以全基因組關聯性分析(GWAS)策略提升育種親本與重要基因座探勘效率

農試所作物組 吳東鴻

一、前言

在品種選育與研發中，仰賴收集來自豐富歧異度的優良種原，並掌握種原內各項遺傳訊息與性狀表現，係作物種苗研發中最重要的關鍵資訊，如何活化運用作物遺傳資源，從過去種原基因庫保存、繁殖等例行性作業，提升至解碼其有益基因與深入利用，儼然成為當前對於活化種原遺傳資源最重要的課題之一。過去在流行病學上，因人類不似植物可創造特定雜交組合，而以關聯性定位 (Association mapping) 在目標族群內尋找特定性狀與某些較高基因頻度的相互關係，探索具有相同性狀的植株，是否帶有相同基因，以此尋找目標群體內是否存有較高特定基因頻度。

相對於QTL定位分析策略，如F_2下單基因分離比1:2:1等平衡族群架構、亦無族群結構的干擾，儘管是種原中稀有的有益基因，也能依靠孟德爾分離比被輕易尋找到，但該策略仍存有部分極限，諸如自交F_2世代等分離族群，產生遺傳重組的機會有限，基因座定位區間的解析度有限 (可能達10-15 cM)，且以二倍體作物而言，一個基因座上僅帶有2個對偶基因，雜交族群上同基因座最多只能偵測到4種對偶基因，對偶基因的種類變異有限。對此，關聯性定位策略則能輕易互補上述瓶頸，因為若將豐富變異的各種原品系視為來自同一個遠古祖先族群，每個種原彼此之間經過上萬次世代的減數分裂，累積了重組次數遠多於自交或回交數次的分離族群，相對定位區間也大幅縮短、提高定位解析度，另外同時分析上百個核心種原，也能同時歸納出特定基因座中各類基因型組成，並從中探勘適宜、有益基因型種類。

隨著次世代定序、核心種原管理以及重要基因探勘技術日漸成熟，以全基因組關聯性定位提升育種親本與重要基因座的探勘效率，將成為育種種原上最佳管理策略之一，進一步更能清楚有益基因型組成來源，在後續規劃雜交組合與分子輔助育種篩檢上都是極其重要的參考依據；有鑑於此，本文將針對關聯性定位原理、全基因組定位影響因子與應用結果進行初步導覽，供作物育種從業人員參考使用。

作　　者：吳東鴻助理研究員
連絡電話：04-23317106

二、連鎖失衡與關聯性定位分析

關聯性定位也可稱連鎖失衡定位，關聯性定位是建立在族群基因型頻度發生連鎖失衡的關係上，意指實際基因型頻度與期望基因型頻度存有顯著差距，因此族群中能符合哈溫平衡情境中數個條件：1. 龐大族群個數，以減少隨機漂變，避免造成部份基因型因同質結合而固定不再分離。2. 需逢機交配，使基因型組成能獨立分配產生。3. 並且無選拔、突變、遷移等情形，不再使基因頻度發生變化。基於上述前提下，能使基因頻度與基因型頻度維持一定，且世世代代不變，因此期望基因型頻度會因獨立分配，而為基因頻度的乘積 ($F_a \times F_b$)。

連鎖失衡即實際基因型頻度(F_{ab})與期望基因型頻度之差距，其實際基因型頻度無法達成期望基因型頻度，如

$$D = F_{ab} - F_a \times F_b \text{ ，} \qquad \text{式1，}$$

而檢定連鎖失衡的統計量有許多種，最直接檢定兩基因座間是否相互獨立，可利用關聯表進行自由度為1的適合度測驗。而另外兩個常用的統計量是平方相關係數 (Squared correlation, r^2) 與標準化連鎖失衡係數 (Standard linkage disequilibrium, D’)，前者是利用相關係數平方轉換所致，如

$$r^2 = \left(\frac{S_{AB}}{S_A S_B}\right)^2 = \frac{(f_A f_B - f_{AB})^2}{(\sqrt{f_A f_a}\sqrt{f_B f_b})^2} = \frac{\hat{D}^2}{f_A f_a f_B f_b}$$

