國　立　交　通　大　學

生物資訊及系統生物研究所

碩士論文

使用智慧型三目標基因演算法選取標籤單核苷酸多型性

Selecting Tag SNPs Using an Intelligent Triobjective Genetic Algorithm

研 究 生：林玉祥

指導教授：何信瑩　教授

中 華 民 國 一 百 年 七 月

使用智慧型三目標基因演算法選取標籤單核苷酸多型性

Selecting Tag SNPs Using an Intelligent Triobjective Genetic Algorithm

研 究 生：林玉祥　　Student：Yu-Hsiang Lin

指導教授：何信瑩　　Advisor：Shinn-Ying Ho

國 立 交 通 大 學

生物資訊及系統生物研究所

碩 士 論 文

A Thesis Submitted to Institute of Bioinformatics and
Systems Biology Department of Biological Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of Master in
Bioinformatics

July 2011

Hsinchu, Taiwan, Republic of China

中 華 民 國 一 百 年 七 月

使用智慧型三目標基因演算法選取標籤單核苷酸多型性

學生：林玉祥　　指導教授：何信瑩

國立交通大學生物資訊及系統生物研究所碩士

摘　要

人類的 DNA 序列中包含各種遺傳變異性，其中最常被發現的遺傳變異為單核苷酸多型性(Single Nucleotide Polymorphism； SNP)。由多個 SNPs 所組成之集合稱為 Haplotype(Haplotype)。國際單型圖譜計畫(International HapMap Project)運用高通量微陣列晶片技術完整解析人類各大族群的 Haplotype 圖譜，並揭露了 SNP 之間有程度不一的連鎖不平衡(linkage disequilibrium)，因此只需鑑定一群具代表性的 SNPs，即足以偵測出一大片段的 Haplotype 資訊。此代表性的 SNPs 稱為標籤單核苷酸多型性(Tag SNPs)。

Tag SNPs 選取問題已經被證明為一 NP-hard 的問題。為了有效解決此問題，本研究提出一套使用三目標最佳化的研究方法來解決 Tag SNPs 選取問題，以選出可辨識的所有 Haplotype 序列樣式之最小 SNPs 部分集合。過去研究多著重於改善求解方法之效率，而未評估所求之最佳解與其它最佳解之間的資訊差異。因此，為了達到選取出 Tag SNPs 最少、Haplotype 相異性最大，且 Haplotype 多樣性最小的三個目標，透過特徵選取(feature selection)的觀念結合智慧型三目標基因演算法(ITOGA)達到最佳化多目標的目的。

本論文使用 ITOGA 具有搜尋全域最佳解與大量參數最佳化之能力，可快速求得多組不受支配的最佳解的方法，並應用到 Tag SNPs 選取問題。其目標特色有(1)選取 Tag SNPs 最小化 (2)Haplotype 相異性最大化(3)Haplotype 多樣性最小化。本文所提出的方法可以確實選擇少量且具影響力的 Tag SNPs，並提高 SNPs 的覆蓋率，避免重複選取相似性較高的 SNPs。另將 ITOGA 的效能和現有多目標演算法 NSGA-2 之效能做比較，以驗證所提的方法具較高效能。

實驗應用於維生素 D 受體基因(Vitamin D receptor； VDR)選取 Tag SNPs，在 977 SNPs 中，文獻記載重要 50 個 SNPs，計算結果顯示，由 ITOGA 選出 18 組重要的 Tag SNPs 最佳解集合。本文所提出的三目標能更精確的找到最佳解。

關鍵字：標籤單核苷酸多型性、單型、連鎖不平衡、多目標基因演算法

Selecting Tag SNPs Using an Intelligent Triobjective Genetic Algorithm

Student：Yu-Hsiang Lin Advisor：Dr. Shinn-Ying Ho

Institute of Bioinformatics and Systems Biology
National Chiao Tung University

ABSTRACT

Human DNA sequence contains several kinds of genetic variation and the single nucleotide polymorphism (SNP) that was found in highest frequency. Haplotype consists of a collection of the SNPs. International HapMap Project used high-throughput microarray chip technology to completely analyze human haplotype map of the major ethnic groups, and revealed varied degrees of linkage disequilibrium between SNPs. Therefore, to identify a representative group of SNPs just to detect a large fragment of Haplotype information. The representation of the SNP is called Tag SNPs.

The problem of selecting Tag SNPs was proved to be a NP-hard problem, so heuristic methods may be useful to effectively solve this problem. This study proposes a Triobjective optimization to solve the problem of selecting Tag SNPs and to identify all the minimum pattern of SNP Haplotype set. The past researches put emphasis on improving the efficiency of solution finding method, but didn't evaluate the distinctions among optimal solutions. In order to obtain the three goals which are minimizing the total amount of Tag SNPs, the dissimilarity of Haplotype and the diversity of Haplotype, we combined the feature selection method with Intelligent Triobjective Genetic Algorithm (ITOGA) to achieve the purpose of multi-objective optimization.

This study use ITOGA to search the global optimal solution, to optimize a large number of parameters and has the ability to obtain the multiple non-dominated optimal solutions quickly. Moreover we apply it in the problem of selecting Tag SNPs. The characters of ITOGA are 1) minimizing the total amount of Tag SNPs, 2) maximizing the dissimilarity of Haplotype, and 3) minimizing the diversity of Haplotype. The proposed method can indeed choose small amount but influential Tag SNPs, It also can improve SNPs coverage rate and avoid choosing the duplications of SNPs. Moreover, compare the performances of ITOGA with existing NSGA-2 to verify the efficiency of the proposed method.

The experiment was applied to select the vitamin D receptor gene Tag SNPs. In 977 SNPs, 50 important SNPs were reported in prior references. The results show that 18 SNPs selected by the ITOGA were the most significant group of Tag SNPs. The three goals presented in this paper can find optimal solution more accurately.

誌　　謝

在研究所兩年半中的過程中，首先我要感謝我的指導老師何信瑩教授，每次與老師的會議中，您總是孜孜不倦的指導學生該如何做好研究，鼓勵我們從不同的角度不同的觀點去判斷問題，解析問題，並期待我們最終能夠獨立的解決問題。此外，除了課業以外，在您身上最大的收穫，就是學會如何自我的銷售。老師不僅重視學生的方法及做事態度，而還常常藉由身邊日常生活發生的事情來，教導我們除了研究外做人做事的應有的態度，使我們懂得要如何更圓融地與其他人相處，真的很感覺老師全方位的訓練學生，使得我們更有信心面對未來的挑戰。

口試期間，感謝，黃憲達教授與黃慧玲教授對於學生論文給予寶貴的建議及指導，提出論文的缺失、不足與需要改進的方向，讓學生的論文更具有內涵與價值。

接著，我還要感謝智慧型計算實驗室(ICLab)的同伴們，感謝慧玲老師在研究上的建言及協助，感謝前實驗室宏銘學長技術協助，感謝義雄學長在本論文上的建議及經驗傳承、還有瓊慧學姊、佳達學長、凱迪在研究的過程中不吝嗇的指教，感謝國慶、明儒、銘鑫(Morris)、韻如、冠維(小羊)、馨云們在我即將畢業的前夕舉辦的畢業旅行，感謝我的戰友泰欽、德芬(小黑)在研究的過程中彼此的鼓勵協助跟討論。謝謝你們塑造出實驗室既熱鬧有不失學術研究氣息的環境，讓我在這兩年多一點點的時間裡很開心地做研究。

最後，最感謝的就是我的父母、弟弟與佳誼，謝謝爸媽在我一路求學的過程中不斷的支持鼓勵，給我一切所需要的幫助，不論是金錢上實質的資助或是精神上的安慰鼓勵，讓我更有勇氣邁步朝理想前進。

目錄

表目錄

圖目錄

一、緒論

1.1 研究背景

人類基因體計畫(Human Genome Project)在 2000 年完成，基因體分析的相關研究剛開始發展的時候，是以如何儲存 DNA、RNA 及蛋白質序列等各式各樣的生物資料庫為當時生物資訊學的發展重點；其後，研究焦點即移轉至如何由資料庫中尋找出有用的資訊並藉由分析及解釋核苷酸序列、蛋白質序列等資訊，以找出序列中影響疾病發生的不正常變異所在。近年來許多學者致力於探究基因體與序列中 SNP(Single Nucleotide Polymorphism；SNP)對人類疾病的發生與特徵性向的表現之影響 [1, 2]。相關文獻中顯示 SNP 是影響疾病與特徵發生的主因，因此本研究將針對 SNP 作探討，為方便闡述本論文及研究主題，以下我們將先針對一些相關的分子生物等專有名詞加以介紹。

1.1.1 DNA、變異(Variation)、突變(Mutation)

去氧核醣核酸(Deoxyribo Nucleic Acid；DNA)是一種呈現雙股螺旋結構的遺傳因子，存在於各種生物細胞的染色體上，其主要功能在於影響生物體的發育及生命機能之運作。這些帶有遺傳訊息的DNA片段稱為基因(Gene)，而生物細胞中所有的遺傳訊息通稱為基因組(Genotype)。DNA的基本單位由四種不同的核苷酸鹼基(locus)組成，分別為腺嘌呤(adenine；A)、胸腺嘧啶(thymine；T)、胞嘧啶(cytosine；C)、鳥糞嘌呤(guanine；G)。這些鹼基互補性的配對，且兩種不同的鹼基對將以不同的氫鍵數目結合。在一般情況下A＝T由兩條氫鍵相連，而C≡G由三條氫鍵相連。由於A與T間的氫鍵數較少，較易受外來因素打斷而發生變異，例如：以T取代C的變異約佔所有生物體變異發生的三分之二。

人類的 DNA 序列約有 99%是一致的，但個體間特徵的差異仍很大，這些差異是由於在生物群體的演化過程中，部分生物體細胞之細胞核內的部份基因發生變異(genetic variation)所致，譬如 DNA 的構造或排列方式發生改變等。倘若這些變異使生物性狀改變，則稱之為突變(mutation)。一般而言，突變通常對生物體是有害的，但若突變的發生使得生物個體更能適應或反應環境的變遷，則會透過「適者生存、不適者淘汰」的篩選過程中被保留下來，使得同物種間有了新的表現個體，經由群體的交配、繁衍造成群體內各個個體之間有所差異，這些差異稱之為生物變異(variation)。簡單來說，突變僅是生物體產生變異的一小部份，在同物種中，發生突變的機率很低，但產生變異的機率卻很高。

1.1.2 單核苷酸多型性(SNP)、單型(Haplotype)

在正常細胞上，染色體(chromosome)以成對(一條來自於父親，另一條來自於母親)的方式存在，因此稱為雙套染色體。染色體由DNA序列構成，不同個體的DNA序列各有不同。在DNA序列上單一核苷酸鹼基對(base pair)發生的變異(亦即鹼基上A、T、C、G的改變)，是一種常見的遺傳變異，稱之為SNP(Single Nucleotide Polymorphism；SNP)，而此鹼基對於染色體上之對應位置稱為基因座(gene locus)。SNP的發生率頻繁，至少佔總DNA序列的0.1%，目前已發現的SNP總共約有400萬個左右，由此可知，處理如此龐大的SNP資料量將消耗不少成本。

在同一族群之中不同個體間存在著相當多的變異，其中SNP所引起的基因變異約佔人類遺傳基因所有變異中的90%。這些變異影響個性性狀的差異，造成生物的多樣性，且物種越接近的，其差異程度越小。舉例來說，不同人種間之SNP的數量及分布較不一致，因此可能造成某些人種或群體較易產生某類疾病；譬如地中海貧血較 易發生在地中海、中東、印度洋及南中國海沿岸一帶之族群，而鐮刀型貧血則普遍存 在非洲黑人部落、印度土著等。

SNP 在基因體的位置與對蛋白質的影響程度，區分成以下不同類型 [3]：

(1) Non-coding SNP：意指在非基因編碼區域之 SNP。

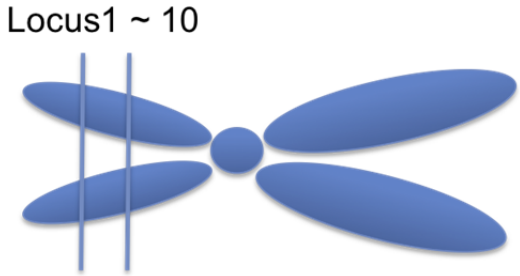
(2) Coding SNP：意指位於基因編碼區域內之 SNP。但在這區域內之突變，並不一定會影響胺基酸型態。所以根據對胺基酸合成的影響程度可再進一步區分為以下兩種型態。

   (a) 同義 SNP (Synonymous SNP)：此種 SNP 上不同之核苷酸，並不會改變胺基酸型態。例：CUU 和 CUC，第三個位置雖然有 SNP，但其合成出的胺基酸依然是 Leucine。

   (b) 非同義 SNP (Non-synonymous SNP)：此種 SNP 上不同之核苷酸，會造成胺基酸的改變，進而可能造成蛋白質結構或是功能之改變。例：CUU 和 CUG，所造成的胺基酸就會不同。若進一步根據胺基酸改變的幅度區分，通常稱單純只造成胺基酸改變之突變型態為錯義(missense)。若會造成 stop codon，使得整條序列無法繼續轉譯成蛋白質之突變，則稱為無義突變(nonsense mutation)。

單套染色體上相鄰的 SNP 鹼基所組成之序列稱之為基因組單型(Haplotype)，以將圖 1 (a)為例，其雙套染色體可拆成兩條單套染色體，假若擷取四個不同個體其在 chromosome 1a 中的 locus 1 到 locus 10，我們可得到圖 1 (b)中的四條 DNA 序列。從圖 1 中可觀察出 locus 2、4、9 在不同個體上之鹼基各有不同，因此該 DNA 序列具有 3 個 SNP，而這些 SNP 可形成一組

Haplotype 序列。舉例來說，圖 1 在右方所顯示的四個 Haplotype 序列分別為{AAA}、{TAC}、{ACA}及{AAA}。



(a)、雙套染色體。

(b)、單套染色體

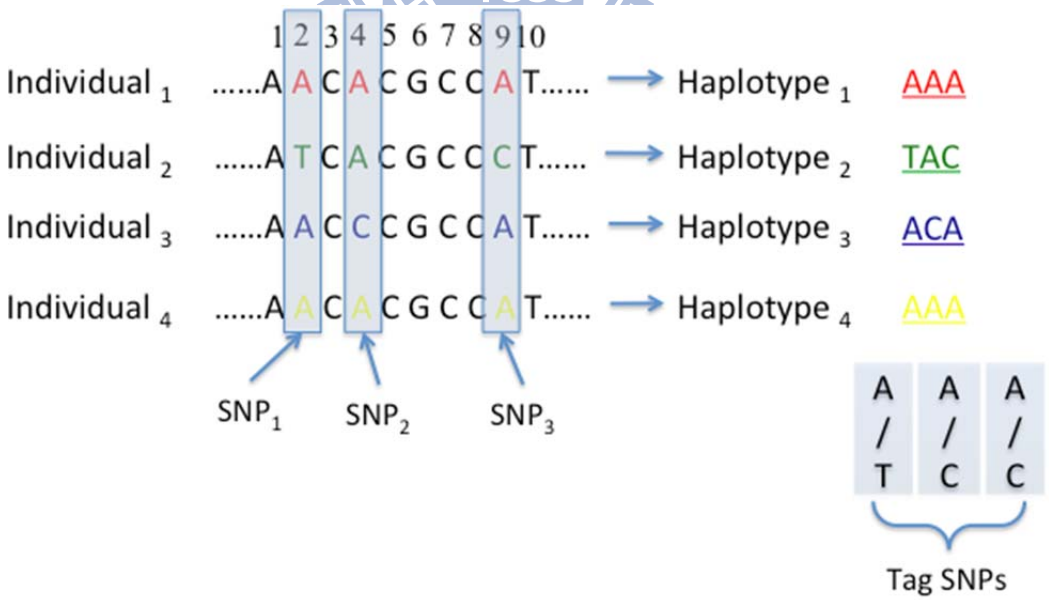


圖 1 SNP、Haplotype 與 Tag SNPs 之說明

1.1.3 連鎖不平衡(Linkage disequilibrium)

連鎖不平衡 [4, 5] 是一種描述不同 SNP，其核苷酸之間的關聯性，會因受到染色體重組而有所影響。如圖 2 所示，SNP_1 與 SNP_2 稱為高連鎖不平衡，即幾乎沒有染色體重組發生。反之，SNP_1 與 SNP_6 則稱為低連鎖不平衡。

