遺傳流行病學

張鴻俊　陳爲堅
國立台灣大學　公共衛生學院　流行病學研究所

前　言

傳統流行病學的研究主要是探討外在環境因子對疾病發生的影響，而傳統遺傳學的研究則在探討遺傳因素控制性狀表現(trait)的機制。遺傳流行病學乃是結合流行病學與遺傳學的一個新興領域，目標在探討遺傳因素與外在環境因子在疾病發生上所扮演的腳色。遺傳流行病學研究可大略分成描述性與分析性兩大類：描述性遺傳流行病學主要在研究某一已知遺傳性狀在族群中的分布，而分析性遺傳流行病學則在探討某一疾病或性狀是如何受到遺傳或環境因子的影響。近十年來，大部分的遺傳流行病學的進展大都是在分析性研究上的創新。

分析性遺傳流行病學的研究策略大概可分成幾個循序漸進的步驟[1]。首先，是利用家族病史詢問(family history approach)或直接測試家族成員的家族研究(family study approach)來評估該疾病是否有家族聚集(familial aggregation)的情形。但疾病有家族聚集的現象，也可能是共同環境因子的暴露所造成，而非必然是由於遺傳的因素。因此，第二個步驟是透過雙胞胎研究來釐清外在環境因子與遺傳因素對疾病發生的影響。同卵雙胞胎罹病的一致性如果較異卵雙胞胎高，則表示有遺傳因素的貢獻；相反的，如果同卵雙胞胎罹病的一致性並未較異卵雙胞胎高，則表示遺傳因素並未有貢獻。雙胞胎研究可能面臨的一個困難是，同卵雙胞胎也可能比異卵雙胞胎有較類似的外在暴露。若要更週全區別外在環境因子與遺傳因素對疾病發生的影響，則需透過領養研究(adoption study)，只是該類研究不易取得合適的樣本。經由雙胞胎或領養研究確定某一性狀確實有來自遺傳因子的貢獻後，接下來的第三個步驟是透過分離率分析(segregation analysis)，來調整外在環境因子的暴露，估計出可能導致疾病發生的遺傳模式(如該疾病是否爲單基因遺傳或多基因遺傳、顯性遺傳抑或是隱性遺傳、導致此疾病的基因在族群中的頻率等)。一旦能辨認出該性狀可能的遺傳模式後，第四個步驟是透過連鎖分析(linkage analysis)來進一步找出致病基因的所在位置。這個方法是收集家族成員的罹病資料，並對家族成員測試事先選定的基因標記，然後分析基因標記與致病基因是否在家族中有連鎖的現象。若有，代表致病基因就在此基因標記的附近，進一步透過分子生物學的方法，可以找出該致病基因。

由於分子生物學的快速進步與人類基因研究計劃所獲致的大量多型性基因標記，連鎖分析在遺傳流行病學上越來越重要，再加上生物統計學與電腦發展的進步，一些新的連鎖分析方法相繼被提出，如：多基因座連鎖分析方法(multipoint linkage analysis)、同宗基因子分析(identical by descent,IBD)、同基因型分析(identical by state,IBS)、半形相對危險性分析(haplotype relative risk,HRR)、基因子傳遞分布不平衡檢定(transmission/disequilibrium test,TDT)等。因此，本文的主旨在介紹這些連鎖分析方法學上的新進展，並分別針對這些研究設計方法的理論基礎與其優缺點加以討論。

Title: Genetic Epidemiology
Authors: Hung C. Chang, Wei J. Chen
Graduate Institute of Epidemiology, College of Public Health, National Taiwan University
Key Words: association, identical-by-descent, linkage, LOD score, sib-pair

