染色體易裂點統計分析方法

台灣大學公共衛生學院流行病學研究所　戴 政

一、易裂點研究的細胞遺傳學背景

細胞遺傳學研究主要是以生物個體之細胞爲單元，透過顯微鏡和化學藥物操作，觀察細胞內各微細組織（細胞核內染色體）的構造、及這些微細組織在不同生命階段的自然變化（細胞體分裂）及受外在因素影響下的反應（國內對輻射屋影響人體染色體產生變異的研究[1,2]）。細胞遺傳學的一個重要研究方向是在了解染色體上存在的一些容易發生斷裂的位置（fragile sites），這些易裂位置經過許多研究發現與癌病[3]精神病[4]等疾病的發生或有關聯，因此若能經由實驗室方法檢定出染色體易裂位置，則對一些特殊疾病的遺傳研究將有對焦的功能，可以精確、快速的進行基因定位工作。

檢定染色體易裂位置是一種相對觀念。人類具有的 22 對體染色體與一對性染色體，易裂位置這個名稱包含二個細胞遺傳學要點：(1)易裂位置的位置（site）如何標示？(2)易裂（fragile）的定義爲何？細胞遺傳學研究可以透過不同染色方法（staining techniques，使用 quinacrine mustard 和 Giemsa 等物質）操作，使得染色體在螢光顯微鏡下呈現出明、暗相間的帶狀分布，這些帶狀分布有寬、有窄，且在每條染色體位置上的分布不相同，因此具有標示不同染色體之不同位置的作用；當染色體上某一位置發生染色體變異（aberrations 即斷裂、缺失、插入、反轉、轉置等現象）時，可藉由染色體帶（chromosome band）來標示出這些變異相對位置，這些發生變異位置常通稱爲斷裂位置（breakage sites）又稱爲斷裂點（breaking points）。易裂的定義在 1960 年代中期細胞遺傳學開始從事這方面研究時，尚不是很清楚。早期對染色體易裂位置的研究，來自於對某些偶發性染病個體（如，智障者）之染色體易於斷裂（liable to break）的觀察。其後的研究進展，則來自於發現細胞在某些特殊的培養基下（缺少 folic acid 和 thymidine），染色體之某些位置表現出易於斷裂的特性。自此之後，利用一些特殊化學物質誘導染色體發生斷裂，觀察斷裂發生之頻率，以爲判定某些位置是否爲易裂位置，成爲一項制式化的研究技術。以下將針對由實驗室所得之斷裂資料（breakage data），介紹易裂位置之定義與統計檢定方法。

二、斷裂資料特性及易裂點定義

斷裂資料的取得來自於隨機由群體中抽取某一數目人數（30 人），每一人再隨機抽取某一數目淋巴細胞（50 個）進行染色體斷裂實驗，直覺上這樣的資料應當是非獨立性資料，但爲了統計處理方便，通常都假設在一次實驗中所有觀察到的所有細胞間獨立。令 M 代表所有觀測到的細胞數，B 代表每一細胞內所有觀察之染色體帶（位置）數目，（如，通常觀察 30 個以上），n_{ij} 代表在第 i 個染色體帶上觀察到第 j 個重複（細胞）的斷裂數目，$i = 1, 2, \ldots, B$，$j = 1, 2, \ldots, M$，$n_{ij} = 0, 1, 2$（每一個體爲雙套，可同時觀察到二個相同染色體帶），在一次實驗中第 i 個染色體帶上觀察到之斷裂數爲 $n_i = \Sigma n_{ij}$，總斷裂觀察數爲 $n = \sum_i n_i$，在實際資料中常出現某些 n_i 爲 0 或數值極小之稀少類資料（sparse data）形式。根據上述斷裂資料，定義易裂點或可以採取決對標準，即約定染色體帶上斷裂數超過某一定數值，即爲易裂點。這樣二分法因爲選取界數值在實際分析時很難獲得實驗者的共識，故不可行。一種折衷的方法是改採以測試樣本爲基礎（test sample-based）的絕對標準，即假設若受測試的每一個染色體帶上發生斷裂的機率相等，則每一染色體帶發生斷裂的數目應爲 n/B，因此當某一染色體帶觀察到斷裂數 n_i 超過 n/B，且達到統計顯著意義，則視爲易裂點。令第 i 染色體帶上之斷裂機率（breakage probability）爲 $\theta_i = 1/B$，

