國立臺灣大學生物資源暨農學院農藝學系

碩士論文

Department of Agronomy

College of Bioresources and Agriculture

National Taiwan University

Master Thesis

探討族群特有基因座與連鎖失衡區塊圖像不完整現象之關聯

Explore the Association between Population-specific Loci and the Incomplete Pattern of Linkage Disequilibrium Blocks

翁兆平

Zhao-Ping Weng

指導教授：陳凱儀 博士

Advisor: Kai-Yi Chen, Ph.D.

中華民國 106 年 3 月

March 2017

doi:10.6342/NTU201700716

誌謝

如何讓人生不是黑白？陳凱儀老師說“要為自己訂定一些目標並且努力完成它。”我想碩士學位就是現階段一個具有一定挑戰性的目標，畢竟是需要做出一點成果的。這段期間跟著老師學做事，不僅僅是在培養學術研究上的態度與邏輯思考的訓練，也同時讓我有機會接觸研究以外的事物。舉凡擔任助教、管理田間、測試儀器等等，老師皆以信任並且開放的態度讓我自由發揮。感謝老師為我們創造這樣良性的學習環境。感謝口試委員胡凱康老師、劉力瑜老師及黃永芬老師，對於論文的建議與指導，使得本篇論文更為完整。本篇所使用的玉米研究材料感謝農試所與台南農改場慷慨提供，尤其陳裕儒學長和詹雅勛學姊提供種原的背景資料與充足的葉片資源。MaizeSNP50 基因型分析資料產生的部分感謝生物科技研究所蔡孟勳老師研究室進行。

感謝實驗室的大家，像小老師般維持秩序的竹茵學姊，在田間及實驗室中解決疑難雜症的亞平學姊和祐祐學姊；毛毛學長的著作給予我在寫作上的靈感；當助教期間互相協助的好夥伴瑋倫；一起討論各種議題的易整；還有為實驗室帶來活潑氣氛的依臻、品堯、昱富和繼勤。謝謝一起進入碩士班時常互相打氣的同學們，岳儒、士庭、子涵、聖凱、弈廷、哲嘉、思妤、謙謙、松翰、佳芸、昱安，簡單的聊聊天說說近況就可以督促彼此，提供充足的精神力量。感謝從大學就加入的國術社，讓我培養出一項運動專長並且認識擁有共同興趣的朋友們。

感謝農藝系的各位老師，在學術知識上的傳播與教學品質上的堅持。最後感謝我的父母一直以來的栽培。即將邁入人生的下一個階段，謝謝臺大讓我知道，知識就是力量，而更重要的是擁有探索未知的勇氣與待人處世的同理心。

中華民國一○六年三月二十一日 翁兆平

doi:10.6342/NTU201700716

中文摘要

全基因體組關聯性分析 (Genome-Wide Association Study; GWAS) 的研究結果時常可見連鎖失衡區塊圖像不完整的現象。本研究欲探討連鎖失衡區塊圖像不完整的成因，是否和參試樣本不同次族群間起源不同的核苷酸變異有關。研究使用 15,919 個具有物理圖譜位置的基因座在 48 個玉米自交系中皆為同型結合基因型且沒有缺值的資料進行分析，此資料取自 MaizeSNP50 基因晶片的分析結果。首先將「參試樣本不同次族群間起源不同的核苷酸變異」定義為族群特有的分子標記位點。參試的玉米自交系不論由種原的來源記錄、STRUCTURE 軟體分析、或是 PCoA 分析方法都確定可以區分為溫帶馬齒種、溫帶甜質種、與熱帶種共三個次族群。藉由 STRUCTURE 軟體分群分析計算所得的 Q 值，建立次族群代表品系，並篩選出次族群特有的分子標記。其次「造成連鎖失衡區塊圖像不完整的分子標記」的篩選。將具完整圖像的連鎖失衡區塊定義為三個以上相鄰分子標記之間任一成對分子標記間相關係數 R^2 大於 0.8 的染色體區間。造成連鎖失衡區塊圖像不完整的分子標記即為連鎖失衡區塊內與兩側相鄰分子標記的 R^2 皆小於 0.8 的分子標記。上述資料分析結果顯示，使用 24 個代表品系可篩選出 6,696 個「族群特有的分子標記」，其中有 195 個分子標記是被定義為「造成連鎖失衡區塊圖像不完整的分子標記」。為了檢測「造成連鎖失衡區塊圖像不完整的分子標記」與「族群特有的分子標記」是否具有關聯性，使用 bootstrap 的概念，重覆進行 10,000 次取樣，每次取樣為自 15,919 個分子標記隨機抽取 6,696 個分子標記，並檢視不同數目的「造成連鎖失衡區塊圖像不完整分子標記」被隨機取得的機率。檢測結果顯示要隨機取得至少 195 個「造成連鎖失衡區塊圖像不完整」分子標記的機率僅為 0.6%。此結果顯示「造成連鎖失衡區塊圖像不完整的分子標記」與「族群特有的分子標記」是具有相關性的。

關鍵字詞：連鎖失衡區塊圖像不完整、族群特有分子標記、STRUCTURE、族群代表品系、bootstrap。

doi:10.6342/NTU201700716

Abstract

In the genome-wide association study (GWAS), incomplete pattern of linkage disequilibrium (LD) blocks were often found. This study is to explore whether incomplete pattern of LD blocks is related to nucleotide variations from different subpopulations. Data containing genotypes of 48 maize inbred lines was obtained by the MaizeSNP50 BeadChip and was used to address the aforementioned question. A number of 15,919 single nucleotide polymorphic markers, also known as the SNP loci, with known physical positions at maize reference sequences, were selected because their genotypes were all homozygous and had no missing value among 48 maize accessions. The nucleotide variations from different subpopulations were defined as "the subpopulation-specific SNP loci". The 48 maize inbred lines used in the current study can be classified as three subpopulations: temperate dent, temperate sweet and tropical. This classification was consensus between the analyses of the STRUCTURE software and the PCoA analysis, as well as the original records attached to these inbred lines. The representatives of different subpopulations were selected based on the three Q values of the STRUCTURE software, which indicate the proportions of genetic components from each of subpopulations. The subpopulation-specific SNP loci were then defined as those showing DNA polymorphism solely in one particular subpopulation. Using the same genotype dataset, "the SNP loci making incomplete pattern of the linkage disequlibrium blocks" were selected independently. The linkage disequlibrium (LD) block was defined as a chromosome region containing more than three flanking SNP loci and the correlation coefficient between at least a pair of the SNP loci in the LD block was greater than 0.8. "The SNP locus making incomplete pattern of the LD blocks" was then defined as the locus in the LD block which had the correlation coefficient less than 0.8 with its flanking loci on both sides.

doi:10.6342/NTU201700716

The data analyses identified 24 representatives for three subpopulations, and a total of 6,696 "subpopulation-specific SNP loci". Among these "subpopulation-specific SNP loci", a number of 195 markers were also idetifined as "the SNP loci making incomplete pattern of the LD blocks". In order to test whether the "population-specific markers" and "the SNP loci making incomplete pattern of the LD blocks"are associated, a total of 10,000 resamplings by the bootstrap approach were made to build the probability mass function for the number of the randomly drawn SNP loci in the LD blocks. From each resampling, a number of 6,696 SNP loci were randomly drawn from 15,919 SNP loci, and then the number of the SNP loci sitting in the LD blocks was recorded. The result showed the cumulated probability to obtain at least 195 random SNP loci sitting in the LD blocks was 0.6%. This result inferred that "the subpopulation-specific SNP loci" and "the SNP loci making incomplete pattern of the LD blocks" are closely associated.

Key words: incomplete pattern of the linkage disequilibrium blocks, the subpopulation-specific markers, STRUCTURE, representatives of different subpopulations, bootstrap.

doi:10.6342/NTU201700716

目錄

doi:10.6342/NTU201700716

圖目錄

doi:10.6342/NTU201700716

表目錄

doi:10.6342/NTU201700716

第一章　前言

第一節　族群特有基因型

物種的演化經歷遷徙、地理隔閡、突變與自然環境的變化等等，不同的天擇壓力和遺傳漂變等現象會造成族群間的外表性狀與遺傳序列具有一定程度的差異。演化歷程造成的族群差異會表現在基因座上對偶基因的頻度上，當具有強烈的天擇壓力下，過少的有效族群個體數可能導致族群中特定對偶基因的消失，呈現特定基因座基因型被固定的情況。而族群分化後發生的突變若不影響存活且可以繁衍，則會造成此族群具有新的對偶基因，這些現象造成族群間基因體結構的差異與族群特有基因型的發生。科學家利用這樣的特性進行物種內不同族群間演化歷程或基因滲入證據的研究。Yamanaka 等人（2004）依據前人對於水稻糯性基因的研究結果，得知在 *waxy* 基因座上， *wx* 為糯質水稻的對偶基因、 Wx^a 為印度型非糯質水稻的對偶基因、 Wx^b 則為日本型非糯質水稻的對偶基因，並發現其後兩者在第一個內含子（intron）內，4 個核甘酸長度的片段上具有單一核苷酸的差異，分別為 AGGT 與 AGTT。研究者以此單一核苷酸多型性 （Single nucleotide polymorphism, SNP） 設計 dCAPS（Derived Cleaved Amplified Polymorphic Sequences）分子標記並掃描 353 個糯質水稻的基因型，結果在所有試驗品系中有 342 個為 AGTT，11 個為 AGGT，因而推論 *wx* 主要為 Wx^b 突變而來，並且認定糯質水稻在演化歷程上與水稻印度型、日本型之間的分化獨立。而 Morrell 等人（2005）以 RFLP 分子標記選出 77 個美洲高粱栽培種內特有的對偶基因型，以此組分子標記測試 5 個不同地區石茅（Johnson grass）族群的基因型，發現有 30 個栽培種特有對偶基因至少存在於一個石茅族群中，證明栽培種高粱與石茅間有基因滲入的現象。

隨著分子標記技術的演進，族群特有基因型的相關研究也將分析標的從單一基因座對偶基因間的序列差異轉變為全基因體 SNP 分子標記在族群間差異的研究，實驗者的研究顯示 SNP 分子標記在族群的尺度上具有與對偶基因相同的特

doi:10.6342/NTU201700716

性，部分 SNPs 的基因型僅在特定族群內存在。Chao 等人（2010）以小麥作為實驗材料，蒐集分別來自 17 個不同的育種計畫的品系，共 478 個北美與中美洲的春小麥與冬小麥，以育種計畫作為分群依據，配合 1536 個分別分布於 A、B、 D 三個不同基因體的 SNP 分子標記，進行依育種計畫來源或基因體分群的族群結構與連鎖失衡的相關分析。在實驗結果上以人為育種計畫作為族群分類的依據，在主成分分析（Principal components analysis, PCA）上並不能明顯的區分。而以冬小麥與春小麥作為分群依據，則可以有良好的分群結果。這樣的現象同樣反映在 STRUCTURE 軟體的分析上，在此實驗中發現分別有 52 和 97 個 SNPs 在春小麥與冬小麥的族群內不具有多型性呈現單一基因型（Monomorphic），表示部分 SNPs 的基因型僅存在於特定族群中。在這篇研究中也同時討論到族群間連鎖失衡的問題。分析結果顯示連鎖失衡衰退的速度，在此篇研究中來自不同育種計畫的小麥族群間沒有明顯的差異，小麥的不同基因體間也沒有明顯差異，但以春小麥與冬小麥族群分別繪製的連鎖失衡區塊，在相同染色體位置上具有不同的圖像。Lam 等人（2010）在大豆的研究中也討論到此種族群間連鎖失衡區塊差異的現象，並提到與族群特有 SNP 的關聯。以野生種與栽培種為分群，兩群皆觀察到各自具有特有連鎖失衡區塊的情形。平均來說栽培種族群具有較大且連續的連鎖失衡區塊、較多被固定的 SNP 且具有較少特有的 SNP，但特有的不良突變（Deleterious mutations）所造成的 SNP 在該族群中的比例卻比較高。作者認為此現象為連鎖失衡造成的突變累積的結果。在族群特有基因型的研究上，依據分群方式的不同會具有不同的研究結果。然而，特定基因型僅在特定族群存在的現象很普遍，從對偶基因的研究到 SNP 分子標記上皆有發生，並從單一位點的討論發展到連鎖失衡區塊圖像的討論。

第二節　連鎖失衡相關利用與研究

分子標記技術的出現，使得遺傳研究不再僅能依靠具有高遺傳率的外表性狀或譜系資料，實驗者能以特定核苷酸序列的特性探討性狀的遺傳，使得遺傳圖譜

doi:10.6342/NTU201700716

的解析度有明顯的進展。隨著技術的演進，高通量的定序技術促使分子標記的大量產生，遺傳研究上的限制因子由分子標記的數量轉變為所使用遺傳材料的數量。傳統的遺傳材料包括人為培育的重組自交系（Recombinant inbred lines, RILs）與近同源系（Near isogenic lines, NILs），其基因座多為同質結合、具明確的譜系資料且有充足的後代可供驗證，然而由於試驗個體的染色體互換次數與其同一基因座上對偶基因個數的侷限，使得遺傳圖譜的解析度無法再進一步提升。為了解決這個問題，前人提出兩個主要的方式。第一種方式為建立多親本的重組自交系族群（Multi-parent advanced generation inter-cross, MAGIC）。Yu 等人（2008）以玉米為材料提出由 26 個品系所建立，共 5,000 個重組自交系組成的 Nested Association Mapping （NAM） 族群。Kover 等人（2009）則以 19 個阿拉伯芥品系建立 527 個多親本雜交的自交系族群。第二種方式為利用蒐集系進行關聯性定位（Association mapping）。相較於建立多親本的重組自交系，此種策略在時間與金錢上都具有更高的優勢。然而由於樣本間沒有明確的譜系關係，以連鎖圖譜（Linkage map）定位基因的方式不再合適，而以連鎖失衡（Linkage disequilibrium）為基礎進行的關聯性定位成為重要的研究領域。

統計學上兩獨立事件同時發生的機率為兩獨立事件發生機率之乘積，連鎖失衡即利用此觀念評估兩基因座是否具有關聯性，目前最主要的評估方式為 D' 與 R^2。D' 為估算族群中兩基因座對偶基因重組的機率，展現族群發展歷史上基因座互換的程度。R^2 則是估計族群中兩基因座對偶基因的相關性，其包含了遺傳重組與核苷酸突變兩種因子，可以更真實的反應試驗材料的基因型狀況。育種家判斷遺傳分析所需的分子標記數量時，會參考連鎖失衡衰退（Linkage disequilibrium decay）的狀況。然而，連鎖失衡的大小在相同物理距離或遺傳距離但不同的成對基因座間會有很大的變異，使得連鎖失衡衰退的距離不易判斷。Remington 等人（2001）提出以非線性回歸方程估計 R^2 期望值的方式。據此，實驗者可以在試驗族群中計算特定物理圖譜距離下的 R^2 期望值，並以其作為判斷試驗族群連鎖

doi:10.6342/NTU201700716

失衡衰退的程度，進一步評估連鎖失衡在染色體上的範圍。在同篇文獻中 Remington 等人以 102 個來自溫帶與熱帶區域的玉米自交系為材料，計算其連鎖失衡衰退的距離，其結果為在物理圖譜上距離約 1,500 bp 時 $R^2 < 0.1$。Hamblin 等人（2005）以高粱為研究材料，於世界各地蒐集 22 個地方種、8 個亞種及 2 個自交系，其物理距離約 15 kb 時 $R^2 < 0.1$。Mather 等人（2007）分別使用 21 個 *indica*、18 個 *tropical japonica* 和 22 個 *temperate japonica* 品系的水稻計算連鎖失衡，當 $R^2 < 0.1$ 時三個亞種具有不同的物理距離，*indica* 為 75 kb、 *tropical japonica* 為 150 kb、 *temperate japonica* 則大於 500 kb。Hyten 等人（2007）以大豆為實驗材料，蒐集 4 種不同類型的材料，分別為 26 個亞洲 *Glycine soja*、 52 個亞洲地方種、 17 個美洲現行品種的先祖和 25 個美洲現行栽培種，以 $R^2 < 0.1$ 為標準，*Glycine soja* 的連鎖失衡大小為 77 kb、 亞洲地方種為 90 kb、 美洲先祖為 212 kb ，而現行栽培種則為 574 kb。Wu 等人（2014）以 367 個中國的優良自交系（elite inbred lines）玉米為試驗材料，其連鎖失衡衰退的距離達到 391 kb。綜合以上研究可以發現，在自然環境下，自交作物的連鎖失衡衰退的距離大於異交作物，而經人為選育後，連鎖失衡的衰退會大幅減緩，且其幅度大於物種間的差距。藉由連鎖失衡衰退的距離與基因體大小的資訊，可以評估在 GWAS（Genome-wide association study）研究時適當的分子標記數量。以連鎖失衡衰退快速的實驗材料、配合高密度的分子標記，可以進行高解析度的遺傳分析。

在遺傳研究上普遍觀察到染色體上不同區域的互換率是不均等的，因此觀察特定染色體範圍內連鎖失衡關係時，可以發現部分染色體片段具有多個分子標記連續高連鎖失衡的現象。Daly 等人（2001）以密度為 5 kb 的分子標記分析大小約 500 kb 的人類 5q31 染色體片段，發現連鎖失衡的現象可以在 5 kb 的距離內驟降，也可以在長達 100 kb 範圍內的分子標記讀值僅有 4 種排列組合。Daly 等人將這樣的現象稱為單倍型結構（haplotype structure）。Goldstein（2001）以島

doi:10.6342/NTU201700716

嶼狀描述此連鎖失衡現象（islands of linkage disequilibrium），並認為連鎖失衡的驟降源自於染色體中互換率高的區域（hot spots）。隨著 SNP 分子標記的大量產生，分子標記的數量與遺傳分析的資訊量不再是線性成長。單倍型圖譜（haplotype map, Hapmap）藉由分析 SNP 間的連鎖失衡關係，將多個具高度連鎖失衡的 SNP 視為單一單倍型，並以少數標記 SNP（tag SNP）代表整個單倍型，大量除去不能提供更多資訊量的 SNP（redundant SNP）。早期的研究並不區分單倍型區塊與連鎖失衡區塊的差異，然而 Olivier（2003）探討單倍型區塊（Haplotype block）與連鎖失衡的關係時，提出連鎖失衡區塊（Linkage disequilibrium block）不等於單倍型區塊，而應為單倍型區塊的組成成分。互換率高的熱點除了形成單倍型區塊的邊界之外，也會在族群演化的歷程中打破原始的單倍型（ancestral haplotype）並形成 SNP。這些 SNP 決定這個單倍型區塊中單倍型的種類個數，而連鎖失衡區塊則是單倍型區塊中穩定具連鎖失衡的區域。單倍型圖譜的研究與連鎖失衡區塊密不可分，單倍型區塊的界定也依不同研究而有不同的見解。然而，連鎖失衡區塊本身的研究與討論卻比較少。連鎖失衡區塊的圖像並非一定完整，其內部的分子標記間不一定呈現連續的高度連鎖失衡，會發生有部分分子標記與區塊內的其他分子標記為低度連鎖失衡的現象。依據前人的研究（Goldstein, 2001; Olivier, 2003），連鎖失衡區塊是染色體中低互換率的區域，故此區塊圖像的不完整現象，應有其他的影響因子可以探討。

第三節　族群結構

在育種領域上除了精準符合需求的育種目標、適當的選拔工具與嚴謹的田間試驗之外，具有足夠遺傳多樣性的遺傳材料是重要的決定因子，因此種原的選擇與保存十分重要。考量保存種原與引種的成本，育種家希望在保持多樣性的同時，降低所需要的種原個數，而族群的區分是常見的參考指標。分群方式從外表型、譜系關係發展到以分子標記作為主流的判斷依據，Inghelandt 等人（2010）甚至認為判斷玉米自交系間的遺傳關係時，利用分子標記計算遺傳距離比依靠譜

doi:10.6342/NTU201700716

系資料更為合適。在基因研究上，關聯性分析與蒐集系的使用也加深了試驗材料分群的重要性，由於基因座上不同族群間對偶基因的頻度會因為族群的演化歷程上經歷瓶頸效應（bottleneck）或創始者效應（founder effect）等等而發生遺傳漂變（genetic drift）的現象，進而形成差異，試驗族群間基因頻度的不均等會造成當直接以性狀與分子標記的讀值進行關聯性分析時容易產生偽陽性（false positive），而造成實驗的誤判（Knowler et al. 1988）。為此，研究者提出族群結構的概念，藉由族群結構矩陣以族群為單位進行校正，降低誤判的機率。

Pritchard 等人（2000）將分群的方式整理成兩個策略，第一種為依距離分群（distance-based methods）泛指利用某種特定參數計算兩兩個體間的距離，再將此距離以樹狀圖或是多維的散布圖，視覺化呈現分群結果的方式，第二種為依模型分群（model-based methods）指建立一個參數模型，將參數代入後計算出個體應屬於何種族群。Pritchard 等人並進一步提出一個模型分群的運算模型成為 STRUCTURE 軟體的理論依據，其利用貝式定理（Bayesian method）並假設一個族群內的基因型應符合哈溫定律，即追求最低的連鎖失衡現象，最後用馬可夫鏈蒙地卡羅法（Markov chain Monte Carlo methods, MCMC）計算以獲得結果，研究者僅需試驗個體分子標誌的基因型讀值，加上事先假設總族群個數，即可獲得每個試驗個體分別為哪些族群以何種比例組成的資訊，相對於依距離分群的方式，在族群劃分上提供更明確且具有統計意義的界線。Evanno 等人（2005）模擬不同演化情境形成的族群資料，提出以 STRUCTURE 軟體估計總族群個數的方式，使得此軟體的使用更為廣泛。Camus-Kulandaivelu 等人（2006）以此軟體應用於美洲與歐洲的玉米族群研究。Jin 等人（2010）應用於目的為進行關聯性分析的中國水稻族群結構的研究。Wang 等人（2012）應用於中國小米地方種的族群結構研究，足見此軟體已經應用在不同作物的族群研究上。

然而隨者 SSR 分子標記逐漸被 SNP 分子標記取代，Price 等人（2006）指出在全基因體關聯性分析時，SNP 分子標記數量的大量提升，放大了

doi:10.6342/NTU201700716

STRUCTRUE 軟體對於硬體運算資源需求高的問題，提出將 PCA（Principal components analysis）應用於族群結構分析的想法，以第一、二特徵值（eigenvalue）繪製的散布圖中也成功看到分群的趨勢，並以 PCA 的結果校正個體的基因型值進行關聯性分析。Inghelandt 等人（2010）評估 SSR 與 SNP 在族群結構研究上的效益，提出在研究中 SNP 分子標記的使用個數在 SSR 的 7 到 11 倍之間時，在族群結構與遺傳多樣性分析上可以得到相近的結果。Inghelandt 等人發表文章的當下，單一 SSR 分子標記的價格已經為 SNP 的 23 倍，而隨著次世代核酸定序技術的進步，SNP 的價格還會繼續降低。這些研究成果，使得依距離分群的方式成為族群結構應用在關聯性分析時主流的策略，而模型分群的應用方向則以族群劃分的研究為主。

