人類遺傳疾病研究統計方法評介

戴　政

一、遺傳 —— 今年花勝去年紅 (歐陽修、浪淘沙)

遺傳疾病是現代社會常可接觸且令人深以爲苦的一種生命折磨。爲人母者，十月懷胎, 臨盆憂見智障兒, 已爲常事; 聽障, 只要我們多關心週遭社會, 亦非少見; 侏儒症或雖不多見, 耳聞目睹總略知一二; 我們壯年時擔憂糖尿病纏身, 下半生將與藥爲伍; 臨老憂愁身染老年癡呆症, 恐生活之尊嚴喪失殆盡。試著了然生、老、病、死的玄機, 或正是人類心靈深處那一點似有又無的禪知，推動了文明的演化，也鼓動了人類探索生命無窮變化的科學進步。於是我們開始見到在白種人易發的纖維性囊腫, 被證明其控制基因何在, 肌肉萎縮症如此, 亨亭頓氏症如此, 遺傳性聽障的一部分與性染體有關；導致侏儒症的一種軟骨發育不全症的基因略在何處，糖尿病病因雖複雜, 但我們漸瞭解遺傳影響的方式, 而歐茲海默症受基因控制，表現出老年癡呆也終將明朗。人類一己的生命或如蜉蝣之寄於天地, 瞬息而已，但遺傳訴說的卻是父（母）傳子(女), 子（女）傳孫, 藉著形體的複製, 無盡綿延, 誠所謂: 自其不變者而觀之, 則物與我皆無盡也。死生之間, 盡在那似無而有的莫言禪意中。

二、哈帝-溫柏定律 —— 天下難事, 必作於易 (老子)

遺傳學以統計爲工具進行研究，肇因於以族群爲觀點審視遺傳現象。生物學研究的範圍很廣, 它可以研究生物的本體如解剖、生理、生化反應，也可以研究生物與生物間或生物與環境間的關係, 如分類、病理、生態、生物氣象等。遺傳學的研究溯自1900年對孟德爾所做創世紀的基本遺傳分離規律重新肯定起，至今將滿百年。回顧這百年間的發展, 從混沌未明，到逐漸清楚，到期待生命的起源終有一日可被解釋，統計在遺傳學的發展過程中一直佔有重要影響。在孟德爾對遺傳所做研究工作被認知後不久，哈帝和溫柏在1908年即歸納出族群遺傳學最基本也是最重要的一個規則: 以一個含有二對偶基因 (A,a) 的基因座爲例，族群中存有三種雙套基因型 (AA,Aa,aa), 若以 (p, q) 表示 (A,a) 在族群

中出現的頻率, 則當族群是隨機交配, 且無外力使 (A,a) 間比率 $p:q$ 發生變化, 則可以證明族群中 (AA,Aa,aa) 間的比率可以從上一代傳到下一代, 維持 $(p^2:2pq:q^2)$ 的固定數值, 此恰爲 $(p+q)^2$ 的展開式, 這種穩定的狀態稱爲哈溫平衡態。族群遺傳學主要是以基因頻度 (如上述 p,q) 爲參數去描述一個族群在穩定狀態下的特徵 (如某一遺傳疾病致病基因率大小), 或在受外力影響 (如族群移入, 移出, 發生突變) 時, 族群處於不穩定狀態下的動向 (如形成生物演化), 以此分法, 哈溫定律雖極簡易, 但卻是談論任何更複雜遺傳問題前, 都無法不碰觸的基本要件。以統計方法檢定一個族群之某一基因座是否處於哈溫平衡比率, 以判定族群是在平衡狀態或否, 是應用很簡單的所謂適合度檢定 (即看資料出現 AA,Aa 和 aa 的數值和期望在平衡比率 $p^2:2pq:q^2$ 下的理想數值間的標準平方差異和)。它的重要性在於一般實驗性或調查性資料, 主要目的都希望將所得結果發布供學術界或社會參考, 如對某一遺傳疾病而言, 若由統計檢定顯示其控制基因不在穩定狀態, 則該資料計算之基因頻度便可能隨每代族群交配行爲而改變, 故本次實驗結果不一定適用於更廣層面的解釋。

