分析大量表現序列標籤重組基因結構之研究

Gene Structure Assembly through Analysis of Large-Scale ESTs

游竣棠
亞洲大學資訊工程所
cfp@ms27.url.com.tw

許芳榮
逢甲大學資訊工程學系
frhsu@fcu.edu.tw

夏偉中
逢甲大學資訊工程學系
p9432656@webmail.fcu.edu.tw

摘要

表現序列標籤的分群是發現未知基因和已知基因功能的研究中的一個重要課題。唯一基因序列收集是傳統上著名的分群方法，早期人類基因體尚未定序完成時，此方法成效非凡。但現今人類基因體定序已接近完成，對齊演算法也非常優良，因此表現序列標籤直接對齊至基因體，重疊形成的群，理論上應該會比較好。

我們根據表現序列標籤對齊至基因體之結果，找出重疊形成的群，再利用群中包含的表現序列標籤來重組基因結構。根據我們的分析，在群的方面，發現選擇性裁切比 Unigene 序列收集有意義；在結構的方面，外顯子和內含子結構有六成以上的準確率。而其他方析，如尋找未知基因和基因融合。在未知基因方面，我們的結果顯示至少有六個可能是未知基因；另外[1]在基因融合方面，發現至少有五種現象，而其中有一種現象已被證實。

關鍵詞：表現序列標籤、群、基因結構、未知基因、基因融合

一、 前言

由於人類基因體計畫[1]已接近完成，美國國立生物技術資訊中心（NCBI）已經提供了相當大量的序列資料可供分析下載。而其中對於基因體研究最重要的資料即是表現序列標籤(Expressed Sequence Tag, EST)[2]。

在基因體中，DNA 經過轉錄成為

* This work is supported in part by the National Science Council, Taiwan, R.O.C, grant NSC 93-3112-B-468-001

pre-mRNA 後，由於選擇性裁剪的發生，pre-mRNA 會被分為外顯子(Exon) 與內顯子 (Intron)，並且保留 Exon 形成 mRNA。mRNA 和 EST 是來自基因體的已知或未知基因，而 EST 是 mRNA 的片斷，所以 EST 可以拼成全長或接近全長的 mRNA。然而基因結構與轉譯蛋白質、基因的功能、疾病的影響、基因與基因的相互作用和尋找未知基因是有關係的，所以基因的結構組成是非常重要的。

在早期基因重組的研究有 TAP [13]，ESTGenes [3]，ECgene [7]等。這些基因結構的準確度已有不錯的結果，然而為何我們仍要重組基因結構呢？主要是因為使用的對齊程式不同、分群方法不同和重組基因結構的方法不同，所以導致基因結構有所不同。更重要的是我們有不同的發現，如尋找未知基因、基因融合等。

二、 方法

本論文利用大量表現序列標籤重建完整的基因結構，重建的步驟可分成四個部分。

第一個部分：資料收集。資料有人類染色體的序列 (版本編號為 Homo_sapiens genome Build 35.1)、表現序列標籤(dbEST 2005/01/27) 和 mRNA。