第四節　MaizeSNP50 BeadChip

21 世紀初物種的定序計畫陸續發表，為生物研究立下重要的里程碑。然而單以參考序列的資訊不能滿足研究者對於分子生物或遺傳學研究的需求，在其研究上必須實際獲取基因型才可進行基因功能或遺傳的分析，因此需要高可靠性、有效率、大規模且符合經濟效益的方式偵測個體的基因型（Sapolsky et al. 1999）。基因晶片為因應此需求而發展出的產品，以微陣列（microarray）的方式將一組欲檢測之分子標記設計於晶片上，並借由螢光與影像技術，進而同步獲取多個基因型資料。其晶片設計的方式，早期為利用互補 DNA（Complementary deoxyribonucleic acid, cDNA），從 cDNA 基因庫選出目標的序列經 PCR 擴增後固定於晶片上，進行雜合。其後發展出利用寡核苷酸（oligonucletide）的策略，省去 cDNA 資源的需求與 PCR 擴增的時間，並且可以更精準的設計目標片段。Affymetrix 公司以原位合成（*in situ* synthesised）的方式將寡核苷酸直接製作於晶片上（Schulze and Downward. 2001）。Illumina 公司則發展出 Bead array 技術，可在每個小於 3 微米的微圓珠 (bead)上接合數十萬個寡核苷酸序列，形成一個檢測單一分子標記基因型的信號點，再將微圓珠置於具有微孔的載體上製作

doi:10.6342/NTU201700716

成晶片，提高製程上的彈性（Shen et al. 2005），同時一次實驗所能檢測的總 SNP 個數僅局限於在單一晶片上所能放置的微圓珠總類數。另外，在微圓珠上所設置的寡核苷酸序列，依其檢測基因型所使用的原理，又可區分出不同的策略。目前主流的方式為 Infinium assay，其微圓珠上所使用的探針（寡核苷酸序列）為目標 SNP 位點旁 50 個鹼基對所組成的序列。樣本的 genomic DNA 經過擴增（amplification）、片斷化（fragmentation）與變性（denaturation）為單股後，進行雜合反應（hybridization），經過單一鹼基對延長（single-base extension）後即可對 SNP 位點的鹼基進行螢光染色，並以光學掃描獲得 SNP 基因型，大幅降低 DNA 合成過程中的錯誤率。其檢測策略的第一款產品為 Sentrix Human-1 Genotyping BeadChip，具有 100,000 個 SNP 位點，宣稱具有 99.4% 的 call rate 與 100% 的再現率（Steemers and Gunderson. 2005）。MaizeSNP50 BeadChip 即為 Illumina 公司在 Bead array 平台上利用 Infinium assay 的產品。

由於基因晶片為人為選定的一組分子標記，其泛用性與其所選定的分子標記緊密相關，因此在選擇上主要注意兩個要點。第一為分子標記的來源：分子標記探勘所使用的實驗材料其遺傳多樣性越廣，該標記可使用的範圍就越廣。第二為降低分子標記間的連鎖失衡現象：具有高連鎖失衡的兩分子標記並不能給予研究者更多的遺傳資訊，為提高基因晶片的使用效率，需降低任意兩分子標記傳達相同遺傳資訊的現象。依據 Ganal 等人（2011）的研究報告，MaizeSNP50 BeadChip 的 SNP 位點組成上即符合這兩個要素，其來源分為五個部分：（1）Panzea 機構所提出玉米第一代單倍型圖譜（First-generation haplotype map of maize）；（2）Syngenta 公司提出由 B73 與 Mo17 品系雜交族群所探勘之 SNP 分子標記，其代表硬稈（Stiff stalk）與非硬稈雜種優勢群的差異；（3）法國農業研究院（Institut National de la Recherche Agronomique, INRA）比較 B73 與 F2 ESTs 的差異所得之 SNPs；（4）TraitGenetics 公司比較 14 個歐洲與北美玉米品系桑格定序的資料所得之 SNPs；（5）已公開之玉米 SNPs，蒐集共 839,350 個

doi:10.6342/NTU201700716

SNPs，最後篩選至 56,110 個 SNPs。MaizeSNP50 BeadChip 組成最主要的來源為 Panzea 的玉米第一代單倍型圖譜，為 Gore 等人（2009）所建立，其利用 NAM（Nested Association Mapping）族群的親本，分別經三組限制酶切後，以 sequencing-by-synthesis （SBS）的方式取得 SNPs，再經連鎖失衡評估建立單倍型，定義標籤 SNPs（tag SNPs）而得。研究報告中的 NAM 族群為 Yu 等人（2008）為了提升連鎖分析與關聯性分析效益所建構之雜交族群，其親本包含 16 個熱帶玉米品系與 10 個溫帶玉米品系，具有廣大的遺傳多樣性。而 MaizeSNP50 在降低連鎖失衡的部分，除了單倍型圖譜的 SNP 本身即具有低連鎖失衡的特性之外，其篩選過程包括除去重複或具有相同毗鄰序列（flanking sequence）的 SNPs，以及轉換成 Infinium assay 平台後的品質篩選。此外，MaizeSNP50 單倍型圖譜 SNP 篩選過程也比對 Schnable 等人（2009）整理的 filtered gene set ，並除去在基因位置上有重疊 SNPs，在 SNPs 的取捨上以嘌呤與嘧啶轉換的突變為優先保留。MaizeSNP50 BeadChip 最後的 SNPs 密度約為每 40 Kb 到 50 Kb 有一個分子標記，相比 Remington 等人（2001）以 102 個玉米自交系評估出連鎖失衡距離約 1 Kb 的實驗結果，此基因晶片應具有較低的連鎖失衡現象。MaizeSNP50 的實驗團隊以 274 個玉米蒐集系測試此晶片，56,110 個位點中有 49,585 個位點可正常讀取並符合品質標準。在可正常讀取基因型的 49,585 位點中，B73 的 call rate 為 0.9987，而大芻草（teosinte）的 call rate 也達到 0.9187。表示 MaizeSNP50 晶片應用於大部分的玉米品系時都應至少可得到約 45,000 個高品質的 SNP 基因型讀值。

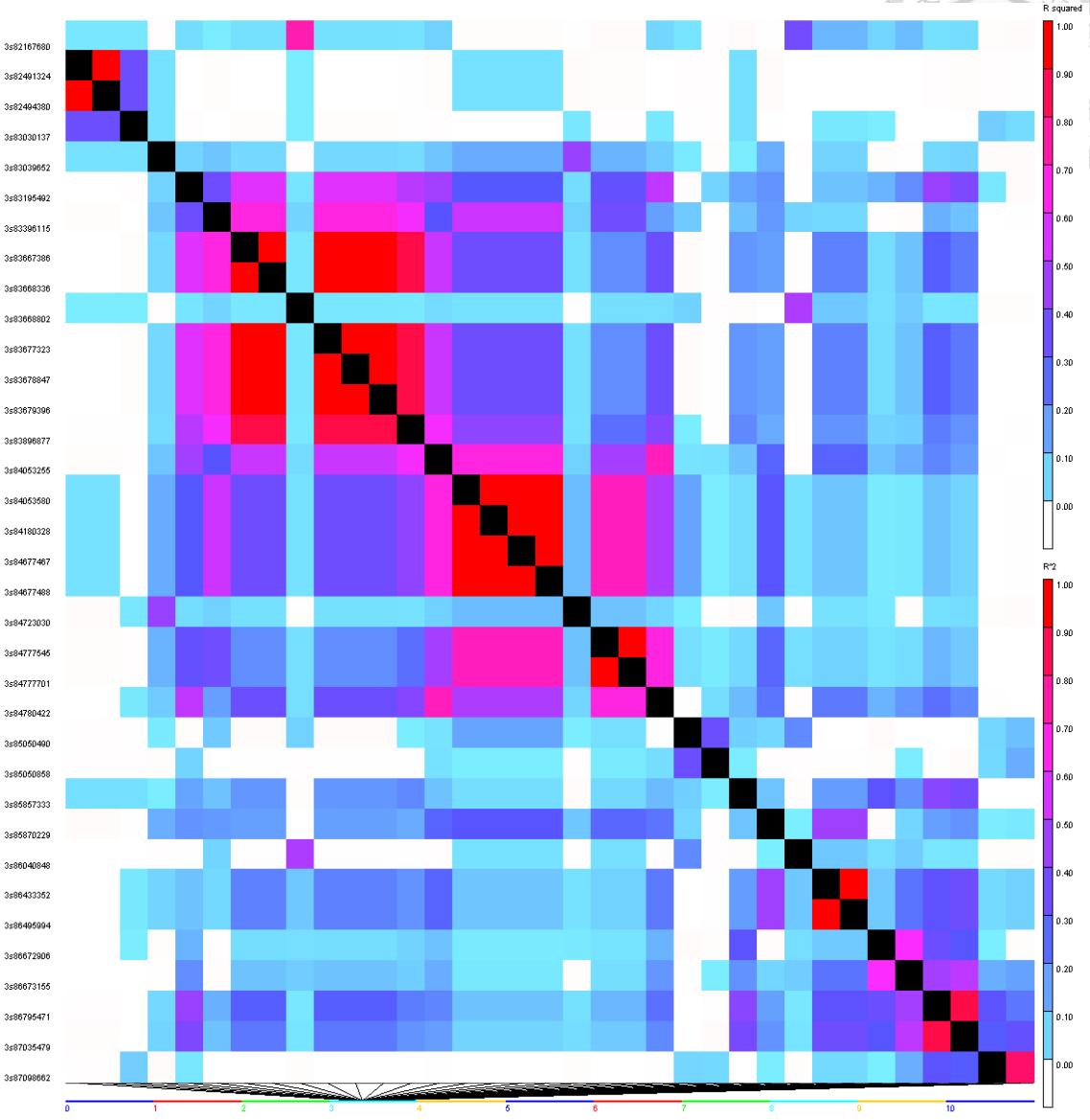
doi:10.6342/NTU201700716

第二章　研究目的

在全基因體關聯性分析或連鎖失衡的相關研究上，可以發現連鎖失衡區塊在不同的試驗族群下會展現不同的圖像，栽培種族群相對於野生種族群的連鎖失衡距離較大並且區塊圖像的完整性較高。然而，鮮少有學者討論在栽培種族群的連鎖失衡區塊中，依然存在部分位點與相鄰位點間相關係數驟降或不具相關性而造成連鎖失衡區塊圖像不完整的現象 (圖一)。

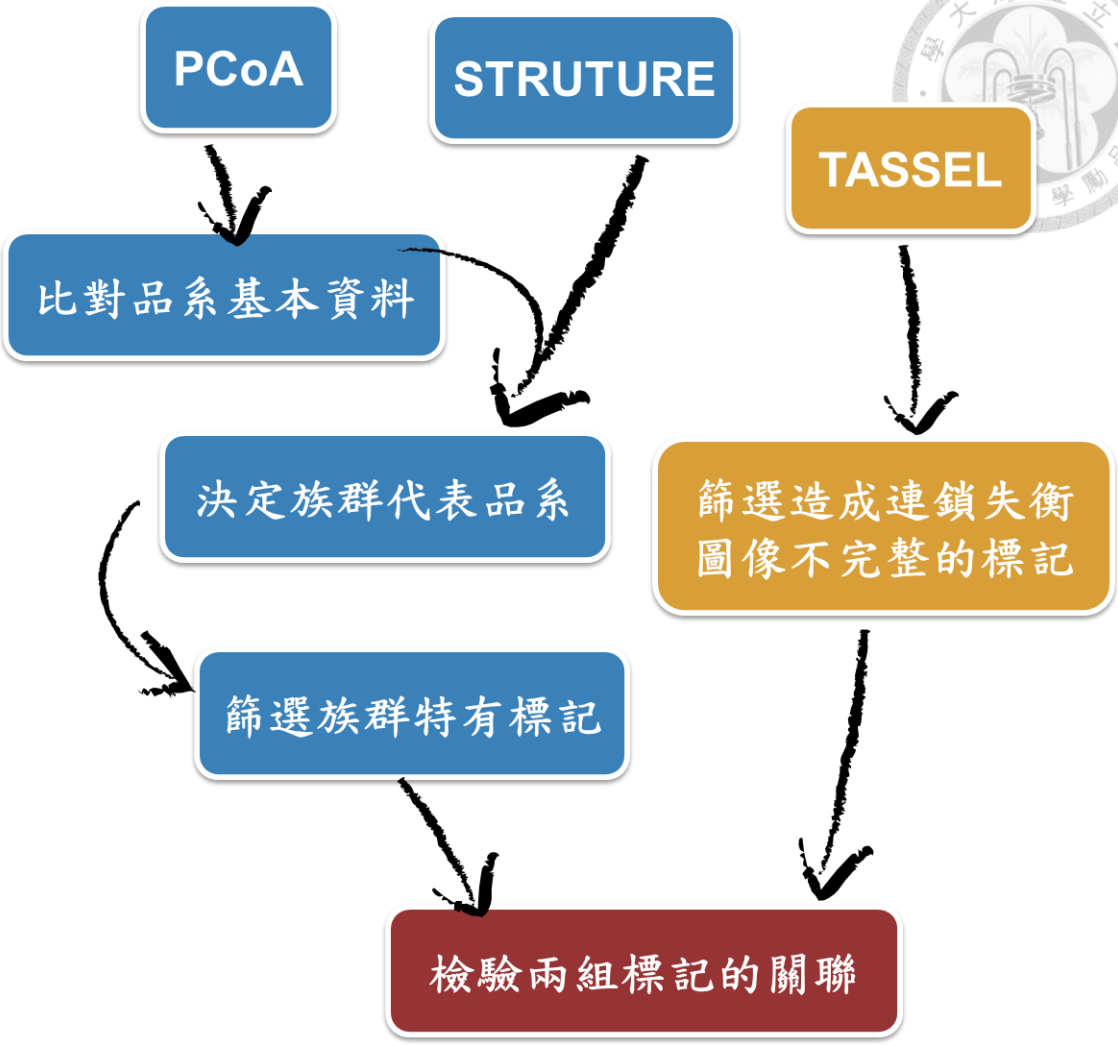
本研究欲探討連鎖失衡區塊圖像不完整的成因，是否和參試樣本不同次族群間起源不同的核苷酸變異有關。使用由農業試驗所與台南區農業改良場因應氣候變遷需求由國際玉米及小麥研究中心引種並蒐集台灣常用玉米品系所選之 48 個自交系種原，並以 MaizeSNP50 BeadChip 獲取大量 SNP 基因型。利用 STRUCTURE 2.3.4 進行族群分群，依分群結果篩選遺傳背景來自單一次族群的品系作為族群的代表品系，並依此選出次族群內特有具多型性之分子標記以代表不同次族群間起源不同的核苷酸變異。最後以 TASSEL 與統計軟體 R 評估整體連鎖失衡的情形，篩選出造成連鎖失衡區塊圖像不完整的分子標記，探討族群特有分子標記與連鎖失衡區塊圖像不完整現象的關聯性（圖二），期望能以演化的角度部分解釋此現象。

doi:10.6342/NTU201700716



圖一、連鎖失衡區塊圖像不完整現象的實例，為本研究資料經 TASSEL 4 輸出而得。在第 3 條染色體上 83,667,386 bp 到 83,896,877 bp 的位置，具有大小約 300 Kb 的連鎖失衡區塊。此由 7 個分子標記組成區塊中，位於 83,668,802 bp 的分子標記與區塊中的其他標記不具有連鎖失衡的現象，造成在視覺上連鎖失衡區塊圖像的不完整，此連鎖失衡陡降後又陡升的現象不符合在染色體上相鄰片段較容易一起傳遞的假說，為本研究主要探討的問題。

doi:10.6342/NTU201700716



圖二、分析流程架構圖，本研究利用 PCoA 與 STRUTURE 兩種分群工具對試驗品系進行分群，並比較兩種方式的差異，依分析結果選定族群代表品系，篩選族群特有分子標記。以 TASSEL 軟體計算連鎖失衡的狀況，並以統計軟體 R 篩選造成連鎖失衡區塊圖像不完整之分子標記。比對上述兩組標記的資料，驗證其相關性。

doi:10.6342/NTU201700716

第三章　材料及方法

第一節 試驗材料

本研究使用 48 個玉米自交系作為試驗材料（表一）：其中 24 個由農業試驗所提供，包括現今玉米參考序列所定序之 B73 品系，甜質種玉米品系金蜜、新吉士、彩白、Su963106、Su963107 等，糯質種玉米台農 5 號和台農 6 號的雜交親本，爆裂種 HP301 與來自國際玉米和小麥改良中心(International Maize and Wheat Improvement Center, CIMMYT)的(亞)熱帶品系。另外 24 個則為臺南區農業改良場朴子分場提供，全為來自 CIMMYT 的(亞)熱帶品系。試驗品系適應的氣候帶、來源地區和胚乳的特性等資訊，分別在美國農業部 (United States Department of Agriculture, USDA)、國際玉米和小麥改良中心和行政院農業委員會農業試驗所的台灣農業研究期刊中獲得 (劉紹國、謝光照，2012；謝光照、盧煌勝，2006；謝光照，2010，2015)。本次所使用之試驗品系，皆為各場認定具有優良特性可做為未來育種計畫的遺傳材料。農業試驗所提供之品系，部分為台灣自行引種並選育而得，故在遺傳背景上應會有溫帶種與熱帶種玉米混雜的狀況。

doi:10.6342/NTU201700716

表一、48 個試驗品系相關資料整理。品系基本資料來自 CIMMYT、USDA 與台灣農業研究期刊。其中適應地的分類方式依據 CIMMYT 的區分原則，非洲、亞洲與拉丁美洲地區需標註來源地， MA：Midaltitude、ST：Subtropical、T：Tropical。基因型資料的欄位中，金蜜 8 和 CML475 同質結合率明顯較低。 STRUCTURE 欄位表示在 K = 3 分群中，各品系 Q 值的計算結果（平均值）。

編號	品系名	品系基本資料				MaizeSNP50 基因型資料		STRUCTURE		
		來源地	氣候帶	適應地	胚乳	Call rate（%）	同質結合率（%）	溫帶馬齒種	溫帶甜質種	(亞)熱帶種
1	B73	USA	溫帶	Temperate	馬齒種	98.12	97.58	1.000	0.000	0.000
2	B73.meth4	USA	溫帶	Temperate	馬齒種	98.06	97.49	1.000	0.000	0.000
3	B73.meth6	USA	溫帶	Temperate	馬齒種	98.02	97.48	1.000	0.000	0.000
4	GEMS0067	USA	溫帶	Temperate	硬粒種	96.19	96.71	0.311	0.153	0.536
5	GXIM.10	Asia	未知	Unknown	硬粒種	95.40	93.16	0.021	0.000	0.978
6	Hi31	Oceania	溫帶	Temperate	馬齒種	96.26	96.39	0.320	0.147	0.532
7	CML161	Mexico	熱帶	Lowland	硬粒種	95.56	96.27	0.000	0.001	0.999
8	CML172	Mexico	熱帶	Lowland	硬粒種	95.47	96.16	0.000	0.001	0.999
9	CML312	Mexico	亞熱帶	Subtropical	硬粒種	95.44	96.24	0.000	0.000	0.999
10	CML444	Africa	亞熱帶	Africa MA/ST	馬齒種	95.46	96.22	0.000	0.000	1.000
11	CML548	Africa	亞熱帶	Africa MA/ST	硬粒種	95.15	95.44	0.001	0.058	0.942
12	Hi	Germany	溫帶	Temperate		94.80	96.57	0.072	0.488	0.440
13	HP301	USA	溫帶	Temperate	爆裂種	95.54	96.22	0.072	0.382	0.546
14	PO4.TNG5.父本	Asia	亞熱帶	Asia Subtropical	糯質種	95.27	96.22	0.031	0.198	0.772
15	PE31.2.TNG5.母本	Asia	亞熱帶	Asia Subtropical	糯質種	95.19	95.17	0.038	0.156	0.806
16	Wx10.84.TNG6.父本	Asia	亞熱帶	Asia Subtropical	糯質種	96.88	96.32	0.504	0.141	0.355
17	Wx12.28.2.TNG6.母本	Asia	亞熱帶	Asia Subtropical	糯質種	95.48	95.82	0.048	0.143	0.808
18	Su963106	Asia	溫帶	Asia Temperate	甜質種	95.07	94.73	0.000	1.000	0.000

doi:10.6342/NTU201700716

表一、48 個試驗品系相關資料整理 (續)

編號	品系名	品系基本資料				MaizeSNP50 基因型資料		STRUCTURE		
		來源地	氣候帶	適應地	胚乳	Call rate（%）	同質結合率（%）	溫帶馬齒種	溫帶甜質種	(亞)熱帶種
19	Su963107	Asia	溫帶	Asia Temperate	甜質種	95.15	95.53	0.000	1.000	0.000
20	彩白 20	Asia	未知	Unknown	甜質種	95.54	95.42	0.003	0.225	0.772
21	彩白 3A	Asia	未知	Unknown	甜質種	95.45	95.44	0.004	0.260	0.735
22	MTC	USA	溫帶	Temperate		95.74	96.58	0.203	0.336	0.461
23	新吉士.600.37	Asia	溫帶	Asia Temperate	甜質種	94.95	95.89	0.000	1.000	0.000
24	金蜜.8	Asia	溫帶	Asia Temperate	甜質種	94.51	85.91	0.052	0.792	0.157
25	CML121	Mexico	亞熱帶	Subtropical	馬齒種	95.48	96.15	0.000	0.000	1.000
26	CML122	Mexico	亞熱帶	Subtropical	馬齒種	95.57	96.25	0.000	0.000	1.000
27	CML124	Mexico	亞熱帶	Subtropical	硬粒種	95.57	96.21	0.002	0.000	0.998
28	CML112	Mexico	亞熱帶	Subtropical	硬粒種	95.52	96.37	0.022	0.076	0.903
29	CML348	Mexico	熱帶	Lowland	硬粒種	95.55	96.24	0.000	0.000	0.999
30	CML426	Asia	熱帶	Asia Lowland	硬粒種	95.28	95.77	0.000	0.000	1.000
31	CML440	Africa	亞熱帶	Africa MA/ST	硬粒種	95.39	96.31	0.000	0.035	0.965
32	CML441	Africa	亞熱帶	Africa MA/ST	硬粒種	95.47	96.38	0.017	0.086	0.897
33	CML449	Mexico	熱帶	Lowland	硬粒種	95.36	96.24	0.000	0.000	0.999
34	CML452	Mexico	熱帶	Lowland	馬齒種	95.47	96.31	0.000	0.000	1.000
35	CML470	Asia	熱帶	Asia Lowland	硬粒種	95.43	96.18	0.000	0.000	1.000
36	CML475	Asia	熱帶	Asia Lowland	馬齒種	94.76	82.40	0.000	0.000	1.000
37	CML494	Mexico	熱帶	Lowland	馬齒種	95.37	96.36	0.000	0.000	0.999

doi:10.6342/NTU201700716

表一、48 個試驗品系相關資料整理 (續)