哈溫定律雖有80多年歷史 (這也差不多是主要統計學發展的歷史時間), 目前仍有些基本問題不清楚, 其原因在哈溫定律論的是族群在進行隨機交配時 (即族群中個體通婚並無特殊偏好, 如身材高男性與身材高女性通婚率高), 可導致族群恆久處於平衡狀態, 但這並不表示族群若進行非隨機交配, 就不能達到平衡。首先質疑這個問題的是中央研究院生物組院士李景均先生 (1988), 他在理論上證明了這種非隨機但可使族群達成平衡的交配方式是可能的。這樣的交配方式如確有可能存在, 而其結果又可能影響演化的速度, 在學術上就非常新人耳目了。目前, 在這方面研究, 發展出一個統計方法能夠測出這種特殊的交配方式, 應是迫切的。

三、分離分析 —— 剪不斷, 理還亂 (李煜, 相見歡)

孟德爾由他所做實驗歸納出的遺傳法則是對一個雙套個體基因座而言 (如 Aa 基因型), 在形成精子或卵子之單套過程中是獨立分離的 (即或 1/2 是 A, 或 1/2 是 a), 當族群是依隨機交配方式通婚, 則從群體觀點而言, A 或 a 單套精、卵子再依其在族群中頻率獨立組合成雙套的 AA、Aa 和 aa。這個規律在1910年代後, 被倒轉過來問而成爲統計上的一個難題。問題是: 對一疑受遺傳控制的人類隱性基因疾病, 如何收集一組資料, 進行判定? 這個問題有二點困難: (1) 由於一般遺傳疾病在群體中的發生率都很低 (如萬分之一以下), 故不可能經由一般統計上的取樣方法, 獲得足夠的有病家庭資料, 再估計有病子女的比率, (2) 由於 (1) 的限制, 所以取得有病子女家庭方式得改爲先找到有病個體, 再透過這個體找到該有病史家族, 再經驗證過程 (如抽血檢驗), 估算所有子代中有病個體所佔比率 (即父、母基因型均爲 Aa, 子、女中爲 aa 者), 這個'分離率'如被證明

