the nontransmission group (or the pseudocontrol group). Denote the four possible haplotypes in the transmission group by $\underline{aB}_1$, $\underline{aB}_2$, $\underline{AB}_1$ and $\underline{AB}_2$, and the two possible marker alleles in the nontransmission group by $\overline{B}_1$ and $\overline{B}_2$, where the "underline" labels the alleles transmitted from

parents to the affected children and the "upper bar" labels those nontransmitted. Ott (1989) derived the conditional joint distribution of transmitted and nontransmitted marker alleles for one parent in CPT families for a recessive disease. Extending Ott's result to the general MOI, Knapp, Seuchter and Baur (1993) derived the conditional joint distribution of transmitted and nontransmitted marker genotypes in CPT families. Following their derivations, we can immediately write down the conditional joint distribution of transmitted haplotypes and nontransmitted marker alleles for one parent in CPT families for the general MOI. Because in practical studies we can only observe marker genotypes of affected children (i.e., we can not distinguish between $\underline{aB}_1$ and $\underline{AB}_1$ and between $\underline{aB}_2$ and $\underline{AB}_2$), combining the undistinguishable haplotypes together we obtain the conditional joint distribution of transmitted and nontransmitted marker alleles for one parent for the general MOI (Table 1). We call the probabilities in Table 1 the parent-based transmission- nontransmission probabilities (P-based TNTP).

According to the marker genotype of the affected child in the family, the four parental marker alleles can be identified as either transmitted or nontransmitted. Let $\underline{B}_1/\underline{B}_1$, $\underline{B}_1/\underline{B}_2$ and $\underline{B}_2/\underline{B}_2$ represent the three possible assorted types of the two transmitted parental alleles, and $\overline{B}_1/\overline{B}_1$, $\overline{B}_1/\overline{B}_2$ and $\overline{B}_2/\overline{B}_2$ represent the three possible assorted types of the two nontransmitted parental alleles. Based on such a (transmission-nontransmission) classification of the four parental alleles, the parental matings in a population are classified into nine (3×3) categories. Under the assumptions: (i) no selection or mutation, (ii) Hardy-Weinberg equilibrium for marker-disease genotypes, and (iii) random mating with respect to marker and disease genotypes, the corresponding conditional joint probabilities for these nine categories were derived. We call these probabilities the parental mating-based transmission-nontransmission probabilities (PM-based TNTP).

In practical analysis, a 3×3 table may be condensed to a 2×2 table so that some simple statistics such as t, χ^2, log-likelihood ratio statistic G^2 or odds ratio can be used. Combining the two rows $\underline{B}_1/\underline{B}_2$ and $\underline{B}_2/\underline{B}_2$ in Table 3 into one category and the two columns $\overline{B}_1/\overline{B}_2$ and $\overline{B}_2/\overline{B}_2$ into one category as well, the observed numbers and conditional joint probabilities of the four condensed parental mating types are also derived. These probabilities will be called condensed parental mating-based

transmission-nontransmission probabilities (CPM-based TNTP).

The P-based, PM-based and CPM-based TNTPs are composed of four parts: (1) the allele frequencies r and s at the marker, (2) the allele frequencies p and q and the penetrance parameters f_2, f_1 and f_0 at the disease locus, (3) the association parameter D and (4) the recombination fraction θ. Let $\tau' = (f_2-f_1)-(f_1-f_0)=f_2-2f_1+f_0$, which is defined as an MOI index of the disease. If the MOI of the disease is recessive ($f_1=f_0$), $\tau' = f_2-f_1= f_2-f_0$. If the MOI of the disease is dominant ($f_1=f_2$), $\tau'=-(f_2-f_0)=-(f_1-f_0)$. If the MOI of the disease is additive ($f_1=(f_2+f_0)/2$), $\tau'=0$. If the MOI of the disease is multiplicative ($f_1=\sqrt{f_2 f_0}$), $\tau'=(f_2+f_0)-2\sqrt{f_2 f_0}$. It can be verified that τ' is an unambiguous index that can completely distinguish the four genetic models. Also, let $\beta' = p(f_2 - f_1) + q(f_1 - f_0)$, which can be regarded as the average gene substitution effect of the disease locus.