編號	品系名	品系基本資料				MaizeSNP50 基因型資料		STRUCTURE		
		來源地	氣候帶	適應地	胚乳	Call rate（%）	同質結合率（%）	溫帶馬齒種	溫帶甜質種	(亞)熱帶種
38	CML521	Africa	亞熱帶	Africa MA/ST	硬粒種	95.54	96.12	0.002	0.041	0.957
39	CML522	Africa	亞熱帶	Africa MA/ST	馬齒種	95.53	96.22	0.000	0.000	1.000
40	CML523	Africa	亞熱帶	Africa MA/ST	馬齒種	95.72	96.43	0.088	0.002	0.910
41	CML538	Africa	亞熱帶	Africa MA/ST	馬齒種	95.58	96.32	0.005	0.047	0.948
42	CML541	Africa	亞熱帶	Africa MA/ST	硬粒種	95.54	96.32	0.025	0.083	0.892
43	CML544	Africa	亞熱帶	Africa MA/ST	馬齒種	95.71	96.45	0.066	0.002	0.932
44	CML545	Africa	亞熱帶	Africa MA/ST	馬齒種	95.62	96.25	0.050	0.001	0.950
45	CML547	Africa	亞熱帶	Africa MA/ST	硬粒種	95.63	96.21	0.045	0.076	0.879
46	CML19	Mexico	熱帶	Lowland	馬齒種	95.40	96.39	0.000	0.000	0.999
47	CML550	Latin Am	熱帶	Latin Am. Lowland /T	硬粒種	94.82	94.47	0.000	0.000	0.999
48	CML551	Latin Am	熱帶	Latin Am. Lowland /T	硬粒種	95.18	96.28	0.000	0.000	1.000

doi:10.6342/NTU201700716

第二節 玉米全基因體組 DNA 萃取

全基因體組 DNA 的萃取方式，使用 Fulton 等人 (1995) 提出的 CTAB 法並作修改。於液態氮中進行均質化步驟以減少 DNA 降解反應的發生，萃取片段長度大於 2,000 bp 的 genomics DNA。以符合 MaizeSNP50 晶片進行基因型定型時 illumine 產品網站上建議之片段大小，並以 QIAGEN DNeasy Blood & Tissue Kit 純化：

1. 製備萃取液：

萃取液的製備以抽取 40 個樣品為例，先將 114 mg sodium bisulfate. 充分溶解於 12.5 mL 的 DNA extraction buffer，然後依序加入 12.5 mL 的 nuclei lysis buffer 及 5 mL 的 5% Sarkosyl solution 並於實驗當天進行配置。其中三樣溶液可以事先製備保存。DNA extraction buffer 以製備 500 ml 為例，取 31.885 g Sorbitol、6.055 g Tris base 和 0.931 g EDTA，加入 ddH_2O 至接近 500 ml。攪拌溶解後以 HCl 調整 pH 值至 8.26，再以 ddH_2O 補滿體積至 500 ml。Nuclei lysis buffer 的製備以 500 ml 為例，先混合 100 ml 1M Tris、100 ml 0.25M EDTA 與 200 ml 5m NaCl 溶液，再加入 10 g CTAB，以 ddH_2O 補體積至接近 500 ml。磁石攪拌過夜讓 CTAB 完全溶解，再以 HCl 調整 pH 值至 7.5，最後以 ddH2O 補滿體積至 500 ml。5 % Sarkosyl 則一次配置 50 ml，將 2.5 g N-lauroylsarcosine sodium salt 加入 50 ml ddH_2O 中攪拌溶解。

2. 粗萃取 DNA：

首先為葉片的均質化，於 -80 ℃ 凍箱中取出玉米葉片並置於液態氮中保持低溫，同時以液態氮預冷研缽。取約 2 x 10 cm^2 面積大小的葉片於研缽中磨碎，加入 0.7 ml 當天製備的萃取液，持續攪拌至均質液回復液態後，移入 1.5 ml 微離心管中。將離心管置於 65 ℃ 水浴 30 分鐘，待反應完成後進行萃取步驟。先以氯仿混合液去除溶液中的葉綠素。將離心管移置抽氣通風櫃中加入 0.6 ml chloroform：isoamyl alcohol = 24：1 混合液，劇烈震盪至乳化狀態。以 10,000

doi:10.6342/NTU201700716

rpm 離心 5 分鐘，吸取上清液 0.55 ml 置於新的離心管中。再加入等倍體積 0.55 ml isopropanol，緩慢倒轉離心管至看到 DNA 沉澱。以 10,000 rpm 離心 5 分鐘，將 DNA 收集至離心管底部，再將上清液倒出。加入 0.2 ml 70 % ethanol 以 10,000 rpm 離心 5 分鐘清洗鹽類。倒出上清液後，靜置風乾 30-60 分鐘。最後加入 0.2 ml TE buffer 靜置溶解，獲得粗萃取液。

3. DNA 純化：

使用 QIAGEN DNeasy Blood & Tissue Kit 進行 DNA 的純化，實驗步驟依造 QIAGEN 產品說明手冊。首先在 0.2 ml DNA 粗萃取液中加入 2 µl QIAGEN Proteinase K (> 600 mAU/ml) 與 2 µl QIAGEN RNase A (100 µg/µl)，混合均勻後靜置 2 分鐘。再加入 200 µl Buffer AL，混合均勻後置於 56 ℃ 水浴 10 分鐘。反應完成後以 10,000 xg 離心 5 分鐘，將上清液移置離心管中。加入 200 µl 100 % ethanol，混合均勻後移置 DNeasy kit 的 Mini spin column 中。接者依序以 500 µl Buffer AW1 和 500 µl Buffer AW2 在 6,000 xg 條件下離心 1 分鐘清洗 DNA。接者以 20,000 xg 離心 3 分鐘除去殘餘溶液，再加入 30 µl Buffer AE 於 spin column 中靜置 1 分鐘，使 DNA 回溶，最後以 6,000 xg 離心 1 分鐘收取純化後之 DNA 溶液。

4. 品質檢測：

純化後的 DNA 以 1 % 洋菜膠體配合 TBE 緩衝液進行電泳，確認 DNA 片段長度是否大於 2,000 bp，並以 NanoDrop 檢測濃度。依生技中心建議，進行 MaizeSNP50 基因型定型之樣品，濃度應為 100 ng/µl，DNA 絕對量達 1.5 µg。

本研究所使用之葉片為進入生殖生長期的老葉，以乾冰冷凍運送，置於攝氏 -80 度凍箱保存。抽取效率上，50 平方公分的葉片約可至少抽出濃度 1,000 ng/µl，體積 100 µl，絕對量 100 mg 的粗萃取 DNA。然而，經 QIAGEN DNeasy Blood & Tissue Kit 純化後，回收率低且樣本間變異大。以 30 µl 回溶純化後之 DNA，其濃度範圍約介於 20 到 600 ng/µl，絕對量為 0.6 到 18 mg 之

doi:10.6342/NTU201700716

間。多數樣本位於 120 ng/μl 左右，絕對量為 3.6 mg。每一樣本皆以濃度約 100 ng/μl，體積 20 μl，DNA 絕對量約 2 μg，送往生技中心，進行 MaizeSNP50 的讀取。其中有 11 個樣品由於回收率低的問題有併管的動作，非同一次抽取之 DNA（附錄表一）。

第三節 MaizeSNP50 資料的產生與處理

個別樣品的基因型定型以 MaizeSNP50 晶片進行。MaizeSNP50 晶片共有 56,110 個單核苷酸多型性 (single nucleotide polymorphism, SNP) 位點，此部分的試驗委託生物科技研究所蔡孟勳老師研究室進行。取得的基因型資料由 Genome Studio 商用軟體 (Illumina Inc., San Diego, USA) 查看結果，並以 GenomeStudio 內建功能整理基本統計數據，將資料以 csv 格式儲存作為初始資料檔案。以統計軟體 R 讀取資料並做整理，利用 subset() 指令篩選出於 48 個試驗品系中皆無缺值或讀取錯誤的分子標記共 40,359 個，並除去於部分品系中具有異質結合的位點，共獲得 22,227 個同質結合分子標記。依據 Illumina 產品介紹文件 “ TOP/BOT ” Strand and “ A/B ” Allele 將基因型資料的表示格式由 A/B 對偶基因形式改為以 A、 T、 C、 G 表示。轉換方式為利用 R 軟體依據個別 SNP 類型以 gsub() 指令進行符號的替換，最後以 rbind() 指令合併資料，完成基因型表示方式的轉換。分子標記的分布以 R 繪製直方圖展示。以第一號染色體為例，首先以 subset() 指令取出第一號染色體的資料，並且除去未知物理位置的分子標記。以染色體上最遠分子標記的物理位置除以 5,000,000 作為組數，以 hist() 指令繪製直方圖，橫軸設定為物理位置，縱軸則為組內分子標記的計數。（附錄程式碼一）

第四節 族群結構的評估與代表品系的決定

族群結構的評估以兩種方式進行：第一種為以統計軟體 R 進行主座標分析（Principal Coordinate Analysis, PCoA），以近似主成分分析的方式觀察樣品的分布情形。第二種則以 STRUCTURE 2.3.4 (Pritchard et al. 2000) 軟體，利用貝氏方

doi:10.6342/NTU201700716

法基於模型分群，依照 Evano 等人 (2005) 提出的方法決定最佳分群數，並決定各次族群的代表品系。

1. 主座標分析

主座標分析首先將單一核苷酸多型性分子標記的 A、B 基因型資料，以統計軟體 R 中 gsub() 指令轉換為 1 和 0 的資料形式。再以 dist() 函數內建之 Euclidean 方法計算歐式距離，並將其除以 2 倍位點數的平方根以轉換成 Modified Rogers' distance (MRD)，使距離的表示數值侷限在 0 到 1 之間，並以此代表遺傳距離。利用 cmdscale() 函數可計算距離資料的特徵值以代表主座標，最後依據第一、二主座標的特徵值繪圖，並附上該特徵值所能解釋變異的比例，依造基本品系資料給定顏色分類，觀察品系分布的情形。（附錄程式碼二）

2. STRUCTURE 分析

模式分群的部分，將基因型資料以 R 改寫成 STRUCTURE 2.3.4 可以讀取的格式（附錄程式碼二），以 .txt 檔儲存，於 STRUCTURE 軟體中進行族群結構的評估。分析條件使用預設 lenth of burnin period = 20,000、number of MCMC reps after burnin = 20,000。分群數 K 設為 2 到 7 群，每個 K 值均計算 20 次。將相同 K 值設定下，分析結果中的“Ln P(D)”數值進行平均，以代表該分群數所建立之模式配適的優劣。依各分群數間“Ln P(D)”差值大小的變化決定何分群數為最佳值(Evanno et al. 2005)。最後以最佳分群數之分析結果中各樣本的“Inferred ancestry of individuals”數值決定該樣本是否能視為代表該族群的品系，以下以 Q 值作為簡稱。代表品系的決定選用兩種 Q 值篩選標準，分別為 Q = 1 與 Q = 0.99，由此篩選兩組代表品系。以 Q = 1 的篩選標準為例，當樣本的 Q 值在某特定族群下為 1 時，表示在所有測試樣本分成 K 群的分析中，該樣本基因組的遺傳成分完全屬於該特定族群，可被視為該族群的先祖。據此各次族群代表品系的判定方式為在 20 次獨立的重複計算中 Q 值為 1 的次數占總重複次數九成以上的樣本將被選為代表品系。

doi:10.6342/NTU201700716

第五節 區分族群內特有之分子標記

以統計軟體 R 整理代表品系的基因型資料。依序分為三個步驟，首先將各代表品系依各自所屬的族群分類。第二步驟利用 subset() 指令配合 == 的判斷式，可以篩選出在單一表格中特定欄位間讀值相同的分子標記，據此篩選出在個別族群內所有代表品系基因型讀值皆相同的分子標記。最後利用 %in% 判斷式比對不同表格間的分子標記資料，再以 setdiff() 指令將表格間的資料做差集。將特定族群內具多型性之分子標記與其他族群內篩選出無多型性分子標記的集合取交集，篩選出僅在特定族群內個體基因型具多型性的分子標記。在此研究中將此僅在特定族群內具有多型性的分子標記視為該族群內特有之分子標記。特有分子標記於染色體上的分布以直方圖表示。不同族群特有之分子標記給定不同顏色並以 adjustcolor() 指令調整顏色的透明度，運用 par(new=TRUE) 指令進行疊圖，將相同染色體上不同族群特有標記在一張直方圖中呈現。（附錄程式碼三）

第六節 連鎖失衡衰退評估

1. 資料格式轉換

連鎖失衡的計算運用 TASSEL v4.3.15 軟體，需將資料檔轉換為該軟體要求的格式。其讀取的檔案具有固定的欄位格式，依序為 ” rs# ”、 ” alleles ”、 ” chrom “、 ” pos “、 ” strand “、 ” assembly# ”、 ” center “、 ” protLSID ”、 ” assayLSID”、 ” panel “、 ” QCcode “、 ” genotype value “，並且分子標記的排序需依照其物理圖譜上的位置。將經 R 篩選後之同質結合分子標記資料檔，以 Excel 軟體更改欄位名稱，並增加部分欄位以符合 TASSEL 之要求。需增加的欄位資訊 ” strand “、 ” assembly# ”、 ” center “、 ” protLSID ”、 ” assayLSID”、 ” panel “、 ” QCcode “ 不影響基因型資料的判讀，可填入” NA ”。以 TAB 分隔檔儲存，並且將附檔名改為” .hmp “後，即可使基因型資料在 TASSEL v4.3.15 中運行。

2. 連鎖失衡圖像

doi:10.6342/NTU201700716

使用在 TASSEL 中內建的連鎖失衡分析工具，以 window size 為 100 個 SNPs 為範圍計算分子標記間連鎖失衡的情形。並使用 TASSEL 內建之圖像化工具，以成對分子標記間的相關係數 (correlation between pair of loci, R^2) 表示連鎖失衡的強度並繪製熱圖，相關係數的計算公式如下：

$$R^2 = \frac{D^2}{p_1 p_2 q_1 q_2}$$

$D = pq_{11} - p_1 q_1$、p_1 和 q_1 分別為兩個分子標記上主要等位基因的頻度，而 p_2 和 q_2 為次要等位基因的頻度，pq_{11} 為在兩個分子標記上皆帶有主要等位基因的頻度。

3. 連鎖失衡衰退曲線

繪製兩組連鎖失衡衰退曲線，設定 window size 為 1 和 10 計算連鎖失衡的情形，分別以兩組設定所計算出分子標記間的 R^2 進行評估。其方式參照 Remington 等人 (2001) 提出估計試驗樣本 R^2 期望值的運算公式：

$$E(R^2) = \left[\frac{10 + C}{(2 + C)(11 + C)}\right]\left[1 + \frac{(3 + C)(12 + 12C + C^2)}{n(2 + C)(11 + C)}\right]$$

n 為樣本數，C 表示整個族群的互換參數 (population recombination parameter) 為 $4Nc$，其中 N 代表有效族群的個數，c 則為位點間經一個世代的互換率。本試驗資料為蒐集系的 SNP 資料，故無法計算互換率。因此參照 Marroni 等人 (2011) 利用統計軟體 R 中的 nls() 函數，以蒐集系之 SNP 資料與分子標記間的相關係數，經非線性最小平方法 (Nonlinear least squares) 估計 C 值。再將所估計出的 C 值代入 R^2 期望值的運算公式估計連鎖失衡的期望值，最後以經運算獲得之 R^2 期望值表示連鎖失衡衰退的情形。（附錄程式碼四）

第七節 特有分子標記假說驗證

將 15,919 個具有物理圖譜資訊的同質結合分子標記作為評估連鎖失衡的資料，並查看族群內特有具多型性之分子標記，於染色體的分布情形是否為非隨機的位於不完整的連鎖失衡區塊內。先將所有分子標記資料依是否為造成連鎖失衡

doi:10.6342/NTU201700716

區塊不完整的分子標記分類，視為成功或失敗。隨機抽取與特有分子標記數量相同的位點，記錄成功的個數並多次抽取建立二項分布。依據中央極限定理視其為常態分布，查看特有分子標記的成功個數位於常態分布上的位置。

1. 篩選造成連鎖失衡區塊圖像不完整之分子標記

將分子標記資料於 TASSEL 中以 window size = 2 計算相鄰三個分子標記間 R^2 值的大小，並於 R 軟體中進行篩選。以 if() 函數建立判斷式，並以 && 連結篩選條件。以三個分子標記為一組，取出相鄰分子標記間 $R^2 < 0.8$，且間距一個分子標記間 $R^2 > 0.8$ 的區域。將其中與另外兩相鄰分子標記間 $R^2 < 0.8$ 的標記視為造成連鎖失衡區塊圖像不完整的分子標記。（附錄程式碼五）

2. 驗證族群特有標記

利用 R 軟體中 %in% 判斷式比對族群特有標記與造成連鎖失衡區塊圖像不完整分子標記兩組資料，配合 length() 指令計算兩組資料中相同標記的個數，並以此代表族群特有標記資料中“成功”的個數。

以 sample() 指令對同質結合分子標記資料檔中具有物理圖譜位置資訊的 15,919 的分子標記進行隨機取樣。取樣個數為族群特有標記的總數，並比對造成連鎖失衡區塊圖像不完整的分子標記，紀錄“成功”的個數。以迴圈的方式重複取樣 10,000 次，以“成功”的個數建立二項分布，並檢測族群特有分子標記資料中“成功”的個數位於此分布上的位置。同時將此二項分布視為常態分布，以 pnorm() 指令計算族群特有分子標記在常態分布上的百分比。（附錄程式碼五）

doi:10.6342/NTU201700716

第四章　結果

第一節 MaizeSNP50 資料讀取

MaizeSNP50 晶片的讀取於 GenomeStudio 軟體中呈現。全部的位點總共 56,110 個，544 個位點有亂碼的現象，共 55,566 個分子標記可以正常讀取，佔全部位點的 99%。其中有 970 個位點於 48 個品系中皆無讀值，以品系的觀點來看，各品系能讀取的位點量大致相同。金蜜 8、CML475、Hi、CML550 等品系的讀取位點量稍低，但仍在 94 % 以上，其中 B73 與其近似同源品系幾乎可完全讀取（表一）。然而，在能正常讀取的位點中有 14,237 個在少數品系中存在缺值。在 48 個品系中皆沒有缺值的位點共 40,359 個，比較所有位點於各染色體的分布和無缺值位點於各染色體的分布，兩者幾乎一致（附錄圖一）。以統計軟體 R 整理各無缺值位點於 48 個品系的讀值，其中有 18,132 個位點其讀值具有異質結合的現象，其中大量位點屬於在 48 個品系中有 1 或 2 個品系具有異質結合的狀況，僅在 1 個品系中為異質結合的位點有 11,258 個，在 2 個品系中為異質結合的位點有 2,983 個，其總和佔所有具異質結合位點總數的 78.5 %（附錄圖二）。在 GenomeStudio 中可以查看單一品系經 56,110 個位點讀取後，基因型呈現 AA、BB 兩種同質結合或異質結合的分布狀況。從品系的觀點看，試驗材料中 43 個品系其同質結合率在 95 % 以上，僅有少量異質結合的狀況，11 個有併管的樣本皆屬於此類。金蜜 8 和 CML475 兩個品系則有比較大量的位點呈現異質結合，其同質結合率低於 90 %（表一）。在最後進行族群結構分析的同質結合位點部分，為將無缺值位點資料扣除所有具有異質結合的位點，得到 22,227 個於 48 個品系皆為同質結合的位點作為後續分析的資料。分子標記對應物理圖譜的部分，illumina 所提供的位置資料中有 11,436 個位點缺乏對應的位置資料，經篩選後的 22,227 的位點中有 15,919 個位點具有位置資料。

doi:10.6342/NTU201700716

第二節 族群結構的評估與代表品系的決定

經由主座標分析，視覺化的了解試驗品系間的遺傳距離與族群結構的概況。藉由比對前人紀錄的品系基本資料，獲得相關資訊與分群結果的關聯，以了解現行玉米的分類方式與遺傳背景間的關係。而 STRUCTURE 軟體分析則可以提供研究者個別試驗樣品遺傳背景的量化資訊。依特定篩選標準，選擇來自單一遺傳背景的品系作為族群的代表品系，以利於篩選出演化歷程上在群化後於特定族群內突變產生的基因型。

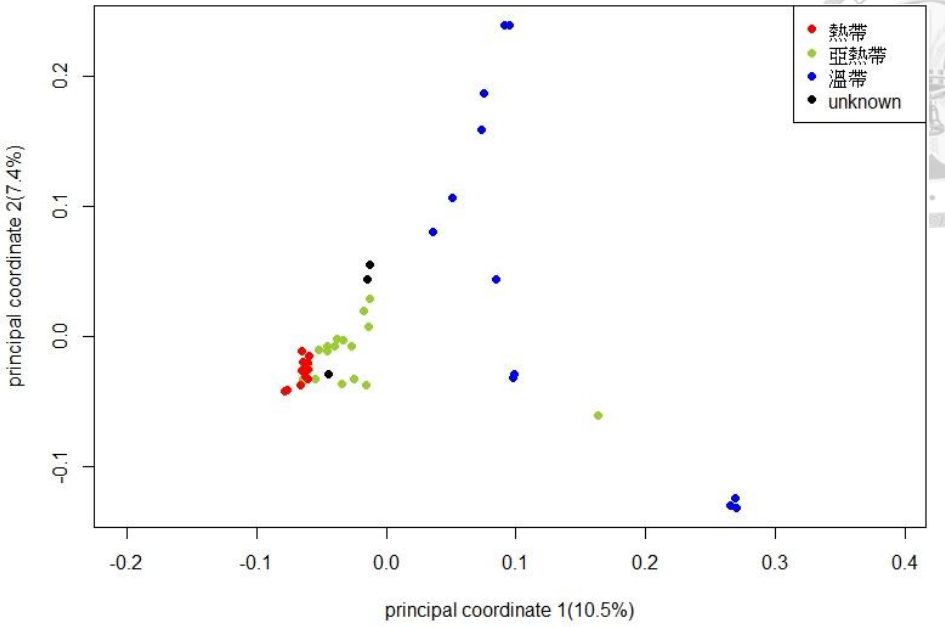
由於欲仿造族群演化的基因型結果，將具有異質結合的基因座，視為因現代人為育種過程中，經由族群間的引種雜交且未經自交固定而產生的結果，故將其刪除，僅以同質結合的資料進行族群結構的評估。在主座標分析中前三個主座標所能解釋的遺傳變異分別為 10.5％、 7.4％ 和 3.9％。以第一主座標與第二主座標為軸繪圖可以發現，品系的分布近似於三角形，分別比對已知的適應氣候、地區、胚乳特性等資料。以適應氣候帶來說(圖三)，在第一主座標上可以明顯看到熱帶、亞熱帶與溫帶品系間的區隔，熱帶與亞熱帶的區域有少部分重疊。然而具有一個例外，台農 6 號的親本，源自 Wxpop 10 的 Wx10.84 為來自越南的白糯玉米，應屬亞熱帶品系，而由主座標分析中卻位於溫帶玉米的區域。熱帶與亞熱帶區域重疊的部分在第三主座標上可以部份區分，但仍有混雜現象。