爲 1/4, 則可說在統計上有證據指出該病受單一隱性基因 (a) 控制。統計難處在於: 因爲過程中是先認定一個染病者, 而後經由這個體去得到該有病家族, 故所得資料不是一個隨機樣本 (即捨棄了許多家庭中未染病者, 或某些染病者少的家庭被選入樣本的機會相對減少), 這種有偏誤的取樣方式導致分離率的估計值遠大於 1/4, 故易誤認遺傳方式或當遺傳控制基因確實存在而做了不存在的推論。這樣的結果就好像在醫科學生班上 (有偏誤的樣本) 估計男生比率必會大於 1/2 (即使男生在族群中確實佔了 1/2), 統計上爲校正這種偏誤, 大約花了60年時間才釐清所有相關問題。以簡單例子說明它的作法, 正如以醫科學生班級爲基礎 (有病個體), 去收集這些學生的同胞中男、女人數 (有病個體親屬中有病、無病人數), 以這些同胞兄弟姊妹人數爲新的樣本 (即這個新樣本不包含我們先知其有病, 而後透過他們找到其家族的那些個體), 再去估計男生比率, 則可得到正確數值 (分離率)。以上的問題, 還有因資料收集方式變化有更難的情形, 在遺傳學研究上通稱爲簡單分離分析, 其中最主要的貢獻者之一是前面提過的李景均院士 (1968)。這方面問題被稱爲簡單分離分析是因爲它所處理的遺傳疾病僅受一個基因座影響, 對一些其它疾病研究 (如糖尿病、心臟血管疾病、精神病等) 則顯示這些疾病的發生可能同時受到遺傳和環境影響, 而環境 (如人文、社會、飲食等) 的影響方式有可能造成擬似遺傳行爲, 亦即由於在同一家庭中接受同一類環境因子, 而這種因子亦可誘發疾病, 故表現出親子、同胞間發生率高於族群中一般值, 而被猜測該病或與遺傳有關, 這種因有共同環境因子而干擾正確遺傳基因推論的研究, 稱爲複雜分離分析。具有這種複雜分離特性的性狀, 又稱爲數量性狀, 在遺傳學研究中, 一直未有完善的作法 (故是未來研究的重點), 有兩項重要成果, 都是在1970年代發生的。第一項成果是由 Elston 和 Stewart(1971) 解決了如何處理人類譜系資料 (多代家庭) 成員間相依情形的概似函數計算方法, 相依在統計學上有很多種情形, 人類譜系資料之相依是屬於結構性, 亦即相依來自子代獲得雙親遺傳物質, 而遺傳控制外表性狀, 故代代間可尋一定路徑溯源。這種結構性的相依不同於一般統計學上所稱二個隨機變數存在相依, 但卻僅能設定參數描述的模糊觀念不同。結構性相依雖較清楚, 但要能用概似函數描述的清楚, 如從祖父母、父母、伯叔、兄弟姊妹、子女, 將垂直的三代及平行的每代衍生出的平輩間親屬關係, 用函數關係寫出一張如樹突狀的複雜相依族譜, 並實際提出一個流程可供計算, 是遺傳統計學家在1970年代前做不到的。這個問題的突破在於二個想法上的改變, 一是認識到親屬間相依的結果源自於個體遺傳型縱向的傳遞, 而對每個體所觀察到的是外表型而非基因型 (如設 aa 基因型表現病徵, AA, Aa 表現正常), 因此如對外表型與基因型間可以一個表現函數簡單描述 (如 aa 對應函數值1時表示有病, 函數值爲 0 時表示 aa 不可能爲正常), 則重點在如何描述親屬間的結構性相依, 這種描述方式最簡捷的是將一個家族中最年輕 (不是年紀, 是輩分最低者) 的個體定好,

由此往上推一代, 形成一個含有父、母、子女的所謂核心家庭, 也就是說先將多代家庭由最基層一代切割出一個核心家庭, 這時因父母親基因型並不知道, 故猜測所有可能基因型配對 (如一個體有三種基因型, 則共九種配對), 每一種配對可先寫出父母親的概似函數, 然後在給定父母基因組合配對條件下, 子代間依孟德爾獨立分離率可以採同胞間的概似函數, 依獨立方式寫出, 如此聯結了二代間的垂直相依關係及同胞間的平行相依關係, 然後再將這個核心家庭概似函數依附 (註記) 到倒數第三代與它有直接血緣關係的父母上, 此時可以倒數第二代平行聯結同胞往倒數第三代上述父母切出另一個核心家庭, 依此類推, 往上推衍, 完成整個家族中的概似函數; 第二個觀念的突破是上面第一點中已提到的, 每一核心家庭溯及父母時, 並不確知九種基因型配對爲何, 故當九種都猜測時, 以其各別基因型頻度來做權數 (比重), 在實際知道個體外表型有病或正常時, 九種配對中的某些不可能, 在寫概以函數中自會被移掉 (如猜父母正常之配對 AA 和 AA 時, 不可能生出有病子女 aa)。

