知識天地

淺談有關基因定位的統計方法

高振宏 (統計所副研究員)

自從奧地利人孟德爾在 1866 年發表他對豌豆種皮為圓滿或有縐紋、子葉為黃或綠色、花為紫或白色、豆莢為平莢或縐莢、豆莢為綠或黃色、花的著生位置，及莖的長短（高或矮）這七個性狀的遺傳試驗結果後，gene（基因）這個廣為一般人認知的觀念才逐漸浮現。孟德爾研究的性狀，因很容易分門別類故稱自為質性狀（qualitative trait）。一般相信，質性狀為單一或少數幾個基因所控制且不易受環境影響，此類基因，因對質性狀有完全或主要的操控性，故稱為主基因（major gene）。生物的另一類性狀例如人類的身高、體重、高血壓、糖尿病；水稻株高及產量對疾病的抵抗程度；老鼠的體脂肪百分比；乳牛的乳產量；雞的產卵量，由於其變異性是連續性的，不易分類，且易受環境影響，故稱為數量性狀（quantitative trait）。數量性狀是由多個基因所控制，由於每個基因對數量性狀均有影響，所以每一基因的作用便相對地小。這些控制數量性狀的基因稱為微效基因（polygenes）或又稱為數量性狀基因座(quantitative trait loci，QTL)。生物上許多重要的經濟、生理、生化或與演化機制有關性狀皆為數量性狀。如果能夠了解控制數量性狀的 QTL，進而操控之，自然能夠增加及改良數量性狀的量與質，造福人類，也能回答許多有關生物遺傳與演化上重要問題。故對於 QTL 的研究，一直是生命科學上的重要課題。

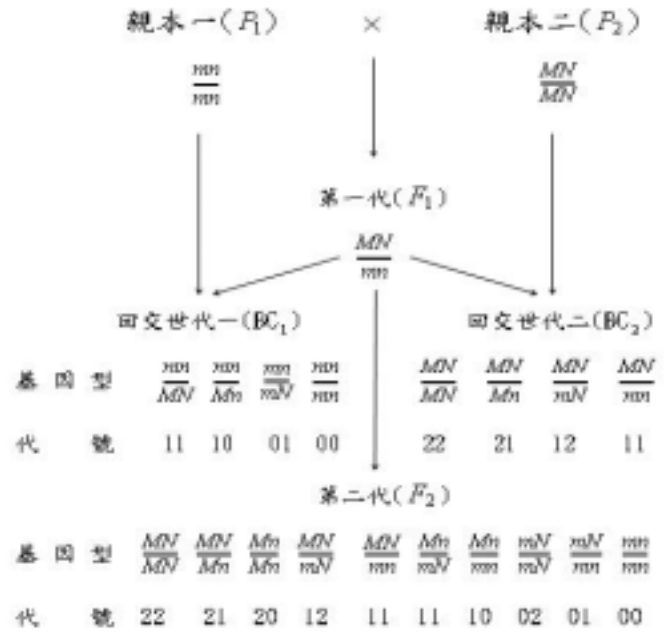
雖然，控制質性狀的主基因和數量性狀的 QTL 遵循相同的遺傳法則，但是由於數量性狀無法明確分類且有易受環境影響的傾向，使得孟德爾對的豌豆性狀研究的方法並不適用於 QTL 的研究上，這使得 QTL 研究較主基因研究困難許多。傳統上對於 QTL 的研究乃利用特殊的雜交試驗，將數量性狀的變異劃分為遺傳和非遺傳成分，而由其中遺傳變異在總變異所佔的比例（遺傳率）來估計對數量性狀的影響程度。這個傳統的方法顯然只能籠統地描述 QTL 對數量性狀的影響程度，並無法提供諸如究竟有幾個 QTL 控制數量性狀，這些 QTL 在染色體上的位置，各個 QTL 作用的大小以及那些 QTL 間有交互作用等，這些長久以來人們一直想得到的重要訊息。於是當時一般人認為要獲得這些個別的 QTL 訊息是不大可能的。近年來，由於分子生物學的快速發展，任何物種的染色體（DNA）經生物技術的處理後可以產生大量的分子遺傳標識（molecular genetic marker），例如 RFLP , RAPD , AFLP , VNTR 等，都是常用的分子遺傳標識。能夠產生這些分子遺傳標識，就如同染色體上可標上記號一般，這種進步提供人類進一步去了解有關 QTL 精確訊息的機會。有了這些資訊自然有助於生物學家們利用各種方法操作 QTL，以更快速、更有效地改進數量性狀，並回答生命科學上的重要問題。而利用分子遺傳標識資料去估計 QTL 訊息的研究，泛稱數量性狀基因座定位 (QTL mapping)。