第二部分：對齊。利用對齊程式，如 Mugup[5]、Blast[9]、Sim4[6]，把人類的表現序列標籤對齊到人類基因組。

第三部分：分群。由於 EST 已對齊到基因組，而 EST 散佈在基因組的各地，因此把有關聯的 EST 分成一群，而群可能是一個已知基因或未知基因的部分或全部。

第四部份：建構基因結構。利用第三部分之分群結果，重組每個群的外顯子和內含子的結構。

在分群和建立基因結構這兩部分，於下面分項詳細說明。

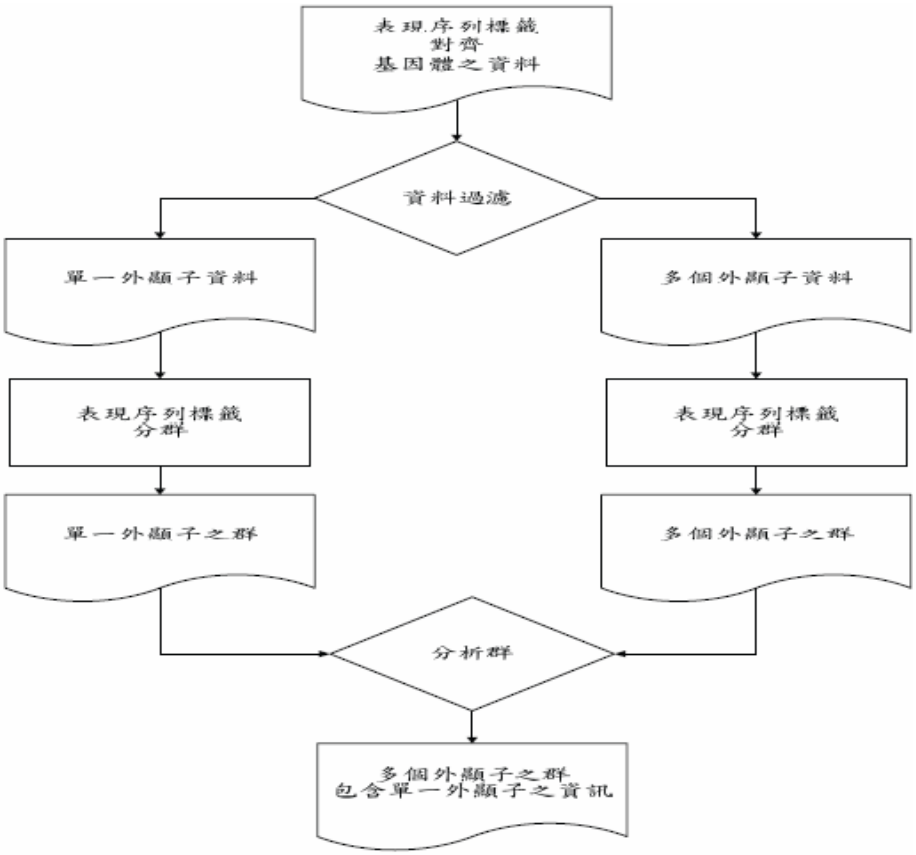

圖 1：分群流程架構圖

分群：分群建立方法分成三個階段並依序說明，流程架構圖如上圖 1 所示：

第一階段：將 EST 對齊到基因體的資料過濾，以建立 Avatar[4]所定的嚴格標準為過濾原則，輔予以保留，例如 EST 對齊至基因體的相似度 94 分以上並且 EST 內含子切位[8]為 GT/AG、GC/AG、CT/AC 和 CT/GC 四種並且佔總內含子數的 80%。再依 EST 的單一外顯子和多個外顯子分成兩類。

第二階段：EST 分群。在單一外顯子和多個外顯子這兩類個別獨立分群。單一外顯子這類分群是利用每一條 EST 在基因體上的坐標，將有重疊的 EST 分成同一群，沒有重疊的 EST 也分成一群，如下圖所示。而多個外顯子這類分群比單一外顯子分群多個，其根據位於在基因體的正股或反股分群，如下圖 2 所示。