文獻中提到，通常距離比較近的 SNP，其中間比較不容易有染色體重組發生，所以有比較高的連鎖不平衡現象。反之，距離較遠的 SNP，其中間比較容易發生染色體重組，所以會有較低的連鎖不平衡現象。

Individual	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	SNP_6
1	C	G	C	C	G	C
2	A	T	A	C	T	T
3	A	T	A	A	G	T
4	A	T	A	C	T	C
5	A	T	A	A	G	C

圖 2 連鎖不平衡示意圖

1.2 研究動機

SNP 是由於人體基因中單一核苷酸改變所造成的，並且被視為人體基因多樣性的主要原因之一。這些單一的核苷酸變異大約每一千個鹼基對就會發生一次。在這些核苷酸位置上通常只會有二種可能的核苷酸會顯現出來，由於 SNP 序列資料的變異有限，加上其資料量十分的豐富，因此很適合拿來當做人類疾病特徵的標誌。一般進行疾病基因之多型性變化分析時會將整個染色體的 SNP 進行 Genotyping 的動作，但是通常這些 SNP 的數量都相當的龐大因此在進行 Genotyping 的過程中必然會花費大量的金錢和時間；因此，我們從所有和疾病相關聯的 SNP 中挑選出部份的 SNP 子集合，並且將這個子集合所蘊含的資訊量能夠與原本 SNPs 集合所包含的資訊量之間的誤差達到最小，而挑選 SNP 的過程則稱作選取 Tag SNPs(Tag SNP selection) [6, 7]。

為了能夠達到獲得少量且蘊含資訊量最大的 Tag SNPs，我們透過特徵選取的觀念提出結合智慧型三目標基因演算法，來挑選最好的 Tag SNPs。此智慧型三目標基因驗算法加入了三個目標特色(1)選取 Tag SNPs 最小化。(2)Haplotype 相異性最大化。(3)Haplotype 多樣性最小化 [8]。本論文將由國際單型圖譜計畫(The International HapMap Project) [9] 所取得的資料集進行測試，並且以重建 Haplotype 的正確率來證明利用智慧型三目標演算法確實能夠達到挑選出較好的 Tag SNPs。

1.3 研究目的及題目定義

本論文研究主要有兩個目的，第一目的使用雙目標及三目標基因演算法挑選出 Tag SNPs，將 Tag SNPs 定義為可辨識所有不同的 Haplotype 樣式的 SNP；此 Tag SNPs 選取問題只在由給定的 SNP 資料中選取最少個數的 SNP 部分集合，以便釋出全部的 Haplotype 樣式，並應用 Huang *et al.* [8] 所提出的四個目標特徵，以雙目標及三目標的方式選出最佳的 Tag SNPs，並證明使用智慧型多目標基因演算法(IMOGA) [10] 選取的 Tag SNPs 性優於非支配排序基因演算法 2(NSGA2)[11]。

第二個目的提出智慧型三目標基因演算法(ITOGA)，使用三個重要目標特徵，選取維生素 D 受體基因 [12-14] 重要的 Tag SNPs，並討論維生素 D 受體基因中挑選出 Tag SNPs 重要性與關係。

1.4 章節概要

本篇論文第一章主要介紹此篇研究的動機與目的，和之前與此論文主題相關的研究。第二章文獻回顧在探討選取 Tag SNPs 之問題，並以四個目標來求解 Tag SNPs 選取問題；目標一、最小化強固型 Tag SNPs，目標二、Tag SNPs 選擇的容錯(Tolerance)，目標三、Haplotype 的相異性計算及目標四、Haplotype 的多樣性計算。第三章為現有方法介紹，在敘述先前已經存在並被使用在本研究中的的一些研究方法，如智慧型多目標基因演算法及演算法流程等。第四章為本研究解決 Tag SNPs 的選擇問題，以本研究使用的方法 IMOGA 以雙目標的方式與 NSGA2 比較所得出的解集。提出三目標基因演算法找出最佳 Tag SNPs 解集合。此章節為本論文之重點部分，一開始即介紹此篇研究所使用的資料來源取得，而後則說明實驗中之各個方法的應用與描述了本研究中所提出來解決問題的新方法。第五章為實驗應用，除了第四章所討論的 IMOGA 與 NSGA2 演算法比較外，本章節將提出使用三目標基因演算法在維生素 D 受體基因找出準確的 Tag SNPs，並進一步分析選出的 Tag SNPs 在維生素 D 受體基因上有哪些生生物意義。第六章為本篇論文的最後一個章節，問題討論與位來展望的探討。

二、相關研究回顧

本章節將介紹相關研究文獻回顧，討論選取 Tag SNPs 之問題，及以多目標選取 Tag SNPs 之問題，並將多目標選取 Tag SNPs 之問題轉換成數學模式，分別用四個目標作為評選準則 [8]，目標一：選取 Tag SNPs 最小化，目標二：Tag SNPs 容錯率最大化，目標三：Haplotype 變異性最大化；目標四：Haplotype 多樣性最小化，再個別以雙目標及三目標方式求得最佳解。

2.1 文獻回顧

目前已可從過去的研究得知Bafna *et al.* [15] 在沒有切割Haplotype區塊(Block)的情況下，以極大化可擷取的Haplotype資訊為目標來求取Tag SNPs；相同情況下，Ke and Cardon [16] 則以最小化Tag SNPs數為目標切入。而在考量切割Block情況下，Patil *et al.* [17]、Zhang *et al.* [18] 以限制Diversity的方式切割Block，在每一Block中求取最小化Tag SNPs之數目。

另一種 Tag SNPs 選取問題則是先限制 Tag SNPs 數目 [19]，再依此限制求取 Block 的切割方式以涵蓋最大範圍的 Haplotype 序列；Zhang *et al.* [20]、Weale *et al.* [21] 等學者則在給定 Tag SNPs 數目下，求解最大化 Haplotype 資訊的 Tag SNPs。除此之外，部分研究者將重點擺在使用 Tag SNPs 預測整體 Haplotype 序列，如 Halperin *et al.* [22] 在給定 Tag SNPs 數目下，選取 Tag SNPs 以使預測 Haplotype 序列之期望誤差最少。在這些相關研究議題中，有的文獻將重點擺放在切割 Block 的方法上，有的將重點放於挑選 Tag SNPs 之方法上，有的則是兩者並重；然而這些文獻之最終目的皆在於縮小原 SNP 資料，使之應用於實務上可獲得較高之執行效率。

然而在Tag SNPs的求解方法上，根據不同Tag SNPs定義之各類Tag SNPs選取問題已有許多學者提出多種不同的演算方法。Avi-Itzhak *et al.* [23] 採用列舉的方式(simple numerical algorithm)求解Block size較小之Tag SNPs選取問題，該研究分別模擬選取45個非洲人與高加索人之6號、21號，以及22號染色體上的Tag SNPs，並提出修正之演算法以處理不完全辨識Haplotype差異之Tag SNPs選取問題，以及模擬兩種情況下(是否完全辨識Haplotype差異)的求解結果。在考慮完整辨識的情況下，非洲人可節省25%的SNP，高加索人則可節省36%之SNP；另外，若在允許遺漏10%的辨識資訊情況下，所需的SNP數更少，非洲人可節省 38%的SNP，高加索人則可節省全部的49%。

Patial *et al.* [17] 先使用貪婪(Greedy)演算法找出切割Block之分界線，使其滿足每一個 Block中至少有80%的Haplotype重複出現(該類Haplotype被定

義為Common Haplotype)；在此種Block條件下求取最少個數之Tag SNPs，此Tag SNPs可辨識每一個Block中80%的Haplotype樣式；其結果將第21號染色體中24,047個SNP切割成 4,135個Block，並從中選取出4563個Tag SNP；Zhang *et al.* [18] 採用Patial *et al.* [17] 的Block定義，用動態規劃(Dynamic Programming Algorithm)切割Block，並在其中求取最少個數之Tag SNPs；其結果將第21號染色體切割成2,575個Block並從中選 取出3,582個Tag SNPs。從上述的結果顯示兩種演算法的效率皆很高，且亦可從結果推論出，不同的Block切割方式也會影響到Tag SNPs的選取結果。

Huang *et al.* [24] 提出兩種Greedy演算法與一種反覆式線性鬆弛法以求解在考慮遺漏部分資訊的情況下的Tag SNPs選取問題。該研究模擬4種不同出處的資料， 比較其兩種Greedy演算法與反覆式線性鬆弛法的求解品質與效率。其結果顯示，反 覆式線性鬆弛法求解品質最佳，而另兩種Greedy演算法的求解品質與效率亦不錯，且在考慮遺漏部分資訊的情況下，第二種Greedy演算法又比第一種Greedy演算法有效率。

Carlson *et al.* [4]亦提出一個 Greedy 演算法，以極大化可獲得之資訊來選取 Tag SNPs。該研究模擬了 47 個獨立個體(24 個非裔美國人與 23 個歐洲人)的 100 個基因，總共有 8,877 個 SNP，於非裔美國人中選取出 3,178 個 SNP 作為 Tag SNPs；歐洲人則選取出 2,375 個 SNP 作為 Tag SNPs；雖然此篇研究主要重點在於闡述其結果數據的生物意義，而非著重於改良其演算法之求解品質與效率，然而若由其最終所選取 Tag SNPs 的縮減模看來，我們亦可推論出該演算法的求解品質其實並不差。

2.2 Haplotype 資料簡介

資料來源由國際單型圖譜計畫(International HapMap Project) [9] 所取得資料來源，因 HapMap 資料格式都是基因型，為了方便計算，將找到的資料轉成二進制格式，基因型格式定義，野生型純合子(homozygous wild type)為 0，突變型純合子(homozygous mutate)為 1，異型合子(heterozygous)為 2 [7] 。如圖 3，G_1 為異型合子，G_2 為同型合子。表 1 為國際 Haplotype 圖譜計畫資料。

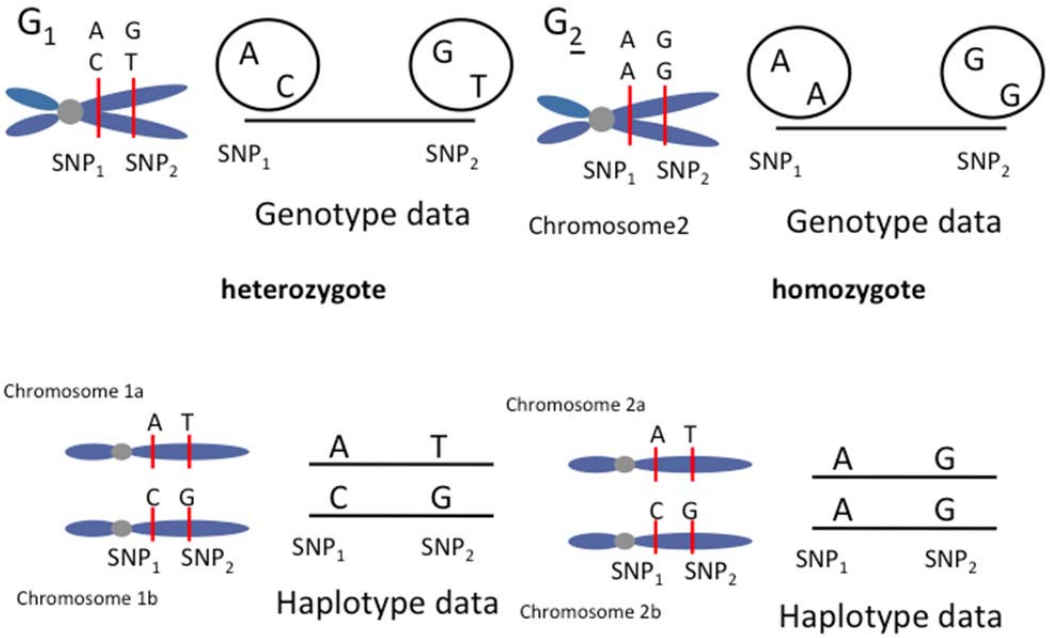


圖 3 SNP 其基因型同型合子與異型合子示意圖

表 1 HapMap 網站所取得的基因型資料

	$Inidvidual_{1a}$ **NA12718_A**	$Inidvidual_{1b}$ **NA12718_B**	$Inidvidual_{2a}$ **NA11843_A**	$Inidvidual_{2b}$ **NA11843_B**	$Inidvidual_{3a}$ **NA12383_A**	$Inidvidual_{3b}$ **NA12383_B**
SNP_1 rs4079417	C	C	C	C	C	C
SNP_2 rs11063263	A	A	A	A	A	A
SNP_3 rs16931854	T	T	T	T	T	T
SNP_4 rs4980929	T	T	T	T	T	T
SNP_5 rs16930883	C	C	C	C	C	C
SNP_6 rs11064561	G	G	G	G	A	G

表 2 所示基因型轉換成對應數值，野生型純合子為 0，突變型純合子為 1，異型合子為 2。將整理好的資料由透過 Haplotype inference(SDPHapInfer 與 PHASE 軟體 [7, 25])轉換成二進制編碼，如圖 4。

表 2 基因型轉換對應數值

	$Individual_1$	$Individual_2$	$Individual_3$
SNP_1 rs4079417	0	0	0
SNP_2 rs11063263	1	1	1
SNP_3 rs16931854	1	1	1
SNP_4 rs4980929	0	0	0
SNP_5 rs16930883	0	0	0
SNP_6 rs11064561	1	1	2

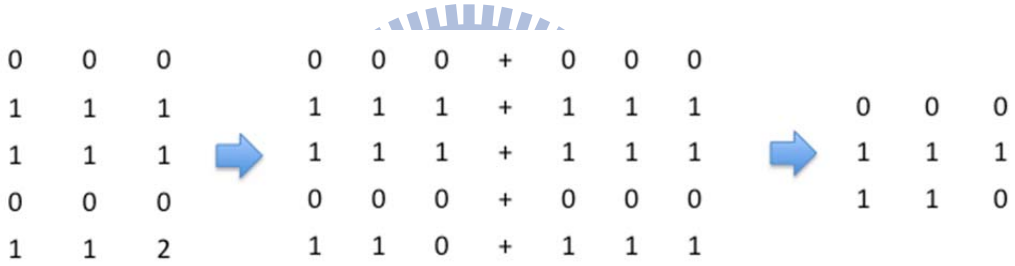


圖 4 基因型轉換二進制編碼

2.3 集合覆蓋問題(Set Covering Problem)

標籤核苷酸多型性的選取問題可視為一個集合覆蓋問題(Set Covering Problem；SCP)。由於集合覆蓋已被證明是 NP-hard complete 問題 [25]，因此可以推論 Tag SNPs 的選取問題亦為一個 NP-hard complete 問題，無法在多項式時間內求得最佳解；因此多種啟發式方法被提出來，常用的啟發式演算法如貪婪演算法 [26, 27]，此種演算法通常是在每一遞迴步驟中選取對目標式最有利之集合，或是選取某集合替換已被選取之集合中的子集合，使其可以改善最多目標值。Chvatal *et al.* [26] 演算法的通式，改變挑選的規則，不是由階度最大的集合開始挑選，而是改用一種隨機的概念大於最大階度值的某一比例開始挑選即可，除此之外，尚有其他演算法被提出，如 Hifi *et al.* [28] 以類神經網路法求解；Ghasem Mahdevar *et al.* [29] 則採用基因演算法求解；Huang *et al.* [24] 採用二次貪婪演算法求解；Huang *et al.* [8] 也提出以多目標基因演算法的方式來求解。本論文則採用智慧型三目標基因演算法實作求得

最佳解。

2.4 Tag SNPs 選取問題

Tag SNPs選取問題，由於DNA單點發生突變的機率很低，可以推論同一點發生兩次突變的可能性十分渺小，因此學者通常會假設單點上僅可能發生一次突變($0 \rightarrow 1$)，故我們可將Haplotype的SNP鹼基用0、1表示，標記為0代表主要等位基因(major allele)，標記為1代表次要等位基因(minor allele) [17]。一列Haplotype是由多個SNP鹼基組成之序列，因此m個Haplotype$\{h_i: i = 1, \dots, m\}$內含n個SNP$\{S_j: j = 1, \dots, n\}$可表示為一個$m \times n$的0、1資料矩陣，如圖5矩陣$H = [H_{i,j}]$所示。