在解決家族內概似函數計算後, 另一重要成果是 Morton 和 MacLean 在1974年以實驗設計學上的混合型模式來描繪一個包含遺傳、擬遺傳環境, 純粹環境的複雜數量性狀。混合型模式在統計學上的大意是指一個反應變數包含了固定因子, 隨機因子及二者間的交換作用和純環境因子。對數量性狀而言 , 所謂固定因子是指存在一個主要基因型的作用力, 如具 aa 基因型者表現出有病, 是因爲某一生化過程反應受到阻礙, 阻礙的程度對所有 aa 者都是一常數 (固定); 隨機因子是指多個微量基因合成的力量, 每個力量都很小, 無法描述, 故在統計處理上, 爲了簡化問題複雜, 假設這些力量的大小是依從常態分布, 這麼假設從生物學方面講是很強烈的逃避問題本身, 因爲一如此假設, 一個複雜的隨機因子描述變成只需要估計常態分布的平均和變異數, 因此當一個數量性狀假設存在一個主要基因是 AA,Aa 和 aa 三種固定作用力, 多個微效基因是常態分布, 擬環境因子和純粹環境因子也都是常態分布, 則在統計上該數量性狀成爲三個以 AA,Aa 和 aa 固定作用力爲中心的三個常態分布的混合分布, 通常這種混合分布是扭曲的分布形式, 故若收集某家族資料, 而呈現出扭曲形式但又非眞受遺傳控制, 則有可能導致複雜分離分析作出遺傳基因存在的錯誤結果, 這是現知混合型模式的最大缺點。撇開這些缺點不談, 混合型模式所以能夠在某些程度推論出是否有單一的主要因子, 是因爲如存在這個因子, 則從家族資料中相當程度會顯出父傳子、子傳孫的遺傳現象, 也就會呈現前述混合分布的結果。

分離一個數量性狀具有主要遺傳基因重要性在於知道這個因子存在後, 才能利用如下述的連鎖方法繼續去探索該基因何在。這方面研究未來需要發展的方向, 我認爲首要的是能 (也要有勇氣) 完全摒棄混合型模式, 去嚐試一個更有遺傳味道的作法 (如不一定硬要用譜系資料, 避開龐大的概以函數計算,

不用統計模式的想法), 次要的是用混合型想法, 但不用現有架構去做問題, 最後才是固守混合型模式架構, 改正其缺點 (如不用常態分布假設)。

四、連鎖分析 —— 衆裡尋他千百度 (辛棄疾, 青玉案)

人類遺傳學的核心問題是在了解每一個受遺傳控制的特質位於某一條染色體上的某一位置, 這個工作稱爲基因定位。解決了這個問題, 才能去進一步分析該遺傳基因在染色體上的遺傳密碼排列, 也才可能知道某些遺傳疾病是那些密碼發生突變, 也才可能進一步治療。

在人類遺傳疾病研究上, 目前最主要的作法是先借由一般流行病學方法 (如病例–對照法, 或世代研究法) 先去尋找粗略證據, 說明該疾病可能與遺傳有關, 再利用前節敍述的分離分析方法進一步確定疾病在家族資料上呈現代與代間基因傳遞 (即所謂的家族聚集), 最後再以連鎖方法去尋找基因位置。連鎖是指二個不同基因座位於同一染色體的相鄰現象, 人類有 23 對染色體, 所謂'對'是因爲每一對包含了'兩'條控制同一些性狀的染色體, 一條來自父方, 一條來自母方, 換言之, 任一個體不能完全與上一代相同。生物體演化出這樣複雜的異性交配方式, 造成群體隨'代'保持個體基因間的交流, 自有生物演化上的意義, 暫且不論, 有趣的是由於這種異性交配方式須先在個體行複製後再減數分裂, 而產生極特殊的互換現象。互換是指在同一對染色體上的兩個相鄰的基因座, 在行交配過程時, 先將每一對中的兩條染色體複製成四條染色體分體 (二條父方, 二條母方), 此時因四條染色體聚在一起, 而可能發生父方一條染色體在某處斷裂後接到母方對等位置, 而母方的轉接到父方上, 這樣的互換並未造成互換部位基因數量有任何改變, 但確是使得原在父方和母方的某些基因去和母方和父方原不在一起的基因在一起, 最後四條染色分體形成僅帶二份父方染色體, 二份母方染色體的四個單套染色體 (這是對某一特定染色體而言, 對 23 對染色體, 父方和母方染色體是任意組合)。上述互換現象是兩個連鎖基因特有性質, 且互換發生機率與他們在同一染色體上基因距離遠近成正比, 基於這兩項特質早在 1920 年代族群遺傳學家便開始研究利用連鎖測定基因位置的方法, 這些研究主要在尋找測度互換率的方法並建立連鎖互換 (機) 率的基因間相對距離染色體圖 , 以供細胞遺傳學家在染色體上標定基因。雖然兩個基因間互換率與它們實際距離成正比, 但這並不表示當以互換率爲測量標準, 則相鄰的三個基因位置 A−B−C, AC間互換率等於 AB 間互換率加上 BC 間互換率, 這種不可累加性質源自於 ABC 間可能發生雙互換現象, 即 AB 間發生互換後 (故 AC 間也發生互換), BC 間又再發生互換 (故 AC 間又再發生一次互換), 則 AC 間被還原成無互換狀態, 此時 AC 間互換率所測值小於 AB 與 BC 間二者和。這種不可累加特性並不能符合遺傳學家所需要的染色體基因圖特質, 故發展一種可將互換率轉換成可累加距離, 成