LOD Score 連鎖分析

染色體在有絲分裂的過程中可能發生互換(cross-over)，因而形成重組型染色體。重組比率(recombination fraction)，θ，指的是兩條染色體上不同的兩基因座之基因子(allele，也稱作對偶基因)因爲互換而形成重組型的比率。形成重組型的比率與兩基因座間的距離有關：兩基因座間若相距較遠，則兩條染色體間比較容易發生交換；若一定有交換，則其後代有一半是重組型。因而，最大的重組比率是 1/2。相反的，兩基因座間若相距較近，則其子代屬於重組型的比率小於 1/2，我們稱此兩基因座之間有連鎖(linkage)。此外，透過估計兩基因座間的重組比率，經由換算可以推估兩基因座間在染色體上的實際距離。通常兩基因座間的重組比率如爲 1% ，約等於在染色體上相距 1 centimorgan(cM)，也就是約一百萬氮鹽基(base-pair)的距離。因爲存在有多次互換(multiple cross-over)與干擾互換(interference)的現象，重組比率與實際距離並非簡單的線性關係，必須作某些調整[2]。

典型的連鎖分析是利用 LOD score 來檢定是否有連鎖的情形，LOD score 的定義是：

$$LOD(\theta) = \log_{10}\frac{likelihood(\theta)}{likelihood(1/2)}, \theta < 1/2,$$

也就是對基因重組率爲 θ 之可能性與基因重組率爲 1/2 的可能性的比例所取的 10 的對數值。我們可以利用最大概率估計法 MLE(maximum likelihood estimation)同時估計出 θ 及 LOD score。通常以最大 LOD score 大於 3 代表基因重組率顯著異於 1/2，也就是兩基因座間的確有連鎖存在 ，並且可以依估計的基因重組率推估兩基座間的距離。然而在連鎖分析中，如果對疾病的遺傳模式、致病基因的頻率或穿透性(penetrance)的估計不確實，則會干擾 θ 及 LOD score 的正確估計。

多基因座連鎖分析

隨著分子生物學的進步，可以測定之基因標記已迅速增加。在獲得家族各成員之多種基因標記資料後，如欲分析基因間的連鎖關係，除了考慮兩兩基因座間的連鎖外，我們亦可同時考慮多基座間的連鎖關係。首先依照傳統連鎖分析，利用 LOD score 來檢定兩兩雙基因座間是否有連鎖的情形。經過傳統連鎖分析，我們可以獲得初步多基因座間的距離與排列順序。再經由 joint likelihood，以 MLE 同時估計多組基因座間的距離與排列順序，如下所示：

$$LOD = \log_{10}\frac{likelihood(\theta_{12},\theta_{23},...,\theta_{n-1n})}{likelihood(1/2,1/2,...,1/2)},$$

其中 θ_{12} 代表第一、二基因座間的基因重組率，$\theta_{n-1\ n}$ 代表第 n-1 與第 n 基因座間的基因重組率。此外，也可利用多基因座連鎖分析，推估某新的基因標記在一組已知排列順序的基因標記間可能的位置，如下所示：

$$LOD = \log_{10}\frac{likelihood(\theta_{01},\theta_{12},...,\theta_{n-1n})}{likelihood(1/2,\theta_{12},...,\theta_{n-1n})},$$

其中 θ_{01} 代表新的基因標記與第一基因座間的基因重組率，θ_{12} 代表已知排列順序之第一、二基因座間的基因重組率，$\theta_{n-1\ n}$ 代表已知排列順序之第 n-1 與第 n 基因座間的基因重組率。目前有許多可以用來進行多基因座連鎖分析的電腦軟體，如 LINKAGE[3]。

多基因座連鎖分析相較於傳統雙基因座連鎖分析的優點包括：1.有較大的檢力(power)發現連鎖的存在。2.可充份利用家族資料；隨著基因座測定數目的增加，家族資料對連鎖分析有助益(informative)的可能性越高。3.可推估染色體上基因座位置的排列順序。但多基因座連鎖分析對 θ 及 LOD score 的估計，比較容易因爲分析中所假定的疾病遺傳模式參數不正確而有偏差。