為了比較Haplotype間之差別，我們可將任兩個Haplotype h_{i_1}與h_{i_2}配對成$E_{i_1 i_2} = (h_1, h_2)$，而令$E = \{E_{i_1,i_2}: i_1 = 1, \dots, m;\ i_2 = 1, \dots, m\}$代表所有可能的Haplotype配對所構成之集合。舉例來說，h_{i_1}與h_{i_2}可配對形成$E_{1,2}$；h_{i_1}與h_{i_3}可配對形成$E_{1,3}$。圖5有四個Haplotype，總共形成$C_2^4 = 6$種配對方式，因此集合$E = \{E_{1,2},\ E_{1,3},\ E_{1,4},\ E_{2,3},\ E_{2,4},\ E_{3,4}\}$。另外，針對每一個$S_j$及$E_{i_1,i_2}$可定義一個指標函數$I(E_{i_1,i_2}, S_j) \in \{0,1\}$，其中若$(h_{i_1,j}, h_{i_3,j}) \in \{(0,0),\ (1,1)\}$時，$S_j$將無法被用來辨識該組Haplotype配對$E_{i_1,i_2}$，因此$I(E_{i_1,i_2}, S_j) = 0$；相反地，假若$(h_{i_1,j}, h_{i_3,j}) \in \{(0,0),\ (1,1)\}$時，$S_j$可被用來辨識該組Haplotype配對$E_{i_1,i_2}$，而$I(E_{i_1,i_2}, S_j) = 1$。如圖2.1以$h_{i_1}$、$h_{i_2}$與$S_3$之鹼基交集位置為例，由於$(h_{1,3}, h_{2,3}) = (0,0)$，代表$S_3$不能被用來辨識$E_{1,2}$，因此$I(E_{1,2}, S_3) = 0$；同理，由於$(h_{3,3}, h_{4,3}) = (1,1)$，代表$S_3$不能被用來辨識$E_{3,4}$，因此$I(E_{3,4}, S_3) = 0$；反之，$(h_{2,3}, h_{3,3}) = (0,1)$，代表$S_3$可被用來辨識$E_{2,3}$，因此$I(E_{2,3}, S_3) = 1$。以此類推，可得$S_3$可否被用來辨識$E$中所有Haplotype配對的狀態，如圖5所示。

$$H = \begin{array}{c} \\ h_1 \\ h_2 \\ h_3 \\ h_4 \end{array} \begin{array}{c} \begin{array}{ccccc} S_1 & S_2 & S_3 & S_4 & S_5 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ h_1 \\ h_2 \\ h_3 \\ h_4 \end{array} \begin{array}{c} \begin{array}{c} S_3 \end{array} \\ \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \end{array} \quad E_{1,2},\ E_{2,3},\ E_{3,4},\ E_{1,3},\ E_{2,4},\ E_{1,4}$$

$$(h_{1,3},h_{2,3})=(0,0)\Rightarrow I(E_{1,2},S_3)=0$$
$$(h_{1,3},h_{3,3})=(0,1)\Rightarrow I(E_{1,3},S_3)=1$$
$$(h_{1,3},h_{4,3})=(0,1)\Rightarrow I(E_{1,4},S_3)=1$$
$$(h_{2,3},h_{3,3})=(0,1)\Rightarrow I(E_{2,3},S_3)=1$$
$$(h_{2,3},h_{4,3})=(0,1)\Rightarrow I(E_{2,4},S_3)=1$$
$$(h_{3,3},h_{4,3})=(1,1)\Rightarrow I(E_{3,4},S_3)=0$$

圖 5 SNP 及其所能辨識的 Haplotype 配對

針對所有的SNP(S_j)重複其與所有E_{i_1,i_2}的比對步驟，可建構一個Haplotype配對與SNP間的比對矩陣ES，該矩陣內的元素即為其所對應的$I(E_{i_1,i_2},S_j)$之值，如圖6之矩陣H經轉換後可得比對矩陣ES。

$$H=\begin{array}{c} \\ h_1\\ h_2\\ h_3\\ h_4\end{array}\begin{array}{c}\begin{array}{ccccc} S_1 & S_2 & S_3 & S_4 & S_5\end{array}\\ \begin{bmatrix}1&0&0&1&1\\0&1&0&1&0\\1&1&1&0&0\\1&1&1&1&1\end{bmatrix}\end{array}$$

$$ES=\begin{array}{c} \\ E_{1,2}\\ E_{1,3}\\ E_{1,4}\\ E_{2,3}\\ E_{2,4}\\ E_{3,4}\end{array}\begin{array}{c}\begin{array}{ccccc} S_1 & S_2 & S_3 & S_4 & S_5\end{array}\\ \begin{bmatrix}1&1&0&0&1\\0&1&1&1&1\\0&1&1&0&0\\1&0&1&1&0\\1&0&1&0&1\\0&0&0&1&1\end{bmatrix}\end{array}$$

圖 6 比較 Haplotype 配對差異與轉換矩陣

若將ES視為一個bipartite網路圖$G=(N,A)$的相鄰矩陣(adjacency matrix)，我們可將所有的SNP(S_j)與Haplotype配對(E_{i_1,i_2})視為S節點與E節點，亦即$N=S\cup E$；針對所有$I(E_{i_1,i_2},S_j)=1$之關係，我們可將其所對應之S、E節點連結，亦即$A=\{(S_j,E_{i_1,i_2}):\forall S,\ E_{i_1,i_2}\in E\}$；反之，對所有$I(E_{i_1,i_2},S_j)=0$之關係，其所對應之$S$、$E$節點將不被連結。如此一來即可將SNP與Haplotype配對

的比對關係以一個bipartite網路圖$G=(N,A)$表示，並將各節點$k \in N$的degree(deg(k))記錄於其旁。舉例來說，圖7即為圖6的ES矩陣所對應之bipartite網路圖$G=(N,A)$。

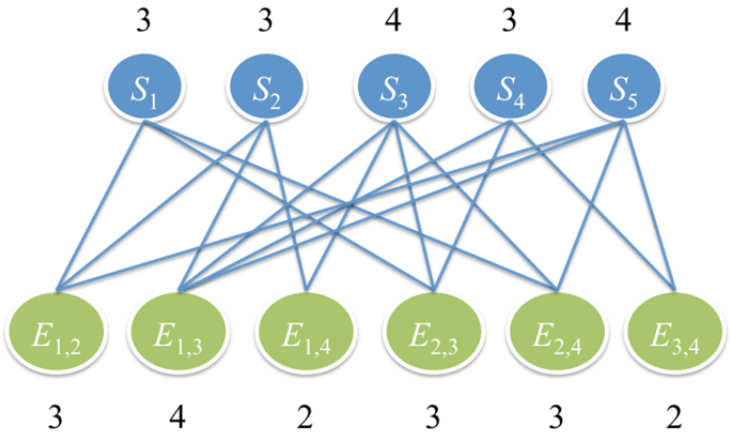


圖 7SNP 與 Haplotype 配對辨別關係 bipartite 網路圖

由圖7可觀察出，當S_3被挑選時，$E_{1,2}$、$E_{2,3}$、$E_{2,4}$皆可被其所辨識。選取Tag SNPs之主要限制在於所選出之SNP集合必須保證可以辨識E集合內的所有元素；此外，在所有滿足此限制條件的候選SNP集合中，Tag SNPs所包含的SNP個數必須為最少。以圖8與圖9為例，選擇$\{S_1, S_2, S_4\}$可以辨識出所有的E元素;同理，選擇$\{S_3, S_5\}$也可以 辨識出所有的E元素;由於$\{S_3, S_5\}$的個數為所有可辨識E集合全部元素之SNP集合中最小者，因此$\{S_3, S_5\}$將被選為Tag SNPs。值得注意的是，在Tag SNPs的選取問題中，其最佳的Tag SNPs解可能不只一個。

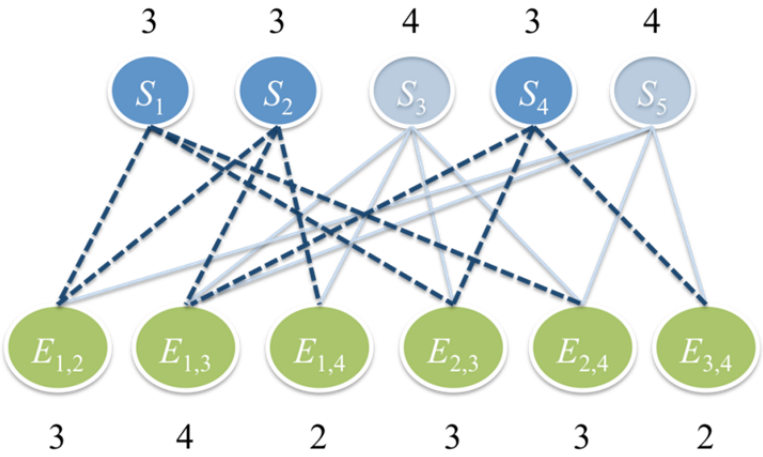


圖 8 選取 S_1、S_2、S_4 Tag SNPs 之圖示

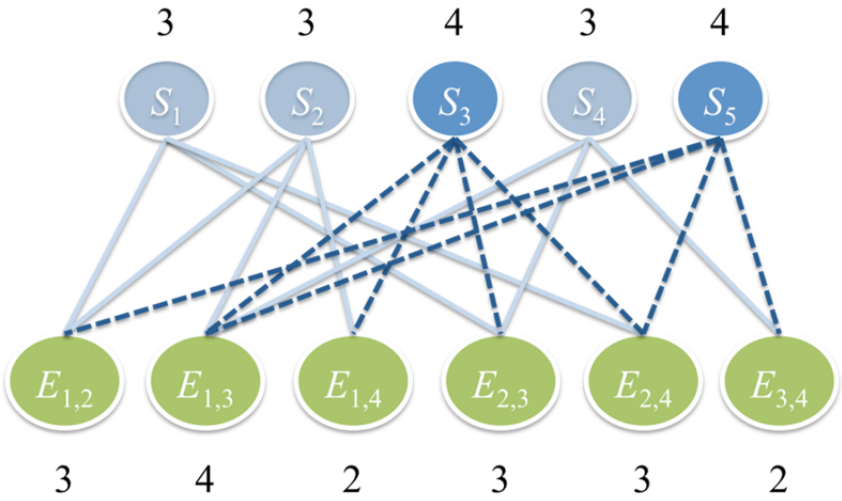


圖 9 選取 S_3、S_5 Tag SNPs 之圖示

2.5 以多目標求解Tag SNPs選取問題

上一章節介紹了Tag SNPs選取問題是一種集合覆蓋問題，在大量的單核苷單酸多型性中選取最少的強固型Tag SNPs解集和。本研究制定最小化強固型Tag SNPs問題作為二進制目標問題，另外還包含兩個新目標，考慮單倍型之間的區別。

數學模式定義Tag SNPs [8, 24]：給一組集合N，SNP$\{S_1, \dots, S_N\}$和M個等位基因類別$\{P_1, \dots, P_M\}$。$P_{i,k}$表示等位基因P_i中有k的元素。S^T為Tag SNPs的集合，S為辨識任何兩個等位基因類別的子集合。也就是說，對於任何兩個等位基因P_i與P_j，至少存在一個Tag SNPs$S_k \in S^T$，這樣$P_{i,k} \neq P_{j,k}$。
四個目標方程式定義如下 [8]：

2.5.1 最小化強固型Tag SNPs

第一個目標找出Haplotype配對上涵蓋所有重要的SNP。假設一個集合S^T表示所選擇的Tag SNPs，$P_{(i,j)}$表示Haplotype樣式P_i與P_j則代表Haplotype矩陣中的維度；其中要找到最少的Tag SNPs，目標方程式如下 [24]：

$$\text{minimize } \|S^T\| \tag{1}$$

$$\|S^T\| \geq [log_2 h],\ P_{(i,j)} \geq 1, 1 \leq i < j \leq h \tag{2}$$

該Haplotype配對維度是指可以識別並區分SNP的個數，由於每個SNP包含兩個等位基因，Tag SNPs可以視為一個二進制的Haplotype，因此集合S所

代表的最小基數至少為$\log_2 h$。

2.5.2 容錯(Tolerance)

第二個目標主要計算Haplotype樣式中SNP的遺失資料部分，我們又可稱為計算Tag SNPs的容錯。經由SNP微陣列所獲得的資料，而每個微陣列中都會有SNP的遺失資料部分，為了防止資料遺失所造成計算Tag SNPs的複雜度，容錯應的越大越好。目標方程式如下 [24]：

$$\text{maximize}\left(\min_{i,j}\left\|D_{ij}(S^T)\right\|\right) \tag{3}$$

方程式中，$D_{ij}(S^T)$表示S^T為Tag SNPs一個集合，可以識別等位基因P_i與P_j。

2.5.3 Haplotype的相異性

第三個目標主要是為了避免相似的Haplotype中獲得Tag SNPs。因Tag SNPs已經轉成二進制編碼，本研究採用漢明距離量化Haplotype樣式之間的相似性，並對所有Haplotype配對計算平均的漢明距離。由於在演化式計算中會求出不同的Tag SNPs解，因此Haplotype配應對Tag SNPs的每個解都必須正規化。目標定義方程式如下：

$$S^T = \bigcup_{k\in K^T} S_k \tag{4}$$

其中，K^T表示S^T的索引集合。兩個等位基因P_i與P_j在漢明距離表示如下：

$$H(P_i,P_j) = \sum_{k\in K^T}\left|P_{i,k} - P_{j,k}\right| \tag{5}$$

每個等位基因配對最大平均漢明距離方程式如下：

$$\text{maximize } \bar{H} \tag{6}$$

$$\bar{H} = \frac{1}{\binom{M}{2}}\sum_{0\le i<j\le M} H(P_i,P_j) \tag{7}$$

2.5.4 Haplotype的多樣性

除了上面敘述的計算Haplotype的相異性外，第四個目標主要計算平衡SNP檢測在每個等位基因類別。這樣的現象可以獲得漢明距離的最小變異值，在全部的Haplotype配對背景中定義一個標籤的解。因此，識別Tag SNPs結果是對每一個等位基因都有充足的採樣。目標定義方程式如下：

$$\text{minimize Var}(H) \tag{8}$$

$$\text{Var}(H) = \frac{1}{\binom{M}{2}} \sum_{0 \le i < j \le M} \left(H(P_i, P_j) - \bar{H} \right)^2 \tag{9}$$

三、最佳化演算法之應用

3.1 智慧型多目標基因演算法

為解決選擇 Tag SNPs 的大量參數最佳化問題，必須藉由一強而有力的最佳化演算法。本文使用智慧型演化式演算法 [10] 做為最佳化工具。智慧型基因演算法結合了基因演算法與直交實驗兩種方法之特性，具有高精確度、收斂速度快的優點。

3.1.1 直交表與因素分析

智慧型基因演算法(Intelligent Genetic Algorithm；IGA) [10] 和一般基因演算法最大的不同之處在於，IGA 在交配(crossover)的過程使用直交表(Orthogonal Arrays)挑選好的參數，使得 IGA 能克服參數過多染色體過長的問題。直交表(OA)與因素分析常用在品質控制的方法中 [30]，也可運用於很有效率地改善交配操作。假設有兩水準的 N 個因素的直交表，那總共會有2^N種不同的組合。當四組(1,1)、(1,2)、(2,1)及(2,2)出現在全部實驗時，那麼兩個因素在直交表的行是彼此正交的。當任兩個因素在實驗集合是正交時，這個集合可被稱之為直交表(OA)。為了建立兩水準 N 個因素的直交表，我們可以得到一個整數$\mathrm{n} = 2^{[log(N+1)]}$，來建立$n$列和$(n-1)$行的$L_n(2^{n-1})$的直交表，然後選擇 N 行來使用。舉例來說表 3 則表示$L_8(2^7)$直交表。

因素分析可以評估在評估函數中的因素效果，排名最有效果的因素以及決定每個因素的最好水準組合，進而促使評估函數有最佳的結果。直交實驗設計可以縮減因素分析的實驗次數。直交表的實驗次數在單一因素分析時只需n次。假設y_t是第t次實驗的函數評估值，我們定義S_{jk}是j因素在水準k的主效果：

$$S_{jk} = \sum_{t=1}^{n} y_t \times F_k \tag{10}$$

其中F_k為一個旗標值，若第t次實驗中第j個因素選用水準為k，則F_k為 1；若否，則F_k為 0。若評估函數值為望大，則較大的主效果值表示對評估函數具有較佳的貢獻度；反之若評估函數望小，則主效果值小者貢獻度較佳。

主效果可以顯示因素中水準的個別影響。例如主效果$S_{j1} > S_{j2}$則表示在參數最佳化的問題中，第j個因素水準值 1 對於整體最佳化函數的貢獻大於水準值 2。如果相反的情形$S_{j1} < S_{j2}$，則表示水準值 2 較佳。在各因素間無交互作用的前提下，主效果的分析可以用來推測出全實驗的最佳解。

