行政院國家科學委員會專題研究計畫成果報告

計畫名稱：(中文) 候選基因/範圍內密集單點核甘酸多型性之相關性研究

(英文) Candidate Gene/Region Association Study Using a Dense Array of Single Nucleotide Polymorphisms

計畫編號：NSC 90-2320-B-002-173-

執行期限：90/08/01~91/07/31

主持人：李文宗　台灣大學公共衛生學院流行病學研究所

計畫參與人員：張晉豪研究生　台灣大學公衛學院流行病學研究所

TEL: (02) 23123456-8357
FAX: (02) 23511955
E-Mail: wenchung@ha.mc.ntu.edu.tw

一、中文摘要：

為探索複雜性疾病之遺傳因素，流行病學中的相關性研究方法會越來越扮演重要的角色。近來，以分析病例雙親三合體資料的傳遞不平衡檢定法（以下簡稱 TDT 法）頗廣為採用。在基因標記和疾病易感受性基因兩者有相關的前提下，TDT 可用來檢定兩者間的連鎖關係。近年來發現在很短的染色體距離內，會有相當數量的單點核甘酸基因多形性（以下簡稱 SNP）。由於在候選基因範圍內多重 TDT 的檢定彼此不獨立，使用 Bonferroni 校正多重比較的問題會使檢力過於保守。對此，本研究發展出新的分析方法(APRICOT)。新法有下列特點：1. 不需知道基因半形性；2. 不須假定族群歷史及結構；3. 可很容易決定統計檢定的水準。在虛無假設下（沒有連鎖不平衡）的電腦模擬結果顯示，無論是在多始祖突變或多結構族群，APRICOT 都有正確的顯著水準。比起傳統的 Bonferroni 校正，也有比較高的檢力。

關 鍵 字：病例雙親三合體、流行病學方法、遺傳流行病學、主成份分析、單點核甘酸多形性、傳遞不平衡檢定

二、緣由與目的

複雜人類疾病(complex human diseases)的易感受性基因(susceptibility gene)將會越來越倚重流行病學「相關」

類型的研究(association approach)作基因定位[1,2]。其中一個最被重視的方法，即為以傳遞不平衡檢定(transmission/disequilibrium test, 簡稱TDT)[3]做為基因定位的工具。

TDT的原理係檢定異合子（heterozygous）親代傳遞給有病子代某個標記基因的機率（transmission probability)，是否有偏離孟德爾遺傳定律的1/2。TDT可以視為傳遞基因與未傳遞基因間的配對研究，吾等可用McNemar test來做統計檢定。TDT是同時檢定疾病基因座與標記基因座間是否有「連鎖」（linkage）以及「相關」（或「不平衡」, disequilibrium）的一種檢定。在同質族群（homogenous population）中，連鎖不平衡（linkage disequlibrium）通常只存在非常近距離（小於1 centi- Morgan, cM以內）的兩基因座間。因此現行TDT的研究，多係針對「候選基因/範圍」(candidate gene/region)進行連鎖檢定。

近年來，由於分子生物學的快速進步以及人類基因體研究計畫（Human Genome Project）的開展，大量的標記基因可以很快速的測定。其中尤以「單點核甘酸多形性」(single nucleotide polymorphism, 簡稱SNP)標記最受注目[4]。據估計染色體上每1kb大約即可找到一個SNP。因此可以合理預期未來之候選基因/範圍的研究，會越來越倚重測定該範圍內之長串密集的SNP標記(a dense array of SNPs)。

針對靠近的SNP所進行的TDT檢定，統計學而言是不獨立的。因此傳統上Bonferroni多重檢定的校正，將造成統計檢定的保守(conservative)趨向，以及檢力(power)的損失[5]。

近年來，多位學者致力於進行所謂「多基因座TDT」(multilocus TDT)的研究[6-14]，以充分利用多個緊密連鎖標記的資訊。這些方法必須事先知道基因半型(haplotype)或是半型相位模糊(haplotype phase ambiguity)要用統計的方法解開。這種作法在面臨一長串密集的SNP標記時將會施展不開。因為吾人同時考慮很多個SNP，可能的基因半型的個數將會龐大無比，以致於利用其他方法來化解半型相位的模糊將是困難甚或不可能的。Morris及Whittaker的方法[15]則是針對基因型資料，因此並無上述問題。然而這個方法假定一個參數模型(parametric model)，以及單一「始祖突變」(ancestral mutation)。因此該法對於「族群結構」(population structure)，比如「族群分層」(population stratification)以及「族群融合」(population admixture)所可能造成連鎖檢定偽陽性率(false positive rate)過高的現象並無法免疫。該法也無法應付較複雜的「族群歷史」(population history)，比如多個始祖突變，在不同時期發生在不同的「始祖基因半型」(ancestral haplotype)之上的情況。而這與原先TDT的無母數(nonparametric)以及對於族群結構的耐受性(robustness)的精神是有違背的。

本計劃採取多變量統計中的主成份分析(principal component analysis)[16]的數學技術，發展出針對候選基因/範圍內長串密集SNP標記的多基因座TDT分析方法，並簡稱為APRICOT。

三、結果與討論

本研究計畫所發展的APRICOT有下列特點：1. 不需知道基因半型性；2. 不須假定族群歷史及結構；3.