爲首要工作。到1950年代前, 有關連鎖方法論的研究一直沒有重要突破, 一直到1955年 Morton 因研究人類基因連鎖, 提出了用逐次分析法來估計連鎖互換率, 而啓動了這個領域的進展。

遺傳學的研究以對象分, 可在微生物、植物、動物、人類等方面。在人類以外的研究因較無須考慮倫理因素, 故可任意交配試驗對象, 以實驗設計方式來減少一部分統計複雜度。但如以人爲對象, 則不能控制交配行爲, 要研究任何遺傳疾病只能由觀察而非設計實驗方式來得到資料, 故人類遺傳學研究特質在以核心家庭爲單位下, 必須面臨相依資料, 這點已在前節提過這種相依是屬於結構性而非一般統計學所處理的概念性相依。Morton的研究具有歷史性開端的意義, 是他雖用逐次分析法工具研究人類連鎖, 但這個方法對後來研究並無衝激 (人類資料收集不適合逐次分析想法), 倒是他較 Fisher 等人在1930年代以前之研究, 更清楚詳盡的引入概似函數觀念來處理不同交配情形下核心家庭相依資料的連鎖推論, 自他的研究後, 直至1971年才由 Elston 和 Stewart(見前節所述) 再將二代資料推展到三代資料。

現代研究人類遺傳疾病的進展可溯自 19- 70 年代, 一方面是分子生物學上發展出限制碲 (酵素) 切割染色體的技術, 故實驗者可以依需要切割染色體爲片斷來研究, 染色體不復已往令人茫然不知從何著手的龐然大物; 另一方面則是 Elston 和 Stewart 提出的方法可以處理譜系資料。這兩方面成果的結合, 使得1980年代著名的文獻 Botstein, White, Skolnick 和 Davis 等人提倡以限制碲切割出的片段長度 (簡稱 RFLP) 來構建人類連鎖基因圖成爲可行。他們的想法是: 要標定出任一遺傳疾病的控制基因, 可以用連鎖方法求證該控制基因與某一個已知位置在某一染色體上何處的標識基因是否連鎖, 如連鎖則二者間距離爲估計之互換率值, 亦即該致病基因位置大致被標定。這個構想的成功取決於是否有足夠的標識基因足以撐開整個人類基因庫。這就如我們要從台北開車到台中, 路中隨時要知道自己所在位置有賴於沿路樹起路標, 如每隔一公里樹一路標, 可稱爲低密度路圖, 如每隔100公尺樹一路標, 則是高密度路圖, 路標密度愈高, 車行導航正確度愈高。換成尋求基因位置問題, 這個譬喻成爲整個人類基因庫約有30億個氮基對 (腺嘌呤與胸腺嘧啶一對, 或鳥嘌呤與胞嘧啶爲一對), 這是相當台北到台中距離, 以氮基對爲丈量單位是絕對距離, 而前述的人類連鎖基因圖是以互換率爲丈量單位, 根據細胞遺傳學家的估計約有33個互換率單位, 因爲遺傳學家稱一個互換率單位爲一個摩根, 故亦可說人類基因庫有約有33個摩根長, 相對地一個摩根長度約爲一億的氮基對。若設計一個低密度基因路標圖, 每一標識間距離爲 0.2 摩根, 則約要165個標識基因方能在給予任一欲待定位基因時標示出其位置平均距某一路標 0.1 摩根; 若提高到爲高密度路標圖, 使得每二標識基因間距離爲 0.02 摩根 (故任一須被定位基因, 平均約距某一標識基因 0.01 摩根), 則約需1650個標識基因。密度 0.2 摩根 (即互換率 0.2) 或 0.02 摩根 (互換率