直交因素實驗為一種部分因素實驗方式，可以有效減少參數設計時的實驗次數，並同時考慮實驗因素之間的交互作用。將直交因素實驗後的數據經過主效果分析，便可以將每個因素對於設計目標的貢獻優劣計算出來，推論出最佳解的實驗參數。

表 3 $L_8(2^7)$直交表

	實驗因素							
實驗編號	1	2	3	4	5	6	7	評估值
1	0	0	0	0	0	0	0	y_1
2	0	0	0	1	1	1	1	y_2
3	0	1	1	0	0	1	1	y_3
4	0	1	1	1	1	0	0	y_4
5	1	0	1	0	1	0	1	y_5
6	1	0	1	1	0	1	0	y_6
7	1	1	0	0	1	1	0	y_7
8	1	1	0	1	0	0	1	y_8
S_{j1}	S_{11}	S_{21}	S_{31}	S_{41}	S_{51}	S_{61}	S_{71}	
S_{j2}	S_{12}	S_{22}	S_{32}	S_{42}	S_{52}	S_{62}	S_{72}	
MED	MED_1	MED_2	MED_3	MED_4	MED_5	MED_6	MED_7	

3.1.2 智慧型交配運算

智慧型基因演算法與傳統基因演算法最大的不同，乃是已智慧型交配取代一般的單點交配或多點交配。傳統的交配方式無法評估染色體中參數個別的優劣，加上交配點是隨機方式產生，得到的後代染色體品質不容易提昇。

事實上，我們可以將染色體交配過程視為一種實驗因素；將來自父代的兩個染色體已切割好欲交配的片段，做為直交實驗的因素，並以染色體片段「互換」或「不換」作為兩種水準值。如此，以兩水準值交實驗產生出優良品質的染色體的機率便可大幅提昇，進行步驟如下：

步驟一：令產生染色體中的由交配點所切割出的基因片段為實驗因素；假設因素數目為n欄作為實驗之用，其中$\beta = 2^{[\log(n+1)]}$。

步驟二：令因素j的水準 1 與水準 2 分別表示來自父代染色體P_1與P_2第j個基因片段。

步驟三：根據直交表，計算個因素組合實驗的評估值y_t，$t = 1,2,\dots,\beta$。

步驟四：計算個因素之主效果S_{jk}，$j = 1,2,3,\dots,N$；$k = 1,2$。

步驟五：決定各因素的最佳水準。在評估函數望大時，則選擇主效果值較大之水準；在評估函數望小時，則各因素的最佳水準為主效果值較小之水準。如評估函數望大且$S_{j1} > S_{j2}$，則因素j的最佳

水準為 1；反之則最佳水準為 2。

步驟六：根據各因素的最佳水準，選擇對應父代染色體中的基因片段，組合成第一個子代染色體。

步驟七：將各因素的主效果差值$(|S_{j1} - S_{j2}|)$排名，差值越大者排名越高。

步驟八：以類似第一個子代染色體的方式來組合因素，將差值排名最差的因素，選擇與第一個子代相反的水準，則可產生第二個子代染色體。

3.2 方法流程

本論文以智慧型多目標基因演算法做為挑選 Tag SNPs 參數的核心演算法，因此在本章節便以多目標基因演算法的各步驟來介紹我們所提出的方法。

3.2.1 選擇運算

我們選用二元競爭法作為基因演算法的選擇運算，每次進行選擇時由族群中隨機挑選出兩條染色體，評估值較佳的染色體則可以進入基因重組運算(即交配與突變運算)。傳統的輪盤法選擇運算，在基因運算法演化至接近最佳解時，由於各染色體評估值變化不大時，造成每個染色體被選擇至下一代的機率都差不多，無法有效選出優秀的染色體。

而二元競爭選擇法的優點是，即使在染色體評估值變化不大時，仍可透過競爭方式取出表現比較優秀的染色體，避免基因演算法演化過早收斂。

3.2.2 智慧型交配

智慧型交配運算在本論文所提的方法中扮演一個相當重要的角色。智慧型交配運算結合了兩水準直交實驗與交配運算，能夠有效率的產生具有優秀評估值的新染色體。假設我們使用$L_{N+1}(2^N)$直交表來做智慧型交配運算，本染色體交配運算的詳細進行步驟如下：

步驟一：假設即將進行交配運算的兩條染色體為P_1、P_2，比對P_1、P_2內的參數基因，並將重複出現在兩條染色體內的參數基因移動至染色體末端，其相對應的控制基因。

步驟二：隨機將參數基因切割成$[N/2]$個基因片段。每個基因片段及代表直交表的一個因素。P_1中第j個基因片段，及代表因素j的第一個水準值；P_2則代表第二的水準值。

步驟三：計算直交表中每個染色體排列組合方式的適應值f_t，

$t = 1, 2, 3, \ldots, N+1$。

步驟四：計算個因素之主效果S_{jk}，$j = 1, 2, 3, \ldots, N$；$k = 1, 2$。

步驟五：決定的因素的最佳水準；在評估函數望大時，則選擇主效果值較大之水準；在評估值望小時，則各因素的最佳水準為主效果值較小之水準。如評估函數望大且$S_{j1} > S_{j2}$，則因素j最佳水準為 1；反之則最佳水準為 2。

步驟六：根據各參數的最佳水準，選擇對應父代染色體中的參數，組合出第一個子代染色體。

步驟七：將各參數的主效果差值排名，差值越大者排名越高。

步驟八：以類似第一個子代染色體的方式組合參數，除了排名最差的參數選擇的水準與第一個子代染色體相反，可產生第二個子代染色體。

3.2.3 突變運算及演化中止條件

突變運算使用任意合理之亂數運算。假設一條染色體編碼總長為N個位元，突變率為P_m，則每次的突變運算，隨機由染色體中選出$[N * P_m]$的位元，然後以亂數變化所選中的 SNP 參數。

基因演算法所需的演化時間必須視問題的複雜度而定，較一般的作法是設定演算法的評估次數與問題中的參數數目成正比。一但終止條件達到設定值，演化即停止，並輸出所搜尋過的最佳解，用最佳解來當模型參數計算出所選的 Tag SNPs 的值。

3.3 智慧型多目標基因演算法

IMOGA 在演化過程中基本上還是依循基因演算法的過程，但是在評估函數方式不像是 IGA 是直接將目標函數當成評估函數而作演化判斷之根據，而是使用基於 Pareto 理論，使用通適化且不受尺度因素影響的評估函數(Generalized Pareto-based Scale-Independent Fitness Function，GPSIFF)作為評估函數[10]。

3.3.1 基於 Pareto 理論通適化且不因尺度影響之評估函數

函數

為了分辨出各染色體之間的優劣，本論文應用了基於 Pareto 理論為基礎的記分方式以避免受到尺度因素的影響，並且對於被支配解(dominated)和未

被支配解(non-dominated)給於具有區分能力的適應函數值，用以取代傳統有失準確性的排名法和距離方式，稱之 GPSIFF。

GPSIFF 使用類競爭式(Tournament-Like)的記分方式來評估 Pareto 解集中染色體個體 x 的適應值，GPSIFF 的數學式如下：

$$\text{Score}(x) = p - q + c \tag{11}$$

其中p表示在目前欲評估的解集中x所支配的個體數目，q表示在目前欲評估的解集中能夠把x支配的個體數目，c是一個較大的正整數，以保證求出的適應值為一正整數。通常以目前參與評估運算的所有個體的數目作為正整數c的值。

GPSIFF 的優點如下：

(a) 不需調整權重值：基於 Pareto 理論來評估解的好壞，沒有權重加總法需決定權重值的困難，也不會受到人為主觀判斷的影響。

(b) 不需考量尺度因素：由於各目標函數值的尺度適應值不盡相同，在權重加總法中需考慮到尺度因素，以免使得權重設定失之準確。

(c) 以積分方式有效辨識不同解的優劣程度：取代傳統排名法可能將不同的解給予相同的排名，以及距離法有尺度因素影響的缺點，以精確的記分評估解的優劣程度。

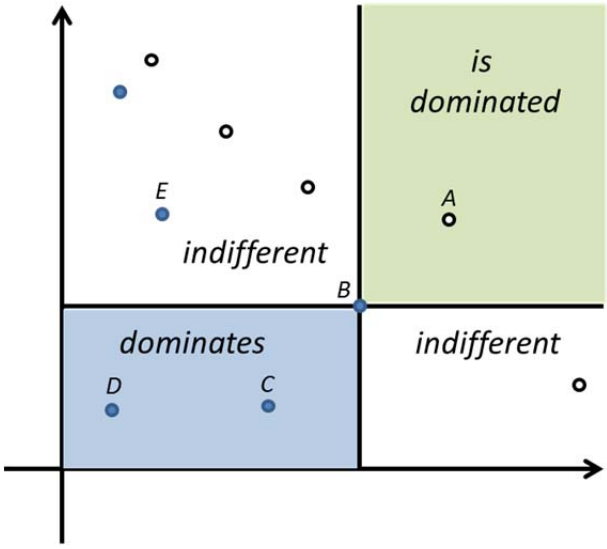


圖 10 支配與被支配關係示意圖

我們用圖 10 表示出在兩個目標中同時做最小化問題之說明，在 Pareto 解集中所有個體在雙目標軸的關係。我們以點 B 為例，點 B 之$f1$和$f2$同時都

比點 A 小，所以點 A 被點 B 所支配；同理，點 C 與點 D 位於點 B 之左下角，點 B 就被點 C 與點 D 所支配，所以點 B 將不會被收在 Pareto 解集中。

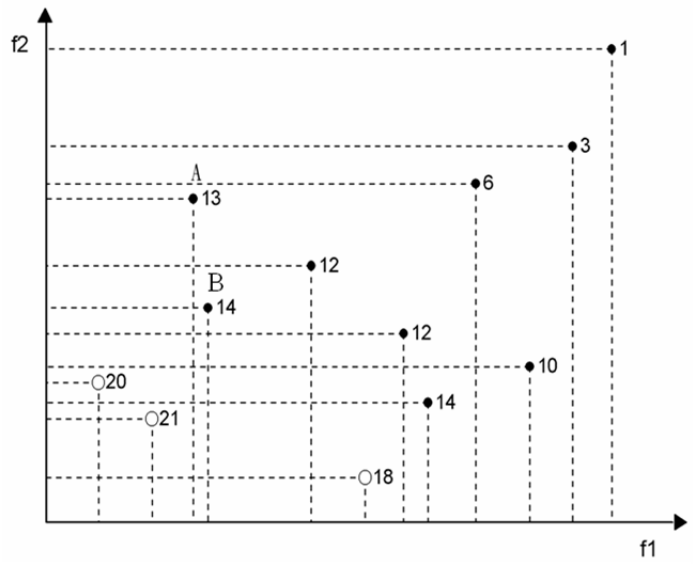


圖 11 GPSIFF 之示意說明圖[10]

同樣的最小化問題，在圖 11 中點上的數字為適應值c。以 A 為例，其$c = 12$，○ 是未被支配解 ● 是被支配解，$p = 3$，$q = 2$，所以適應值為 13。

3.3.2 演算法流程

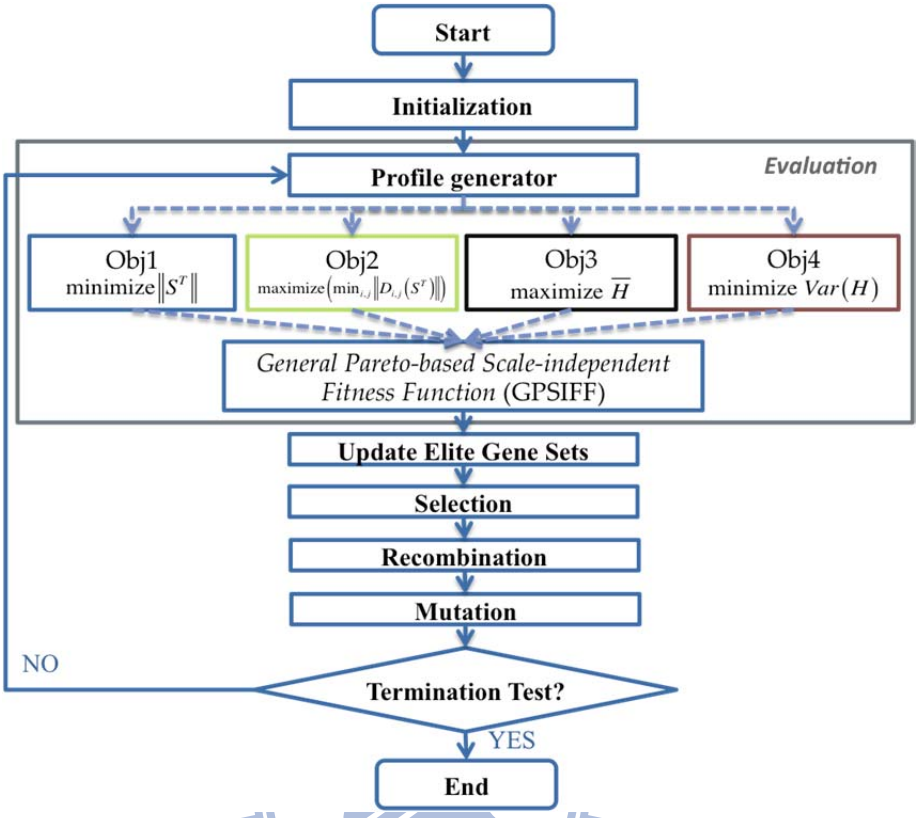


圖 12 智慧型多目標基因演算法流程圖

輸入： N_{pop} 族群大小

N_E 最大容量的菁英集合

p_s 選擇比例

p_c 交配機率

p_m 突變機率

γ IGC 運算因素個數

輸出： E 未被支配解集

智慧型多目標基因演算法之流程，其步驟詳細敘述如下：

步驟一：(初始化，Initialization)亂數產生初始用的族群數量N_{pop}個染色體以及兩個空的優秀基因集合，一個是E；一個是E'。

步驟二：(評估初始值，Fitness Evaluation)計算族群裡全部染色體的兩項目標函數值，並藉由 GPSIFF 分配每條染色體一個評估值。

步驟三：(更新優秀基因集合，Elite sets)將未被支配(non-dominated)的染色體同時丟入E和E'，然後清空E'。考量E中的所有染色體，將被支配(dominated)的染色體移除。若N_E的數量大於原本的所設

定數量，則將從亂數去除超過的部分。

步驟四：(挑選，Selection) 從族群裡用 binary tournament selection 挑選出$N_{pop} - N_{ps}$個染色體，並從E中亂數挑選出N_{ps}個染色體形成一組新的群組。其中$N_{ps} = N_{pop} \times P_s$，若$N_{ps} > N_E$，則令$N_p = N_E$。

步驟五：(重組，Crossover) 藉由 Intelligent Gene Collector (IGC)運作從$N_{pop} \times P_c$選擇親代。每次 IGC 皆是由 OA 重組因子(副產物)找出未被支配的染色體以及兩個子代加入至E'。

步驟六：(突變，Mutation)根據P_m對整個族群進行突變機制。

步驟七：(終止條件，Termination Test)假設已滿足停止條件即可停止演算，反之回到步驟二。

四、實驗結果與討論

4.1 實驗資料蒐集

本實驗數據來源為國際單型圖譜計畫網站(International HapMap Project)中獲得，如表 4 所示。

表 4 測試資料 [8]

Released by HapMap Bulk data Perlegen

染色體	3 號
人種	漢人 (CHB)
區塊編號(連續)	1034217~1034270
SNP 個數	1032(722 useful SNPs)

將 53 個區塊內的基因型透過 Haplotype inference(SDPHapInfer 與 PHASE 軟體 [7, 31])機制轉換為二進制編碼，資料編碼方式，主要等位基因(major allele)編碼為 0，次要等位基因(minor allele)編碼為 1，圖 13 為編碼示意圖，藍色為主要等位基因，黃色為次要等位基因；其中樣品個體數為 48 個，SNPs 個數 1032 個。經過編碼後會得到 Haplotype 樣式矩陣圖。