由於基因標記測定變得較爲容易，目前對於複雜疾病之連鎖分析，通常採用整個基因體掃瞄(genome-wide scan)的方式。首先在染色體上每隔 10cM 找一個基因標記，整個基因體總共大約需要使用 300 個基因標記。兩基因座連鎖分析之結果中若有某一標記之最大 LOD score 顯示有連鎖存在之潛能，再於此標記附近作更密集的基因標記測定，然後進行多基因座

連鎖分析。此種基因體掃瞄的研究方法，由於牽涉多重比較的問題，因此，作爲判斷是否顯著的LOD score閾值應該調整。至於如何調整，目前仍存有爭議[4]。

同宗基因子分析及其衍生的方法

Penrose 在 1935 年首先提出利用配對手足(sib-pair)的資料來判定兩個基因之間是否有連鎖存在。假定兩基因間未有連鎖，則配對手足在甲基因表現型(phenotype)的一致性機率應與其在乙基因表現型是否一致無關。透過檢定其獨立性可用來判斷兩基因座是否有連鎖存在。舉例來說，如果控制血型的基因與導致糖尿病的基因並無連鎖，則手足是否有相同血型並不影響手足是否同爲糖尿病患。因此，族群中並不會特別容易見到有相同血型且同樣患有糖尿病的手足，或是有不同血型且不同糖尿病罹病狀態的手足。反之，若族群中配對手足爲相同血型且相同糖尿病罹病狀態或不同血型且不同糖尿病罹病狀態的比率比預期高，則控制血型的基因與導致糖尿病的基因之間可能有連鎖存在。

Suarez 等人[5]延續此種想法，進一步提出以配對手足間共享同宗基因子(identical by descent, IBD)的程度來判定基因標記是否與未知的致病基因有連鎖。共享同宗基因子是指兩個人之基因子來自親代相同一條染色體，利用檢定資料中配對手足的某一基因標記的同宗基因子分布是否偏離期望分布，可以間接推估此一基因標記是否與致病基因連鎖。一般來說，配對手足在某一基因標記上共享兩個(IBD=2)、一個(IBD=1)、零個(IBD=0)同宗基因子的機率應是 1/4,1/2,1/4。若疾病基因與所測定基因標記並無連鎖，則配對手足的罹病狀態並不影響手足間共享同宗基因子的機率分布，應符合 IBD=2、IBD=1、IBD=0 的機率分別是 1/4,1/2,1/4 的期望分布。如果發現配對手足其同宗基因子分布明顯偏離期望分布，則表示基因標記與致病基因有連鎖的情形。一般而言，配對手足同爲罹病狀態較非同爲罹病狀態之配對手足有較大的檢力。跟傳統連鎖分析比起來，配對手足的優點主要是不需假定特定的遺傳模式(如該疾病是否爲單基因遺傳或多基因遺傳、顯性遺傳抑或是隱性遺傳、導致此疾病的基因在族群中的頻率等)，而且只需收集配對手足資料，而不似傳統連鎖分析需費盡心力收集龐大的家族資料。然而其檢力較傳統連鎖分析弱，並且未能直接估計基因重組率。

配對手足間共享的同宗基因數的判定有時並不容易，例如：父親爲同基因子合子(homozygous)，則配對手足縱使爲相同基因型也難以判斷共享同宗基因子的情形，況且有時父母親的基因型無法獲得。因此，Lange[6]提出共享同基因型分析法(identical by state, IBS)，意指基因型相似的程度；IBS=2 代表手足的基因型有 2 個相同的基因子，IBS=1 代表手足有 1 個相同的基因子，而 IBS=0 則代表手足相同的基因子個數爲 0。其如同共享同宗基因子的分析方法，比較配對手足間在某些基因標記的共享同基因型的機率分布與其期望分布，可以推估基因標記是否與未知的疾病基因相連鎖。這個方法克服共享同宗基因子不容易判定的問題。Lange 同時發現，所用的基因標記越是多型性(polymorphic)，則共享同宗基因子的結果與共享同基因型的結果就越接近。