可以很容易決定統計檢定的水準。此外，在虛無假設下（沒有連鎖不平衡）的電腦模擬結果顯示，無論是在多始祖突變或多結構族群，APRICOT 都有正確的顯著水準。比起傳統的 Bonferroni 校正，也有比較高的檢力。

四、計畫成果自評：

「多基因座 TDT」(multilocus TDT)的研究多數必須事先知道基因半型(haplotype)，若不然，至少半型相位的模糊(haplotype phase ambiguity)必須能用統計的方法化解。這種作法在面臨一長串密集的 SNP 標記時將會施展不開。Morris 及 Whittaker 的方法無法應付連鎖檢定偽陽性率(false positive rate)過高的現象，並且違背原先 TDT 的無母數(nonparametric) 以及對於族群結構的耐受性(robustness)的精神。

本研究計畫所發展的 APRICOT，則有效地解決上述問題。在後基因體時代，流行病學家將會面臨越來越多的長串密集 SNP 標記。對於長串密集 SNP 標記的資料分析而言，相信 APRICOT 將是一個較佳的分析方法。本計畫的研究成果已獲 "Epidemiology" 雜誌接受發表。

五、參考文獻

1. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996;273:1516-7
2. Khoury MJ, Yang Q. The future of genetic studies of complex human diseases: an epidemiologic perspective. Epidemiology 1998;9:350-4
3. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52:506-16
4. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbel E, Robinson E, Mittman M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson T, Lipshutz R, Chee M, Lander ES. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 1998;280:1077-82
5. McIntyre LM, Martin ER, Simonsen KL, Kaplan NL. Circumventing multiple testing: a multilocus Monte Carlo approach to testing for association. Genet Epidemiol 2000;19:18-29
6. Wilson SR. On extending the transmission/disequilibrium test (TDT). Ann Hum Genet 1997;61:151-161
7. Chiano MN, Clayton DG. Fine genetic mapping using haplotype analysis and the missing data problem. Ann Hum Genet 1998;62:55-60
8. Clayton D. A generalization of transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 1999;65:1170-7
9. Clayton D, Jones H. Transmission/disequilibrium tests for extended marker haplotype. Am J Hum Genet 1999;65:1161-9
10. Bourgain C, Genin E, Quesneville H, Clerget-Darpoux F. Search for multifactorial disease susceptibility genes in founder populations. Ann Hum Genet 2000;64:255-65
11. Dudbridge F, Koeleman BPC, Todd JA, Clayton DG. Unbiased application of the transmission/

disequilibrium test to multilocus haplotypes. Am J Hum Genet 2000;66:2009-12

12. MacLean CJ, Martin RB, Sham PC, Wang H, Straub RE, Kendler KS. The trimmed-haplotype test for linkage disequilibrium. Am J Hum Genet 2000;66:1062-75

13. Zhao H, Zhang S, Merikangas KR, Trixler M, Wildenauer DB, Sun F, Kidd KK. Transmission/ disequilibrium tests using multiple tightly linked markers. Am J Hum Genet 2000;67:936-46

14. Morris AP, Whittaker JC. Fine scale association mapping of disease loci using simplex families. Ann Hum Genet 2000;64:223-237

15. Hodge SE, Boehnke M, Spence MA. Loss of information due to ambiguous haplotyping of SNPs. Nat Genet 1999;21:360-1

16. Johnson RA, Wichern DW. Applied Multivariate Statistical Analysis. 2nd ed, Prentice-Hall, Englewood Cliffs, New Jersey, 1992

ABSTRACT

The future of genetic studies of complex human diseases will rely more and more on the epidemiologic `association' paradigm, in particular the application of the transmission/disequilibrium test (TDT) to detect linkage disequilibrium in a case-parents study. With the rapid progress in genomic studies, many single nucleotide polymorphisms will be identified and genotyped within a very short physical distance. Testing multiple tightly linked markers within a candidate gene/region with Bonferroni correction will lead to a conservative test and hence a power loss. On the other hand, current methods of multilocus TDT have shortcomings/limitations. I propose a new method to test for possible linkage between a dense array of SNPs and a disease-susceptibility gene. The new method has the following properties:(1) it does not need haplotype information; (2) it is nonparametric—it does not make specific assumptions about the population history or population structure; and (3) the calculation of the test statistic and the determination of its significance level are simple and straightforward. Monte-Carlo simulation reveals that the new method maintains the nominal significance level under the null hypothesis of no linkage disequlibrium, even under complex situations of multiple ancestral haplotypes and structured populations. It provides a substantial power asvantage over the conventional Bonferroni approach. The new method is a promising method for candidate gene testing using single nuleotide polymorphisms.

Keyword: case-parents triads; epidemiologic methods; genetic epidemiology; principal component analysis; single nucleotide polymorphism; transmission/disequilibrium test.