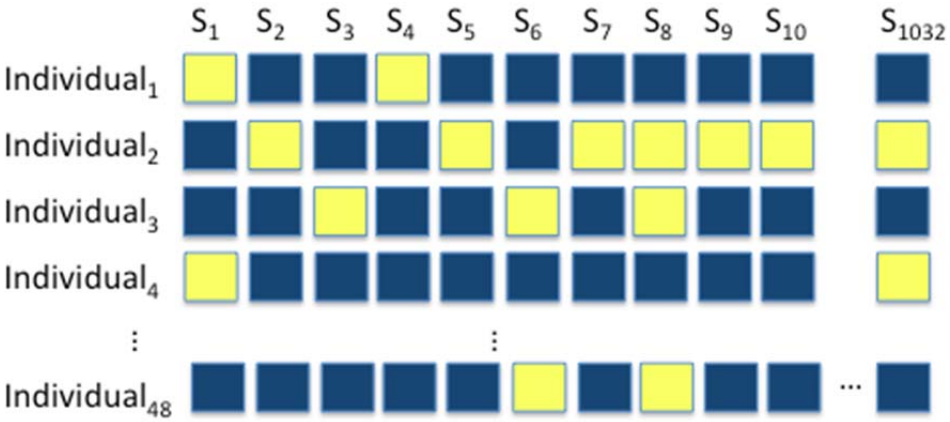


圖 13 Haplotype 樣式矩陣圖

(Individual 表示個體；S 表示 SNP)

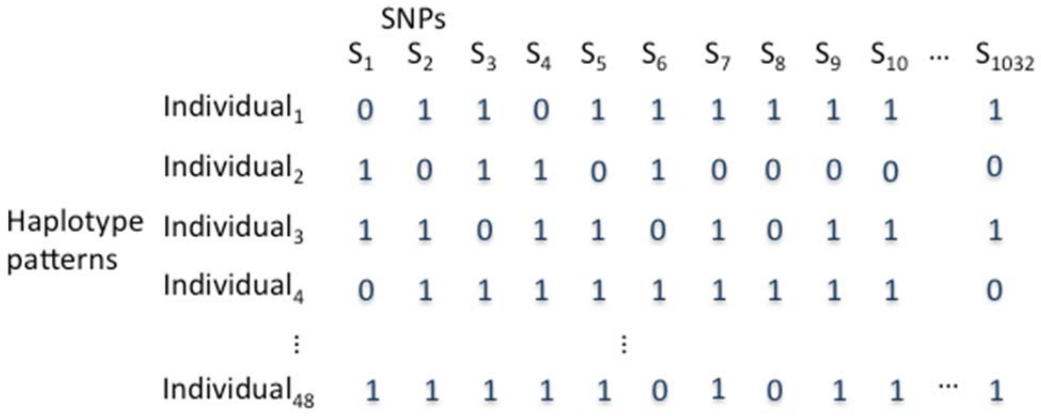


圖 14 資料編碼

(Individual 表示個體；S 表示 SNP)

4.2 利用智慧型多目標基因演算法取得最佳解集合

經由 4.1 節將資料編碼後，本章節使用四個目標參數，將比較 IMOGA 與 NSGA2 演算法所求選取最少的 Tag SNPs 解集合；參數設定如表 5。目標一：最小化 Tag SNPs，目標二：Tag SNPs 容錯率最大化，目標三：Haplotype 的變異性最大化；目標四：Haplotype 的多樣性最小化。

表 5 IMOGA 與 NSGA2 參數設定

參數	**值**
編碼表示(Representation)	二進制編碼 (選或不選)
族群大小(Population size)	200
交配率(Crossover)	$P_c = 0.7$
突變率(Mutation)	Bit-flip 突變；$P_m = 1/l$ (其中 l 代表可用的 SNPs 個數)
終止條件(Termination)	500 次迭代
評估值(Evaluation)	25000

由圖 15 得知目標一與目標二互不為衝突，當目標一求最少 Tag SNPs 解集時，目標二 Haplotype 樣式中 SNP 遺失資料也會增加，即計算最少 Tag SNPs 的容錯也會增加，所以這兩個目標示互不相衝突的。Tag SNPs 解集要望小，Haplotype 樣式中 SNP 遺失資料容錯要望大。

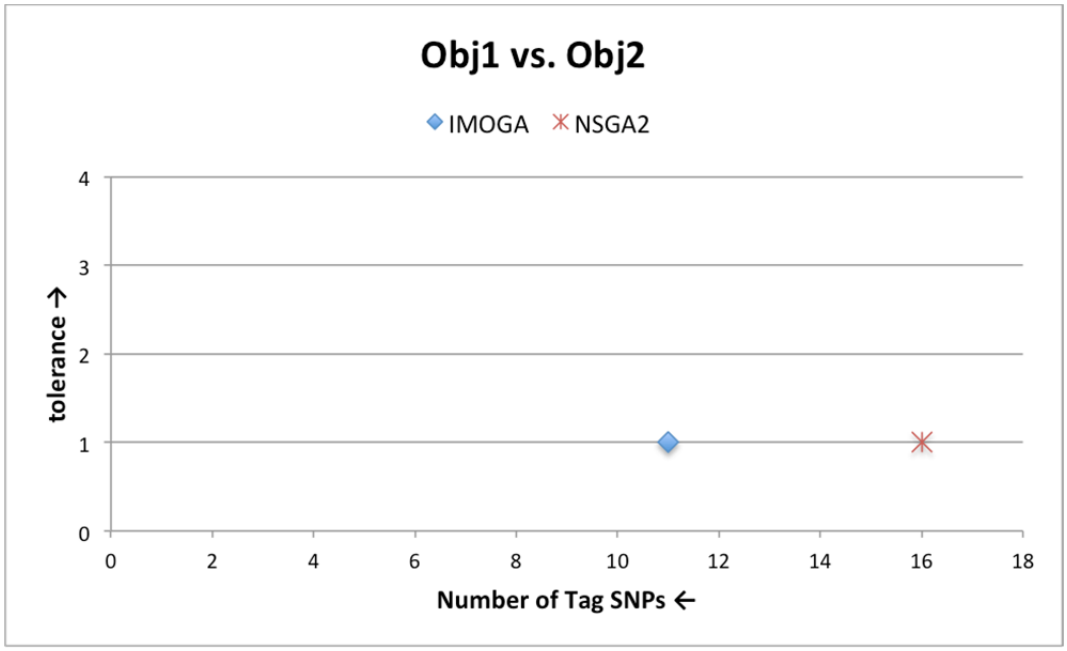


圖 15 IMOGA 與 NSGA2 實驗比較，X 軸為目標一、Tag SNPs 個數，Y 軸為目標二、Tag SNPs 容錯

圖 16 目標一與目標三互為衝突，當目標一求最少 Tag SNPs 解集時，Haplotype 樣式之間過於相似時，無法識別出最佳的 Tag SNPs，所以這兩個目標示互相衝突的。Tag SNPs 解集要望小，Haplotype 平均漢明距離要望大。

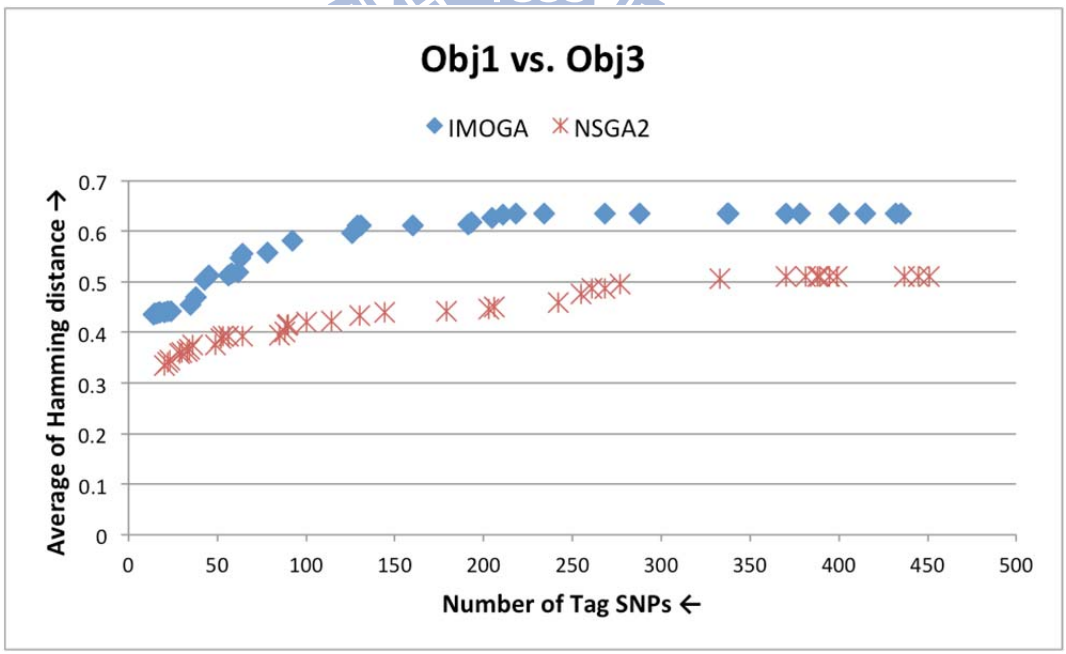


圖 16 IMOGA 與 NSGA2 實驗比較，X 軸為目標一、Tag SNPs 個數，Y 軸為目標三、Haplotype 平均漢明距離

圖 17 目標一與目標四互為衝突，目標一求最少 Tag SNPs 解集時，Haplotype 樣式上的 SNP 每個等位基因需平衡，如果每個等位基因變異性大，無法識別出最佳的 Tag SNPs，因此，Tag SNPs 解集要望小，Haplotype 變異漢明距離要望小。最佳解集合會落在圖 12 淡藍色區域。

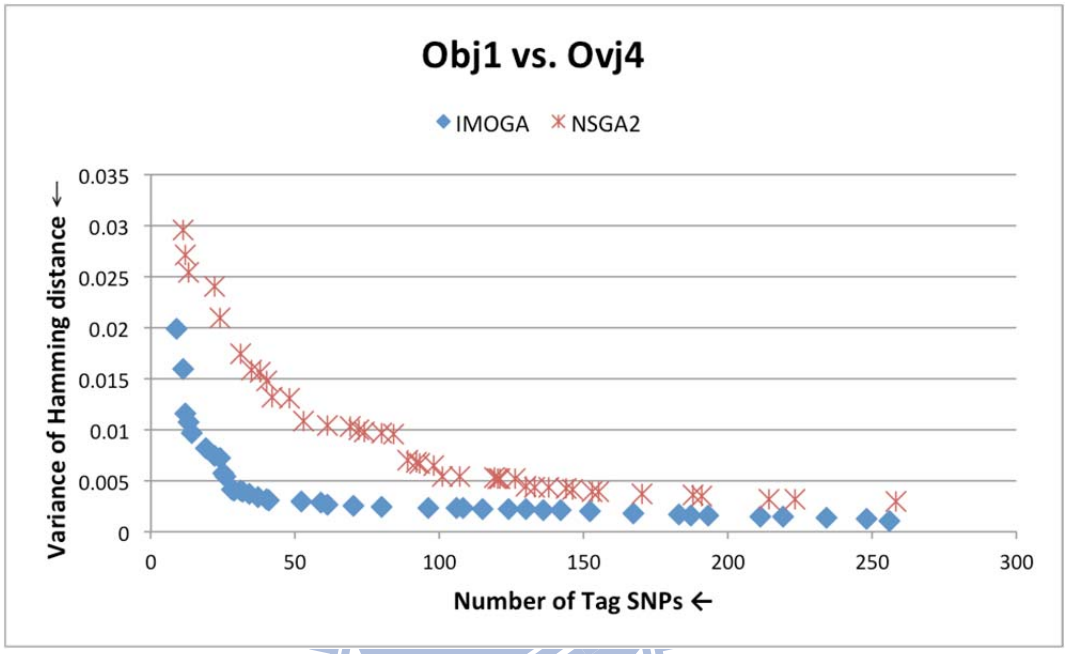


圖 17 IMOGA 與 NSGA2 實驗比較，X 軸為目標一、Tag SNPs 個數，Y 軸為目標四、Haplotype 變異漢明距離

圖 18 目標二與目標三互為衝突，目標二為 Haplotype 樣式中 SNP 遺失資料容錯，目標三 Haplotype 樣式之間的 SNP 遺失的資料越大時，整的 Haplotype 樣式就會越複雜，計算上所需要的成本也就越大，所以 SNP 能容錯的值愈大時，則仍要考慮到 Haplotype 樣式中 SNP 彼此之間的的相似度。因此，SNP 容錯解集要望大，Haplotype 平均漢明距離要望大。

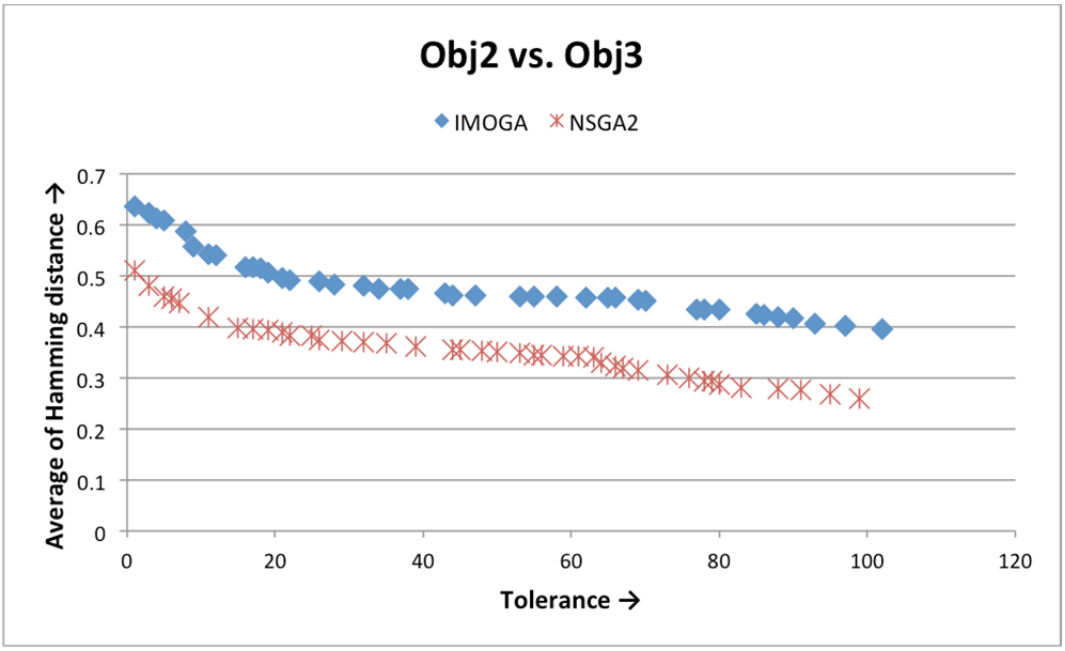


圖 18 IMOGA 與 NSGA2 實驗比較，X 軸為目標二、Tag SNPs 容錯，Y 軸為目標三、Haplotype 平均漢明距離

圖 19 目標二與目標四互為衝突，目標二為 Haplotype 樣式中 SNP 的容錯，目標四 Haplotype 樣式上的 SNP 每個等位基因需平衡，如果每個等位基因變異性大，加上有 SNP 資料遺失，會影響整個 Haplotype 樣式求 Tag SNPs 的最佳解，所以 SNP 能容錯的值愈大時，則仍要考慮到 Haplotype 樣式中 SNP 等位基因之間的的平衡。因此，SNP 容錯解集要望大，Haplotype 變異漢明距離要望小。

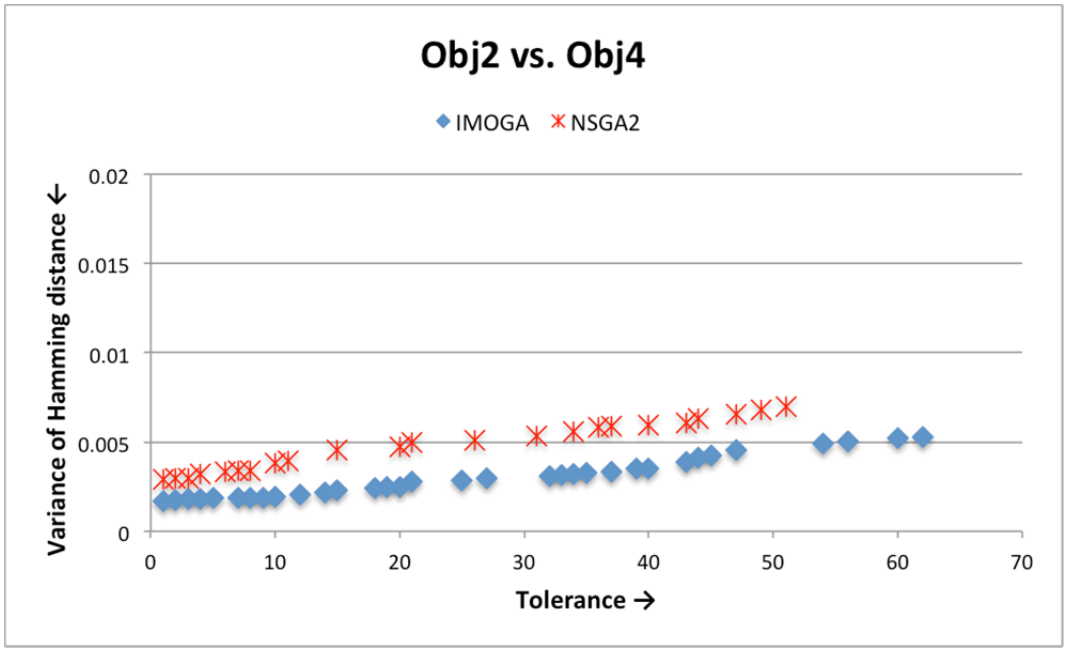


圖 19 IMOGA 與 NSGA2 實驗比較，X 軸為目標二、Tag SNPs 容錯，Y 軸為目標四、Haplotype 變異漢明距離

目標三與目標四互為衝突，當目標三 Haplotype 樣式中的獲得相似 Tag SNPs 越多，則會無法找出 Tag SNPs 最佳解，加上每個 SNP 的等位基因相似性要低，變異性要高，才能求得 Tag SNPs 最佳解。其中，Haplotype 平均漢明距離解集要望大，Haplotype 變異漢明距離要望小。