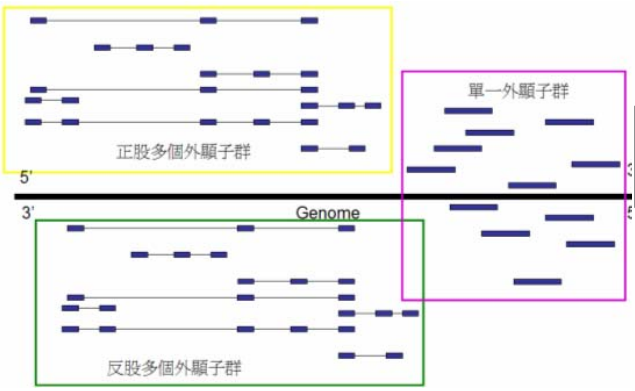

圖 2：分群示意圖

第三階段：分析分群資料。根據單一外顯子和多個外顯子這兩類所建立的群，將兩類的群作交集且加入可以利用的單一外顯子群至多個外顯子群，如下圖 3 所示。

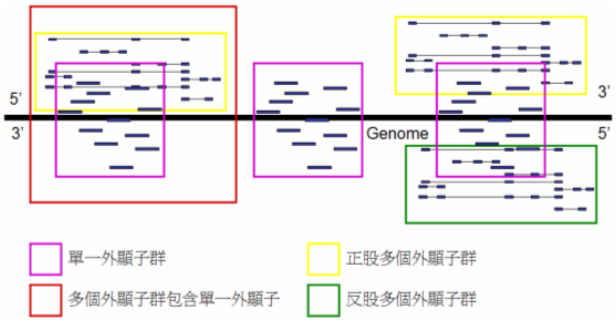

圖 3：分析群資料之示意圖

建立基因結構：基因表現之結構，由分群所包含的表現序列標籤建立，而結構建立可分成五個階段，如下依序說明，建立流程架構圖如下圖 4 所示：

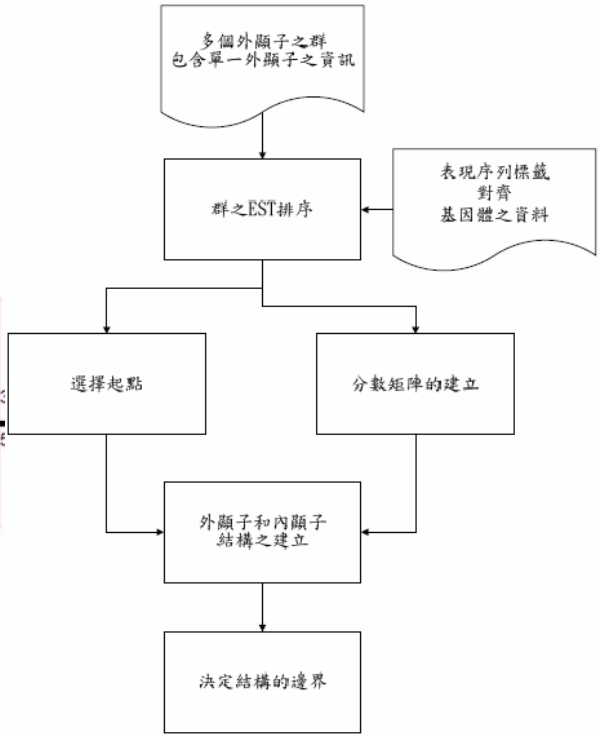

圖 4：外顯子和內顯子建立流程架構圖

第一階段：群之 EST 排序。以每一條 EST 的第一個外顯子在基因體上的位置由小到大排序，如下圖 5 所示。

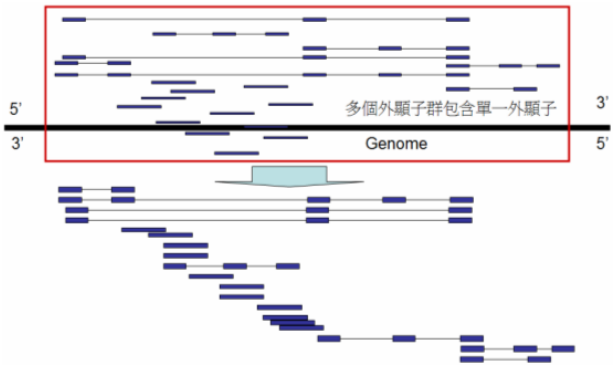

圖 5：群之 EST 排序之示意圖

第二階段：選擇起點。在 EST 已排序的群中，主要以多外顯子的 EST 為主，挑選可能的路徑起點，挑選為起點的多外顯子 EST 的第一個內含子並不與群中任何多外顯子 EST 的內含子相同，如下圖 6 紅色圓點所示。