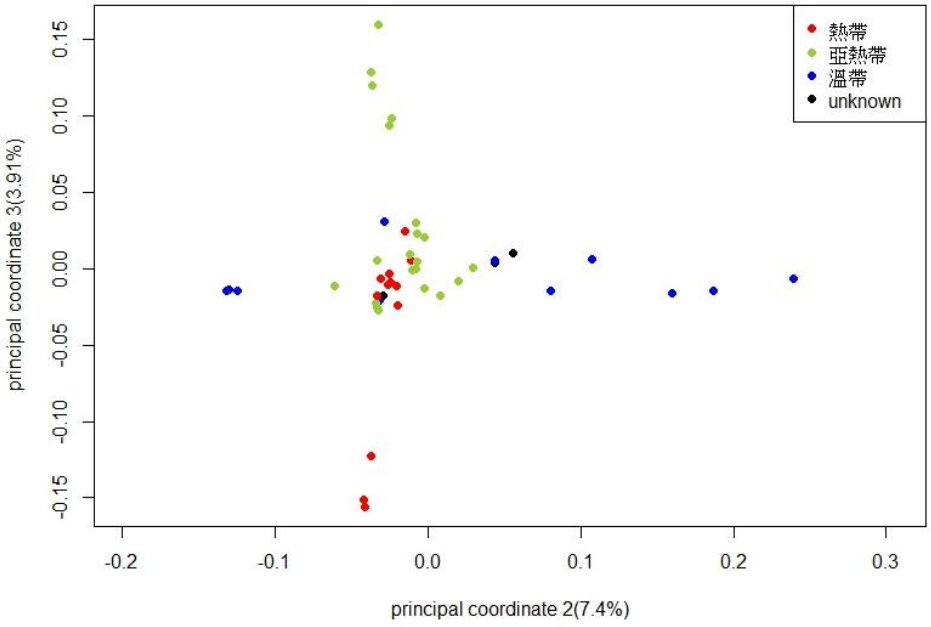
以單以地區來分類並沒有明顯的區分效果。而以胚乳特性分類也無法良好的區分，僅在第二主座標上可看到甜質種有比較聚集的現象(圖四)；同時比對氣候帶與來源地資訊時，可以有比較優良的結果。在第一主座標上，來自亞洲、墨西哥、拉丁美洲的熱帶玉米，可以看到位置上有重疊但可以區分出亞洲和美洲兩群。在亞熱帶玉米上，美洲與非洲的亞熱帶玉米有重疊的現象，而與亞洲的亞熱帶玉米有區隔；而第二主座標則可以區分出亞洲與美洲的溫帶玉米；在第三主座標上非洲的亞熱帶玉米與亞洲的熱帶玉米明顯與其他品系區隔，但其他品系在第三主座標上仍呈現混雜的狀態。總體而言，在第一與第二主座標的繪圖中所呈現

doi:10.6342/NTU201700716

的三角形的三個極端點分別是亞洲的熱帶玉米、亞洲甜質種溫帶玉米與美洲馬齒種溫帶玉米(圖五)。

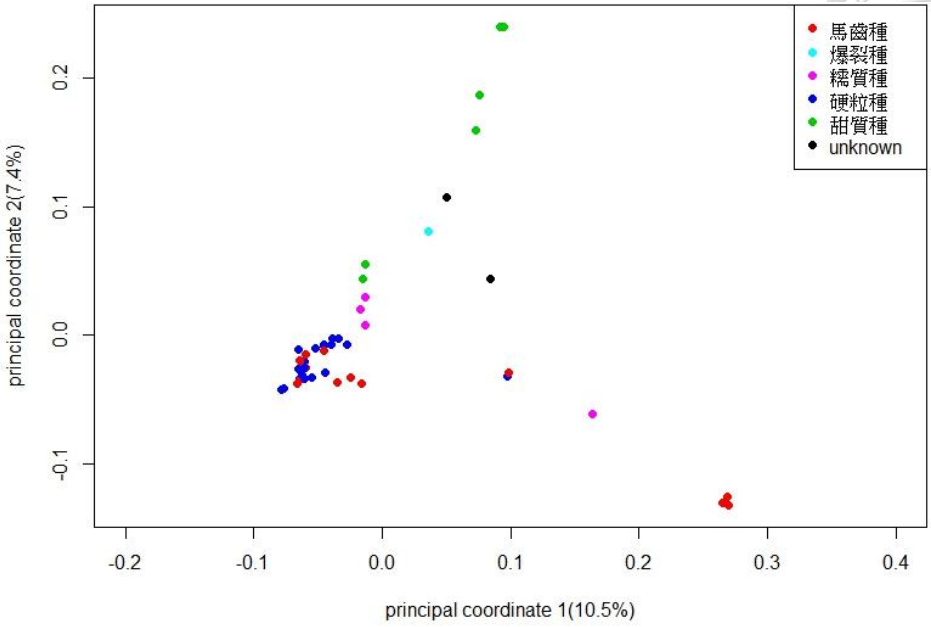
doi:10.6342/NTU201700716

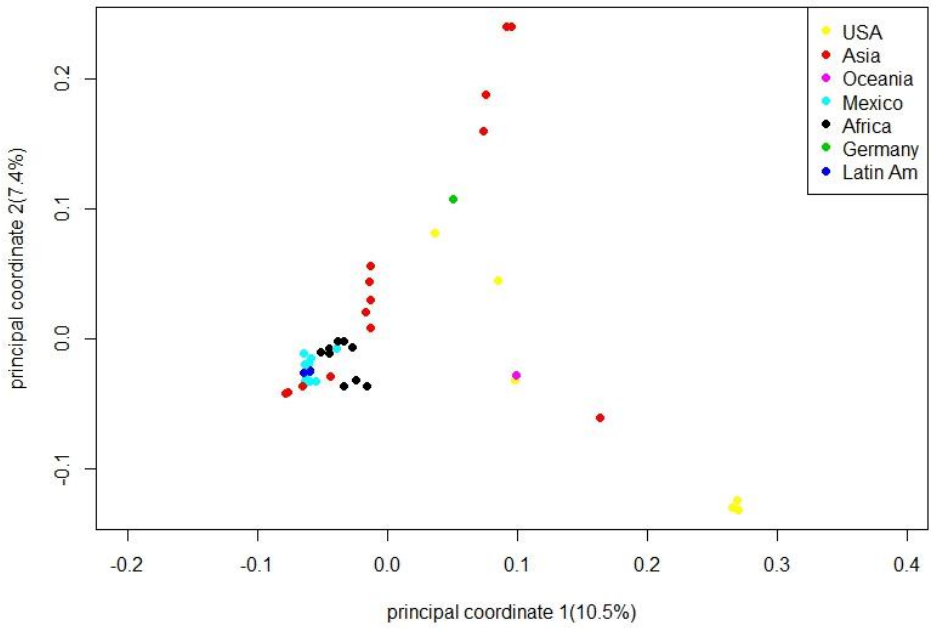




圖三、48 個試驗品系的主座標分析圖，以品系資訊（表一）中的適應氣候帶分類，紅色為熱帶品系、綠色為亞熱帶、藍色為溫帶，而黑色表示未有適應氣候帶的資料。在此分析結果中氣候帶的分類與玉米的基因型具有相當程度的關聯，在第一主座標上有良好的區分效果。

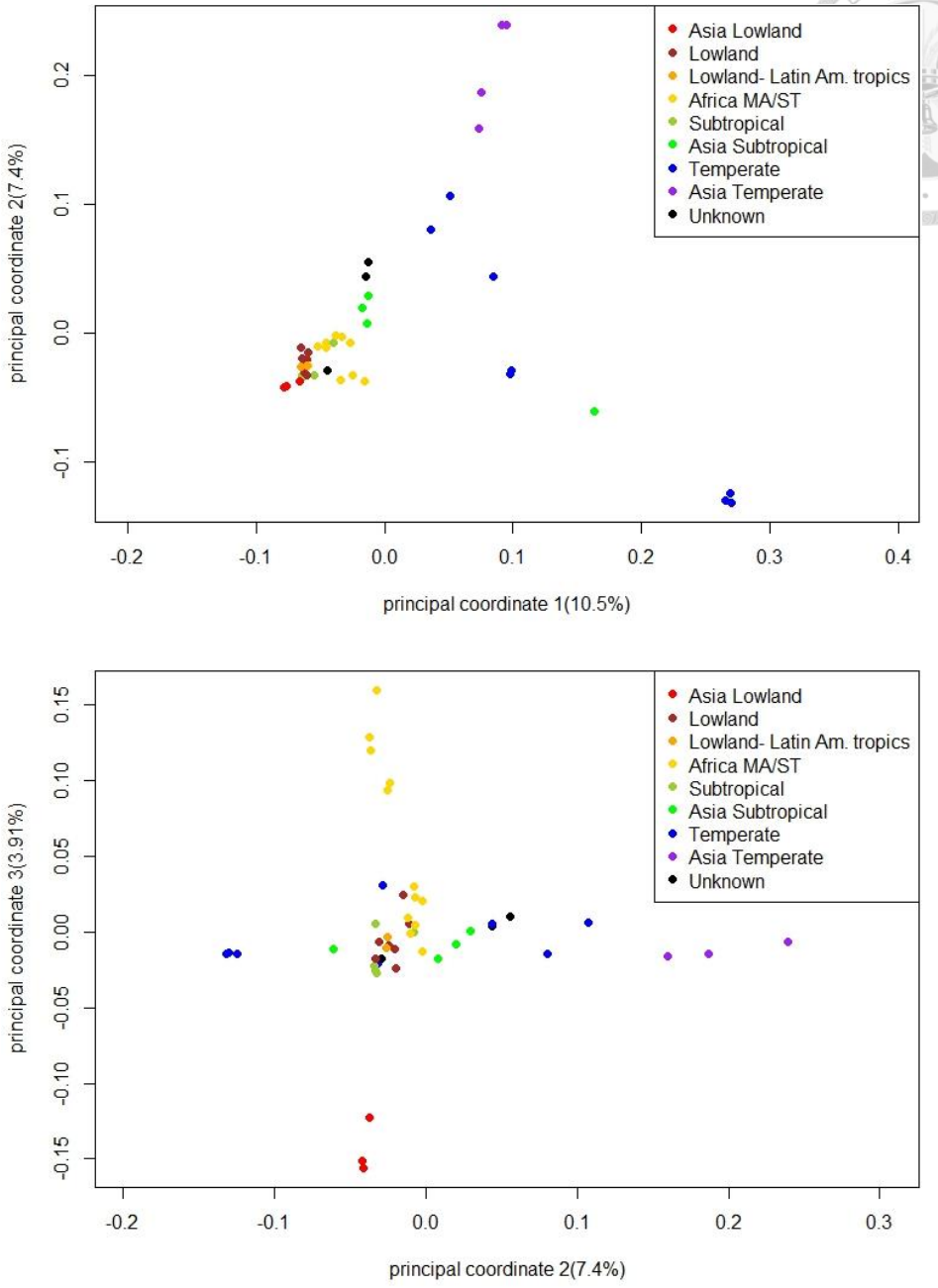
doi:10.6342/NTU201700716





圖四、48 個試驗品系的主座標分析圖，以品系資訊（表一）中的胚乳特性(上圖)和來源地(下圖)做分類，沒有明顯的分群狀況。在胚乳特性上僅第二主座標能將甜質種與其它胚乳做區分；在來源地的部分墨西哥、非洲、拉丁美洲的品系雖具有明顯的群聚現象，然而地區間的混雜嚴重，表示僅以來源地作為區分無法良好分別出玉米的基因型差異。

doi:10.6342/NTU201700716



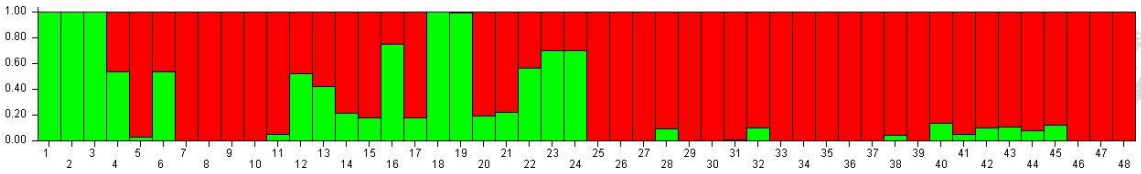
圖五、48 個試驗品系的主座標分析圖，綜合品系資訊（表一）中氣候帶與來源地區的適應地資料進行分類，分類方式依據國際玉米和小麥改良中心對於其發展出之品種的分類方法。第一主座標可區分溫帶與(亞)熱帶玉米，第二主座標區分亞洲與其他來源地的溫帶玉米，第三主座標則在(亞)熱帶玉米中區分出亞洲的熱帶玉米。相較於單純利用氣候帶區分具有更好的區分效果。

doi:10.6342/NTU201700716

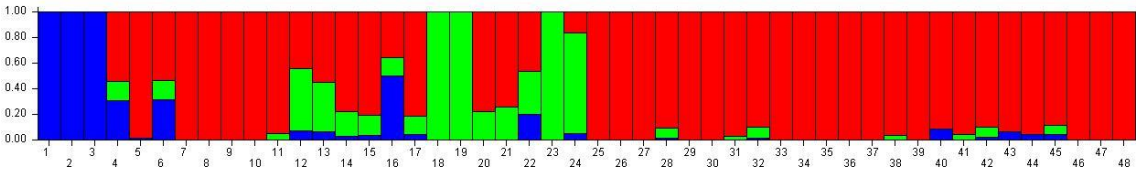
STRUCTURE 的族群結構分析方法由於是以貝氏方法基於模式分群而具有演化的含意。以設定 K = 2 為例，此表示認定所有樣本來自於兩個不同的始祖，在哈溫定律的假設下，評估每個試驗材料遺傳背景的狀況，並以“Inferred ancestry of individuals”表示，以下以 Q 值作為簡稱。本實驗資料的分析結果中，可以區分出溫帶玉米與(亞)熱帶玉米。當設定 K = 3 時區分出馬齒種溫帶玉米、甜質種溫帶玉米與(亞)熱帶玉米，與主座標分析的繪圖結果吻合。而設定 K = 4 時馬齒種溫帶玉米與甜質種溫帶玉米依然各自屬於一個始祖，(亞)熱帶玉米則是呈現兩個始祖各佔一部份遺傳背景的狀況為主。比對 Q 值，可以發現是非洲的亞熱帶玉米與亞洲的熱帶玉米兩群，且有少數非洲亞熱帶玉米與亞洲熱帶玉米群混雜。K = 5 時非洲亞熱帶玉米內區分為兩群，且不與亞洲熱帶玉米混雜（圖六）。而 K 值再往上後就沒有明顯的區分結果，並且重複計算間出現分群差異。在計算結果中，由 K = 2 與 K = 3 之間 Ln P(D) 的差值最大為 70,947.06， K = 3 和 K = 4 的差值就降為 27,873.62，K 值再提高的差值皆約在兩萬多，因此 K = 3 應為最佳分群數（圖七）。據此在代表品系的選擇上：有 3 個美洲馬齒種溫帶玉米品系，分別為 B73、 B73.meth4、 B73.meth6；2 個亞洲甜質種溫帶玉米品系， Su963106、 Su963107；與 3 個亞洲熱帶玉米品系 CML426、 CML470、 CML475，在 20 次的計算上 Q 值皆為 1。1 個亞洲甜質種溫帶玉米品系，新吉士.600.37，有 19 次計算上 Q 值為 1，因而決定出 9 個代表品系（圖八）其座落於主座標分析散布圖上極端的位置。除此之外，發現在(亞)熱帶玉米的族群中有另外 15 個品系在 20 次的計算上 Q 值皆大於 0.99：包含 7 個源自墨西哥的熱帶玉米、 2 個源自拉丁美洲的熱帶玉米、 4 個源自墨西哥的亞熱帶玉米和 2 個源自非洲的亞熱帶玉米。加上原本選出的 9 個代表品系後，比對這 24 個品系在主座標分析上的位置（圖九），在第一主座標與第二主座標的平面圖上三個族群的代表品系依然分布在三個端點，而在第三主座標上 18 個(亞)熱帶玉米彼此被區分開。

doi:10.6342/NTU201700716

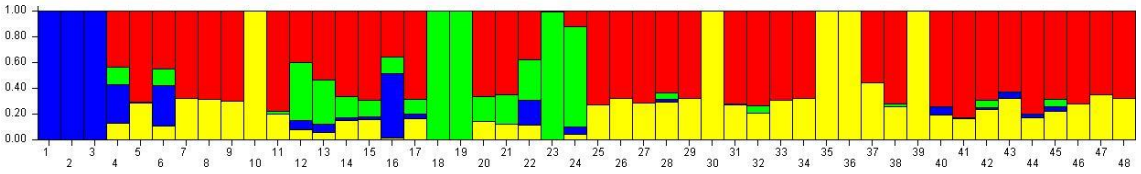
K = 2



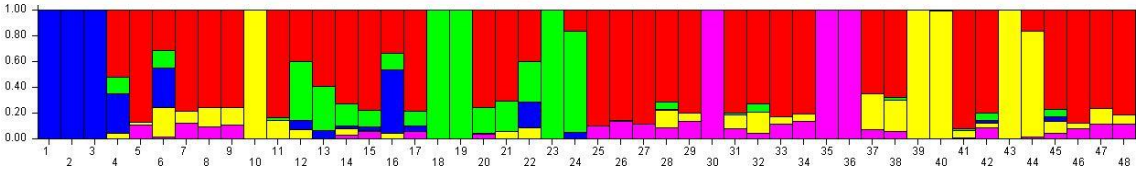
K = 3



K = 4

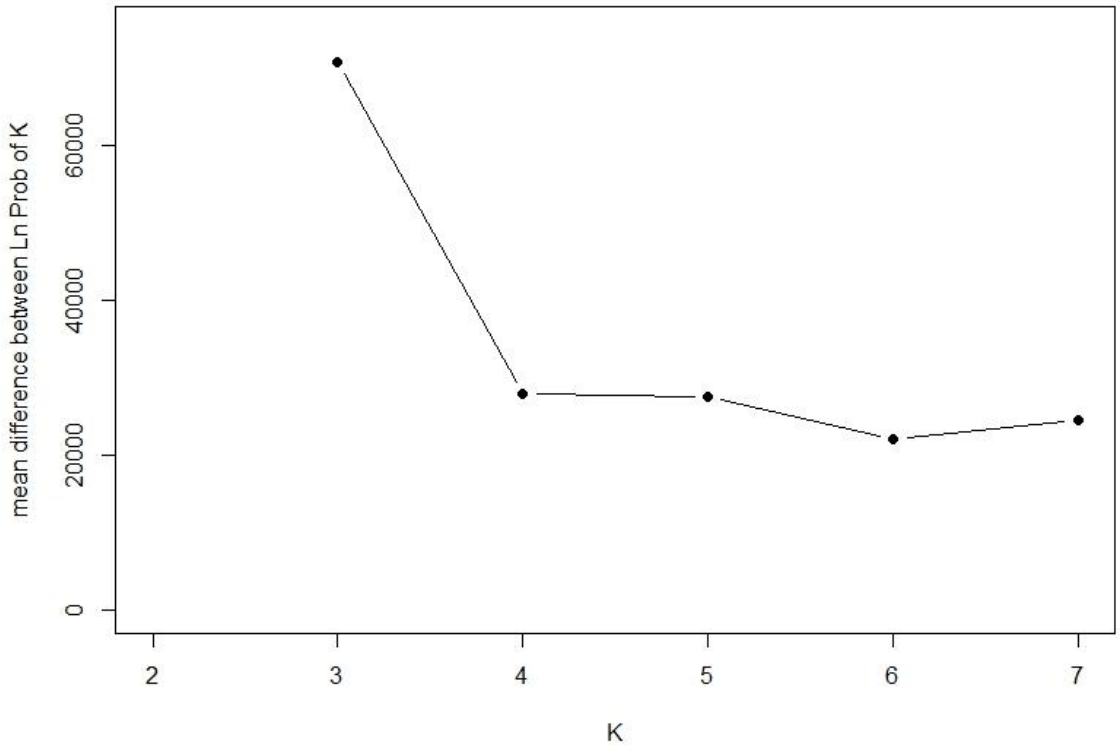


K = 5



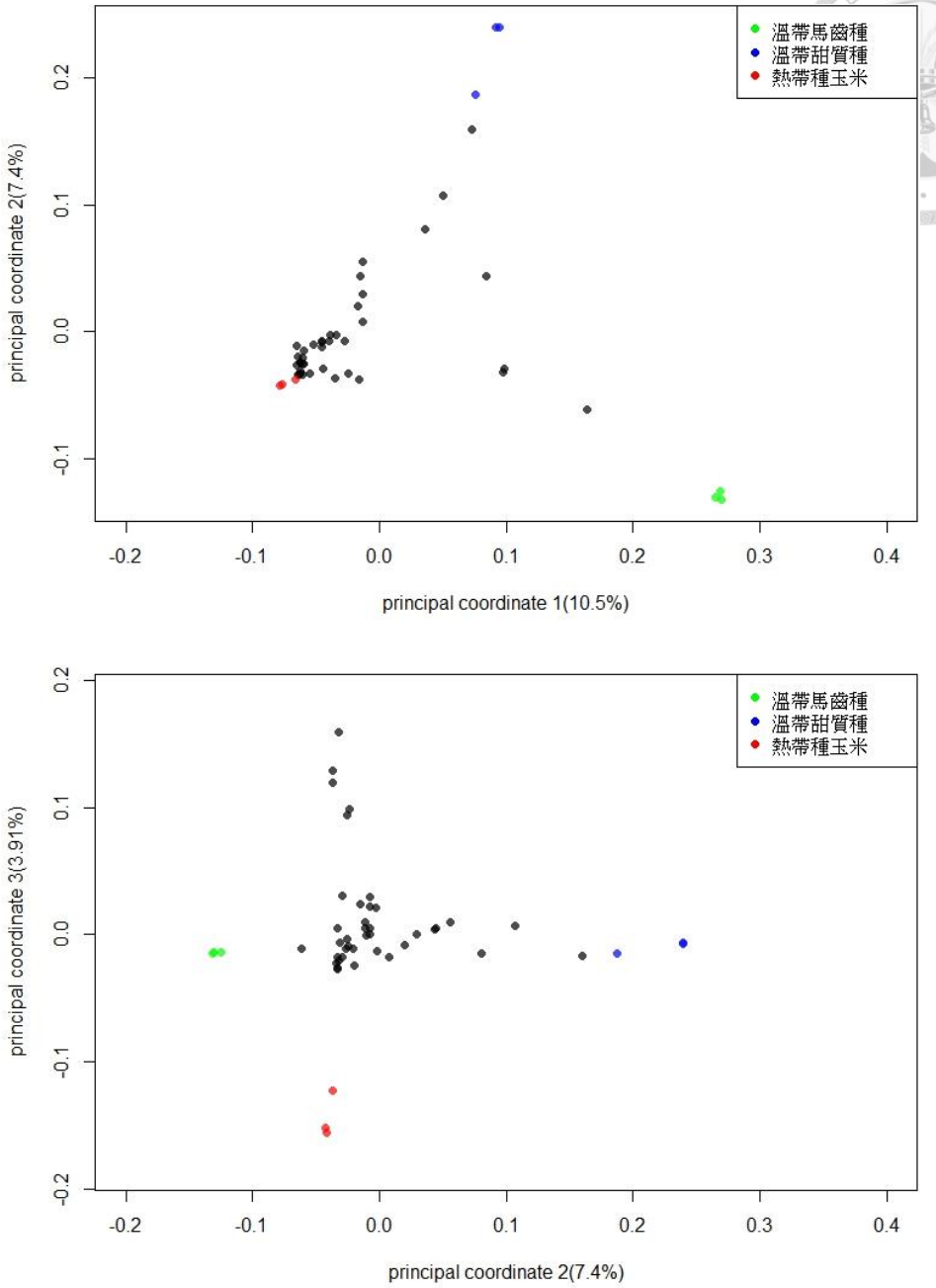
圖六、STRUCTURE 分群結果，品系編號同表一，K = 2 圖中綠色為溫帶玉米、紅色為熱帶玉米；K = 3 圖中藍色為溫帶馬齒種、綠色為溫帶甜質種；K = 5 紫色為亞洲熱帶種。K 值的上升可以更細微的區分出品種的差異，並且與適應地的資訊相近。然 K = 4 圖中黃色的部分包含亞洲熱帶種與非洲亞熱帶種，而在 K = 5 圖中雖能將亞洲熱帶種區分開但仍有部分非洲熱帶種並非以黃色背景為主。本資料部分品系間相近的遺傳背景使得 K 值的上升並不能得到更好的區分效果。