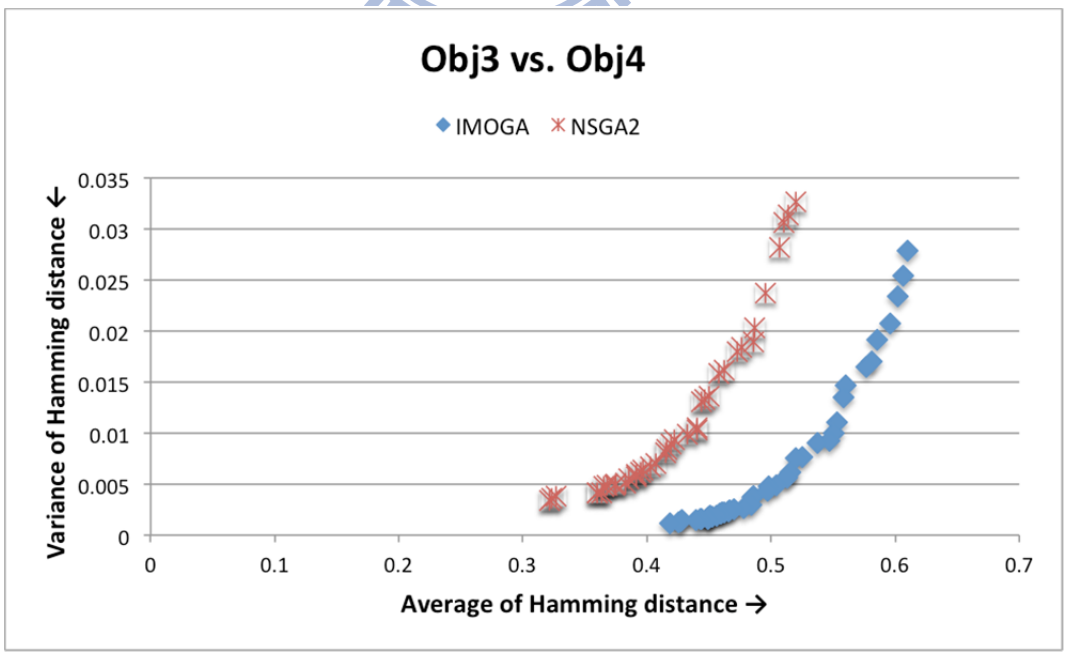


圖 20 IMOGA 與 NSGA2 實驗比較，X 軸為目標三、Haplotype 平均漢明距離，Y 軸為目標四、Haplotype 變異漢明距離

4.3 覆蓋測量

IMOGA 與 NSGA2 針對每個雙目標測試資料在相同的評估次數之獨立執行 30 次實驗，並利用覆蓋測量(coverage metric) [32] 比較兩個演算法在每次實驗所得的最佳解集合的品質，計算方法如下：

$$C(A，B) = \frac{B\text{的不被支配解集被 A 弱支配的數目}}{B\text{的不被支配解的個數}} \tag{12}$$

$C(A，B) = 1$表示所有在 B 的不被支配解集被都 A 弱支配，$C(A，B) = 0$則表示所有在 B 的不被支配解集皆沒有被 A 弱支配。將 30 次實驗結果的 *C(IMOGA, NSGA2)*和 *C(NSGA2, IMOGA)*繪製成統計上常用的箱型圖(box plot)圖。從箱型圖圖 21 與圖 22 可以觀察出 *C(IMOGA, NSGA2)*趨近於 1，而 *C(NSGA2, IMOGA)*趨近於 0，表示 NSGA2 的不被支配解集被都 IMOGA 弱支配，IMOGA 的不被支配解集皆沒有被 NSGA2 弱支配。IMOGA 所求得的解集品質依舊優於 NSGA2 解集。

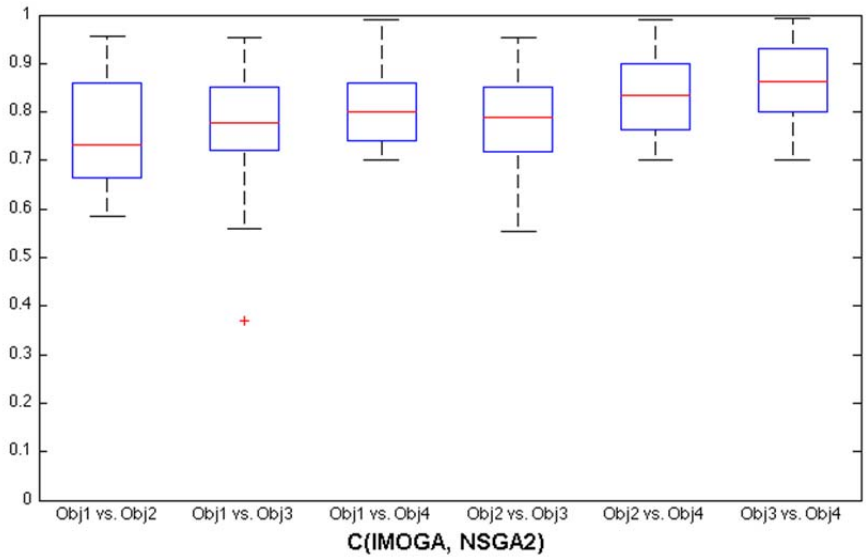


圖 21 箱型圖 *C(IMOGA, NSGA2)*覆蓋測量

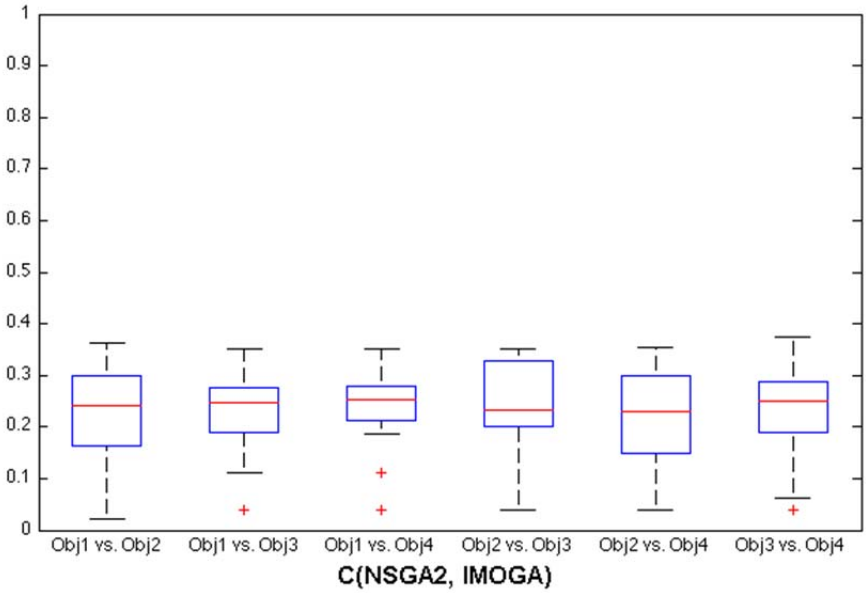


圖 22 箱型圖 *C(NSGA2, IMOGA)*覆蓋測量

五、維生素 D 受體基因上 Tag SNPs 特徵選取之應用

5.1 維生素 D 受體基因簡介

維生素 D 受體(Vitamin D receptor；VDR)屬於核受體(nuclear receptors)超家族的類固醇及甲狀腺激素受體，VDRs 的細胞表現於器官中，包括腦、心臟、皮膚、性腺、前列腺癌和乳癌等。VDR 參與細胞增生和細胞分化，也影響免疫系統，進一步激化 T 細胞與 B 細胞；VDR 也活化腸道、骨骼、腎臟、甲狀旁腺細胞，血液中鈣、磷修復和維護骨骼 [33]。基因功能如表 6。

表 6 維生素 D 受體介紹

基因名稱	基因功能	染色體位置	Ensembl 基因編號
維生素 D 受體 VDR: Vitamin D 1,25 dihydroxyvitamin D_3 Receptor(NR111)	維生素 D3 編碼核激素受體，主要讓礦物質代謝通過受體，及調節其他各種代謝途徑	12q13.11	ENSG0000011424

5.2 維生素 D 受體基因實驗資料蒐集

實驗數據來源為國際單型圖譜計畫網站(International HapMap Project)，如表 7 所示。

表 7 HapMap3 Release #2 (Phase 3)

染色體	12 號
基因	維生素 D 受體(VDR)
染色體位置	46521589..46585081
人種	西歐(CEU)
族群人數	121
SNPs 個數	977

同 4.1 章節，將 HapMap 資料中的基因型透過 Haplotype inference (SDPHapInfer 與 PHASE 軟體 [7, 31])機制轉換為二進制編碼，資料編碼方式，主要等位基因(major allele)編碼為 0，次要等位基因(minor allele)編碼為 1；其

中樣品個體數為 121 個，SNPs 個數 977 個。

驗證資料部分，S.Karami *et al.* [13]、J.B.Egan *et al.* [14] 及 J.Ahn *et al.* [12] 文獻中提到重要的維生素 D 受體基因上 Tag SNPs，為了方便記錄文獻中所提到的重要 Tag SNPs，將 S.Karami *et al.*、J.B.Egan *et al.*及 J.Ahn *et al.*以 Ref.1、Ref.2 及 Ref.3 代表。圖 23 及表 8 為三篇文獻中所提到重要的 Tag SNPs，Tag SNPs 個數總共有 70 個，Ref.1、Ref.2 交集部分有 11 個，Ref.1、Ref.3 交集部分有 3 個，Ref.2、Ref.3 交集部分有 2 個，Ref.1、Ref.2、Ref.3 全交集有 2 個，未重複有 32 個。

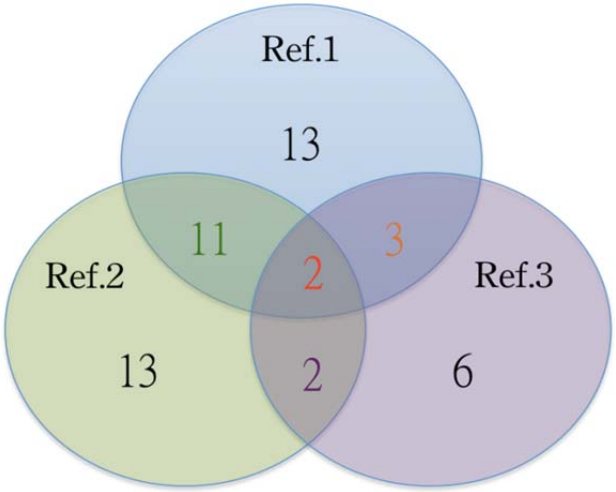


圖 23 維生素 D 受體基因上的 Tag SNPs 交集

Ref.1、Ref.2、Ref.3 全交集中有兩個重要的 Tag SNPs，分別是 **rs11574026** 與 **rs1544410**；如表 8 所呈現上述三篇文獻中所提到的重要 Tag SNPs。

表 8 文獻中 Tag SNPs 交集

文獻	Tag SNPs	rs#
Ref.1、Ref.2、Ref.3 交集	2	rs11574026、rs1544410
Ref.1、Ref.2 交集	11	rs4760648、rs2853564、rs2254210、rs2239186、rs3819545、rs2189480、rs886441、rs2239179、rs2107301、rs2239182、rs11574077
Ref.1、Ref.3 交集	3	rs4516035、rs731236、rs10875695

Ref.2、Ref.3 交集	2	rs11574143、rs7299460
Ref.1 未交集	13	rs10459228、rs10783219、rs11168292、rs11574077、rs11168287、rs222857、rs3782905、rs12717991、rs2239180、rs2248098、rs2239185、rs3847987、rs12721364
Ref.2 未交集	13	rs121721364、rs7968585、rs757343、rs7967152、rs2283342、rs2239181、rs1540339、rs10735810、rs2238136、rs4328262、rs4334089、rs4237855、rs3890733
Ref.3 未交集	6	rs11168293、rs4760655、rs4760658、rs11568820、rs7310552、rs7970314

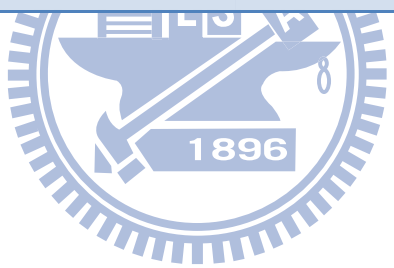
表 9 中粗體標示的參照 SNP 認證碼(rsID)為上述文獻中提到最中重要的 Tag SNPs。

表 9 重要維生素 D 受體基因上的 Tag SNPs

VDR	Total	Tag SNP
Analysis of SNPs and Haplotypes in Vitamin D Pathway Genes and Renal Cancer Risk [13] (Ref.1)	29	rs10459228 rs4516035 rs10783219 rs11168292 rs10875695 rs11574026 rs11574027 rs11168287 **rs4760648** **rs2853564** **rs2254210** rs2228570 rs2239186 rs3782905

		rs3819545 rs2189480 rs886441 **rs12717991** **rs2239179** rs2239180 rs2107301 rs2239182 rs2248098 rs11574077 rs2239185 rs1544410 rs3847987 rs731236 rs12721364
Vitamin D related genes, serum vitamin D concentrations and prostate cancer risk [12] (Ref.2)	28	rs121721364 rs7968585 **rs11574143** **rs757343** **rs1544410** rs7967152 rs11574077 rs2239182 rs2107301 rs2283342 rs2239181 rs1540339 rs2239179 rs886441 rs2189480 rs3819545 rs2239186 rs10735810 rs2254210 rs2238136 rs2853564 rs4760648 rs4328262

		rs4334089 rs4237855 rs11574026 rs3890733 rs7299460
Genetic Polymorphisms in Vitamin D Receptor VDR/RXRA Influence the likelihood of Colon Adenoma Recurrence [14] (Ref.3)	13	**rs11574143** **rs731236** **rs1544410** **rs11574026** **rs10875695** **rs11168293** **rs4760655** **rs7299460** **rs4760658** **rs4516035** **rs11568820** **rs7310552** **rs7970314**



5.3 智慧型三目標基因演算法

由 4.2 節得知目標一與目標二互不衝突，加上 Huang *et al.* [8] 所提出的多目標演算法還是屬於雙目標演算法求得最佳解。我們將 IMOGA 加以改良，以一次可以考量三個目標的多目標演算法來求得最佳解集。圖 24 為智慧型三目標基因算法(ITOGA)的流程圖。評估初始值，將一次考慮目標一、目標三及目標四。