Symmetry and Marginal Homogeneity Properties of the Three Types of TNTP

Based on the P-based TNTP and under the $H_0^{(1)}: (1 - 2\theta)D\beta = 0$, overall we can find three properties for the development of valid statistics for test of linkage: property (i): $u_{12} = u_{21}$ (symmetry), property (ii): $u_{1.} = u_{.1}$ (marginal homogeneity), property (iii): $u_{2.} = u_{.2}$ (marginal homogeneity). Note that as $\theta D \beta = 0$, $u_{ij} = u_{i.}\, u_{.j}$ (independence). Hence, a P-based statistic developed for testing independence (e.g., Pearson's χ^2 for 2×2 tables) is invalid for test of linkage.

For the PM-based TNTP, we have the following results: (1) under $H_0^{(2)}: (1 - 2\theta)(rE + F) = (1 - 2\theta)(rD\beta + D^2\tau) = (1 - 2\theta)D(r\beta + D\tau) = 0$, $v_{12} = v_{21}$; (2) under $H_0^{(3)}: (1 - 2\theta)[2rsE + (s - r)F] = (1 - 2\theta)[2rsD\beta + (s - r)D^2\tau] = (1 - 2\theta)D[2r(1-r)\beta + (1 - 2r)D\tau] = 0$, $v_{13} = v_{31}$; (3) under $H_0^{(4)}: (1 - 2\theta)(sE - F) = (1 - 2\theta)(sD\beta - D^2\tau) = (1 - 2\theta)D((1-r)\beta - D\tau) = 0$, $v_{23} = v_{32}$; (4) under $H_0^{(5)}: (1 - 2\theta)(2rE + F) = (1 - 2\theta)(2rD\beta + D^2\tau) = (1 - 2\theta)D(2r\beta + D\tau) = 0$, $v_{1.} = v_{.1}$; (5) under $H_0^{(6)}: (1 - 2\theta)[(s - r)E - F] = (1 - 2\theta)[(s - r)D\beta - D^2\tau] = (1 - 2\theta)D[(1 - 2r)\beta - D\tau] = 0$, $v_{2.} = v_{.2}$; (6) under $H_0^{(7)}: (1 - 2\theta)(2sE - F) = (1 - 2\theta)$

$(2sD\beta - D^2\tau) = (1 - 2\theta)D(2(1\text{-}r)\beta - D\tau) = 0$, $v_{3.} = v_{.3}$; (7). In summary, under the respective $H_0^{(k)}$, $k = 2, 3, \ldots, 7$, there are six properties that can be used for the development of valid statistics for test of linkage: property (iv): $v_{12} = v_{21}$ (symmetry), property (v): $v_{13} = v_{31}$ (symmetry), property (vi): $v_{23} = v_{32}$ (symmetry), property (vii): $v_{1.} = v_{.1}$ (marginal homogeneity), property (viii): $v_{2.} = v_{.2}$ (marginal homogeneity), property (ix): $v_{3.} = v_{.3}$ (marginal homogeneity). Note that a PM-based statistic developed for testing independence is also invalid for test of linkage.

For the CPM-based TNTP it can be shown that under $H_0^{(8)}$: $(1 - 2\theta)(2rE + F)$ $= (1 - 2\theta)(2rD\beta + D^2\tau) = (1 - 2\theta)D(2r\beta + D\tau) = 0$, $w_{12} = w_{21}$, there are three properties that can be used for the development of valid statistics for test of linkage: property (x): $w_{12} = w_{21}$ (symmetry), property (xi): $w_{1.} = w_{.1}$ (marginal homogeneity), property (xii): $w_{2.} = w_{.2}$ (marginal homogeneity). Again, the independence test using the CPM-based TNTP is invalid for test of linkage.