QTL mapping 研究有數個重要環節，其中兩個重要環節是資料的產生及 QTL 的估計。在資料產生階段，適當的試驗設計相當重要，如親本的選擇及樣本數大小的決定，往往在研究上具有關鍵地位。在親本的選擇上有幾項原則，如儘量選擇親緣關系較遠，且性狀值差異較大者為親本。例如若要定位與控制水稻產量有關的 QTL，傾向找一非常豐產與另一非常低產的品系作為親本。若要定位與抗某疾病有關的 QTL，則會選定

一非常抗病與另一非常易染病的品系作為親本。若要研究果蠅精子競爭(sperm competition)的遺傳機制，會選擇一非常有競爭力與另一競爭力非常弱的品系作為親本。這樣選取親本的主要原因，是儘量使父母本帶有不同的對偶基因(allele)，使後代基因型可有不同組合，而進一步利用這些差異性來進行研究。親本選定後，一般經由回交（backcross）或 F2 設計來產生資料(樣本)供進一步分析(例如樣本的數目可能為數百個)。在回交的過程中（圖一），F1 世代中每一個體在每個基因座上皆為異質結合體（heterozygote），此異質結合體的一個對偶基因來自 P1，另一對偶基因來自 P2。若以小寫英文字母表示 P1 的對偶基因，以大寫英文字母表示 P2 的對偶基因，則 F1 個體的每個基因座上有一大寫字母與一小寫字母之對偶基因。若再將 F1 回交 P1 或 P2 就產生所謂回交族群。回交族群中之個體的每一基因座上的基因型有兩種類型，不是同質（MM 或 mm）就是異質（Mm）結合體，其比例為１：１。若將 F1 個體自交或互交，則產生第二代（F2）族群。F2 個體的每個基因座上有三種可能的基因型，即 P1 同質結合體 (mm)，異質結合體 (Mm) 和 P2 同質結合體 (mm)。此三種基因型的期望比例為 1：2：1。若同時考慮兩個基因座，則 P1 之基因型為 MN/MN，P2 之基因型為 mn/mn。P1 所產生的配子（gamete）全為 mn 型，而 P2 所產生的配子全為 MN，使得 P1 和 P2 所產生的後代 F1 只有一種基因型 MN/mn。F1 所可能產生的配子由於染色體會重組的關係，有四種（並非兩種）型態。其中的兩種 MN 和 mn 為非交換型，另兩種為 Mn 和 mN 為交換型。若 F1 回交 P1，族群中的個體有四種可能的基因型 (圖一)。若 F1 個體自交或互交, 共可得如圖一所示之十種不同的基因型 (其中 MN/mn 與 Mn/mN 無法辨別，一般歸為九種)。若以代碼 2 己表示某基因之基因型為 P2 同質結合體，以 1 代表異質結合體而以 0 代表 P1 同質結合體。則回交族群之各基因座上之基因型代碼為 1 或 0 (1 或 2) 共兩種，而 F2 族群之代號可有 2，1 或 0 共三種（圖一)。以上是以兩個基因為例，來說明回交及 F2 族群的遺傳結構。當同時考慮 k 個基因時 則在回交及 F2 族群各有 2^K 和 3^K 種可能的基因型。若將整個染色體組（genome）上許多基因座的基因型皆以代號表示，則任一個體之染色體組成可用一組簡單的數字來表示（表一)。例如，表一中的回交族群之第一個個體的前九個標識之基因型皆為異質結合體(代碼 1)。F2 族群第一個體的第八、九個標識之基因型皆為 P2 同質結合體 (代碼 2) 。在實際的情況下，可能有上百個標識。每一個體除了標識基因外，也都伴隨有一個(或多個)數量性狀值 y 。理論上帶有越多對 y 有正向影響的基因型(x)的個體，其 y 值越大。