， 式2，

因為會受遺傳重組與發生突變的時間點的影響，主要用於偵測原始世代所發生的突變，若突變基因是在很久遠前出現在族群內，即使分子標記與目標性狀基因座有連鎖存在，經過多世代的逢機交配，其失衡程度也可視為0，但若有連鎖失衡存在，此分子標記必與控制目標性狀基因座緊密連鎖。而另一個標準化連鎖失衡係數因為D的值域太廣泛，一般人無法直接判斷其相關強度，只能概略且抽象地比喻其強度，故將連鎖失衡係數標準化至-1～1之間，如

$$|D'| = \begin{cases} \dfrac{D}{\min(ps, qr)}, if\_D > 0 \\ \dfrac{D}{\min(pr, qs)}, if\_D > 0 \end{cases}$$

， 式4，

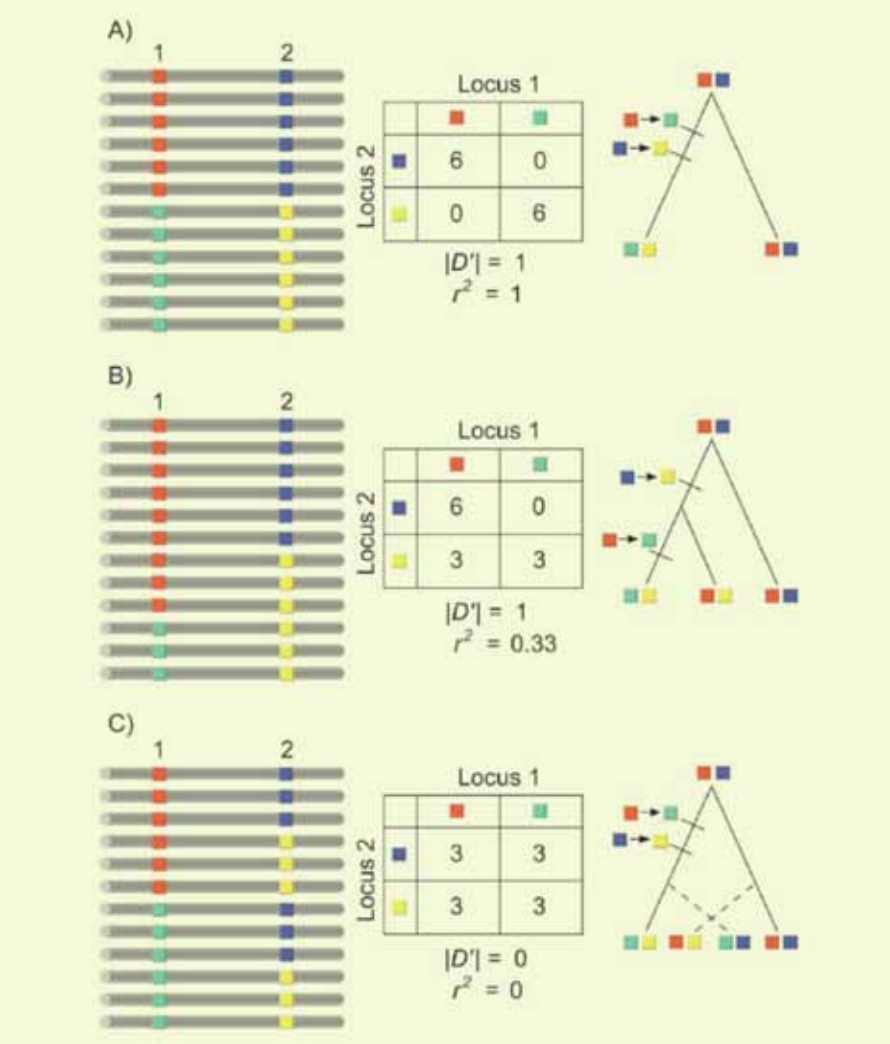
藉由實際頻度的最小乘積作為分母的校正項，當D’絕對值很接近1時，可知兩基因座相關程度很強，反之若值很接近0時，表示關聯性程度很弱，此式並不受突變時間點的影響，僅受遺傳重組影響，如圖一 (Flint-Garcia et al., 2003)。

全基因組關聯性定位分析 (Genome wide association study, GWAS) 便是基於上述原理從單點擴展至全基因組通盤掃描上，針對全體試驗族群利用全基因組高密度分子標誌建立基因型組成資料，再分析所得基因型資料與外表型性狀之關聯性，充分利用族群演化過程中數千世代發生的重組事件，透過打破連鎖失衡或是特定染色體區域內多型性位點的關聯性來增加目標性狀的定位效率。