doi:10.6342/NTU201700716



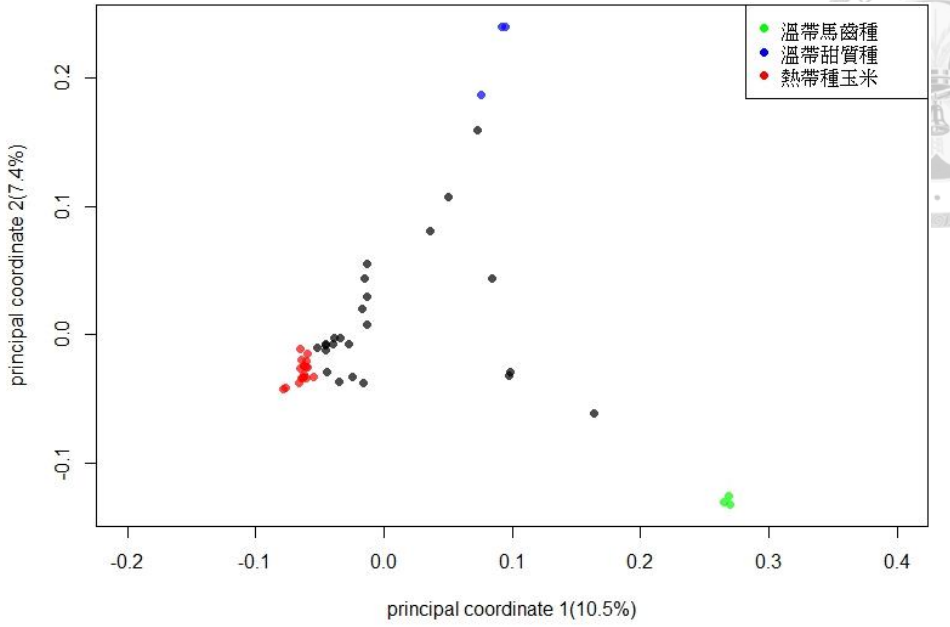
圖七、不同 K 值設定下 Ln P(D) 數值的差異，依據 Evanno 等人於 2005 年文獻中的模擬結果，Ln P(D) 差值的峰值為的最佳 K 值。在此分析中，K = 3 為最適合的分群設定。

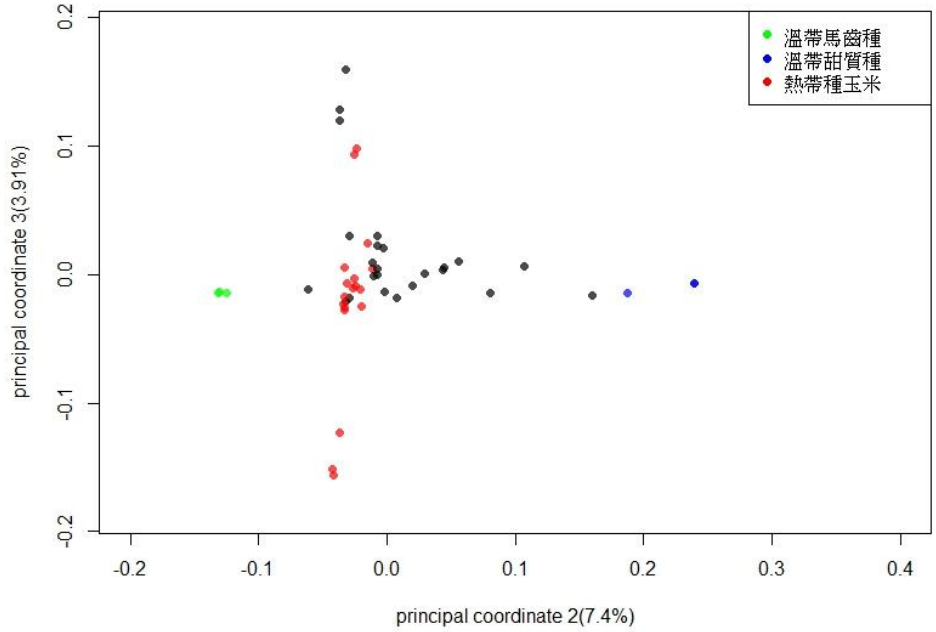
doi:10.6342/NTU201700716



圖八、48 個試驗品系的主座標分析圖，依 STRUCTURE 分析所得之 9 個代表品系分類。在 Q = 1 的篩選標準下，所得之代表品系在主座標分析上位於圖的三個極端點，分布集中並且沒有混雜現象，兩個分群方式的結果十分吻合。

doi:10.6342/NTU201700716





圖九、48 個試驗品系的主座標分析圖，依 STRUCTURE 分析所得之 24 個代表品系分類。在 Q > 0.99 的篩選標準下，所得之代表品系在第一、第二主座標分析圖上位於極端點並且集中，然而熱帶種玉米在第三主座標上分散。

doi:10.6342/NTU201700716

第三節 族群特有之分子標記

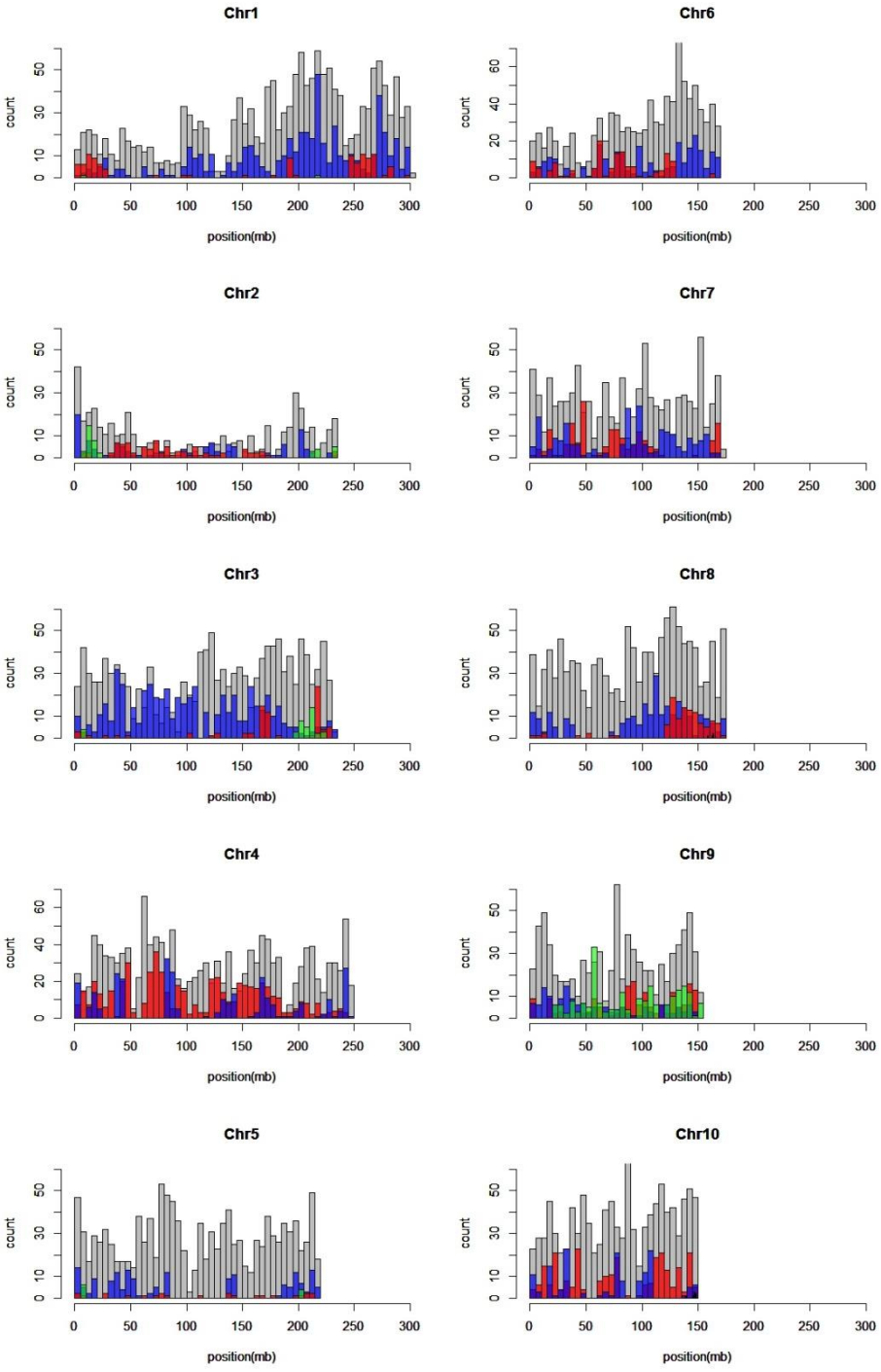
在 22,227 個同質結合分子標記中 15,919 個標記具物理位置資訊，以三個次族群，溫帶馬齒種、溫帶甜質種與熱帶種玉米的 9 個代表品系進行篩選。溫帶馬齒種內特有之分子標記共 275 個，主要集中在 9 號染色體，溫帶甜質種中有 2,447 個分子標記，(亞)熱帶種中有 1,530 個，共有 4,252 個族群內特有分子標記，非族群內特有的分子標記則為 17,975 個，扣除物理圖譜上位置未知（MaizeSNP50 晶片資料中物理位置資訊為 0 的標記）或未知位於幾號染色體的資料點後分別剩下 194、 1,772、 1,123、 3,089 和 12,830 個。而若以 Q 值 0.99 以上的 24 個品系作為篩選標準，溫帶馬齒種特有之分子標記共 83 個，溫帶甜質種中有 636 個分子標記，(亞)熱帶種中有 8,319 個，共有 9,038 個族群內特有分子標記，非族群內特有的分子標記則為 13,189 個，扣除物理圖譜上位置未知或未知位於幾號染色體的資料點後分別剩下 65、 454、 6,177、 6,696 和 9,223 個，其於染色體上之分布見圖十、圖十一。

在本研究試驗資料中可以發現族群內特有分子標記造成連鎖失衡區塊圖像不完整現象的例子（圖十二）。為了驗證「造成連鎖失衡區塊圖像不完整的分子標記」與「族群特有的分子標記」的關聯性，首先以 TASSEL 估算試驗材料的連鎖失衡狀況，依其基因型計算之 R^2 值表示。並以 $R^2 < 0.8$ 為標準篩選出 405 個造成連鎖失衡區塊圖像不完整的 SNP 分子標記。其依染色體區分從 1 號染色體開始依序是 48、 8、 48、 79、 30、 35、 32、 37、 36、 52 個位點。並使用 bootstrap 的概念，重覆進行 10,000 次取樣，計算不同數目的「造成連鎖失衡區塊圖像不完整分子標記」被隨機取得的次數以建立抽樣分布。在 9 個代表品系篩選出 4,252 個族群特有分子標記中具有確切物理位置資訊的 3,089 個族群特有分子標記。其中有 85 個分子標記為造成連鎖失衡圖像不完整的 SNPs。15,919 個具物理位置資訊的同質結合分子標記資料，以 3,089 個個數隨機抽樣所建立的分布中，中位數為 79，位於 95% 的個數為 92，85 約位於分

doi:10.6342/NTU201700716

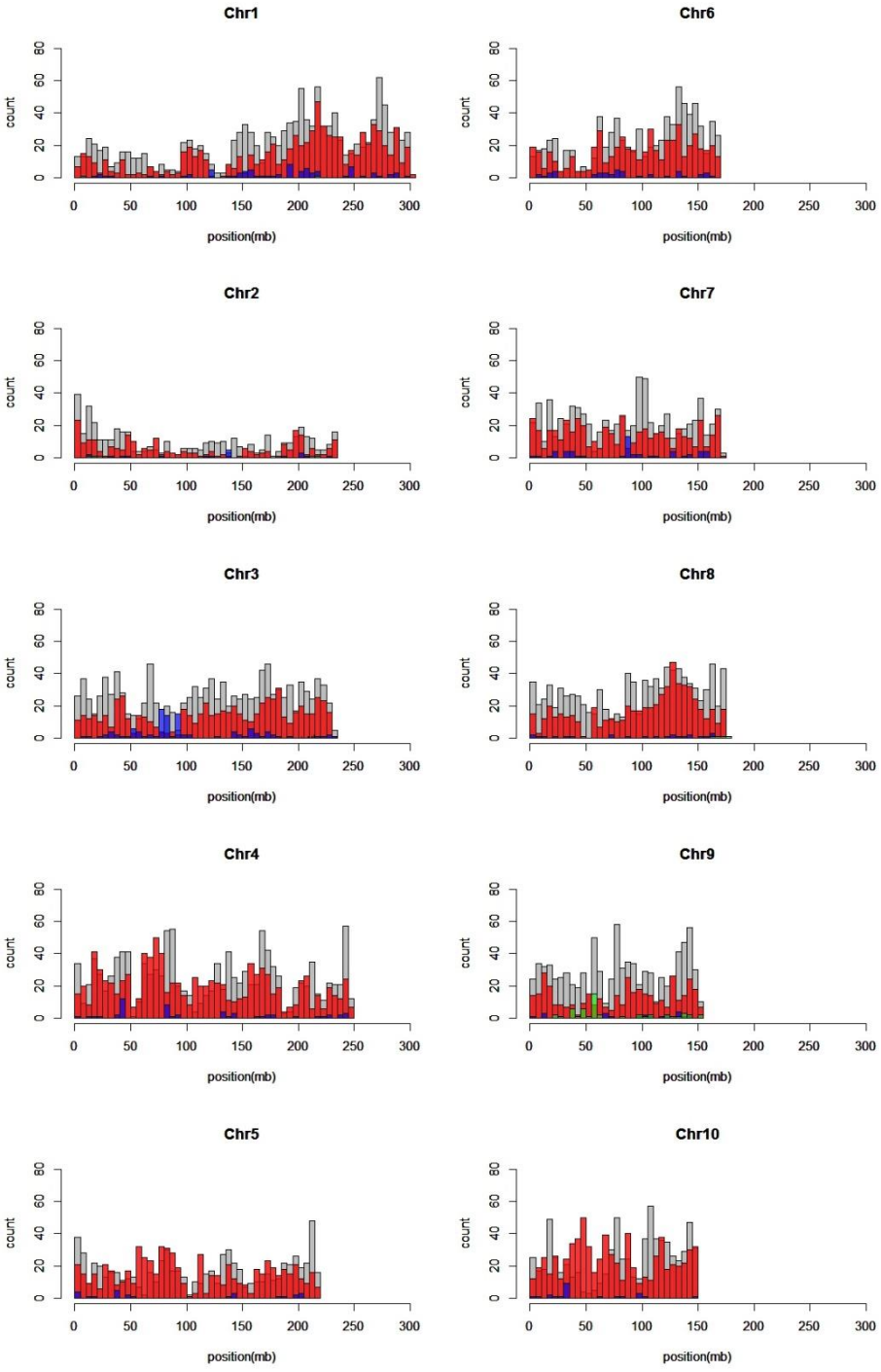
布中 79.7% 的位置（圖十三）；而以 24 個代表品系所建立的資料中，9,038 個特有分子標記中 6,696 個具有位置資訊。隨機抽樣的分布中，其中位數為 170，95% 的個數為 186，特有分子標記的資料中則有 195 個造成連鎖失衡圖像不完整的 SNPs（表二），位於分布中 99.4% 的位置（圖十三）。

doi:10.6342/NTU201700716



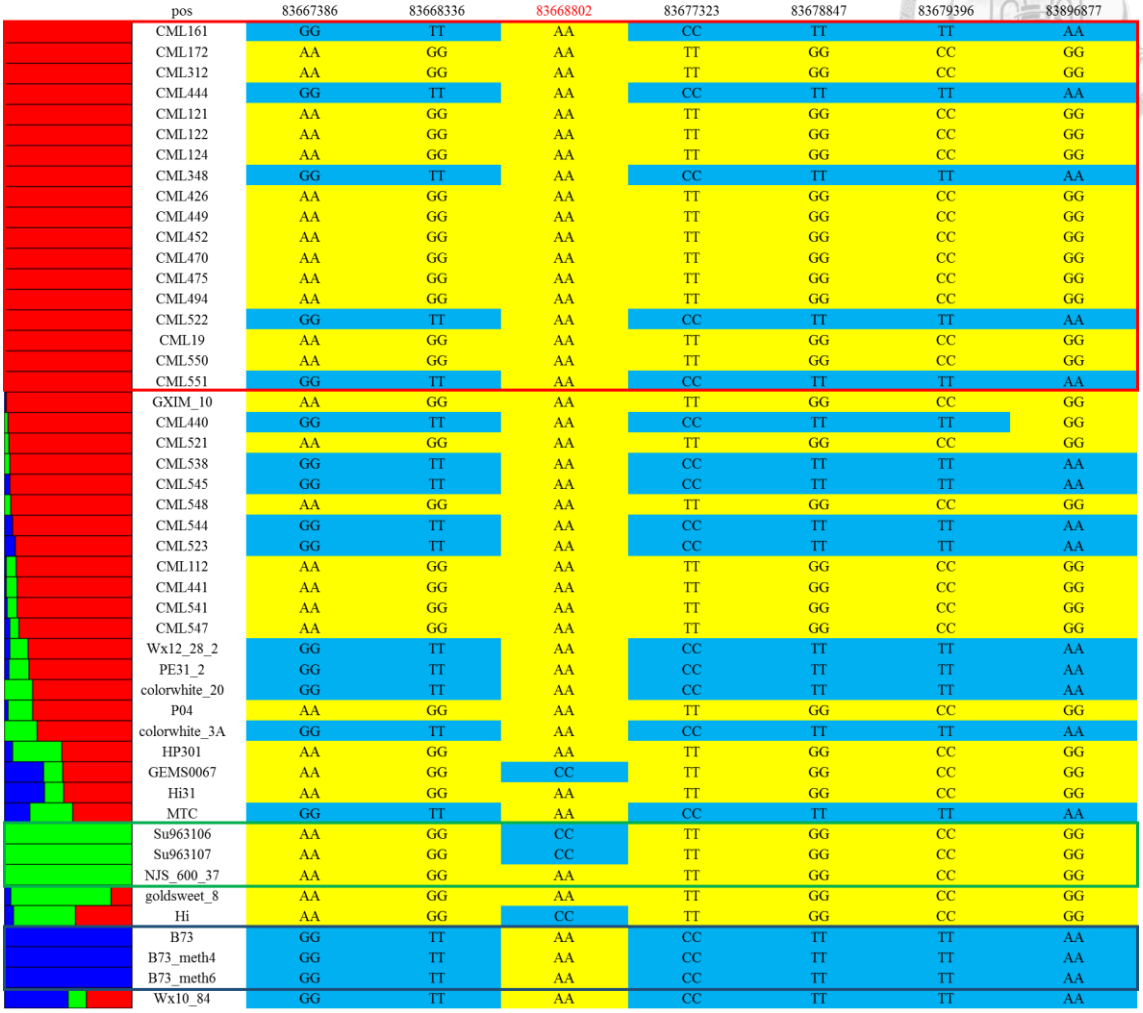
圖十、以 9 個代表品系篩選之族群特有分子標記分布，紅色為熱帶種、藍色為溫帶甜質種、綠色為溫帶馬齒種而灰色為非族群特有。溫帶馬齒種的特有分子標記主要集中在 Chr9 ，整體而言族群特有標記在染色體上沒有明顯的分布特性。

doi:10.6342/NTU201700716



圖十一、以 24 個代表品系篩選之族群特有分子標記分布，紅色為熱帶種、藍色為溫帶甜質種、綠色為溫帶馬齒種而灰色為非族群特有。由於熱帶種的代表品系數量明顯提高，造成該族群特有標記的數量大量提高，而在染色體上的分布依然沒有觀察到明顯的特性。