P-Based, PM-Based and CPM-Based Transmission Linkage Disequilibrium Tests

Since the linkage disequilibrium parameter D interferes with all tests of linkage under the eight null hypotheses $H_0^{(k)}$'s, tests developed under these hypotheses are exactly to test linkage and linkage disequilibrium rather than linkage only. Based on the twelve properties demonstrated in the above section, nine linkage disequilibrium tests are derived :

$$T_1 = \frac{(m_{12} - m_{21})^2}{(m_{12} + m_{21})},$$

$$T_2 = \frac{(m_{1.} - m_{.1})^2}{(m_{1.} + m_{.1})} + \frac{(m_{2.} - m_{.2})^2}{(m_{2.} + m_{.2})},$$

$$T_3 = \frac{m_{1.}/m_{2.}}{m_{.1}/m_{.2}} = \frac{m_{1.}m_{.2}}{m_{2.}m_{.1}} \quad \text{(new)},$$

$$T_4 = \frac{(n_{12}+n_{13}+n_{23}-n_{21}-n_{31}-n_{32})^2}{n_{12}+n_{13}+n_{23}+n_{21}+n_{31}+n_{32}},$$

$$T_5 = \frac{(n_{12}-n_{21})^2}{n_{12}+n_{21}} + \frac{(n_{13}-n_{31})^2}{n_{13}+n_{31}} + \frac{(n_{23}-n_{32})^2}{n_{23}+n_{32}} \quad \text{(new)},$$

$$T_6 = \frac{(n_{1.} - n_{.1})^2}{n_{1.} + n_{.1}} + \frac{(n_{2.} - n_{.2})^2}{n_{2.} + n_{.2}} + \frac{(n_{3.} - n_{.3})^2}{n_{3.} + n_{.3}} \quad \text{(new)},$$

$$T_7 = \frac{(o_{12} - o_{21})^2}{o_{12} + o_{21}} = \frac{(n_{12} + n_{13} - n_{21} - n_{31})^2}{n_{12} + n_{13} + n_{21} + n_{31}},$$

$$T_8 = \frac{(o_{1.} - o_{.1})^2}{o_{1.} + o_{.1}} + \frac{(o_{2.} - o_{.2})^2}{o_{2.} + o_{.2}},$$

$$T_9 = \frac{o_{1.} o_{.2}}{o_{2.} o_{.1}}.$$

Simulation Results

Our simulation results indicate that:

(1) In most situations the three P-based statistics T_1, T_2 and T_3 are more powerful than other statistics.

(2) The nine test statistics are more powerful under the models of greater $|E|$ values. When $|E| \leq 0.01$, no matter how small the θ value is, the power of any statistic approaches the α level.

(3) Because the magnitude of E affects the testing power and the magnitude of E is affected by the disease gene frequency, the testing power is affected by q. In our study with the decrease of q value, the powers of the nine tests have the tendency to increase under the recessive and dominant models, but to decrease under the additive and multiplicative models.

(4) When the sample size is 50, no matter under which model all tests show low powers. High powers are observed when the number of CPT families increases to 300 for the recessive and dominant models.

Discussion

The conventional linkage analysis methods collect pedigree data to perform parametric inference of linkage between a disease gene and a marker. These methods are usually involved in complex sampling and cumbersome computation. The methods developed with the use of case-parent triad data avoid these disadvantages and stratification problem, but they have the limitation that their powers may be low