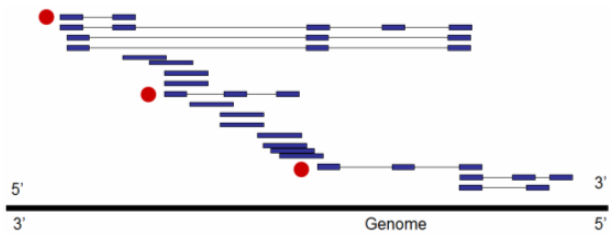

圖 6：選擇起點

第三階段：分數矩陣的建立。結構之建立所依據的分數矩陣。分數規則有五種，說明如下。

第一種：完整重疊。是指兩條 EST 對齊到基因組序列上的結果相同，如下圖 7 所示，第一個內含子要到第二個內含子可經由這兩條 EST，所得分數為 2。

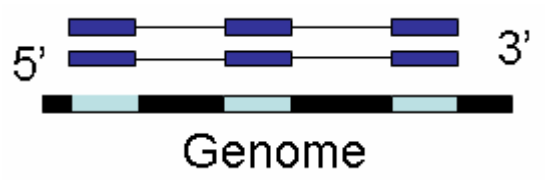

圖 7：分數規則-完整重疊

第二種：互斥。是指兩條 EST 對齊到基因組序列上的結果不能使用，如下圖 8 所示，第一條 EST 的第一個內含子包含第二條的第一個內含子和第二個內含子，因此不能同時使用，所得分數為負無限大。

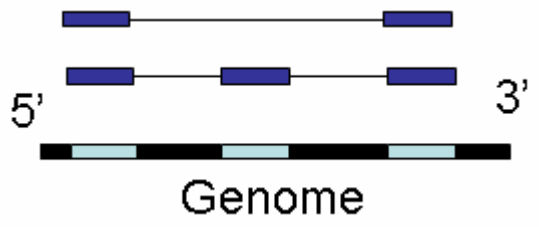

圖 8：分數規則-互斥

第三種：重疊無單一外顯子。是指兩條 EST 對齊到基因組序列上，一條 EST 的最後一個外顯子重疊到另外一條 EST 的第一個外顯子，如下圖 9 所示，第一個內含子要到第二個內含子可經由兩條 EST 的部分重疊，所得分數為 1。

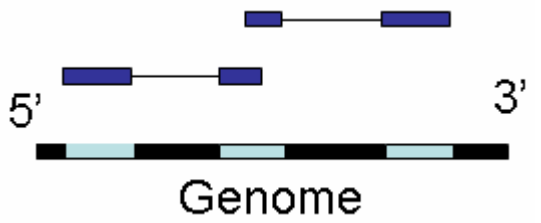

圖 9：分數規則-重疊無單一外顯子

第四種：重疊有單一外顯子。是指兩條 EST 對齊到基因組序列上，一條 EST 的最後一個外顯子無重疊到另外一條 EST 的第一個外顯子，可由某一條單一外顯子的 EST 連接，如下圖 10 所示，第一個內含子到第二個內含子，雖然兩 EST 沒有部分重疊，但有某條單一外顯子 EST 與兩條 EST 部分重疊，所得分數為 1。

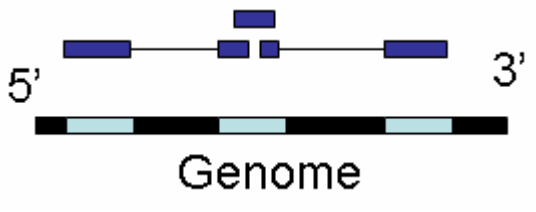

圖 3-10：分數規則-重疊有單一外顯子

第五種：無重疊。是指兩條 EST 對齊到基因組序列上，一條 EST 的最後一個外顯子無重疊到另外一條 EST 的第一個外顯子，如下圖 11 所示，第一個內含子到第二個內含子，兩 EST 沒有部分重疊，且沒有某條單一外顯子 EST 與兩條 EST 部分重疊，所得分數為 0。