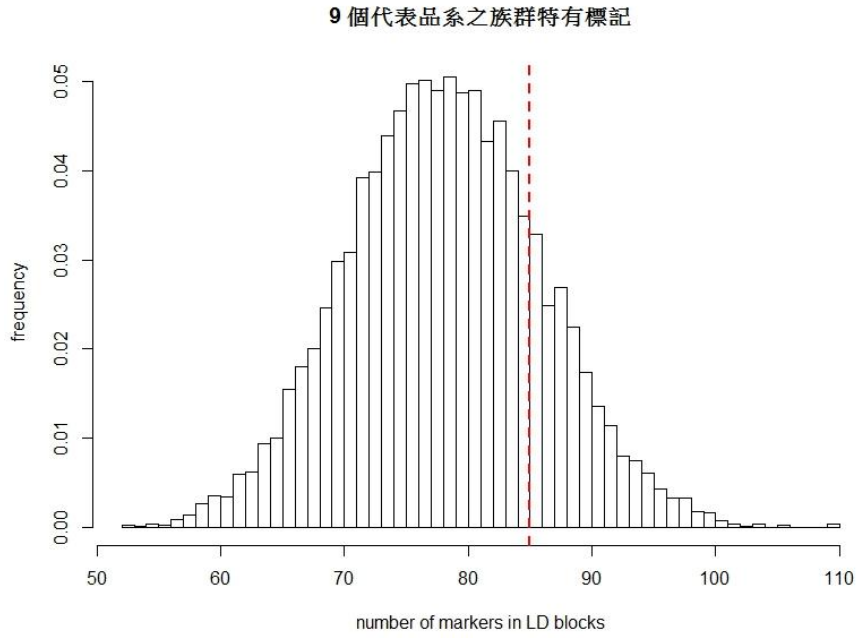
doi:10.6342/NTU201700716

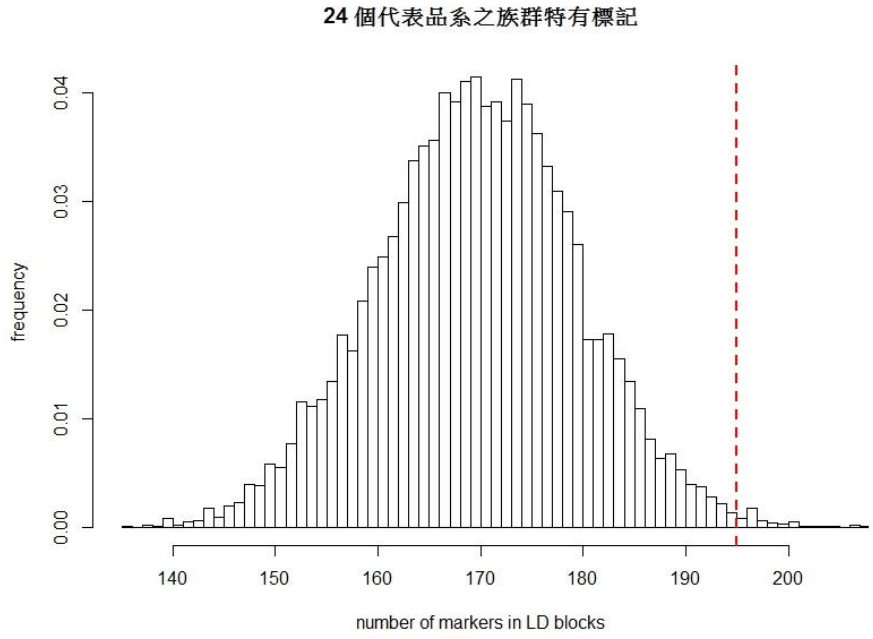


rs#		PZE-103059843	PZE-103059850	PZE-103059853	PZE-103059856	PZE-103059857	PZE-103059858	PZE-103059941
alleles		A/G	T/G	A/C	T/C	T/G	T/C	A/G
chrom		3	3	3	3	3	3	3
pos		83667386	83668336	83668802	83677323	83678847	83679396	83896877
	CML161	GG	TT	AA	CC	TT	TT	AA
	CML172	AA	GG	AA	TT	GG	CC	GG
	CML312	AA	GG	AA	TT	GG	CC	GG
	CML444	GG	TT	AA	CC	TT	TT	AA
	CML121	AA	GG	AA	TT	GG	CC	GG
	CML122	AA	GG	AA	TT	GG	CC	GG
	CML124	AA	GG	AA	TT	GG	CC	GG
	CML348	GG	TT	AA	CC	TT	TT	AA
	CML426	AA	GG	AA	TT	GG	CC	GG
	CML449	AA	GG	AA	TT	GG	CC	GG
	CML452	AA	GG	AA	TT	GG	CC	GG
	CML470	AA	GG	AA	TT	GG	CC	GG
	CML475	AA	GG	AA	TT	GG	CC	GG
	CML494	AA	GG	AA	TT	GG	CC	GG
	CML522	GG	TT	AA	CC	TT	TT	AA
	CML19	AA	GG	AA	TT	GG	CC	GG
	CML550	AA	GG	AA	TT	GG	CC	GG
	CML551	GG	TT	AA	CC	TT	TT	AA
	GXIM_10	AA	GG	AA	TT	GG	CC	GG
	CML440	GG	TT	AA	CC	TT	TT	GG
	CML521	AA	GG	AA	TT	GG	CC	GG
	CML538	GG	TT	AA	CC	TT	TT	AA
	CML545	GG	TT	AA	CC	TT	TT	AA
	CML548	AA	GG	AA	TT	GG	CC	GG
	CML544	GG	TT	AA	CC	TT	TT	AA
	CML523	GG	TT	AA	CC	TT	TT	AA
	CML112	AA	GG	AA	TT	GG	CC	GG
	CML441	AA	GG	AA	TT	GG	CC	GG
	CML541	AA	GG	AA	TT	GG	CC	GG
	CML547	AA	GG	AA	TT	GG	CC	GG
	Wx12_28_2	GG	TT	AA	CC	TT	TT	AA
	PE31_2	GG	TT	AA	CC	TT	TT	AA
	colorwhite_20	GG	TT	AA	CC	TT	TT	AA
	P04	AA	GG	AA	TT	GG	CC	GG
	colorwhite_3A	GG	TT	AA	CC	TT	TT	AA
	HP301	AA	GG	AA	TT	GG	CC	GG
	GEMS0067	AA	GG	CC	TT	GG	CC	GG
	Hi31	AA	GG	AA	TT	GG	CC	GG
	MTC	GG	TT	AA	CC	TT	TT	AA
	Su963106	AA	GG	CC	TT	GG	CC	GG
	Su963107	AA	GG	CC	TT	GG	CC	GG
	NJS_600_37	AA	GG	AA	TT	GG	CC	GG
	goldsweet_8	AA	GG	AA	TT	GG	CC	GG
	Hi	AA	GG	CC	TT	GG	CC	GG
	B73	GG	TT	AA	CC	TT	TT	AA
	B73_meth4	GG	TT	AA	CC	TT	TT	AA
	B73_meth6	GG	TT	AA	CC	TT	TT	AA
	Wx10_84	GG	TT	AA	CC	TT	TT	AA

圖十二、族群特有分子標記造成連鎖失衡區塊圖像不完整現象的實例。此 7 個分子標記圖一舉例之分子標記相同。左圖為 STRUCTURE 的分群結果，藍色為溫帶馬齒種、綠色為溫帶甜質種而紅色為熱帶種。右圖中品系的排序與左圖對應，在框線內的品系為次族群代表品系。造成連鎖失衡區塊圖像不完整的分子標記為 PZE-103059853 以紅字標示，其基因型的比對結果為溫帶甜質種次族群內特有之分子標記。

doi:10.6342/NTU201700716





圖十三、族群特有分子標記於隨機抽樣標記分布上的位置，紅線表示特有分子標記中為造成連鎖失衡區塊圖像不完整分子標記的個數，分別位於 79.7% 與 99.4% 的位置。以 24 個代表品系篩選的族群特有分子標記中為造成連鎖失衡區塊圖像不完整分子標記的機率顯著不同於隨機抽取的分子標記，顯示族群特有分子標記與造成連鎖失衡區塊圖像不完整的分子標記具有相關性。

doi:10.6342/NTU201700716

表二、族群特有分子標記中為造成連鎖失衡區塊圖像不完整分子標記的個數

	溫帶馬齒種	溫帶甜質種	熱帶種	全部
9 個代表品系	2	47	36	85
24 個代表品系	0	16	179	195
共有個數	0	16	36	52

doi:10.6342/NTU201700716

第五章　討論

第一節　MaizeSNP50 資料讀取與利用

在 Ganal 等人（2011）對於 MaizeSNP50 的測試中，B73 的 call rate 達到 99.87 %，teosinte 的 call rate 也達到 91.87 %。本研究的實驗結果與之相符，B73 的 call rate 最高為 98.12%，最低的金蜜 8 為 94.51 %，平均 call rate 為 95.61 %（表一），以單一品系來看具有相當高的可用性，每個品系平均約有 53,000 個 SNP 位點具有可使用的基因型資料。然而，無法正常讀取的分子標記在品系間並不一致，造成在 48 個品系皆具有讀值的位點數為 40,359，佔總 SNP 個數的 71.93 %。這些會在部分品系無法正常讀取的分子標記在染色體中沒有明顯的分布，這種缺值隨機分布的情況會在試驗族群的來源複雜時會造成可用位點數的快速下降。未來如果使用大量蒐集系進行全基因體關聯性分析或是 Genomic Selection 時，對於缺值的統計處理將會十分重要。

在同質結合率的部分，B73 的同質結合率為 97.58 %，在試驗品系中為最高。在所使用的品系中，有 40 個品系其同質結合率位於 97 % 到 95 % 之間，約相當於自交作物 F6 族群，但有部分品系的同質結合率較低。金蜜 8 的同質結合率為 85.91 %，而 CML475 僅有 82.40 %，約相當為自交作物的 F4 族群（表一），足見玉米在育成自交系的過程中，即便在專業的研究機構也並不容易固定基因座，此結果支持了在遺傳研究材料上雙單倍體族群建立的重要性。在本研究中為了探討族群特有基因型與演化歷程的關聯，探討突變發生在特定族群時連鎖失衡區塊圖像的變化，故需除去族群間雜交的影響，在簡化變因的需求下僅使用同質結合的位點。由於異質結合在染色體中的隨機分布，雖然大部分的位點在 48 個品系中皆僅有一或二個品系有異質結合的現象，但僅有一個品系為異質結合的位點就具有上萬個，經篩選後 22,227 個位點為在此 48 個品系中皆被固定，應用於後續研究，僅佔總位點數的 39.61 %。

doi:10.6342/NTU201700716

第二節　連鎖失衡的衰退距離

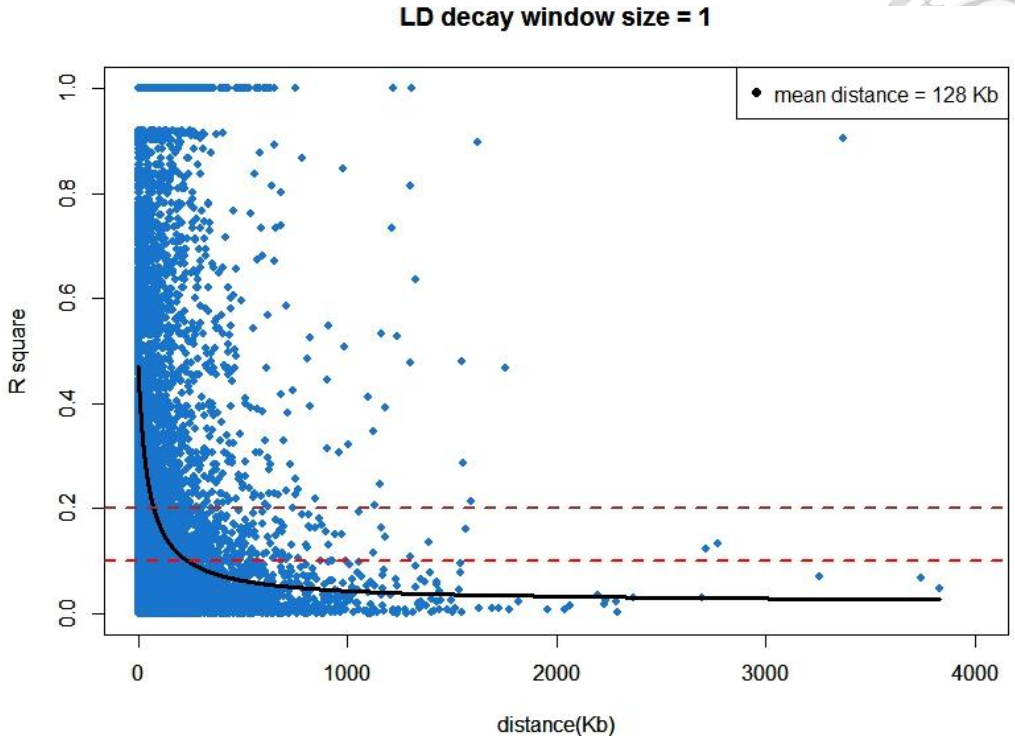
本研究利用 MaizeSNP50 獲得基因型，此基因晶片的設計概念與此晶片主要的 SNP 來源玉米第一代單倍型圖譜，皆是以降低連鎖失衡為主要的訴求。以篩選後 22,227 個分子標記，實際計算相鄰兩位點之間的距離，最遠為 8,012 Kb，最近的距離僅為 2 bp，平均距離約為 128 Kb，依據 Remington 等人（2001）以玉米自交系為材料評估出 1 Kb 的連鎖失衡衰退距離，所使用分子標記的間距可能過大而無法偵測。然而，整理兩位點間的距離大小的分布情形，得到前三分之一的分子標記間距在 6 Kb 以下，後三分之一的間距在 150 Kb 以上，可知分子標記分布並不均勻，具有分布較緊密的區域。另外由 Wu 等人（2014）以優良玉米自交系（Elite inbred lines）所評估連鎖失衡距離得到 391 Kb 的結果，可以得知在現代育種過程會大幅度的改變連鎖失衡的範圍，本研究所使用的自交系多為各研究機構經現代在育種程序育成，連鎖失衡的範圍應較廣，故認定此分子標誌系統足夠評估本試驗族群。

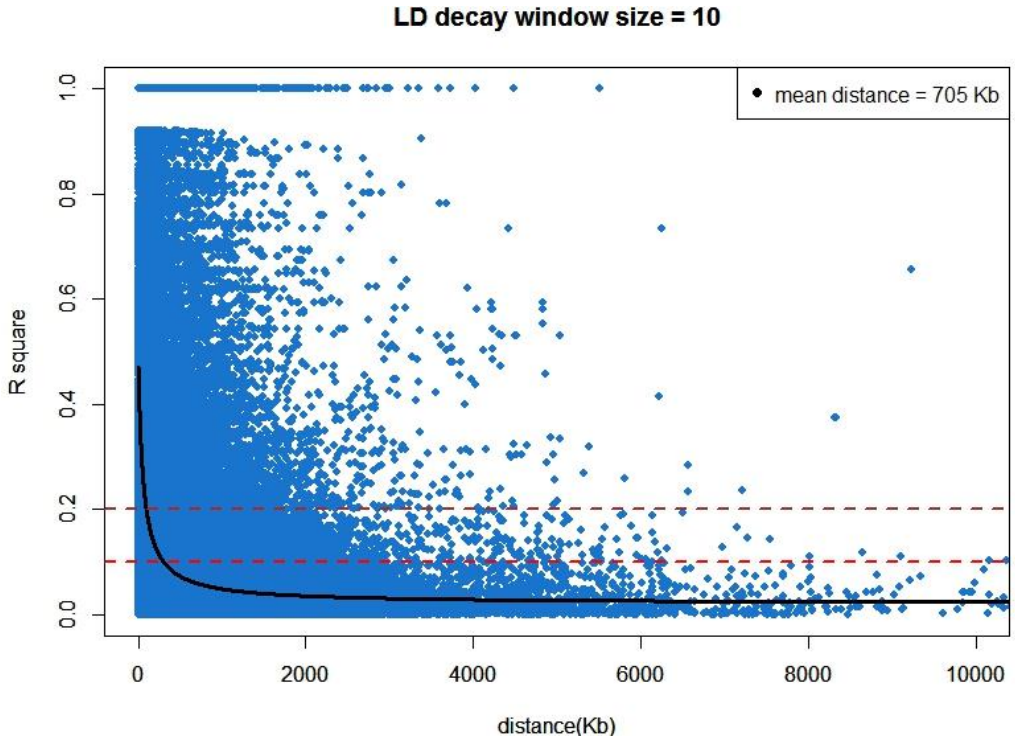
在以 Remington 等人（2001）所提出之 R^2 估計式評估本資料的連鎖失衡程度時，由於使用的分子標記眾多，計算所有兩兩位點間的 R^2 列入評估需要極大的運算資源，故簡化列入計算的位點。以 TASSEL 程式計算相鄰位點間的 R^2，代入估計式的分析結果中 R^2 在距離 74 kb 時小於 0.2，距離 234 Kb 時小於 0.1。當以鄰近 10 個分子標記間的 R^2 代入計算時，R^2 在距離 95 kb 時小於 0.2，衰退至 0.1 的距離會達到 300 Kb，並且可以看到相距 5 Mb 以上的分子標記間 R^2 仍大於 0.8 的現象（圖十四），由此可知依據列入計算的位點數不同會估計出不同的衰退距離，並且當逐步增加列入估計式的位點資訊時會使估算的連鎖失衡衰退距離增加。依據 Wei 等人（2007）利用由 B73 與 Mo17 雜交育成的 IBM 族群（Lee et al. 2002）比較物理圖譜與遺傳圖譜的結果，1 cM 的物理距離介於 1.8 Mb 和 10 Kb 之間，平均為 182Kb，足見在本試驗族群中，存在因染色體結構不易互換產生的大範圍連鎖失衡之外，其他造成此現象的因素。由於

doi:10.6342/NTU201700716

連鎖失衡現象在物理圖譜上具有極大的變異，未來進行相關評估時，以較少人為選擇並可產生更大量位點數的 GBS（Genotyping by sequencing）分子標記或許是更好的策略。在觀察連鎖失衡區塊時，若有分子標記的密度偏低的現象，會有三種可能結果：(1) 沒有偵測到實際存在的連鎖失衡區塊，(2) 觀察到的區塊範圍大小可能會被低估，(3) 在視覺化上每個區塊的組成被簡化，前兩點不影響本研究的目標，而第三點的部分，由於區塊內的位點數下降，連鎖失衡區塊圖像不完整的程度會有一定程度的簡化。

doi:10.6342/NTU201700716





圖十四、連鎖失衡衰退情形。window size = 1 時，R^2 在距離 74 kb 時小於 0.2、234 Kb 時小於 0.1；window size = 10 時，R^2 在距離 95 kb 時小於 0.2、 300 Kb 時小於 0.1，當 window size 提高時能觀察到距離數 mb 仍具有高度連鎖失衡的資料點。

doi:10.6342/NTU201700716

第三節　玉米族群結構探討

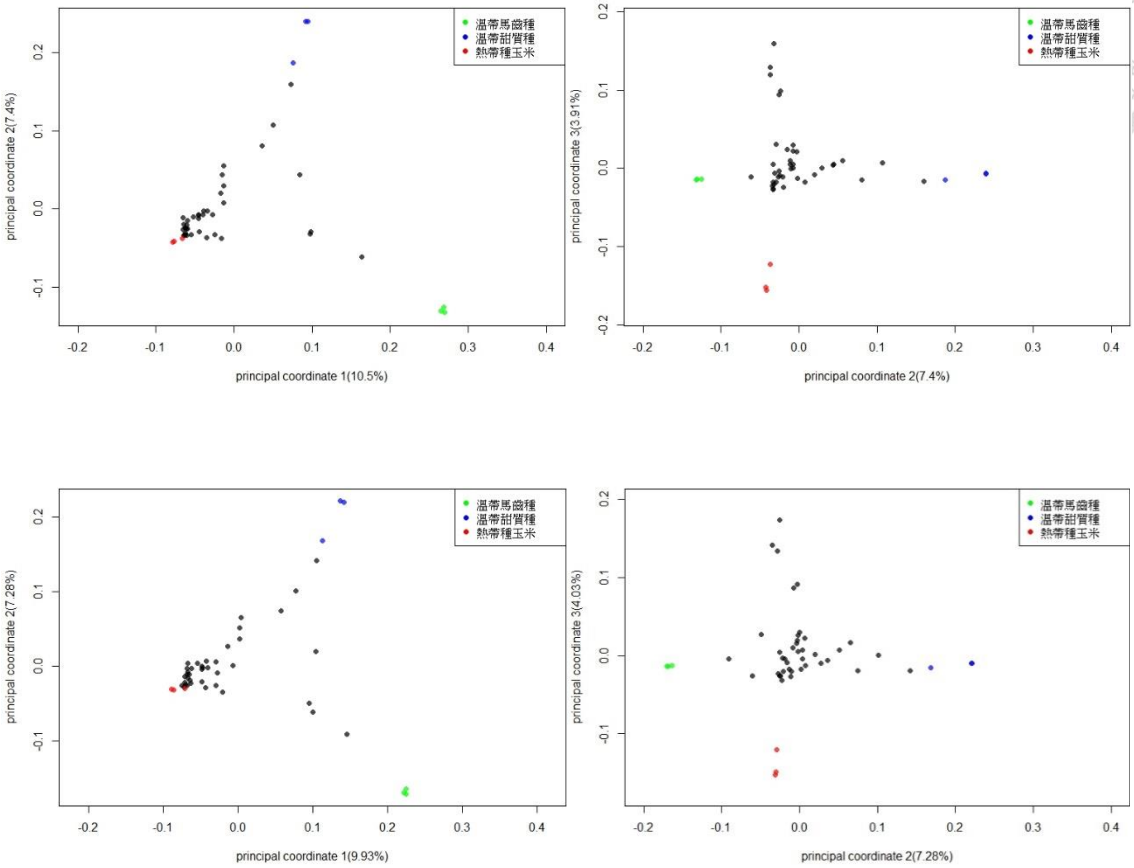
使用分子標記基因型資料估計遺傳距離時，MRD（Modified Roger's distance）是常用的計算方式，由於其同時符合歐基里德距離的性質，在這樣的前提下主成分分析與主座標分析會得到近乎相同的結果。本研究以 MRD 計算遺傳距離，並以此進行主座標分析做為主要的分析方法。Ganal 等人（2011）提出 MaizeSNP50 由於其組成中部分來源僅來自於溫帶型玉米的多型性，因此在進行分析時會放大溫帶型玉米間的差異，僅使用其中來自 Panzea 的分子標記可以除去此疑慮。在篩選後的 22,227 個分子標記中具有 14,570 個 Panzea 的分子標記，分別以此兩組標記進行主座標分析，期望後者可以較凸顯（亞）熱帶玉米間的差異。然而，分析結果上在主座標的解釋力和二維散布圖的分布上皆無明顯差異（圖十五），推論其原因可能是在 MaizeSNP50 中第二主要的分子標記來源為 B73 與 M17 的雜交族群。此兩個玉米品系皆為溫帶馬齒種，而在本研究所使用的品系中僅有 4 個溫帶馬齒種品系，並且其中 3 個為 B73 的近同源系（Near-isogenic line, NIL），故不影響品系在散布圖上的整體狀況，因此後續也以 22,227 個分子標記進行分析。在品系資料比對散布圖的結果上，單以地區、胚乳特性分類均不能看到明確的區分界線，僅有依氣候分類得到明確的區分。在第一主座標上溫帶玉米的特徵值在 0.035 以上，（亞）熱帶玉米則在 -0.013 以下，亞熱帶與熱帶玉米的交界約在 -0.059 的位置。在同時利用氣候與地區資訊時，地理位置的差異就可以被區分出來。Suwarno 等人（2014）提到利用分子標記評估遺傳距離可能不是作為區分雜種優勢群的良好策略，但在其研究中僅使用 1,536 個分子標記，有分子標記不足的疑慮。本研究以 CIMMYT 品系配合 22,227 個分子標記進行主座標分析，並比對 CIMMYT 對其品系的雜種優勢群分類驗證此假說，在主座標分析圖上無法看到分群界線，支持前人的研究結果（圖十六）。