則在上述之均等斷裂機率模式下（equiprobability model，簡稱 EPM），檢定擬說設定爲 $H_0 : \theta_i = \theta_{i0} = 1/B$（隨機斷裂）versus $H_1 : \theta_i > \theta_{i0}$（非隨機斷裂）。另一種方法仍是以測試樣本爲基礎，但採取相對觀念，即受測試的每一個染色體帶上發生斷裂的機率與染色體帶的寬度（band width）成正比。令每一染色體帶的寬度爲 w_i，$i = 1, 2, \ldots, B$，所有受測染色體帶寬度和爲 $W = \sum_i w_i$，則第 i 個染色體帶斷裂機率爲 $\theta_i = w_i/W$，當第 i 個染色體帶觀察斷裂數超過 θ_i，且達到統計顯著意義，則認爲第 i 個染色體帶爲易裂點。在上述比例機率模式（proportional probability model，簡稱 PPM）定義下之易裂點檢定擬說爲 $H_0 : \theta_i = \theta_{i0} = w_i/W$ versus $H_1 : \theta_i > \theta_{i0}$。

三、易裂點的檢定方法

易裂點的統計分析方法發展，主要是循著單一位置（single-site）和多位置（multiple-site）檢定二個方向。

(一)單一位置檢定方法

1. 一般方法

斷裂資料在 1970 年代的分析方法都只是一般簡易統計分析方法（卡方統計量），這樣形成的適合度檢定方法，不能得到正確的結論。注意到斷裂資料的大量多重檢定問題且爲稀少類資料特性而提出較合適的統計檢定方法者爲 Smith [5]，他的想法是以 $\sqrt{\chi^2}$ 來改進 χ^2 統計量之缺點。其後所提出的方法，包含使用二項分布 [6]、卜瓦松分布 [7] 和負二項分布 [8] 等。Tai et al. [9] 綜合以往討論，提議使用二項分布與 F 分布之間的尾部機率轉換關係 [10] 來設立檢定統計量

$$F^* = \frac{1-\theta_{i0}}{\theta_{i0}} \frac{n_i}{n-n_i+1} \qquad (1)$$

其中，θ_{i0} 是在 EPM 或 PPM 定義下之 H_0，F^* 爲具 $2(n - n_i + 1)$ 和 $2n_i$ 自由度的 F 分布。F^* 的好處是處理了稀少類二項分布資料分析的計算困難（以往研究如要精確計算二項分布尾部機率，得採用蒙地卡羅方法 [6]），統計量(1)的提出，幫助了實驗室的細胞遺傳學家在以單一位置分析爲想法下，可以有個簡單且較爲可靠檢定易裂點的方法。

2. 確認分析方法

實驗室研究的一項特點是同一（或相似）問題在世界各地不同實驗室中都在進行，較小的實驗室受到經費、材料限制，樣本數往往不夠多，因此若能合併相似實驗室資料，對同一問題的分析可以有更綜合客觀的結論。根據這樣的想法，Tai, Hou and Wang-Wuu [11] 利用貝氏分析事前訊息融入現存資料的做法，將不同實驗室相似的斷裂資料融合，推導出類似(1)的檢定統計量

$$F^{**} = \frac{1-\theta_{i0}}{\theta_{i0}} \frac{n_i + t - 0.5}{n + k - n_i - t + 1.5} \qquad (2)$$

其中，k 和 t 可視爲融合參考資料（reference data，其他實驗室資料）和主題資料（objective data，實驗者自有資料）的權數。這個方法被特稱爲確認方法（confirmation method [11]），它的優點是實務上，讓細胞遺傳學家可以分析其資料在與其他相似資料融合之後，所呈現出的綜合分析結果，另方面也可以藉助確認過程，某種程度解決大量二項分布檢定帶來的多重檢定問題。

(二) 多位置檢定方法

1. 二群分類方法

Bohm et al. [12] 認爲上述二項分布檢定方法疏忽了 B 個染色體帶彼此間觀察到的斷裂關係，分析時若能置入這種相互關係（即，以多項聯合分布爲基礎，而不是以邊際二項分布爲基礎），則結論應該更有意義。他們的想法是易裂點與非易裂點是一種二分法，故應以受測試的 B 個斷裂點爲基礎，區分爲二群：B_1 個非易裂點的斷裂機率定義爲 $P_1 = \ldots = P_{B1}$，每一個 $P_i < 1/B$，$i = 1, 2, \ldots, B_1$，另外（$B - B_1$）個易裂點的斷裂機率則爲 $P_i > 1/B$，$i = B_1+1, \ldots, B$。這樣二分法的分析要點在於找出合理的 B_1 切點，他們根據 Bonferroni 的想法，依靠遞迴技巧，逐次緊縮顯著水準，每一次檢定一群含 B_1 個的斷裂機率的 $H_0 : P_1 = 1/B_1, P_2 = 1/B_1, \ldots,$