unless the detected locus is closely linked to the disease locus. In the previous sections we have extended the recessive disease model to general situations and demonstrated three ways of extracting the transmission and nontransmission information from a CPT family. The PM-based TNTP is the basic one from which the P-based and CPM-based TNTPs evolve. Based on the three types of TNTP and under the null hypothesis of either $\theta = 1/2$ (no linkage), $D = 0$ (no association) or certain nuisance composition structures, twelve properties of symmetry and marginal homogeneity are formulated for the development of linkage disequilibrium test statistics. T_1, T_4 and T_7 are three McNemar-type statistics, which are derived by the symmetric properties of P-based, PM-based and CPM-based TNTPs respectively. T_2, T_5, T_6 and T_8 are developed with the idea of combining two or more binomial statistics. The two odds-ratio-type statistics are T_3 and T_9. It appears that T_3, T_5 and T_6 have not been discussed before. The power performance of the TDT under different modes of inheritance had been investigated by different studies (e.g., Knapp, 1999). Our simulation studies for the testing powers of the nine statistics T_1, T_2,…, T_9 indicate that these statistics may be suitable for the analysis of the recessive and dominant disease models but not suitable for the additive and multiplicative models. Nevertheless, the P-based methods (T_1, T_2, T_3) are generally the best choices under all disease models. Effective designs should be developed to enhance the test powers for the additive and multiplicative models (e.g., Zheng et al., 2002).

計畫成果自評：

傳遞連鎖檢定方法屬於無母數檢定方法一種，由於收集資料較家族資料容易，故常被考慮為人類基因定位時的一種實作方法。本研究討論了九種檢定方法，其中三種屬於新的方法；此外，本研究更清楚推導出干擾這些檢定的干擾參數聯合結構形式。整體而言，這些新的結果都是以往所未知的。非常感謝國科會自然處支持本計畫。

References

Knapp M (1999). A note on power approximations for the transmission/ disequilibrium test. Am J Hum Genet; 64:1177-1185.

Knapp M, Seuchter SA, Baur MP (1993). The haplotye-relative-risk (HRR) method for analysis of association in nuclear families. Am J Hum Genet; 52:1085-1093.

Ott J (1989). Statistical properties of the haplotype relative risk. Genet Epidemiol; 6:127-130.

Zheng G, Freidlin B, and Gastwirth JL (2002). Robust TDT-type candidate-gene association tests. Ann Hum Genet 2002; 66:145-155.

Table 1

Conditional joint probabilities of transmitted and nontransmitted marker alleles for one parent in CPT families for general disease models (P-based TNTP)

Transmitted allele	Nontransmitted allele		
	$\overline{B}_1$ (j=1)	$\overline{B}_2$ (j=2)	Sum
$\underline{B}_1$ (i=1)	$\underline{B}_1\overline{B}_1$ *	$\underline{B}_1\overline{B}_2$	$\underline{B}_1$
	u_{11} **	u_{12}	$u_{1.}$
	$= r^2 + rE$	$= rs + sE - \theta E$	$= r + E - \theta E$
	m_{11} ***	m_{12}	$m_{1.}$
	$= 2n_{11} + n_{12} + n_{21} + n_{221}$	$= n_{12} + 2n_{13} + n_{222} + n_{23}$	$= 2n_{1.} + n_{2.}$
$\underline{B}_2$ (i=2)	$\overline{B}_1\underline{B}_2$	$\underline{B}_2\overline{B}_2$	$\underline{B}_2$
	u_{21}	u_{22}	$u_{2.}$
	$= rs - rE + \theta E$	$= s^2 - sE$	$= s - E + \theta E$
	m_{21}	m_{22}	$m_{2.}$
	$= n_{21} + n_{222} + 2n_{31} + n_{32}$	$= n_{221} + n_{23} + n_{32} + 2n_{33}$	$= 2n_{3.} + n_{2.}$
	$\overline{B}_1$	$\overline{B}_2$	
Sum	$u_{.1} = r + \theta E$	$u_{.2} = s - \theta E$	$u_{..} = 1$
	$m_{.1} = 2n_{.1} + n_{.2}$	$m_{.2} = 2n_{.3} + n_{.2}$	$m_{..} = 2n$

* : marker genotypes with transmitted and nontransmitted information

** : u_{ij}, conditional joint probabilities of parental genotypes

***: m_{ij}, observed numbers of parental genotypes, where n_{ij} is defined in Table 3