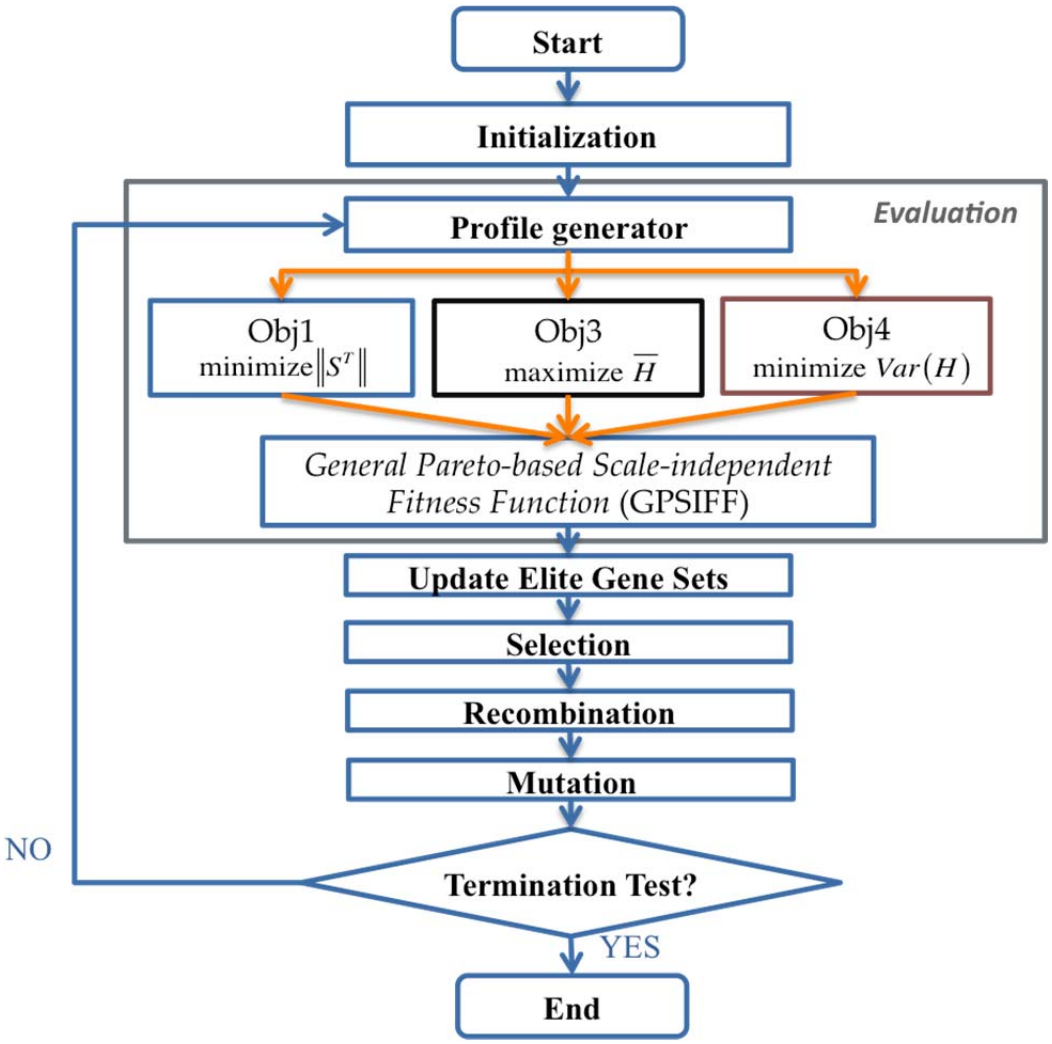


圖 24 智慧型三目標基因演算法流程圖

輸入： N_{pop} 族群大小

N_E 最大容量的菁英集合

p_s 選擇比例

p_c 交配機率

p_m 突變機率

γ IGC 運算因素個數

輸出： E　　未被支配解集

智慧型三目標基因演算法之流程，其步驟詳細敘述如下：

步驟一：(初始化，Initialization)亂數產生初始用的族群數量N_{pop}個染色體以及兩個空的優秀基因集合，一個是E；一個是E'。

步驟二：(評估初始值，Fitness Evaluation)計算族群裡全部染色體的兩項目標函數值，並藉由 GPSIFF 分配每條染色體一個評估值。

步驟三：(更新優秀基因集合，Elite sets)將未被支配(non-dominated)的染色體同時丟入E和E'，然後清空E'。考量E中的所有染色體，將被支配(dominated)的染色體移除。若N_E的數量大於原本的所設定數量，則將從亂數去除超過的部分。

步驟四：(挑選，Selection) 從族群裡用 binary tournament selection 挑選出$N_{pop} - N_{ps}$個染色體，並從E中亂數挑選出N_{ps}個染色體形成一組新的群組。其中$N_{ps} = N_{pop} \times P_s$，若$N_{ps} > N_E$，則令$N_p = N_E$。

步驟五：(重組，Crossover) 藉由 Intelligent Gene Collector (IGC)運作從$N_{pop} \times P_c$選擇親代。每次 IGC 皆是由 OA 重組因子(副產物)找出未被支配的染色體以及兩個子代加入至E'。

步驟六：(突變，Mutation)根據P_m對整個族群進行突變機制。

步驟七：(終止條件，Termination Test)假設已滿足停止條件即可停止演算，反之回到步驟二。

5.4 智慧型三目標基因演算法求得 VDRTag SNPs 最佳解集合

由第 4 章節得知，目標一與目標二互不不衝突，當 Tag SNPs 選取望小時，Haplotype 樣式中 SNP 遺失資料的部份不會增加，隨著選取 Tag SNPs 越大時，需要容錯的值也會愈大，反之，選取最少的 Tag SNPs，所需要的容錯值也會越來越小。

推論，只需三個目標值即可以找到 Tag SNPs 的最佳解集合，分別是目標一、目標三及目標四。經由智慧型三目標基因演算(ITOGA)法可求得 Tag SNPs 的最佳解集合；圖 25 為三目標曲面圖，目標一 x 軸為 Tag SNPs 的數目，望小；目標三 y 軸為 Haplotype 平均漢明距離，望大；目標四 z 軸為 Haplotype 變異漢明距離，望小。黃色圈圈部分為 Pareto 最佳解集合分布。

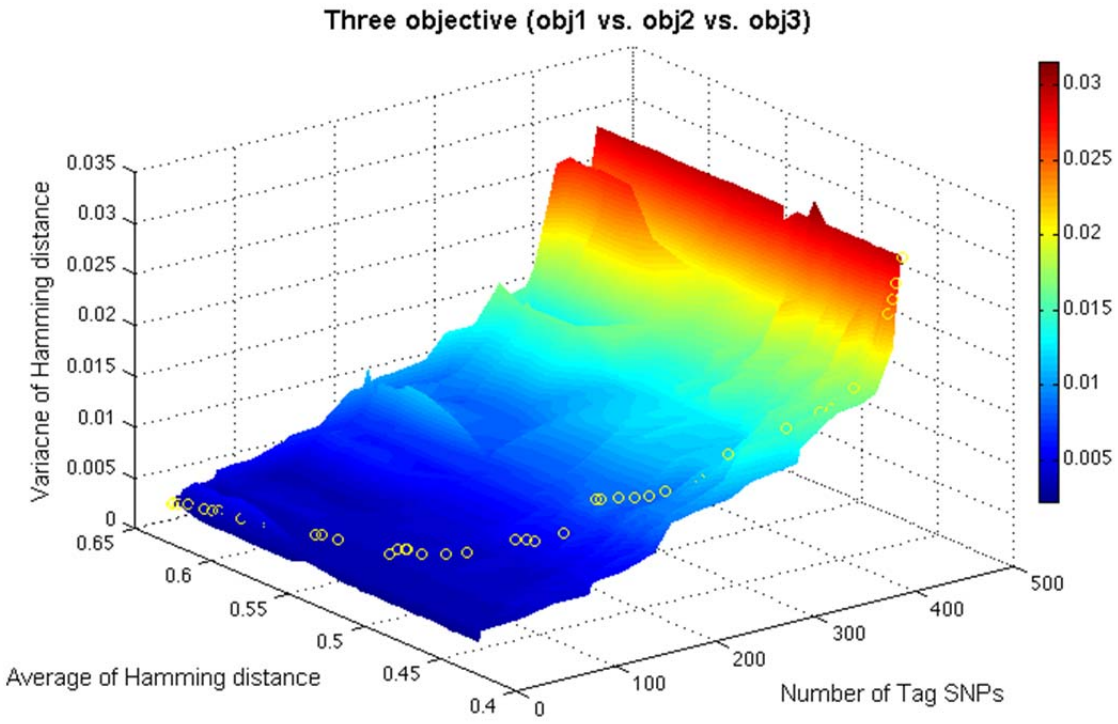


圖 25 智慧型三目標基因演算法實驗 3D 曲面圖

所得出的 Tag SNPs 的最佳解集合為 14、15、16、17、18、20、24、28，如表 10 列出選取到的 Tag SNPs 集合；斜體字型部分為 Tag SNPs 最佳解集，以 14 解集為例，最佳解集合中{*rs4760648*、*rs2853564*、*rs2254210*、*rs12717991*、*rs2239179*、*rs11574143*、*rs757343*、*rs1544410*、*rs11574143*、*rs731236*、*rs1544410*、*rs11574026*、*rs10875695*、*rs7299460*}，與 5.2 章節中文獻中提到的重要 Tag SNPs，經還原比對後，所得到的 Tag SNPs rsID 符合文獻中所提到的 Tag SNPs；在考量三目標最佳化的情況下，我們所提出的 ITOGA 以三目標的方式確實可以找到 Tag SNPs 最佳解集合。

圖 26 為 ITOGA 所得到的 Tag SNPs 統計次數，統計後選取到的 Tag SNPs 最佳解集合中的解發現，第 14 的解集合的 Tag SNPs 出現頻率最高最為重要。

表 10 ITOGA 選取維生素 D 受體基因 Tag SNPs 最佳解集合

Selection Tag SNPs (ITOGA)	Tag SNPs
14	rs4760648,rs2853564,rs2254210,rs12717991,rs2239179,rs11574143,rs757343,rs1544410,rs11574143,rs731236,rs1544410,rs11574026,rs10875695,rs7299460
15	rs4760648,rs2853564,rs2254210,rs12717991,rs2239179,rs11574143,rs757343,rs1544410,rs11574143,rs731236,rs1544410,rs11574026,rs10875695,rs7299460,rs4516035
16	rs4760648,rs2853564,rs2254210,rs12717991,rs2239179,rs11574143,rs757343,rs1544410,rs11574143,rs731236,rs1544410,rs11574026,rs10875695,rs7299460,rs4516035,rs3890733
17	rs4760648,rs2853564,rs2254210,rs12717991,rs2239179,rs11574143,rs757343,rs1544410,rs11574143,rs731236,rs1544410,rs11574026,rs10875695,rs7299460,rs4516035,rs3890733,rs7136534
18	rs4760648,rs2853564,rs2254210,rs12717991,rs2239179,rs11574143,rs757343,rs1544410,rs11574143,rs731236,rs1544410,rs11574026,rs10875695,rs7299460,rs4516035,rs3890733,rs7136534,rs731236
20	rs4760648,rs2853564,rs2254210,rs12717991,rs2239179,rs11574143,rs757343,rs1544410,rs11574143,rs731236,rs1544410,rs11574026,rs10875695,rs7299460,rs4516035,rs3890733,rs7136534,rs731236,rs1540339,rs2107301
24	rs4760648,rs2853564,rs2254210,rs12717991,rs2239179,rs11574143,rs757343,rs1544410,rs11574143,rs731236,rs1544410,rs11574026,rs10875695,rs7299460,rs4516035,rs3890733,rs7136534,rs731236,rs1540339,rs2107301,rs2239182,rs11168275,rs4760648,rs2238136
28	rs4760648,rs2853564,rs2254210,rs12717991,rs2239179,rs11574143,rs757343,rs1544410,rs11574143,rs731236,rs1544410,rs11574026,rs10875695,rs7299460,rs4516035,rs3890733,rs7136534,rs731236,rs1540339,rs2107301,rs2239182,rs11168275,rs4760648,rs2238136,rs4237855,rs2239179,rs10875693,rs2189480
41	rs4760648,rs2853564,rs2254210,rs12717991,rs2239179,rs11574143,rs757343,rs1544410,rs11574143,rs731236,rs1544410,rs11574026,rs10875695,rs7299460,rs4516035,rs3890733,rs7136534,rs731236,rs1540339,rs2107301,rs2239182,rs11168275,rs4760648,rs2238136,rs4237855,rs2239179,rs10875693,rs2189480,rs4328262,rs381

	9545,rs10735810,rs7299460,rs11168292,rs757343,rs10875695,rs1 1574027,rs731236,rs2239182,rs2254210,rs4328262,rs4760648
43	*rs4760648,rs2853564,rs2254210*,rs12717991,*rs2239179*,rs115741 43,rs757343,rs1544410,*rs11574143*,*rs731236*,*rs1544410,rs115740 26*,rs10875695,*rs7299460*,rs4516035,rs3890733,rs7136534,rs7312 36,rs1540339,*rs2107301,rs2239182*,rs11168275,rs4760648,rs2238 136,rs4237855,rs2239179,rs10875693,rs2189480,rs4328262,rs381 9545,rs10735810,rs7299460,rs11168292,rs757343,rs10875695,rs1 1574027,rs731236,rs2239182,rs2254210,rs4328262,rs4760648,rs2 238136,rs2239179
45	*rs4760648,rs2853564,rs2254210*,rs12717991,*rs2239179*,rs115741 43,rs757343,rs1544410,*rs11574143*,*rs731236*,*rs1544410,rs115740 26*,rs10875695,*rs7299460*,rs4516035,rs3890733,rs7136534,rs7312 36,rs1540339,*rs2107301,rs2239182*,rs11168275,rs4760648,rs2238 136,rs4237855,rs2239179,rs10875693,rs2189480,rs4328262,rs381 9545,rs10735810,rs7299460,rs11168292,rs757343,rs10875695,rs1 1574027,rs731236,rs2239182,rs2254210,rs4328262,rs4760648,rs2 238136,rs2239179,rs3819545,rs2283342
56	*rs4760648,rs2853564,rs2254210*,rs12717991,*rs2239179*,rs115741 43,rs757343,rs1544410,*rs11574143*,*rs731236*,*rs1544410,rs115740 26*,rs10875695,*rs7299460*,rs4516035,rs3890733,rs7136534,rs7312 36,rs1540339,*rs2107301,rs2239182*,rs11168275,rs4760648,rs2238 136,rs4237855,rs2239179,rs10875693,rs2189480,rs4328262,rs381 9545,rs10735810,rs7299460,rs11168292,rs757343,rs10875695,rs1 1574027,rs731236,rs2239182,rs2254210,rs4328262,rs4760648,rs2 238136,rs2239179,rs3819545,rs2283342,rs7299460,rs2239186,rs1 1574077,rs10875693,rs1540339,rs2107301,rs2853564,rs4237855,r s2189480,rs10735810,rs11168275
58	*rs4760648,rs2853564,rs2254210*,rs12717991,*rs2239179*,rs115741 43,rs757343,rs1544410,*rs11574143*,*rs731236*,*rs1544410,rs115740 26*,rs10875695,*rs7299460*,rs4516035,rs3890733,rs7136534,rs7312 36,rs1540339,*rs2107301,rs2239182*,rs11168275,rs4760648,rs2238 136,rs4237855,rs2239179,rs10875693,rs2189480,rs4328262,rs381 9545,rs10735810,rs7299460,rs11168292,rs757343,rs10875695,rs1 1574027,rs731236,rs2239182,rs2254210,rs4328262,rs4760648,rs2 238136,rs2239179,rs3819545,rs2283342,rs7299460,rs2239186,rs1 1574077,rs10875693,rs1540339,rs2107301,rs2853564,rs4237855,r s2189480,rs10735810,rs11168275,rs10459228,rs10783219