接著的步驟便是利用表一的資料，借用統計學的原理、原則來估計 QTL 的諸多訊息。以下以圖一為例略述之。如果 M 與 y 無關 (M 非 QTL)，而 N 與 y 有關(N 是 QTL)，則 y 在剔除外在環境的影響後，M 是 1 或 0(2、1 或 0)的改變，並不會影響 y 的改變，而 N 的改變卻會影響 y 的改變。反之亦然。若 M 與 N 皆與 y 無關(M 與 N 皆非 QTL)，則不論 M 與 N 如何改變，都不會影響 y 的改變。若 M 與 N 皆與 y 有關(M 與 N 皆為 QTL)，則 M 與 N 的改變，是會影響 y 的增減。若將標識基因比喻成各類工廠，那麼 y 即為工廠之某類產品之產量，而 BC 或 F2 的設計就類似隨機地讓每個工廠可能形成兩種和三種(相當於兩種和三種代碼)供電狀態。BC 的設計使每個工廠有 100%(代碼 2)或 50%(代碼 1)供電(50%供電(代碼 1)或完全停電(代碼 0))兩種狀態，而 F2 的設計使每個工廠有 100%供電(代碼 2)、50%供電(代碼 1)或完全停電(代碼 0)三種狀態。對 k 個工廠(標識)而言，總共有 2^k 個和 3^k 種不同的供電狀態。每一個樣本可能是這 2^k 個或 3^k 種狀態的一種。如果產品是面板(y 為面板產量)，則那些面板工廠(QTL)是 100%供電(完全停電)的樣本，會有較高(非常低)的面板產量。反之，那些與面板無關之工廠(如紡織或食品工廠)的供電狀況並不影響面板產量。我們是透

過各種不同工廠的供電情形(標識基因型)，與面板產量的變化來找出那些面板工廠(QTL)及估計其產能。在這樣的邏輯下，統計方法可以很有效地找出那些 QTL(面板工廠)及估計其關係的大小(產能)。然而，實際情況會更複雜，而使得 QTL mapping 在統計學的領域上更富有挑戰性。第一是當影響 y 的 QTL 數目變多時，這些 QTL 的改變與 y 的增減關係變得極為複雜。第二是環境效應可能相當大，會遮蓋住基因的效應。第三是 QTL 極可能不正好在標識上。在這些考量下已有更完整、更有效的策略被提出來研究 QTL。由於篇幅有限，無法多做說明，有興趣進一步了解的讀者請參考中國統計學報第三十八卷第一期有關數量性狀基因座定位的統計方法介紹。有關 QTL mapping 的應用研究在生命科學上相當廣泛，例如對於水稻、黃米、蕃薯、松樹、尤加利、高粱、老鼠、雞、豬、果蠅等多種生物的許多性狀的研究隨處可見，也已回答了許多生命科學的重要問題。有關 QTL mapping 研究的流程、步驟可說是經緯萬端，需各領域互相配合(生命、資訊、數學、統計等科學)，為一典型的跨領域課題。其關心的基本上是複雜性狀的遺傳機制，為一有趣且重要、值得進一步投入的研究。



圖一：回交和 F_2 族群的產生流程和遺傳結構

表一：數量性狀基因座定位的資料型式

回交族群

個體	性狀值(y)	遺傳標識(x) 1 2 3 4 5 6 7 8 9 …
1	179.80	1 1 1 1 1 1 1 1 1 …
2	207.80	1 1 1 1 0 0 0 0 0 …
3	240.80	1 1 1 1 0 0 1 1 0 …
4	212.17	0 0 0 0 0 0 1 1 1 …
5	202.17	1 1 1 0 0 0 0 0 0 …
6	197.17	0 0 0 0 0 0 0 0 1 …
⋮	⋮	⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮

F_2 族群

個體	性狀值(y)	遺傳標識(x) 1 2 3 4 5 6 7 8 9 …
1	85.5	0 1 1 1 1 1 1 2 2 …
2	89.0	2 1 2 2 2 2 2 2 2 …
3	94.5	2 2 2 2 2 2 1 1 2 …
4	84.5	1 1 2 2 2 2 2 2 1 …
5	86.5	2 1 0 0 0 1 1 1 0 …
6	86.5	0 1 1 1 1 2 2 2 1 …
⋮	⋮	⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