在 STRUCTURE 軟體分群部分，軟體計算上所得到 Ln P(D) 的變異程度會因為分群數的增加而增加，在 MCMC reps = 20000 的設定條件下，認為有將變

doi:10.6342/NTU201700716

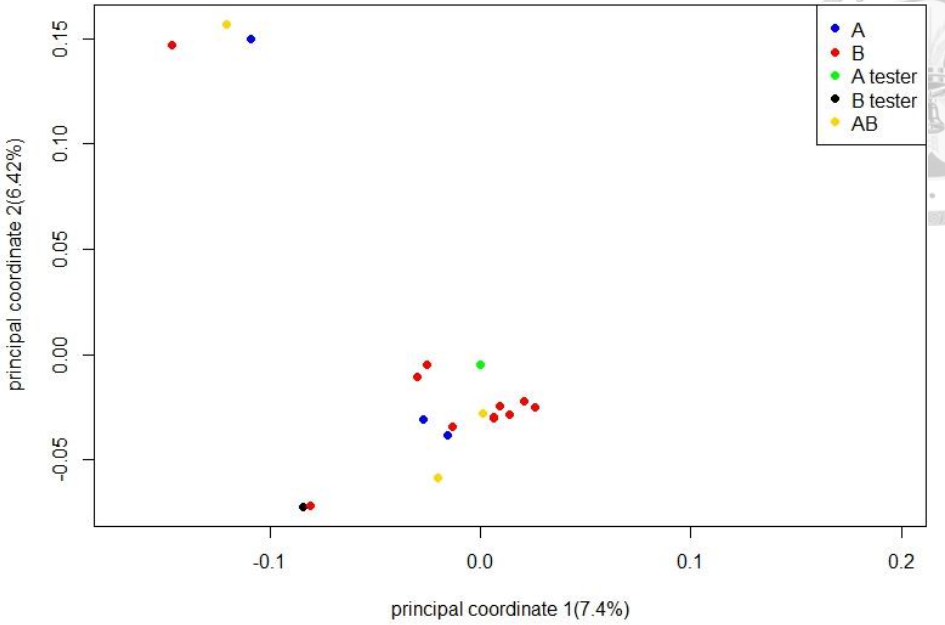
異控制在可接受範圍內。本研究依據 Evanno 等人（2005）提出的方式決定最適分群數，設定 2 到 7 的分群數，進行 20 次的重複計算，依其方法選定的分群數為 3。此分群數設定在分群結果上十分穩定，在每次重複中軟體所選定的起始品系皆相同，其結果僅 Q 值具有些微差異。而在更高分群數設定的分群結果中會出現軟體選定的始祖品系不一致的情形，其嚴重程度依分群數的提高而上升。在 Chao 等人（2010）的小麥研究上也具有相同的現象，此顯示當試驗族群中個體間的差異不足時，軟體在始祖品系的選定上會產生隨機錯誤的現象。在分群數為 3 的設定下，其分析結果的高穩定性也應證前人所使用的估計方式穩定可信賴。另外 STRUCTURE 所選出的最佳分群數與主座標分析的散布圖互相支持，有鑑於 STRUCTURE 在分析上所使用的運算資源較高，在決定分群數的過程中，應可以事先進行主成份或主座標分析，在散布圖上預先判斷可能的分群數，減少 STRUCTURE 候選分群數的範圍，如此應可減少大量運算資源的消耗。

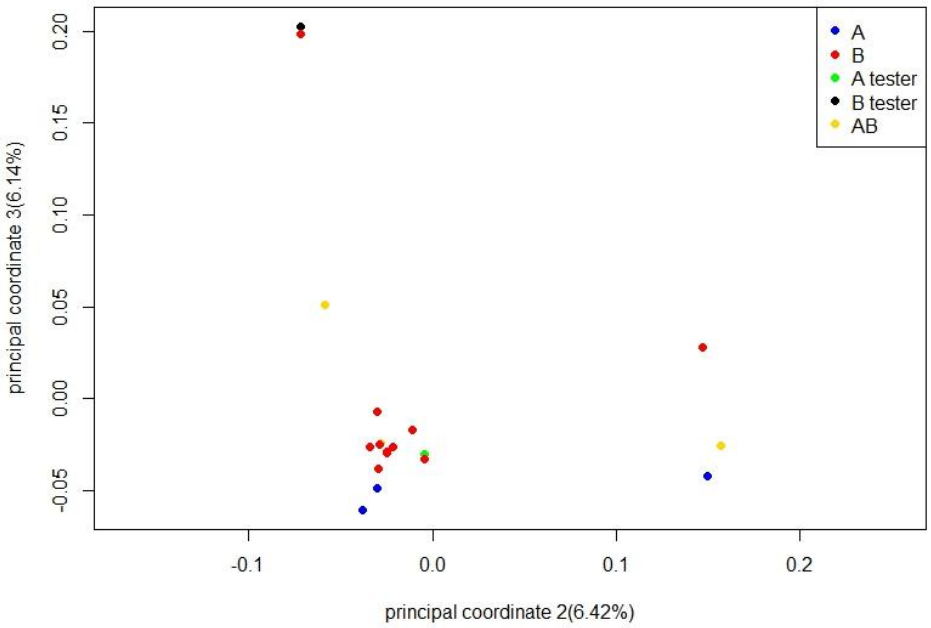
doi:10.6342/NTU201700716



圖十五、48 個試驗品系的主座標分析圖，上圖為以 22,227 個同質結合分子標記的分析結果；下圖為同質結合分子標記中使用來自 Panzea 的 14,570 個標記的分析結果。比較兩組分子標記資料的主座標分析圖，不具有明顯差異，以本資料來說 Panzea 資料的分析結果並不能改善熱帶種玉米族群分布密集的現象。

doi:10.6342/NTU201700716





圖十六、以來自 CIMMTY 的 29 個品系進行之主座標分析，依據其單位公布之雜種優勢群分類。主座標分析的結果與雜種優勢群的分類不能吻合，應證雜種優勢的現象不能單純以遺傳距離或基因型的差異來評估。

doi:10.6342/NTU201700716

第四節　族群代表品系與特有分子標記的決定

STRUCTURE 軟體的分群結果以 Q 值表示，軟體評估所使用的所有試驗品系，依據 K 值設定建構出 K 個始祖品系，並以 Q 值表示各始祖品系佔目標試驗樣品遺傳背景的比例，若試驗樣品的遺傳背景來自特定始祖品系為主，則可認定為來自特定族群。然而，劃分為特定族群的 Q 值標準在各相關文獻中沒有統一，部分學者傾向將所有品系均劃入族群中（Wang et al. 2012），部分為設定特定 Q 值標準（Yan et al. 2009, Jin et al.2010, Zhao et al. 2010, Semagn et al. 2012, Wu et al. 2014），其標準介於 0.5 到 0.8 之間，依據研究者的主觀認定。在本研究中具有 9 個品系其 Q 值為 1，表示其遺傳背景組成與軟體建構的特定始祖品系相同，以單一族群來看，這樣的情況可將 Q 值為 1 的品系認定為族群的起源點。在這些品系間彼此依然會具有部分的多型性，這些具有多型性的位點，是以本研究所使用的品系中假定在族群建立的初期，未經族群間雜交就具有的多型性。其存在具有兩種可能：第一種為在更早的演化歷程中也就是族群分化前產生，第二種為族群分化後經突變產生。若為第一種，假定在後續的演化過程中，該位點並沒有在族群中因遺傳漂變而被固定，則此多型性在每個族群內都應該存在。若為第二種，則表示在更早的演化歷程中此位點應僅具有一種基因型，在族群分化後發生的突變會造成族群特有的基因型的產生，進而在該位點上形成僅在特定族群內具有多型性的現象。

本研究將 Q 值為 1 的品系定為族群的代表品系，並將此種僅在特定族群的代表品系內具有多型性且在其他族群中僅有共同的單一基因型位點視為族群特有的分子標記。另外本研究所使用的 CIMMTY 品系間遺傳背景十分相近，造成在計算上具有 15 個熱帶玉米品系，其 Q 值在 0.99 以上，因此建立兩組依不同標準選定的代表品系。由於本研究所使用的品系為各實驗機構育成之自交系，具有異質結合的位點應為在人為育種多代自交的過程中並未固定下來的位點，認定為族群分化後經族群間雜交而產生，故在特有分子標記的篩選中除去。在標記篩

doi:10.6342/NTU201700716

選的結果上，熱帶種族群的特有標記最多，溫帶馬齒種的特有標記則最少，比對本資料經 STRUCTURE 計算的 Fst（Fixation index）值，熱帶種為 0.205、溫帶甜質種 0.637 而溫帶馬齒種則為 0.948，表示在溫帶馬齒種族群內的變異極少，呼應該群內特有標記最少的結果。

第五節　族群特有分子標記與連鎖失衡區塊

連鎖失衡區塊圖像不完整的現象一直是比較少研究者關注的目標，在前人的研究文獻中發現當評估連鎖失衡所使用的實驗材料來源越單一，則連鎖失衡區塊的範圍越大並且區塊圖像不完整的現象越少。關於連鎖失衡區塊範圍的前人研究中，Goldstein（2001）認為連鎖失衡與染色體的互換率相關，由於部分染色體區域的互換率極高，造成連鎖失衡程度的驟降，區塊間的區域若在物理圖譜上範圍小，即表示該區域為互換率高的位置。目前學者們普遍接受在自交或遺傳背景較窄的族群中，由於其具有多型性的位點不足，或經遺傳漂變固定部分位點的基因型，造成染色體互換的現象有一定程度不會顯現在資料中，在沒有外來品系的影響下會具有較大的連鎖失衡區塊。而連鎖失衡區塊圖像不完整的現象則常被研究者所忽略，由於不論是在 GWAS 或是單倍型圖譜的研究，造成區塊圖像不完整的位點並不是主要影響 QTL 探勘或單倍型圖譜簡化等研究的變因，是以研究者多僅討論連鎖失衡的範圍作為後續實驗的參考。

本研究關心這些造成區塊圖像不完整分子標記的成因，希望能對此現象做出可能的解釋。連鎖失衡區塊為染色體上因物種本質或地理區隔等因素，造成相鄰基因座間對偶基因彼此沒有獨立分配的區域。前人研究（Goldstein, 2001）中推論染色體的互換造成連鎖失衡區塊之間的分界，決定了區塊的範圍，因此本研究在區塊完整性的討論上不考慮染色體互換的因素。Lam 等人（2010）的研究結果顯示試驗材料的組成不只影響了連鎖失衡區塊的範圍也同時影響了區塊的完整性。Chao 等人（2010）的研究中也提到春小麥與冬小麥族群在相同染色體區域上的

doi:10.6342/NTU201700716

連鎖失衡區塊具有不同的圖像。因此本研究推測區塊圖像不完整的現象可能是由於試驗材料中具有族群結構，將不同族群的基因型資料共同計算連鎖失衡而造成。在演化過程中，在族群分化後於特定族群內發生的基因突變而產生的核苷酸變異，可能是在由多個族群共同計算的連鎖失衡區塊內具有斷點的原因。據此利用 bootstrap 的概念，以隨機抽取的方式建立分布，驗證族群內特有具多型性的分子標記中造成連鎖失衡區塊圖像不完整的分子標記的個數是否在統計上顯著不同於隨機抽取的分子標記中造成連鎖失衡區塊圖像不完整的分子標記。實驗結果顯示以 9 個代表品系所選出的族群特有分子標記中造成連鎖失衡區塊圖像不完整的分子標記的個數在抽樣分布上位於 79.7% 的位置，並沒有達到統計上的顯著，但具有可能相關的趨勢。而利用 24 個代表品系的資料所進行的分析結果位於抽樣分布上 99.4%，顯示族群特有分子標記中為造成連鎖失衡區塊圖像不完整的分子標記的機率與隨機選取的分子標記不同，與造成連鎖失衡區塊圖像不完整的分子標記具有相關性。

所選出造成連鎖失衡區塊圖像不完整的族群特有分子標記主要來自於熱帶玉米族群，來自溫帶馬齒種族群則最少。推測此結果應與所使用的品系來源在三個族群的比例不一有關，溫帶馬齒種的代表品系僅有 3 個，本實驗對於探勘造成連鎖失衡區塊圖像不完整分子標記的設定中，以 R^2 小於 0.8 做為不與相鄰分子標記具有高連鎖失衡的標準，在本資料中表示在 48 個品系中有 3 個品系以上的基因型在兩個位點間不相同，因此若兩分子標記的基因型讀值僅在此族群中有差異，則在連鎖失衡的計算上，兩分子標記間的 R^2 會大於 0.8 而不被篩選到。另外若將 48 個品系中以最高 Q 值將所有品系分配到三個族群中，則分配到溫帶馬齒種族群中的品系也僅有 4 個，同理在 9 個代表品系的特有標記篩選中，由於熱帶種族群所使用的代表品系僅有 3 個，會出現相同的問題。推測此問題應為兩組代表品系資料所得之結果具有差異的原因，因此推論若實驗所使用的品系數上升，並且來源分布更平均，族群特有分子標記與造成連鎖失衡區塊圖像不

doi:10.6342/NTU201700716

完整的分子標記會有更顯著的相關性。

本研究結果指出造成連鎖失衡區塊圖像不完整的分子標記應為族群分化後在特定族群內突變產生的核苷酸變異，故推測這些在連鎖失衡區塊內顯示低連鎖失衡關係的分子標記在染色體上應仍然具有低互換率的特性，因此在遺傳關係較相近品種間的品種鑑定，或是檢查育種過程中品系間的譜系關係上這些分子標記可做為參考依據。在 GWAS 的 QTL 的探勘上，當進入較精細的基因定位程序時，若試驗材料具有一定的族群結構，這些族群特有標記應可作為參考進行部分的校正。將這類位於連鎖失衡區塊內，因突變造成基因型讀值與相鄰位點之間相關性驟降的基因座在 QTL 分析中去除。使得當進行遺傳定位時，在具有連鎖失衡現象的區域中減少突變的干擾，或可使定位結果更加精準。

doi:10.6342/NTU201700716

第六章　結論

本研究以 48 個玉米自交系作為遺傳材料，並以 MaizeSNP50 基因晶片建立基因型資料庫。分別以 STRUCTURE 與 PCoA 評估遺傳結構並決定各族群的代表品系，篩選出在特定族群內具有多型性的分子標記。同時以 TASSEL 評估分子標記間連鎖失衡的情形，篩選出連鎖失衡區塊內 R^2 驟降造成區塊圖像不完整的標記，並以所有分子標記資料依 bootstrap 的方式建立抽樣分布，檢定族群特有分子標記相對於隨機抽取的分子標記是否具有較高的比例為造成連鎖失衡區塊圖像不完整的分子標記，以推論其與族群特有具多型性分子標記間是否具有相關性。試驗結果中，STRUCTURE 與 PCoA 皆顯示本資料的最佳分群數為 3，分別為溫帶馬齒種、 溫帶甜質種與熱帶種。其中由於熱帶種族群中具有多個品系遺傳背景十分相近，故在代表品系的選擇上建立兩組不同 Q 值篩選標準的資料，並各自建立族群特有分子標記資料庫，配合所篩選出造成連鎖失衡區塊圖像不完整的分子標記進行檢定。兩組資料皆顯示族群特有分子標記與造成區塊圖像不完整的分子標記具有相關的趨勢，其中以 24 個代表品系篩選的分子標記位於抽樣分布上 99.4% 的位置，與隨機抽取的分子標記間具有顯著的差異，此結果顯示造成連鎖失衡區塊圖像不完整的分子標記，其低連鎖失衡關係應為演化歷程中族群分化後，在特定族群內發生的突變有關。

doi:10.6342/NTU201700716

參考文獻

謝光照，盧煌勝。(2006) 黑糯玉米台農 5 號之育成。台灣農業研究 55(3): 149-163。

謝光照。(2015) 白糯玉米「台農 6 號」之育成。台灣農業研究 64(2): 99-108。

謝光照。(2010) 糯質玉米自交系籽粒種皮厚度之差異。台灣農業研究 59(2): 103-111。

劉紹國，謝光照。(2012) 超甜玉米品種間遺傳距離與產量之相關性。台灣農業研究 61(3):186-195。

Camus-Kulandaivelu L, Veyrieras JB, Madur D, Combes V, Fourmann M, Barraud S, et al. (2006) Maize adaptation to temperate climate: relationship between population structure and polymorphism in the *Dwarf8* gene. *Genetics* 172:2449-2463

Chao S, Dubcovsky J, Dvorak J, Luo MC, Baenziger SP, Matnyazov R, et al. (2010) Population- and genome-specific patterns of linkage disequilibrium and SNP variation in spring and winter wheat (*Triticum aestivum* L.). *BMC Genomics* 11:727.

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. (2001) High-resolution haplotype structure in the human genome. *Nature Genetics* 29:229 - 232

Evanno G, Regnaut S, Goudet J. (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* 14:2611–2620.

Fulton TM, Chunwongse J, Tanksley SD. (1995) Microprep Protocol for Extraction of DNA from Tomato and Other Herbaceous Plants. *Plant Molecular Biology Reporter* 13:207-209.

Ganal MW, Durstewitz G, Polley A, Bérard A, Buckler ES, Charcosset A, et al. (2011) A Large Maize (*Zea mays* L.) SNP Genotyping Array: Development and Germplasm Genotyping, and Genetic Mapping to Compare with the B73 Reference Genome. *PLoS ONE* 6: e28334.

Goldstein DB. (2001) Islands of linkage disequilibrium. *Nature Genetics* 29:109 – 111

Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, et al. (2009) A first-generation haplotype map of maize. *Science* 326:1115-1117

Hamblin MT, Fernandez MGS, Casa AM, Mitchell SE, Paterson AH, and Kresovich S. (2005) Equilibrium Processes Cannot Explain High Levels of Short- and Medium-Range Linkage Disequilibrium in the Domesticated Grass Sorghum bicolor. *Genetics* 171: 1247–1256.

Hyten DL, Choi IY, Song Q, Shoemaker RC, Nelson RL, Costa JM, et al. (2007) Highly Variable Patterns of Linkage Disequilibrium in Multiple Soybean Populations. *Genetics* 175(4): 1937–1944

doi:10.6342/NTU201700716

Inghelandt DV, Melchinger AE, Lebreton C, Stich B. (2010) Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. *Theoretical and Applied Genetics* 120: 1289–1299

Jin L, Lu Y, Xiao P, Sun M, Corke H, Bao J. (2010) Genetic diversity and population structure of a diverse set of rice germplasm for association mapping. *Theoretical and Applied Genetics* 121:475-487

Knowler WC, Williams RC, Pettitt DJ, Steinberg AG. (1988) Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *The American Journal of Human Genetics* 43:520–526

Konishi S, Izawa T, Lin SY, Ebana K, Fukuta Y, Sasaki T, et al. (2006) An SNP caused loss of seed shattering during rice domestication. *Science* 312(5778):1392-1396

Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, Purugganan MD, et al. (2009) A Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in *Arabidopsis thaliana*. *PLoS Genetics* 5: e1000551.

Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL. et al. (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genetics* 42: 1053–1059

Lee M, Sharopova N, Beavis WD, Grant D, Katt M, Blair D, et al. (2002) Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population. *Plant Molecular Biology* 48:453-461.

Marroni F, Pinosio S, Zaina G, Fogolari F, Felice N, Cattonaro F, et al. (2011) Nucleotide diversity and linkage disequilibrium in Populus nigra *cinnamyl alcohol dehydrogenase* (CAD4) gene. *Tree Genetics & Genomes* 7:1011–1023

Mather KA, Caicedo AL, Polato NR, Olsen KM, McCouch S, Purugganan MD. (2007) The Extent of Linkage Disequilibrium in Rice (*Oryza sativa* L.). *Genetics* 177: 2223–2232

Morrell PL, Williams-Coplin TD, Lattu AL, Bowers JE, Chandler JM, Paterson AH. (2005) Crop-to-weed introgression has impacted allelic composition of johnsongrass populations with and without recent exposure to cultivated sorghum. *Molecular Ecology* 14:2143–2154

Olivier M. (2003) A haplotype map of the human genome. *Physiological Genomics* 13:3-9

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38:904-909

Pritchard JK, Stephens M, Donnelly P. (2000) Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, et al.

doi:10.6342/NTU201700716

(2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *PNAS* 98:11479-11484

Sapolsky RJ, Hsie L, Berno A, Ghandour G, Mittmann M, Fan JB. (1999) High-throughput polymorphism screening and genotyping with high-density oligonucleotide arrays. *Genetic Analysis* 14:187-192.

Semagn K, Magorokosho C, Vivek BS, Makumbi D, Beyene Y, Mugo S, et al. (2012) Molecular characterization of diverse CIMMYT maize inbred lines from eastern and southern Africa using single nucleotide polymorphic markers. *BMC Genomics* 13:113

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326:1112-1115

Schulze A, Downward J. (2001) Navigating gene expression using microarrays--a technology review. *Nature Cell Biology* 3:E190-E195

Shen R, Fan JB, Campbell D, Chang W, Chen J, Doucet D, et al. (2005) High-throughput SNP genotyping on universal bead arrays. *Mutation Research* 573:70-82

Steemers FJ, Gunderson KL. (2005) Illumina, Inc. *Pharmacogenomics* 6:777-782.