62	*rs4760648,rs2853564,rs2254210*,rs12717991,*rs2239179*,rs115741 43,rs757343,rs1544410,*rs11574143*,*rs731236*,*rs1544410,rs115740 26*,rs10875695,*rs7299460*,rs4516035,rs3890733,rs7136534,rs7312 36,rs1540339,*rs2107301,rs2239182*,rs11168275,rs4760648,rs2238 136,rs4237855,rs2239179,rs10875693,rs2189480,rs4328262,rs381 9545,rs10735810,rs7299460,rs11168292,rs757343,rs10875695,rs1 1574027,rs731236,rs2239182,rs2254210,rs4328262,rs4760648,rs2 238136,rs2239179,rs3819545,rs2283342,rs7299460,rs2239186,rs1 1574077,rs10875693,rs1540339,rs2107301,rs2853564,rs4237855,r s2189480,rs10735810,rs11168275,rs10459228,rs10783219,rs1116 8287,rs2228570,rs3782905,rs12717991
63	*rs4760648,rs2853564,rs2254210*,rs12717991,*rs2239179*,rs115741 43,rs757343,rs1544410,*rs11574143*,*rs731236*,*rs1544410,rs115740 26*,rs10875695,*rs7299460*,rs4516035,rs3890733,rs7136534,rs7312 36,rs1540339,*rs2107301,rs2239182*,rs11168275,rs4760648,rs2238 136,rs4237855,rs2239179,rs10875693,rs2189480,rs4328262,rs381 9545,rs10735810,rs7299460,rs11168292,rs757343,rs10875695,rs1 1574027,rs731236,rs2239182,rs2254210,rs4328262,rs4760648,rs2 238136,rs2239179,rs3819545,rs2283342,rs7299460,rs2239186,rs1 1574077,rs10875693,rs1540339,rs2107301,rs2853564,rs4237855,r s2189480,rs10735810,rs11168275,rs10459228,rs10783219,rs1116 8287,rs2228570,rs3782905,rs12717991,rs2239180
64	*rs4760648,rs2853564,rs2254210*,rs12717991,*rs2239179*,rs115741 43,rs757343,rs1544410,*rs11574143*,*rs731236*,*rs1544410,rs115740 26*,rs10875695,*rs7299460*,rs4516035,rs3890733,rs7136534,rs7312 36,rs1540339,*rs2107301,rs2239182*,rs11168275,rs4760648,rs2238 136,rs4237855,rs2239179,rs10875693,rs2189480,rs4328262,rs381 9545,rs10735810,rs7299460,rs11168292,rs757343,rs10875695,rs1 1574027,rs731236,rs2239182,rs2254210,rs4328262,rs4760648,rs2 238136,rs2239179,rs3819545,rs2283342,rs7299460,rs2239186,rs1 1574077,rs10875693,rs1540339,rs2107301,rs2853564,rs4237855,r s2189480,rs10735810,rs11168275,rs10459228,rs10783219,rs1116 8287,rs2228570,rs3782905,rs12717991,rs2239180,rs2248098
78	*rs4760648,rs2853564,rs2254210*,rs12717991,*rs2239179*,rs115741 43,rs757343,rs1544410,*rs11574143*,*rs731236*,*rs1544410,rs115740 26*,rs10875695,*rs7299460*,rs4516035,rs3890733,rs7136534,rs7312 36,rs1540339,*rs2107301,rs2239182*,rs11168275,rs4760648,rs2238 136,rs4237855,rs2239179,rs10875693,rs2189480,rs4328262,rs381

	9545,rs10735810,rs7299460,rs11168292,rs757343,rs10875695,rs1 1574027,rs731236,rs2239182,rs2254210,rs4328262,rs4760648,rs2 238136,rs2239179,rs3819545,rs2283342,rs7299460,rs2239186,rs1 1574077,rs10875693,rs1540339,rs2107301,rs2853564,rs4237855,r s2189480,rs10735810,rs11168275,rs10459228,rs10783219,rs1116 8287,rs2228570,rs3782905,rs12717991,rs2239180,rs2248098,rs22 39185,rs3847987,rs12721364,rs121721364,rs7968585,rs7967152,r s2239181,rs4334089,rs11168293,rs4760655,rs4760658,rs1156882 0,rs7310552, rs7970314
92	*rs4760648,rs2853564,rs2254210*,rs12717991,*rs2239179*,rs115741 43,rs757343,rs1544410,*rs11574143*,*rs731236*,*rs1544410,rs115740 26*,rs10875695,*rs7299460*,rs4516035,rs3890733,rs7136534,rs7312 36,rs1540339,*rs2107301,rs2239182*,rs11168275,rs4760648,rs2238 136,rs4237855,rs2239179,rs10875693,rs2189480,rs4328262,rs381 9545,rs10735810,rs7299460,rs11168292,rs757343,rs10875695,rs1 1574027,rs731236,rs2239182,rs2254210,rs4328262,rs4760648,rs2 238136,rs2239179,rs3819545,rs2283342,rs7299460,rs2239186,rs1 1574077,rs10875693,rs1540339,rs2107301,rs2853564,rs4237855,r s2189480,rs10735810,rs11168275,rs10459228,rs10783219,rs1116 8287,rs2228570,rs3782905,rs12717991,rs2239180,rs2248098,rs22 39185,rs3847987,rs12721364,rs121721364,rs7968585,rs7967152,r s2239181,rs4334089,rs11168293,rs4760655,rs4760658,rs1156882 0,rs7310552,rs7970314,rs739837,rs7136534,rs2853559,rs1116826 6,rs2239184,rs11574114,rs6580642,rs7954412,rs11574049,rs1157 4029,rs11574038,rs11574032,rs11574015,rs11574048

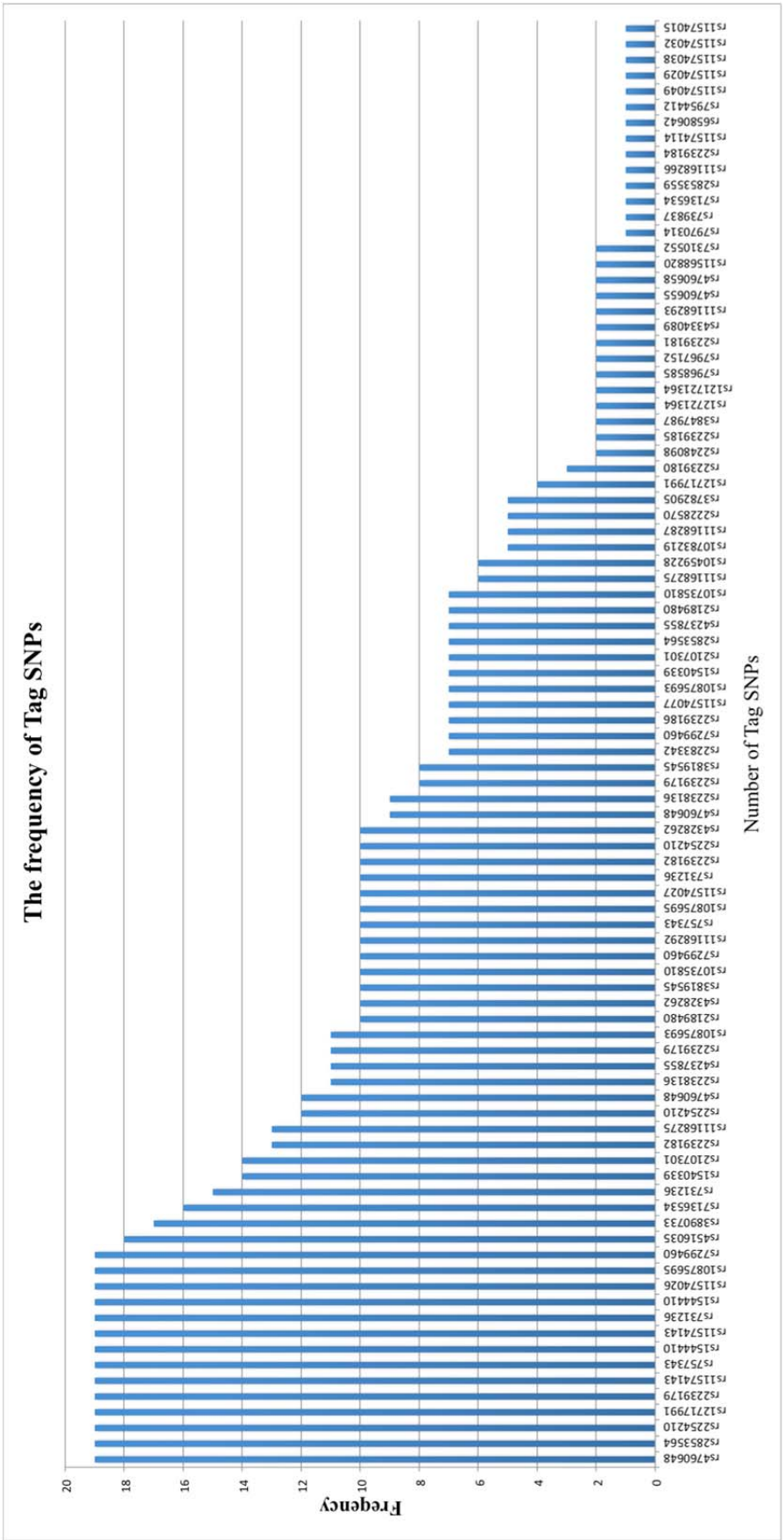


圖 26 維生素 D 受體基因 Tag SNPs 出現次數頻率統計

六、問題討論與展望

6.1 討論

由於 Tag SNPs 選取問題，一般會依使用的目的不同而有不同的模式表現型態，若能同時考慮多個不同目標，則可增加其使用的範圍。在本研究裡我們使用了最佳化的演化式演算法 IMOGA 及 ITOGA 去選取標籤單核苷酸型性，在四個目標函數中，(1)選取 Tag SNPs 最小化。(2)Tag SNPs 資料遺失容錯最大化。(3)Haplotype 樣式相異性最大化。(4)Haplotype 樣式多樣性最小化 [8]。利用我們所提出的 ITOGA 將三個目標函數計算出最佳的參數組合，並比較 Huang et al. [8] 所提出使用 NSGA2 演算法；在第四章中經由覆蓋測量，我們清楚的得知 IMOGA 演算法在四個目標函數計算中遠遠優於 NSGA2 演算法。

我們發現目標一選取 Tag SNPs 最小化，及目標二 Tag SNPs 資料遺失容錯最大化，是沒有互相衝突的，當生物晶片所篩選的 SNP 資料越多時，資料遺失的錯誤率也相對提高，所以在計算選取籤單核苷酸型性資料時，相對的容錯能裡也需提高。

本研究中所使用的資料來源，都是從國際單型圖譜計畫(The International HapMap Project)中蒐集的資料，在無法取得生物、醫學疾病上的實驗資料，只能做單向分析。因所搜集到的資料都是 Haplotype 樣式或是 Haplotype 配對，對於選取出的 Tag SNPs，要還原成 SNP 是有難度的。

6.2 未來展望

目前已發現的 SNP 資料龐大，但微陣列晶片探針的密度有限，不可能將所有的 SNP 資料存入晶片中做應用，再加上相關研究亦指出我們其實不需要儲存如此大規模的 SNP 資料；相反地，僅由挑選出這些資料的一小部分，即可代表原始資料所欲呈現的大部分資訊。此乃由於鹼基的突變會使受驗者鹼基上的表現異於正常個體，而藉由篩選 SNP 的過程來發現這些特殊的鹼基即可找出真正影響疾病的鹼基。Tag SNPs，其功能乃以部分的 SNP 序列來代表原本完整的 SNP 序列，可以有效地減少資料儲存空間，並以更短的時間進行研究分析，如此便能使其在微陣列探針晶片上做應用。

我們的研究中除了可以準確的選出 Tag SNPs 的最佳解集合，在未來如果也可獲得更多全基因體微陣列探針晶片所提供的實驗數據，進行案例控制的關聯性研究，在醫學疾病上對於解析更複雜的疾病基因組成中，SNP 上篩選，

有更大的貢獻及研究。

參考文獻

[1] P. Nowotny, *et al.*, "SNP analysis to dissect human traits," *Curr Opin Neurobiol,* vol. 11, pp. 637-41, Oct 2001.

[2] B. S. Shastry, "SNP alleles in human disease and evolution," *J Hum Genet,* vol. 47, pp. 561-6, 2002.

[3] G. Gibson and S. V. Muse, *A Primer of Genome Science*, Third Edition ed.: Sinauer Associates, Inc, 2009.

[4] C. S. Carlson, *et al.*, "Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium," *Am J Hum Genet,* vol. 74, pp. 106-20, Jan 2004.

[5] B. Devlin and N. Risch, "A comparison of linkage disequilibrium measures for fine-scale mapping," *Genomics,* vol. 29, pp. 311-22, Sep 20 1995.

[6] M. J. Daly, *et al.*, "High-resolution haplotype structure in the human genome," *Nat Genet,* vol. 29, pp. 229-32, Oct 2001.

[7] Y. T. Huang, *et al.*, "An approximation algorithm for haplotype inference by maximum parsimony," *J Comput Biol,* vol. 12, pp. 1261-74, Dec 2005.

[8] C. K. Ting, *et al.*, "Multi-objective tag SNPs selection using evolutionary algorithms," *Bioinformatics,* vol. 26, pp. 1446-52, Jun 1 2010.

[9] G. A. Thorisson, *et al.*, "The International HapMap Project Web site," *Genome Res,* vol. 15, pp. 1592-3, Nov 2005.

[10] S. Y. Ho, *et al.*, "Intelligent evolutionary algorithms for large parameter optimization problems," *Ieee Transactions on Evolutionary Computation,* vol. 8, pp. 522-541, Dec 2004.

[11] N. Srinivas and K. Deb, "Muiltiobjective optimization using nondominated sorting in genetic algorithms," *Evol. Comput.,* vol. 2, pp. 221-248, 1994.

[12] J. Ahn, *et al.*, "Vitamin D-related genes, serum vitamin D concentrations and prostate cancer risk," *Carcinogenesis,* vol. 30, pp. 769-76, May 2009.

[13] S. Karami, *et al.*, "Analysis of SNPs and haplotypes in vitamin D pathway genes and renal cancer risk," *PLoS One,* vol. 4, p. e7013, 2009.

[14] J. B. Egan, *et al.*, "Genetic polymorphisms in vitamin D receptor VDR/RXRA influence the likelihood of colon adenoma recurrence," *Cancer Res,* vol. 70, pp. 1496-504, Feb 15 2010.

[15] V. Bafna, *et al.*, "Haplotypes and informative SNP selection algorithms: don't block out information," presented at the Proceedings of the seventh annual international conference on Research in computational molecular biology, Berlin, Germany, 2003.

[16] X. Ke and L. R. Cardon, "Efficient selective screening of haplotype tag SNPs," *Bioinformatics,* vol. 19, pp. 287-8, Jan 22 2003.

[17] N. Patil, *et al.*, "Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21," *Science,* vol. 294, pp. 1719-23, Nov 23 2001.

[18] K. Zhang, *et al.*, "A dynamic programming algorithm for haplotype block partitioning," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 99, pp. 7335-9, May 28 2002.

[19] K. Zhang and L. Jin, "HaploBlockFinder: haplotype block analyses," *Bioinformatics,* vol. 19, pp. 1300-1, Jul 1 2003.

[20] K. Zhang, *et al.*, "Haplotype block structure and its applications to association studies: power and study designs," *Am J Hum Genet,* vol. 71, pp. 1386-94, Dec 2002.

[21] M. E. Weale, *et al.*, "Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping," *Am J Hum Genet,* vol. 73, pp. 551-65, Sep 2003.

[22] E. Halperin, *et al.*, "Tag SNP selection in genotype data for maximizing SNP prediction accuracy," *Bioinformatics,* vol. 21 Suppl 1, pp. i195-203, Jun 2005.

[23] H. I. Avi-Itzhak, *et al.*, "Selection of minimum subsets of single nucleotide polymorphisms to capture haplotype block diversity," *Pac Symp Biocomput,* pp. 466-77, 2003.

[24] Y. T. Huang, *et al.*, "Selecting additional tag SNPs for tolerating missing data in genotyping," *BMC Bioinformatics,* vol. 6, p. 263, 2005.

[25] M. R. Garey and D. S. Johnson, *Computers and Intractability; A Guide to the Theory of NP-Completeness*. New Yourk, NY, USA: W. H. Freeman & Co., 1990.

[26] V. Chvatal, "A Greedy Heuristic for the Set-Covering Problem," *Mathematics of Operations Research,* vol. 4, pp. 233-235, 1979.

[27] P. Slavik, "A tight analysis of the greedy algorithm for set cover," *Journal of Algorithms,* vol. 25, pp. 237-254, Nov 1997.

[28] M. Hifi, *et al.*, "A neural network for the minimum set covering problem," *Chaos Solitons & Fractals,* vol. 11, pp. 2079-2089, Oct 2000.

[29] G. Mahdevar, *et al.*, "Tag SNP selection via a genetic algorithm," *J Biomed Inform,* vol. 43, pp. 800-4, Oct 2010.

[30] G. i. Taguchi and S. Konishi, *Taguchi Methods Orthogonal Arrays and Linear Graphs: Tools for Quality Engineering*: Amer Supplier Inst, 1987.

[31] M. Stephens, *et al.*, "A new statistical method for haplotype reconstruction

from population data," *Am J Hum Genet,* vol. 68, pp. 978-89, Apr 2001.

[32] E. Zitzler and L. Thiele, "Multiobjective evolutionary algorithms: A comparative case study and the Strength Pareto approach," *Ieee Transactions on Evolutionary Computation,* vol. 3, pp. 257-271, Nov 1999.

[33] M. F. Holick, "Sunlight and vitamin D for bone health and prevention of autoimmune diseases, cancers, and cardiovascular disease," *Am J Clin Nutr,* vol. 80, pp. 1678S-88S, Dec 2004.