三、試驗族群與影響因子

在育種程序中常見的試驗族群，有遺傳種原庫、優良栽培品種、合成品種族群等，皆可直接應用關聯性定位分析策略 (Breseghello and Sorrells, 2006)，這些試驗族群往往累積大量性狀調查數據，是個豐富試驗資料來源，而各試驗族群的遺傳特性皆有長處，其中核心種原能代表整體遺傳變異，且適當族群數下能方便管理，最適於遺傳研究，而栽培品種是遺傳特性最穩定者，能用於多環境區域比較試驗。

種原庫的長處是可代表一個物種的全部遺傳歧異度，利用分層均勻隨機抽樣的方式，藉由過去已經建立種原基礎性狀進行分類排序，利用多個主要性狀換算為特定綜合加權指數，然後將上萬個種原排序後，每均勻間隔取樣全部取出10%，使得在最少試驗族群數下能保有最大遺傳變異，但限制是試驗材料內的遺傳異質結合比率可能過高，尤其是由開放授粉品種組成的天然族群，導致基因型相互混雜，使試驗結果不易區分。此試驗材料可解決過去受限於栽培種僅能

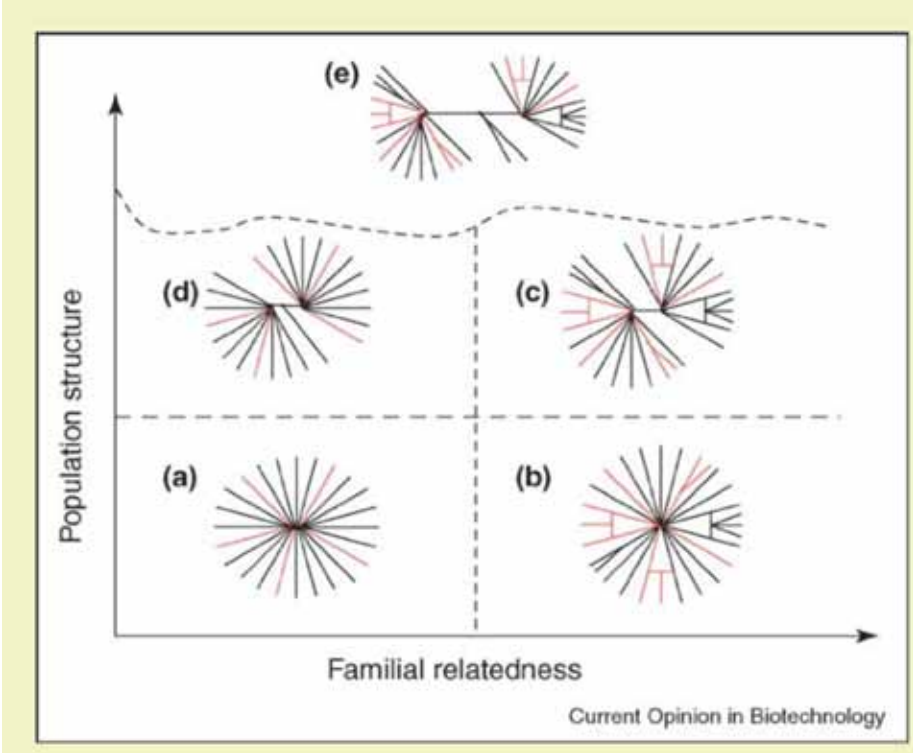


圖一、在兩基因座間不同連鎖失衡程度從完全連鎖失衡(子圖A)至完全平衡(子圖C)中，圖解說明連鎖失衡係數平方相關係數(r^2)與連鎖失衡標準化係數(D')間的差異，其中r^2係數會受遺傳重組與發生突變的時間點的影響，而D'係數僅受到世代間遺傳重組的影響 (修改自Flint-Garcia et al., 2003)。

針對已馴化的性狀，如種子休眠性、開花類型等，最適合針對抗病性或特定質量性狀 (顏色、香味)，不適於分析數量性狀，因非原產地環境可能造成產量性狀表現不佳等。

以優良栽培品種為試驗材料，其所得結果能使分子標記輔助篩選效率發揮至最大，並因對普通栽培環境適應性高，適於評估低遺傳率性狀，產量、產量構成要素、非生物逆境的耐性，且經由多年度、多區域重複試驗下累積大筆外觀性狀紀錄可供分析。但一般栽培品系往往是近代中來自一個基礎族群，並經過高選拔強度，因此LD強度應很高，而與雙親雜交的定位族群相較之下，其關聯性定位的解析度雖未改進，但在目標族群中能得到大量有用對偶基因是一大優點。