此外，共享同宗基因子的分析並不侷限於手足之間的配對關係，其他如父子、甥舅、叔姪或爺孫等配對，也可經由比較兩者間共享同宗基因子的機率分布與期望分布，來檢定基因之間是否連鎖。此種研究分析方法稱爲罹病家族成員分析法(affected-pedigree-member method, APM)。由於罹病家族成員之間要判定共享同宗基因子數遠較於配對手足困難，因此，罹病家族成員通常以共享同基因型進行分析。但共享同基因型容易受到基因子頻率的影響：基因子頻率愈高，則配對家族成員愈容易因爲碰巧而有相同的基因型；相反的，基因子頻率愈低、基因標記越是多型性，則共享同基因型的結果與共享同宗基因子的結果就越接近。此外，若基因標記與致病基因存在連鎖，則越遠的親屬關係配對其共享同宗基因子的機

表一：2n 個親代傳遞基因子給罹病個案的分布情形

傳遞基因子	非傳遞基因子 M1	非傳遞基因子 M2	人數
M1	a	b	a+b
M2	c	d	c+d
人數	a+c	b+d	2n

率分布與期望分布差距越大。因此，越遠的親屬關係檢定是否連鎖的檢力就越大。雖然此種分析方法較傳統分析方法的檢力弱，但由於只需要檢測同爲罹病狀態的家族成員的基因標記，因此常用來作爲連鎖偵測的初步分析。當利用此種方法發現某基因標記與致病基因之間存在連鎖，才進一步檢測所有家族成員的基因標記，進行傳統的 LOD score 連鎖分析[7] 。

半形相對危險性分析

一般流行病學的研究藉由評估暴露與疾病發生的相關性來探討疾病可能的致病因或危險因子。因此，若將測定的基因標記視爲一種暴露，則應可透過類似的相關性研究(association study)來了解基因標記與疾病發生的關係。但此種相關性研究的結果可能代表的意義包括：此基因即是真正致病基因，或者是基因標記與致病基因間有連鎖不平衡(linkage disequilibrium)。另一方面，也可能由於選樣的偏差，造成假性相關，例如：研究個案中病例組與對照組的族群分布不同，會造成兩組之基因頻率有顯著的差異。爲了避免選樣偏差，利用親子對照的研究設計(case-parental study, or internal control)可以控制背景基因分布的差異。

Rubinstein 等人[8,9]採用親子對照的研究設計，提出半形相對危險性分析。其研究分析方法是經由收集罹病的研究個案與其雙親的資料，透過比較雙親將某基因座之甲、非甲兩基因子傳遞給罹病個案的比例，來代表甲基因子的相對危險性。假設該疾病爲體染色體隱性遺傳，而且父母兩人的基因型是互相獨立的(random mating)，則父母將某特定基因子甲與非甲基因子傳遞給罹病子代的比例，與不傳遞特定基因子甲與非甲基因子給罹病子代的比例之比值定義爲半形相對危險性。舉例來說：若有 n 個罹病個案，則此 n 個罹病個案的雙親共有 2n 人。假設在 2n 個親代中，傳遞基因子給罹病個案的情形如表一所示：親代基因型爲 M1M1，即同時傳遞與不傳遞 M1 基因子給罹病個案的人數有 a 人；親代基因型爲 M1M2，只傳遞 M1 而不傳遞 M2 基因子給罹病個案的親代人數有 b 人，而只傳遞 M2 而不傳遞 M1 基因子給罹病個案的親代人數有 c 人；親代基因型爲 M2M2，即同時傳遞與不傳遞 M2 基因子給罹病個案的人數有 d 人。親代將 M1 基因子與 M2 基因子傳遞給罹病子代的比例 =[(a+b)/2n]/[(c+d)/2n]=(a+b)/(c+d)，不傳遞 M1 基因子與 M2 基因子給罹病子代的比例 =[(a+c)/2n]/[(b+d)/2n]=(a+c)/(b+d)，則 M1 基因子的半形相對危險性 = [(a+b)/(c+d)]/[(a+c)/(b+d)]。半形相對危險性的大小與該疾病盛行率、基因子在族群中的頻率、致病基因與基因標記是否有連鎖不平衡及兩基因座間是否有連鎖存在有關。當致病基因與基因標記有連鎖不平衡，並且兩基因座間有連鎖存在，則基因子相對危險性將會不等於 1。因此，利用卡方檢定其基因子相對危險性是否顯著異於 1，可以探討該基因子與疾病的相關性。