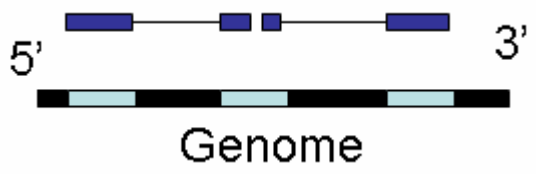

圖 11：分數規則-無重疊

以下圖 12 作為一個範例並說明，此群有 7 條 EST，其編號為 A 到 G，其中有 6 條多個外顯子的 EST 和 1 條單一外顯子的 EST，形成 6 個內含子，其編號為 1 到 6。從 A 這一條 EST 開始走，當要從內含子 1 到內含子 2，只有 B/C 這二條與 A 有重疊的 EST，所以在矩陣 M 的（1,2）位置值為 2，如下表；當從內含子 1 到內含子 3 時只有一條 EST D 與 A 重疊，所以在矩陣 M 的（1,3）位置值為 1；當從內含子 2 到內含子 3，因內含子 2 與內含子 3 互斥，所以在矩陣 M 的（2,3）位置值為負無窮大；當從內含子 2 到內含子 4 只有兩條 EST D 和 C 重疊，所以在矩陣 M 的（2,4）位置值為 2；當內含子 3 到內含子 4，因內含子 3 與內含子 4 互斥，所以在矩陣 M 的（3,4）位置值為負無窮大；當從內含子 3 到內含子 5 只有一條 EST E 與 D 重疊，所以在矩陣 M 的（3,5）位置值為 1；當從內含子 4 到內含子 5 只有一條 EST E，所以在矩陣 M 的（4,5）位置值為 1；當從內含子 5 到內含子 6，雖然 E 與 G 兩條 EST 無重疊，但靠著單一外顯子 F 相連，所以在矩陣 M 的（5,6）位置值為 1。

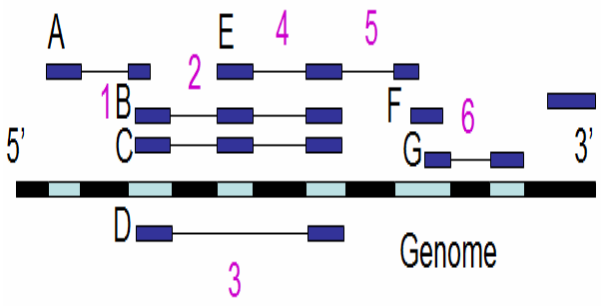

圖 12：分數規則之範例

	Begin	1	2	3	4	5	6	End
Begin		1						
1			2	1				
2				-∞	2			
3					-∞	1		
4						1		
5							1	
6								0
End								

表 1：分數規則之範例矩陣分數表

第四階段：結構之建立：根據分數矩陣，逐一建立外顯子和內含子的結構，並依據所得分數的高低，保留分數高的五個基因結構。如上面範例所建立的分數矩陣來建立結構，這個例子會形成兩個結構，所以兩個結構都予以保留，一個結構建立是 Begin ➔ 1 ➔ 2 ➔ 4 ➔ 5 ➔ 6 ➔ End，分數為 7；另一個結構建立是 Begin ➔ 1 ➔ 3 ➔ 5 ➔ 6 ➔ End，分數為 4。

第五階段：判斷建立之結構的邊界：在結構邊界的判斷，使用統計方法。在 5' 端的邊界，統計 EST 支持的數目，選出

EST 支持最高的邊界為結構 5'的邊界；而在 3'端的邊界，亦使用相同方式。

三、 結果

從 dbEST 資料庫取得六佰萬條人類的 EST 序列，經由 Mugup 對齊基因體並過濾後，符合標準的 EST 序列有四佰萬條，其中多個外顯子有一佰八十萬條而單一外顯子有二佰二十萬條。根據符合標準的資料所作之成果分析分成四節：第一節主要探討本分群方法與 NCBI UniGene 之關係，第二節主要分析重組之基因結構與標準基因結構之關係，第三節主要尋找可能的基因，第四節主要尋找基因融合的情形。

第一節：群之分析

探討 mRNA 分群與 NCBI UniGene[11]之間的關係，來驗證 EST 分群。

為了要了解使用我們的分群方法和 UniGene 分群方法有何不同，我們使用 22952 條 mRNA 作為測試集合，mRNA 的編號包括開頭為 NM、NR、XM、XM 等四種。

令 M={ m_1 , m_2 ,……, m_k } 為 Mugup 所分的群，m_i 代表第 i 個群。

令 U={ u_1 , u_2 ,……, u_n } 為 UniGene 所分的群，u_j 代表第 j 個群，其中 u_0 代表未分群或在 UniGene Build 181 被刪除。M 與 U 關係如表 2 所示。

我們挑選信任的編號為 NM 共有 17816 條，位於 14116 個群中，以分析其與 UniGene 的關係。其中有 85 個群對應到兩個以上的 UniGene 群。剩下的 14031 的群無對應到兩個以上的 UniGene 群。