Suwarno WB, Pixley KV, Palacios-Rojas N, Kaeppler SM, Babu R. (2014) Formation of Heterotic Groups and Understanding Genetic Effects in a Provitamin A Biofortified Maize Breeding Program. *Crop science* 54:14–24

Wang C, Jia G, Zhi H, Niu Z, Chai Y, Li W, et al. (2012) Genetic diversity and population structure of Chinese foxtail millet [*Setaria italica* (L.) Beauv.] landraces. *G3: genes genomes genetics* 2:769-777

Wei F, Coe E, Nelson W, Bharti AK, Engler F, Butler E, et al. (2007) Physical and Genetic Structure of the Maize Genome Reflects Its Complex Evolutionary History. *PLoS Genetics* 3:e123

Wu X, Li Y, Shi Y, Song Y, Wang T, Huang Y, et al. (2014) Fine genetic characterization of elite maize germplasm using high-throughput SNP genotyping. *Theoretical and Applied Genetics* 127: 621–631

Yamanaka S, Nakamura I, Watanabe KN, Sato Y. (2004) Identification of SNPs in the *waxy* gene among glutinous rice cultivars and their evolutionary significance during the domestication process of rice. *Theoretical and Applied Genetics*. 108:1200–1204

Yan J, Shah T, Warburton ML, Buckler ES, McMullen MD ,and Crouch J. (2009) Genetic Characterization and Linkage Disequilibrium Estimation of a Global Maize Collection Using SNP Markers. *PLoS One* 4: e8451

Yu J, Holland JB, McMullen MD, Buckler ES. (2008) Genetic Design and Statistical Power of Nested Association Mapping in Maize. *Genetics* 178:539-551

doi:10.6342/NTU201700716

Zhao K, Wright M, Kimball J, Eizenga G, McClung A, Kovach M, et al. (2010) Genomic diversity and introgression in *O. sativa* reveal the impact of domestication and breeding on the rice genome. *PLoS One* 5:e10780

doi:10.6342/NTU201700716

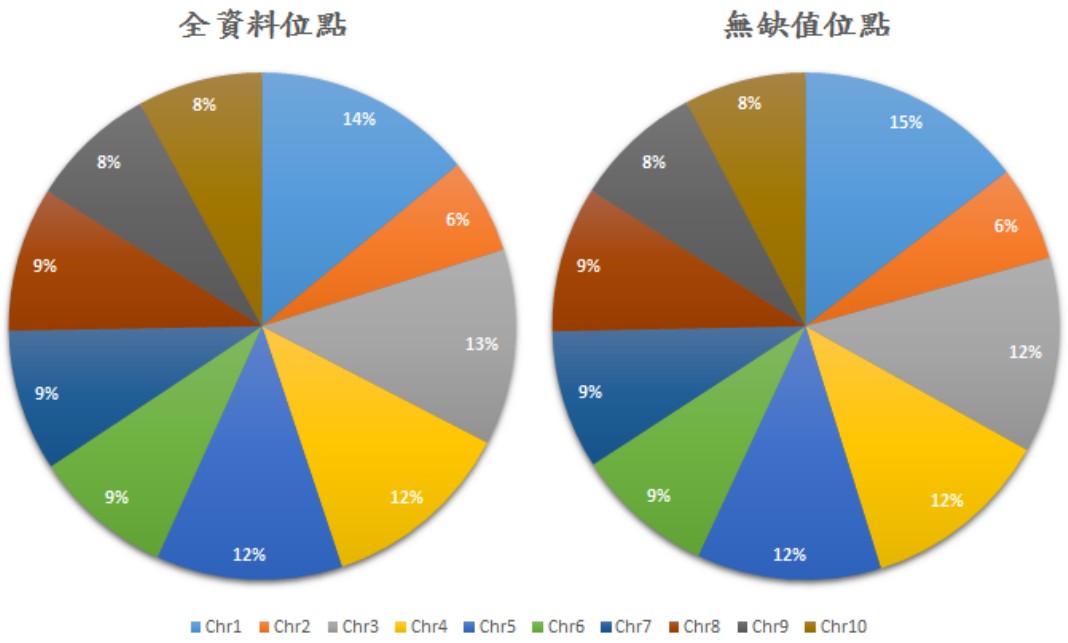
附錄

附錄表一、11 個有併管之 DNA 樣品。

此 11 的樣品在 Call rate 和同質結合率上皆與 48 個樣品的平均值接近，並沒有因為併管而造成資料讀取上的問題。

品系名	Call rate	同質結合率
CML312	95.44 %	96.24 %
Hi	94.80 %	96.57 %
PO4.TNG5.父本	95.27 %	96.22 %
新吉士.600.37	94.95 %	95.89 %
CML112	95.52 %	96.37 %
CML426	95.28 %	95.77 %
CML452	95.47 %	96.31 %
CML494	95.37 %	96.36 %
CML523	95.72 %	96.43 %
CML544	95.71 %	96.45 %
CML547	95.63 %	96.21 %

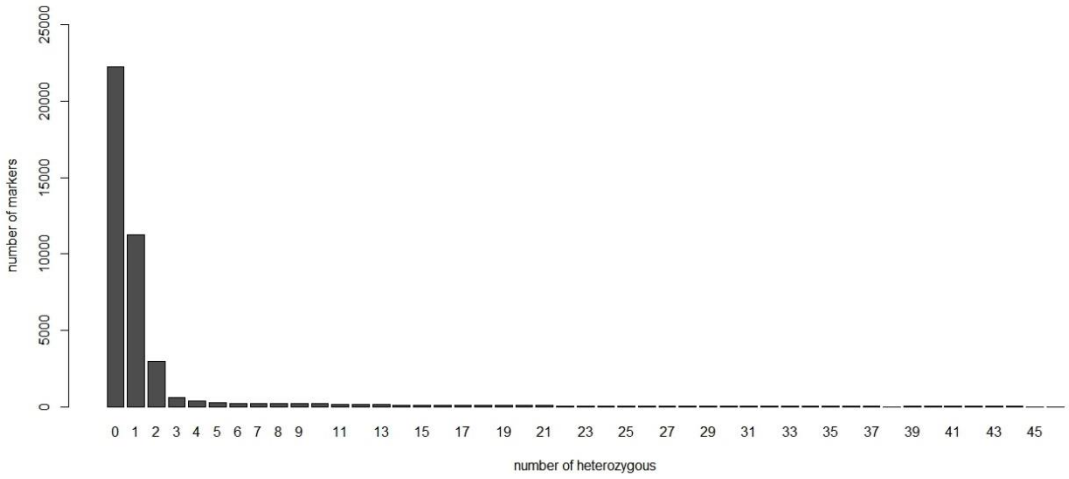
doi:10.6342/NTU201700716



附錄圖一、資料點於各染色體的分布。

兩組資料在不同染色體上的比例近乎一致，表示 MaizeSNP50 晶片的讀取不會在特定染色體上有較容易產生缺值的現象。

doi:10.6342/NTU201700716



附錄圖二、各分子標記於 48 個品系的基因型讀值中為異質結合的個數。
在所有分子標記中，具有異質結合現象的分子標記。其檢測 48 個品系的基因型讀值，大多僅有 1 或 2 個品系的讀值為異質結合基因型。

doi:10.6342/NTU201700716

附錄程式碼一、MaizeSNP50 資料的整理

##去除亂碼與缺值

```
rm(list=ls())
setwd(choose.dir())
raw1=read.csv("maizeSNP50 Full Data Table.csv")
raw2=read.csv("maizesnp50_SNP info1.csv")
```

#maizeSNP50 Full Data Table.csv 為此次試驗的基因型定型結果資料，maizesnp50_SNP info1.csv 為 maizeSNP50 中 SNP 位點的基本資料。前者由 Genome Studio 輸出獲得，後者由 illumina 產品網站獲得之 maizesnp50_SNP info.xls 檔案除去商品介紹資訊後，僅含 SNP 相關資訊並轉換成 csv 格式儲存而得。

```
raw1v=raw1[,c(2,4,5,11,15,19,23,27,31,35,39,43,47,51,55,59,63,67,71,75,79,83,87
,91,95,99,103,107,111,115,119,123,127,131,135,139,143,147,151,155,159,163,167
,171,175,179,183,187,191,195,199)]
```

#將資料簡化成僅有 IlmnStrand、 SNP、 Name、 chr、 position and genotype。

```
data=cbind(raw2[,c(3,4)],raw1v)
for (i in c(6:53)){
  data=subset(data,data[,i]=="AA"|data[,i]=="BB"|data[,i]=="AB")
   }
```

#將 genotype 資料簡化成僅有 AA、 AB、 BB 三種，將其餘缺值或亂碼資料去除。

```
write.table(data, "maizeSNP50 Full Data.csv", sep=",")
```


doi:10.6342/NTU201700716

##去除異質結合基因座

```
Data=read.csv("maizeSNP50 Full Data.csv")
for (x in c(6:53)){
  Data=subset(Data,Data[,x]!="AB")
  }
write.table(Data, "maizeSNP50 Full Data without heterozygous.csv", sep=",")
```

##基因型資料轉換為核苷酸表示

```
rm(list=ls())
setwd(choose.dir())
raw=read.csv("maizeSNP50 Full Data without heterozygous.csv")
TOP=subset(raw,raw$IlmnStrand=="TOP")
BOT=subset(raw,raw$IlmnStrand=="BOT")
#依據 Illumina 產品介紹文件，將資料依 IlmnStrand 欄位分為 TOP、 BOT
兩群。
CG=subset(TOP,TOP$SNP=="[C/G]")
for (i in c(6:53)){
  CG[,i]=gsub("AA",replacement="CC",CG[,i])
  CG[,i]=gsub("BB",replacement="GG",CG[,i])
}
CG=subset(BOT,BOT$SNP=="[C/G]")
for (i in c(6:53)){
  CG[,i]=gsub("AA",replacement="GG",CG[,i])
  CG[,i]=gsub("BB",replacement="CC",CG[,i])
}
```


doi:10.6342/NTU201700716

#依 SNP 欄位決定替換項目，以 SNP 欄位為 C/G 為例。

TOPdata=rbind(AT,AC,AG,TA,TC,TG,CG,GC)

#合併資料，以 TOP 資料為例

##繪製分子標記的分布

achr=subset(data,data$Chr=="1")

achr=subset(achr,achr$Position!="0")

#以第一號染色體為例，以 subset() 指令取出第一號染色體的資料，並且除去未知物理位置的分子標記。

p=max(achr$Position)

k=ceiling(p/5000000)

hist(achr$Position/1000000,breaks=k,main="Chr1",col="gray",

xlim=c(0,300),ylim=c(0,150),xlab="position(mb)",ylab="count")

#以染色體上最遠分子標記的物理位置除以 5,000,000 作為組數，繪製橫軸為物理位置，縱軸為組內分子標記的計數。

doi:10.6342/NTU201700716

附錄程式碼二、族群結構的評估

##主座標分析

```
rm(list=ls())
setwd(choose.dir())
RawA=read.csv("maizeSNP50 Full Data without heterozygous 01 A.csv")
Name=read.csv("maizeSNP50 sample name.csv")
# “ maizeSNP50 Full Data without heterozygous 01 A ”為以 1、 0 表示的基因型資料，“ maizeSNP50 sample name “ 為試驗品系的基本資料，同表一。
y=as.matrix(RawA[,-c(1:4)])
X=t(y)
OTU=names(RawA[-c(1:4)])
#將資料除去多餘欄位，僅留下基因型資訊，以 names 指令取出每行的名稱。
D <- dist(X, method="euclidean")/sqrt(2*length(RawA$Name))
#以 dist() 函數內建之 Euclidean 方法計算距離，除以 2 倍位點數的平方根轉換成 Modified Rogers' distance (MRD)。使距離的表示數值侷限在 0 到 1 之間，以代表遺傳距離。
mds= cmdscale(D,k=4,eig=TRUE)
eig= mds$eig/sum(mds$eig[mds$eig>0])*100
#利用 cmdcale 函數可計算距離資料的特徵值，並計算個別特徵值佔總特徵值的百分比。
plot(c(1:length(eig)),eig,xlab="principal coordinate",ylab="explained
variance[%]")
#繪製各特徵值所能解釋變異的百分比。
plot(mds$points[,1],mds$points[,2],asp=1,pch=16,xlab=paste("principal
coordinate 1(",format(eig[1],digits=3),"%)",sep=""),ylab=paste("principal
```


doi:10.6342/NTU201700716

```
coordinate 2(",format(eig[2],digits=3),"%)",sep=""))
text(mds$points[,1],mds$points[,2],OTU,pos=3,cex=0.67)
#以第一特徵值與第二特徵值繪製主座標分析圖。
col=c()
col[Name$Adaptation=="Africa MA/ST"]="gold"
col[Name$Adaptation=="Asia Lowland"]="red"
col[Name$Adaptation=="Lowland"]="brown"
col[Name$Adaptation=="Lowland- Latin Am. tropics"]="orange"
col[Name$Adaptation=="Subtropical"]="yellow green"
col[Name$Adaptation=="Asia Subtropical"]="green"
col[Name$Adaptation=="Temperate"]="blue"
col[Name$Adaptation=="Asia Temperate"]="purple"
col[Name$Adaptation=="unknown"]="black"
#依不同適應區給定代表顏色。
plot(mds$points[,1],mds$points[,2],asp=1,pch=16,col=col,xlab=paste("principal
coordinate 1(",format(eig[1],digits=3),"%)",sep=""),
ylab=paste("principal coordinate 2(",format(eig[2],digits=3),"%)",sep=""))
legend("topright",pch=16,col=c("red","brown","orange","gold","yellow
green","green","blue","purple","black"),legend=c("Asia
Lowland","Lowland","Lowland- Latin Am. tropics","Africa
MA/ST","Subtropical","Asia Subtropical","Temperate","Asia
Temperate","Unknown"))
```


doi:10.6342/NTU201700716

##轉換資料格式以符合 STRUCTURE 軟體需求

```
raw=read.csv("maizeSNP50 Full Data without heterozygous ATCG.csv")
for (i in c(6:53)){
    raw[,i]=gsub("AA",replacement="1",raw[,i])
    raw[,i]=gsub("TT",replacement="2",raw[,i])
    raw[,i]=gsub("CC",replacement="3",raw[,i])
    raw[,i]=gsub("GG",replacement="4",raw[,i])
    }
raw=raw[,-c(1,2,4,5)]
write.table(raw, "maizeSNP50 Full Data without heterozygous 1234.csv", sep=",")
```

#依照 STRUCTURE 軟體的說明文件“ Documentation for structure software: Version 2.3 ”，基因型資料須以數字表示，並且以對偶基因為單位。將 A、T、 C、 G 分別以 1、 2、 3、 4 取代

```
raw1=read.csv("maizeSNP50 Full Data without heterozygous 1234.csv")
raw2=read.csv("maizeSNP50 Full Data without heterozygous 1234 2.csv")
```

#玉米為雙倍體植物，基因型資料也已經刪減成僅有同質結合個體。故將原本單行的基因型資料複製成兩行，“ maizeSNP50 Full Data without heterozygous 1234 ”與“ maizeSNP50 Full Data without heterozygous 1234 2 ”資料差異為後者的樣品名後方加上“ .12 “。

```
data=cbind(raw1,raw2[,-c(1)])
x=cbind.data.frame(data$Name, data$B73, data$B73.12, data$B73.meth4,
data$B73.meth4.12,……..)
```

#合併資料並將同樣品資料接鄰排列，並除去 raw2 資料第一行的分子標記名稱資料。

doi:10.6342/NTU201700716

```
data=t(x)
#將矩陣轉置以符合 STRUCTURE 軟體資料讀取方向。
write.table(data, "maizeSNP50 without h for structure1234.txt", sep="\t")
#此檔案需在純文字文件中除去存檔後多出的一排編號資料、 " Name "和樣品名後多增加的 " .12 "，即可以 STRCTURE 軟體讀取。
```


doi:10.6342/NTU201700716

附錄程式碼三、區分族群內特有之分子標記

##篩選族群特有分子標記（以 CML 族群為例）

```
Raw=read.csv("maizeSNP50 Full Data without heterozygous ATCG.csv")
test=Raw
test=subset(test,test$CML426==test$CML470)
testcml=subset(test,test$CML470==test$CML475)
#將原資料經由“subset()”指令配合“==”的判斷式，可以篩選出欄位間讀值相同的分子標記。（CML426 、CML470 、CML475 為 CML 族群的代表品系）
test=subset(testB73,testB73$Name%in%testsu$Name)
#以“subset()”指令配合“%in%”的判斷式與分子標記的名稱，篩選出在另外兩個族群內皆不具多型性的分子標記。（“testcml”、“testsu”、“testB73”分別為在個別三個族群內不具多型性的分子標記資料）
test=subset(test,test$B73==test$Su963106)
#以“==”指令確定選出的分子標記，在另外兩個族群中的讀值相同。
x=setdiff(test$Name,testcml$Name)
#利用“setdiff()”指令可將所有分子標記資料與 CML 族群內不具多型性的資料做差集，可選出在 CML 族群內具多型性的分子標記。
data1=subset(test,test$Name%in%x)
#篩選出 CML 特有多型性的分子標記。
write.csv(data1,file="CML separate.csv")
```

##篩選非族群特有之 SNP。

```
test=Raw
ddata=rbind(data1,data2,data3)
```


doi:10.6342/NTU201700716

```
x=setdiff(test$Name,ddata$Name)
Data=subset(test,test$Name%in%x)
write.csv(Data,file="cluster without separate.csv")
```

##繪製族群特有分子標記的分布（以第一號染色體為例）

```
data1=read.csv("B73 separate.csv",head=T)
data2=read.csv("Su separate.csv",head=T)
data3=read.csv("CML separate.csv",head=T)
data4=read.csv("cluster without separate.csv",head=T)
#輸入各族群特有之分子標記與非特有分子標記的資料檔。
gray="gray"
gray<- adjustcolor(gray, alpha.f = 1)
red="red"
red=adjustcolor(red, alpha.f = 0.8)
blue="blue"
blue=adjustcolor(blue, alpha.f = 0.7)
green="green"
green=adjustcolor(green, alpha.f = 0.6)
#以“adjustcolor()”指令調整不同顏色的透明度。
achr=subset(data3,data3$Chr=="1")
achr=subset(achr,achr$Position!="0")
bchr=subset(data2,data2$Chr=="1")
bchr=subset(bchr,bchr$Position!="0")
cchr=subset(data1,data1$Chr=="1")
cchr=subset(cchr,cchr$Position!="0")
```


doi:10.6342/NTU201700716

```
dchr=subset(data4,data4$Chr=="1")
dchr=subset(dchr,dchr$Position!="0")
#每筆資料以“Chr==″1″”指定第一條染色體的資料，並以“Position!=″0″”
除去未知物理位置點的分子標記。
p=max(achr$Position)
#以物理位置最大的分子標記視為染色體長度。
k=ceiling(p/5000000)
#將總長度除以 5,000,000，並以“ceiling()”指令四捨五入到整數，決定直方
圖的組數。
hist(dchr$Position/1000000,breaks=k,main="Chr1",col=gray,
xlim=c(0,300),ylim=c(0,80),xlab="position(mb)",ylab="count")
#繪製直方圖，橫軸為物理位置，縱軸為組內分子標記的計數。物理位置除
以 1,000,000 以轉換成 Mb 表示。
par(new=TRUE)
#利用此指令進行疊圖，並依序繪圖。
hist(bchr$Position/1000000,breaks=k,main="Chr1",col=blue,
xlim=c(0,300),ylim=c(0,80),xlab="position(mb)",ylab="count")
par(new=TRUE)
hist(achr$Position/1000000,breaks=k,main="Chr1",col=red,
xlim=c(0,300),ylim=c(0,80),xlab="position(mb)",ylab="count")
par(new=TRUE)
hist(cchr$Position/1000000,breaks=k,main="Chr1",col=green,
xlim=c(0,300),ylim=c(0,80),xlab="position(mb)",ylab="count")
```


doi:10.6342/NTU201700716

附錄程式碼四、連鎖失衡衰退曲線

##兩兩位點間連鎖失衡資料檔整理

```
D=read.csv("test window size 1.csv",head=T)
# test window size 1.csv 為 22,227 個同質結合的分子標記，經 TASSEL 計算相鄰兩位點間連鎖失衡關係所輸出之 csv 檔。
D$Dist_bp=as.numeric(as.character(D$Dist_bp))
#將資料架構由 factor 轉為 number，如此能使用邏輯句。Dist_bp 欄位表示兩位點間距離
D=subset(D,D$R.2>=0)
#篩選 R^2 值大於等於零。（去除缺值）
D=subset(D,D$Dist_bp>0)
#除去重疊位置點所計算出的資料。
D=subset(D,D$Locus1!=0)
D=subset(D,D$Locus2!=0)
#Locus 欄位表示所在之染色體，若為 0 表示未知位於何染色體上。
D=subset(D,D$Position1!=0)
D=subset(D,D$Position2!=0)
#Position 欄位表示在染色體上所在位置，若為 0 表示未知位置。
write.table(D, "maize LD window size 1.csv",sep=',')
```

##連鎖失衡衰退曲線計算與繪製

```
rm(list=ls())
setwd(choose.dir())
data=read.csv("maize LD window size 1.csv")
distance=c(data$Dist_bp)
```


doi:10.6342/NTU201700716

```
LD.data=c(data$R.2)
n<-48
MD.st<-c(C=0.00001)
#給定非線性最小平方法計算的起始點。
MD.nonlinear<-
nls(LD.data~((10+C*distance)/((2+C*distance)*(11+C*distance)))*(1+((3+C*dist
ance)*(12+12*C*distance+(C*distance)^2))/(n*(2+C*distance)*(11+C*distance)))
,start=MD.st,control=nls.control(maxiter=100))
new<-summary(MD.nonlinear)
new.rho<-new$parameters[1]
#將非線性最小平方法計算出的 C 值取出。
fpoints<-
((10+new.rho*distance)/((2+new.rho*distance)*(11+new.rho*distance)))*(1+((3+n
ew.rho*distance)*(12+12*new.rho*distance+(new.rho*distance)^2))/(n*(2+new.rh
o*distance)*(11+new.rho*distance)))
#將計算出的 C 值代入公式，計算每個點的 R^2 期望值。
ld.df<-data.frame(distance,fpoints)
ld.df<-ld.df[order(ld.df$distance),]
#將資料依位點間距離大小，由小而大排序。
plot(ld.df$distance/1000,ld.df$fpoints,lwd=3,type="l")
#繪製 LD decay 曲線。
```


doi:10.6342/NTU201700716

附錄程式碼五、特有分子標記假說驗證

##篩選造成連鎖失衡區塊圖像不完整之分子標記

```
test=read.csv("maize LD window size 2.csv")
```

#" maize LD window size 2.csv "為以 window size = 2 計算之相鄰三個分子標記間的 R^2 值。

```
k=test[(1),]
```

#以資料的第一排作為迴圈的起始，於迴圈計算完後刪除。

```
for (i in c(1:27539)) {
if(test[(i+1),]$Site1==test[(i),]$Site1&&test[(i+1),]$R.2<0.8&&test[(i),]$R.2>0.8)
```

#迴圈內主要計算式，以 if() 函數建立判斷式，並以 && 連結三個條件：第一個條件為兩列資料具有相同的 site1，表示兩列資料中其一分子標記相同。第二個條件為查看兩列資料的 R.2 ，取出相鄰分子標記間 $R^2 < 0.8$。第三個條件為間距一個分子標記間 $R^2 > 0.8$。

```
k<-rbind(k,test[(i+1),])
```

#將符合條件的為資料與 K 合併。

```
}
k=k[-1,]
```

#將迴圈前創建的起始資料刪除。

```
T=k$Position2
length(unique(T))
```

#取出分子標記資料的物理位置，確定沒有位置重疊的位點。

```
write.table(k, "LD block origin ver.csv",sep=',')
```


doi:10.6342/NTU201700716

##族群特有標記與隨機抽取標記中屬於連鎖失衡區塊斷點的比例差異

#由於 TASSEL 計算連鎖失衡所輸出的資料檔不包含分子標記名稱與樣本基因型值，故篩選出之造成連鎖失衡區塊圖像不完整之分子標記資料檔，必須以物理距離為標準，於原完整分子標記資料檔中取出分子標記名稱。

A=read.csv ("maizeSNP50 Full Data Table 同質結合 for tassel.csv")

#讀取原分子標記資料檔，共 22,227 個位點，包含分子標記的名稱、物理位置、對偶基因類型及樣本基因型等資訊。

data=read.csv("LD block origin ver.csv")

X=subset(A,A$pos%in%data$Position2)

#由於" LD block origin ver.csv "資料中並無物理位置上的重疊，因此可由物理圖譜的位置作為篩選標準，由原分子標記資料檔中獲得位點的完整資訊。

q=read.csv("cluster specific marker.csv")

#由 STRUCTURE 分群選出之族群特有分子標記。

q=subset(q,q$Chr!=0)

q=subset(q,q$Position!=0)

#去除未知物理圖譜位置的位點。

y=c()

for(i in c(1:10000)){

a=sample(A$rs., length(q$Name),replace = FALSE)

#利用 sample 指令對原分子標記資料檔，以分子標記的名稱做隨機選取，選取個數為族群特有分子標記資料的分子標記個數。(已抽取之分子標記不放回)

y[i]=length(which(a%in%X$rs.))

#將取出之分子標記比對造成連鎖失衡區塊圖像不完整之分子標記，並記錄“成功”的個數。

doi:10.6342/NTU201700716

```
}
hist(y,breaks=100)
#將 10,000 次重複取樣的資料繪製直方圖。
quantile(y,0.95)
#取出分布中第 95% 位置的成功個數。
Q=length(which(q$Name%in%X$rs.))
#以族群特有分子標記資料比對造成連鎖失衡區塊圖像不完整之分子標記，
並記錄“成功”的個數。
pnorm(Q,mean(y),sd(y))
#將此二項分布視為常態分布，並計算 Q 在常態分布上的位置。
```


doi:10.6342/NTU201700716