$P_{B1} = 1/B_1$（註：B_1 爲可變動的數字），若拒絕 H_0，則有最高觀察斷裂機率的斷裂位置爲易裂點，剔除這些點後，再以剩餘斷裂點繼續分析。操作同樣程序，直至 H_0 不被拒絕，此時的斷裂點群視爲非易裂點。Bohm et al [12] 的以上做法是以 EPM 爲易裂點定義推出，在 PPM 想法之下因爲虛無擬說會隨著染色體帶寬度變動，故在每一次遞迴檢定過程挑出最大觀察斷裂機率之位置時並無易裂意義，Hou, Chiang and Tai [13] 採用最大標準化斷裂機率來克服這個問題。

2. 多群分類方法

Hou, Chiang and Tai [14]根據多項分布將二群分類方法推展到多群分類方法。他們的想法是若 B 個染色體帶應該分成 K 個群，每一群內的染色帶斷裂機率相同，但群與群間的斷裂機率不同，因此，利用逐次檢定想法[15]可以達到分群目的。分群的結果，在於區分出染色體帶之斷裂機率狀況，而不刻意用二分法區分爲易裂或非易裂，是較符合大數目分析（如，這裡 B 值極大）的自然原則。

四、結論

細胞遺傳資料分析的統計方法應用發展約有 30 年，在 1990 年前都只是單純的二項分布或列聯表分析應用，其後的研究單位，在國際上主要一在美國的德州大學[12, 16]一在國內[9, 11, 13, 14]。二者之間對分析的看法並不完全相同，但彼此之間對這個領域問題的討論[17, 18, 19]，促進了統計方法的發展，所餘下的問題有：(1)易裂點的 EPM 或 PPM 定義，如何由實際資料驗證？(2)如何以逐次檢定爲基礎，估計樣本數？(3)檢定易裂點應以個體爲基礎先行檢定，再綜合所有個體檢定結果再下結論？還是合併所有個體分析，進行判定？(4)如何將這個領域發展方法應用到其他研究領域？

參考文獻

[1] W.L. Chen, C.L. Taur, J.J. Tai, K.D. Wuu and S. Wang-Wuu, *Mutation Research*, **377**, 247 (1997).

[2] S. Wang-Wuu, J.J. Tai, J. Wuu, S.Y. Lin, W.L. Chen and K.D. Wuu, *Journal of Biomedical Science*, **8**, 411 (2001)

[3] K. Ried, M. Finnis, L. Hobson et al., *Human Molecular Genetics*, **9**, 1651 (2000)

[4] C.H. Chen, H.H. Shih, S. Wang-Wuu, J.J. Tai and K.D. Wuu, *Human Genetics*, **103**, 702 (1998)

[5] C.A.B. Smith, *Annals of Human Genetics*, **50**, 163 (1986)

[6] M. DeBraekeleer and B. Smith, *Annals of Human Genetics*, **52**, 63 (1998)

[7] T. Mariani, *Human Genetics*, **81**, 319 (1989)

[8] D.K. Jordan, T.L. Burns, J.E. Divelbiss, R.F. Woolson and S.R. Patil, *Human Genetics*, **85**, 462 (1990)

[9] J.J. Tai, C.D. Hou, S. Wang-Wuu, C.H. Wang, S.Y. Leu and K.D. Wuu, *Cytogenetics and Cell Genetics*, **63**, 147 (1993)

[10] G.H. Jowett, *The Statistician*, **13**, 55 (1963)

[11] J.J. Tai, C.D. Hou and S. Wang-Wuu, *Cancer Genetics and Cytogenetics*, **105**, 1 (1998)

[12] U. Bohm, P.F. Dahm, B.F. McAllister and I.F. Greenbaum, *Human Genetics*, **95**,249 (1995)

[13] C.D. Hou, J. Chiang and J.J. Tai, *Human Genetics*, **104**, 350 (1999)

[14] C.D. Hou, J. Ching and J.J. Tai, *Biometrics*, **57**, 435 (2001)

[15] Y. Houchberg, *Biometrika*, **75**, 800 (1988)

[16] I.F. Greenbaum, J.K. Fulton, E.D. White and P.F. Dahm, *Human Genetics*, **101**,109 (1997)

[17] P.F. Dahm and I.F. Greenbaum, *Cytogenetics and Cell Genetics*, **66**, 214 (1994)

[18] P.F. Dahm, A.W. Olmsted and I.F. Greenbaum, *Biometrics* (2002, to appear)

[19] J.D. Hou and J.J. Tai, *Biometrics* (2002, to appear)