0.02) 所代表的實質距離約分別爲2000萬個或200萬個氮基對的長度, 而通常一個基因約佔1000到10000個氮基對。在染色體上要找到已知位置的自然標識基因遠低於上述數目, 依 Botstein 等人想法, 由於 RFLP 的進展, 可利用這些人造的切割片斷長度當做標識基因, 則人類基因庫定位 (自然包括遺傳疾病基因) 終有一日可以完成 (速度超過以往遺傳學家推測), 這即是所謂的人類基因庫計畫。

由統計角度看連鎖分析, 未來難度高需要發展的主要仍是對數量性狀定位問題 (如避開上節所述複雜分離分析直接定位, 如對每一量化變數不先簡化成存在一個主要基因, 而直接處理, 如是否有簡單快捷的無母數方法等), 另一則是如何設計有效的快速計算機程式處理龐大的基因庫資料。

五、結語 —— 天涯地角有窮時 (晏殊, 木蘭花)

遺傳的特色之一在人種的差別或會導致同一性狀的遺傳控制不同, 表現不同, 這種異質性的特性, 使得遺傳性疾病的研究只要涉及與國內有關問題時, 勢必要自行解決 (雖然國外先進的基礎研究仍可參考), 龐大的研究經費, 人才的培養, 團隊的合作, 重點的選擇, 都是這方面日後國內勢必考慮的問題。放大角度看這方面問題, 絕對可以樂觀的預期人類基因圖終有一日可以完成, 身在此時間點, 總不免再低吟歐陽修的浪淘沙: 今年花勝去年紅, 可惜明年花更好, 知與誰同? 詞意不盡, 眞個江南草長時節!

誌 謝

我對中央研究院院士李景均教授及統計科學研究所前任所長趙民德教授致上我最眞誠的感謝, 由於他們對學術之爲何物的執著, 使我耳濡目染, 十年來身受薰陶, 並感終身受用。

參考文獻

1. Botstein D. White R.L. Skolnick, M.H. and Davis, R.W. (1980), Construction of a genetic linkage map in man using restriction fragment length polymorphism. Am. J. Hum. Genet. 32, 314-331.
2. Elston R.C. and Stewart, J. (1971), A general model for the analysis of pedigree data. Human Heredity 21, 523-542.
3. Li, C.C. and Mantel, N. (1968), A simple method of estimating the segregation ratio under complete ascertainment. Am. J. Hum. Genet. 20, 61-81.
4. Li, C.C. (1988), Pseudo-random mating populations. In celebration of the 80th anniversary of the Hardy-Weinberg Law. Genetics 119, 731-737.
5. Morton N.E. (1955), Sequential tests for the detection of linkage. Am. J. Hum. Genet. 7, 277-318.

6. Morton, N.E. and MacLean, C.J. (1974), Analysis of family resemblance. III. Complex segregation of quantitative traits. Am. J. Hum. Genet. 26, 489-503.

—本文作者任職於中央研究院統計科學研究所—