基因子傳遞分布不平衡檢定

事實上，半形相對危險性分析並未直接驗證連鎖是否存在。當觀察到某一基因子與疾病有相關時，其真正的意義是致病基因與此基因標記有連鎖不平衡存在，而且在由雙親傳遞至

表二：不同連鎖分析研究設計之比較

研究設計	資料型態	理論基礎/分析方法	優缺點	參考資料
LOD score 連鎖分析：				
雙基因座連鎖分析	家族資料	利用最大概率估計法(MLE)同時估計出θ及LOD score	•需完整的家族資料 •疾病的遺傳模式估計不確會干擾θ及 LOD score 的估計	(1)
多基因座連鎖分析	家族資料	經由 joint likelihood 同時估計多組基因座間的θ與排列順序	•較傳統雙基因座連鎖分析有較大的檢力(power) •充份利用家族資料 •可推估染色體上基因座位置的排列順序 •比較容易受到疾病遺傳模式估計不確的干擾	(3)
同宗基因子分析及其衍生方法:				
同宗基因子分析法(IBD)	配對手足資料	檢定某一罹病狀態之手足間之共享同宗基因子分布是否偏離期望分布，推估兩基因間是否有連鎖	•不需假定特定的遺傳模式 •只要收集手足資料 •並未直接估計基因重組率 •檢力較傳統連鎖分析弱 •手足間共享同宗基因子的判定困難	(5)
同基因型分析法(IBS)	配對手足資料	檢定手足其共享同基因型分布是否偏離期望分布，推估兩基因間是否有連鎖	•判定較 IBD 容易 •檢力較 IBD 弱	(6)
罹病家族成員分析法(APM)	罹病親屬資料	檢定罹病親屬其共享同宗基因子或同基因型分布是否偏離期望分布，推估兩基因間是否有連鎖	•不需假定特定的遺傳模式 •只要收集家族中罹病個案資料 •越遠的親屬配對檢力越大	(7)
半形相對危險性分析(HRR)	罹病個案及其雙親資料	透過比較雙親將某基因子傳遞給罹病個案的比例，來代表該基因子的相對危險性	•不需假定特定的遺傳模式 •不受族群基因背景分布差異的影響 •需要罹病個案雙親的資料 •並未直接估計基因重組率 •其相關性反應的是連鎖與連鎖不平衡同時存在	(8)
基因子傳遞分布不平衡檢定(TDT)	罹病個案及其雙親資料	在連鎖不平衡態下，經由檢定異基因合子型親代其傳遞基因子與非傳遞基因子分布是否偏離期望分布，推估兩基因間是否有連鎖	•不需假定特定的遺傳模式 •檢力較同宗基因子分析強 •需假設已存在連鎖不平衡 •需要罹病個案雙親的資料 •並未直接估計基因重組率	(10)

子代的過程中，基因重組的比率低於 1/2(即有連鎖的情形) 。因此，假設已經存在有連鎖不平衡，則上述分析正可以用來評估基因標記與致病基因是否發生連鎖。根據這個想法，Spielman 等人[10]乃提出基因子傳遞分布不平衡檢定，透過檢定其傳遞基因子與非傳遞基因子分布是否偏離期望分布，來推論兩基因間是否有連鎖存在。由於同基因合子型的父母對檢定基因重組率並無助益(non-informative)，基因子傳遞分布不平衡檢定乃是分析異基因合子型的父母，將某特定基因子傳給罹病子代與不傳給罹病子代的比例。如果存在有連鎖不平衡的情形，此比例與是否存在連鎖有關：當比例不等於 1 代表有連鎖存在；相反的，如果此比例並無顯助異於 1，則表示基因並無連鎖存在。以表一的例子來說：異基因合子型的親代將 M1 基因子傳給罹病子代與不傳給罹病子代的比例 =b/c。當基因標記與致病基因並無連鎖，則此比例將等於 1，並且$(b-c)^2/(b+c)$爲卡方分布。因此，利用卡方檢定，判斷此比例是否顯著異於 1，可以用來檢定是否存在連鎖。