另一種仿如合成品種選育模式一般，由多個基本親本互交且進 多世代的循環雜交，然再經過多世代自交固定互交後裔，讓每個品系都充分均勻獲得來自不同親本的遺傳組成，其族群遺傳變 比過去雙親雜交族群要大，多世代互交可打破 鎖重組機會高，且並累積更多親本可以定位 多種類的QTL，人為創造下可視為逢機交配的族群，因在品種建立與維持時均要求減低品系自交率，其族群結構的干擾會最低，並因起始世代的重組少，連鎖失衡會最高，可偵測到染色體大片段與目標性狀具有關聯性，解析度雖較低，



圖二、應用關聯性定位策略中，圖解說明各類定位族群中族群結構與譜系關聯對於族群親緣變化，(a)期望逢機族群不具次族群分化與譜系關聯，(b)顯示樣品間存有譜系關係，(c)同時包含次族群分化與譜系關聯，(d)則顯示僅存次族群分化，(e)表示定位族群中同時包含多種次族群分化與譜系關聯 (修改自Yu and Buckler, 2006)。

而後期世代的遺傳重組現象增加，定位解析度會提高，加上人為選拔促使增加有利基因的組合，更易偵測到優良基因型，優點眾多。

而連鎖失衡會受族群結構、植株授粉特性等影響；以族群結構而言，會因花期早晚、株高高低或親緣譜系關係等影響 (如圖二)，導致試驗材料間具有事先非獨立關係存在，使控制目標性狀的基因頻度在該族群中應該會比無目標性狀族群中的高；族群若有再分化、該次族群個數越多時，關聯性係數也會越高，而呈現偽關聯性，但可藉由非連鎖的分子標記檢測試驗族群結構，即以各次族群分群比較時，存有基因頻度差異，則是因特定族群結構所致，而以目標性狀分群比較時，存有基因頻度差異，才是控制該目標性狀的基因座。授粉特性會導致差異，是因自交作物上的基因座為同質結合體若發生遺傳重組也無意義，因此異交作物的連鎖失衡在長達500 bp之後會發生衰降，但自交作物的連鎖失衡距離則能保持10 kb左右，以此延伸出分子標記對控制目標性狀基因的適切偵測密度，以異交作物玉米而言是100至200 bp，自交作物阿拉伯芥則是50 kb。而現今檢定作物關聯性定位分析大多採用混合線性模式 (Mixed linear model, MLM)，在模式中同時考量族群結構與兩兩個體間的親緣關係，在模式中多以Q矩陣以及K矩陣代表上述結構與親緣關係，比起廣義線性模式 (General linear model, GLM) 未考量上述兩種因子，更能控制第一型與第二型錯誤。

四、應用展望

隨著越來越多重要作物性狀的目標基因已被定位且選殖出，可依此開發功能性分子標記，或自行利用分離族群定位控制目標性狀之重要基因，定位結果往往局限於該族群親本間，不易推延至該作物核心種原族群內，大幅降低育種家藉由分子標記輔助選種、親本評估之效益。因此基於過去應用標誌建構基礎篩檢能力，在初期開發階段針對單點基因座進行性狀選育與評估，隨著基因型分析門檻快速遞減，能針對複雜性狀提升至基因組全面性評估，就各優良作物種原建立重要性狀關聯性定位分析，探討種原內目標基因之分布情形，落實數量性狀基因座定位研究之應用性，建置重要性狀基因型與外表型關聯性之快速分析平台，依各目標性狀提供育種家優良親本候選清單，以提升逆境育種之育種效率。

五、參考文獻

Breseghello, F. and M. E. Sorrells. 2006. Association analysis as a strategy for improvement of quantitative traits in plants. Crop Science 46(3):1323-1330.

Flint-Garcia S. A., J. M. Thornsberry and E. S. 4th Buckler. 2003. Structure of linkage disequilibrium in plants. Annu Rev Plant Biol. 54:357-374.

Yu, J. and E. S. Buckler. 2006. Genetic association mapping and genome organization of maize. Current Opinion in Biotechnology 17:155–160.