基因子傳遞分布不平衡檢定的主要優點是：不需要如傳統連鎖分析要假定特定的遺傳模式，且資料收集較爲簡單，不限定在多發性家族或需完整的家族罹病資料，並且檢力也較同宗基因子分析強。但其缺點則是：並未直接估計用以判斷是否連鎖的參數值—基因重組率，而且必須假設已存在連鎖不平衡。在一些中、晚年才好發的疾病，其罹病個案的雙親多數已故，因而無法獲得雙親的資料。因此，基因子傳遞分布不平衡檢定的研究，常是於一般族群流行病學中已發現基因標記與疾病有明顯相關(即有連鎖不平衡)，而雙親的資料容易獲得的情形下，再進行此類型的研究來確定兩基因間是否有連鎖存在。

結 語

綜合以上所述，本文介紹的幾種連鎖分析方法，在分析檢力與資料收集上，各有其優缺點，如表二所述。研究時該採行何種分析方法，事實上決定於所欲研究之疾病本身的特性，如資料收集的可行性、對遺傳模式了解的程度等(單一基因或多基因遺傳、顯性或隱性遺傳、基因頻率、穿透性，單一基因可解釋的比率等)。雖然過去連鎖分析已經成功地找出一些人類疾病的遺傳基因，如杭丁頓氏舞蹈症(Huntington's disease)、阿耳滋海默氏病(Alzheimer's disease)及遺傳性乳癌等等。但大多數人類疾病的致病因素相當複雜，遺傳因素與環境因素可能同時影響疾病的發生，其彼此間也可能存在有交互作用。因此，未來遺傳流行病學的重要課題是如何發展釐清遺傳因素與環境因素對疾病的影響，並且有效發現遺傳因子的研究方法。而致力於了解遺傳因素與環境因素的作用，一方面可以針對具有遺傳傾向的高危險群，進行有效的預防措施，以減少外在環境暴露，另一方面有可能透過基因治療的方法，來達到預防疾病發生的目的。

推薦讀物

1. Khoury MJ, Beaty TH, and Cohen BH: Fundamentals of Genetic Epidemiology, Oxford University Press, New York: 1993.
2. Ott J: Analysis of Human Genetic Linkage, Revised ed.,The Johns Hopkins University Press, Baltimore and London, 1991.
3. Lathrop, GM, Lalouel JM, Julier C, and Ott J: Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci USA 1984;181:3443-6.
4. Elston RC: Algorithms and inferences: The challenge of multifactorial diseases. Am J Hum Genet 1997;60:255-62.
5. Suarez BK, Rice J, and Reich T: The generalized sib pair IBD distribution: its use in the detection of linkage. Ann Hum Genet 1978;42:87-94.
6. Lange K: The affected sib-pair method using identity by state relations. Am J Hum Genet 1986;39:148-50.

7. Weeks DE, Lange K: The affected-pedigree-member method of linkage analysis. Am J Hum Genet 1988;42:315-26.
8. Rubinstein P, Walker M, Carpenter C, Krassner J, Falk C, and Ginsberg F: Genetics of HLA disease associations. The use of the haplotype relative risk(HRR) and the haplo-delta(Dh) estimates in juvenile diabetes from three racial groups. Hum Imm1981;3:384.
9. Ott J: Statistical properties of the haplotype relative risk. Genet Epidemiol 1989;6:127-30.
10. Spielman RS, McGinnis RE, and 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.