分析那些群有對應到兩個以上的 UniGene 群，發現有 10 種情況，且在多數的情況下，都發現可能是選擇性裁切所照成的，而在 UniGene 群則會分成不同群。由此可知我們的分群方法對發現選擇性裁切和相似基因體序列，優於 UniGene 的分群方法。以上 10 種情形的數量，如表 3 所示。

第二節：基因結構之分析

根據分子生物學的中心法則，大量的表現序列標籤理論應該重組回原先的基因結構，我們使用 HMR195[10]的資料來分析基因結構。HMR195 資料包括人、小鼠、大鼠的資料，其中人類資料有 65 個不同基因當作測試集合。

我們人工分析 65 個基因於 UCSC Genome Brower[12]基因結構、Mugup 基因結構和重組的基因結構。經過分析，我們發現 UCSC 基因結構和 Mugup 基因結構大部分相同只有少部分為邊界和缺少外顯子的情形，而重組的基因結構的分析結果如下表 4 所示。

型態	數量
$m_i \subseteq u_j$ ， $\|m_i\|$>=1	17046 個
$m_i \subseteq u_j$ ， $\|m_i\|$>1	2259 個
$m_i \subseteq u_j$ ， $\|m_i\|$=1	14787 個
$m_i \subseteq u_0$ ， $\|m_i\|$>=1 ， u_0：在 UniGene Build 181 被刪除	114 個
$m_i \subseteq u_0$ ， $\|m_i\|$>1 ， u_0：在 UniGene Build 181 被刪除	2 個
$m_i \subseteq u_0$ ， $\|m_i\|$=1 ， u_0：在 UniGene Build 181 被刪除	112 個
$m_i \subseteq u_0$ ， $\|m_i\|$>=1 ， u_0：未分群	1213 個
$m_i \subseteq u_0$ ， $\|m_i\|$>1 ， u_0：未分群	68 個
$m_i \subseteq u_0$ ， $\|m_i\|$=1 ， u_0：未分群	1145 個
$m_i \subseteq u_0$ ， $\|m_i\|$>1 ， u_0：未分群 和 在 UniGene Build 181 被刪除	3 個
$e_1 \in m_i$ ， $e_2 \in m_i$ ，$e_1 \in u_k$ ， $e_2 \in u_0$ ， u_0：在 UniGene Build 181 被刪除	6 個
$e_1 \in m_i$ ， $e_2 \in m_i$ ， $e_1 \in u_k$ ， $e_2 \in u_0$， u_0：未分群	84 個
$e_1 \in m_i$ ， $e_2 \in m_i$ ， $e_1 \in u_{k_1}$ ， $e_2 \in u_{k_2}$ ， $k_1 != k_2$	191 個

表 2：mRNA 在分群與 UniGene 181 之關係

種類	數量
兩段基因體序列相似	6
暫時移除的 mRNA	9
被取代的 mRNA	7
有問題的 mRNA	4
群中有某一條 mRNA 在 NCBI 對應至單一外顯子	2
群中有 mRNA 在 NCBI 結果顯示距離遠	2
NCBI 的 mRNA 結果分群與 UniGene 群相同	35
NCBI 的 mRNA 結果分群與 UniGene 群不相同	15
NCBI 的 mRNA 結果分群與 UniGEne 群不相同且與分群不想同	1
其它情形	4

表 3：mRNA 在分群與 UniGene 181 之 10 種情況

型態：	數量	百分比	數量	百分比
完全無對應	12	18.46%	12	19.67%
部分無對應	5	7.69%	5	8.20%
內含子正確	41	63.08%	41	67.21%
無群對應	3	4.62%	3	4.92%
其它	4	6.15%		
總數量：	65	100.00%	61	100.00%

表 4：HMR195 基因結構分析

下面將會根據上表的型態一一說明並探討原因。說明順序如下：其他、無群對應、完全無對應、部分無對應、內含子正確。

其他：在此有兩種情況，一種為單一外顯子情形，另一種為非 NM_XXXXXX 情形。

無群對應：會照成這種情況應為過濾可用的表現序列標籤的標準過高，導致沒有任何的表現序列標籤位於此基因的位置上。

完全無對應：在此有三種情況。情況說明如下：

第一種情況：包含於重組基因的範圍內，如下圖 13 所示。出現此種現象主要因為 SKIV2L 的表現序列標籤量太少或無重疊。

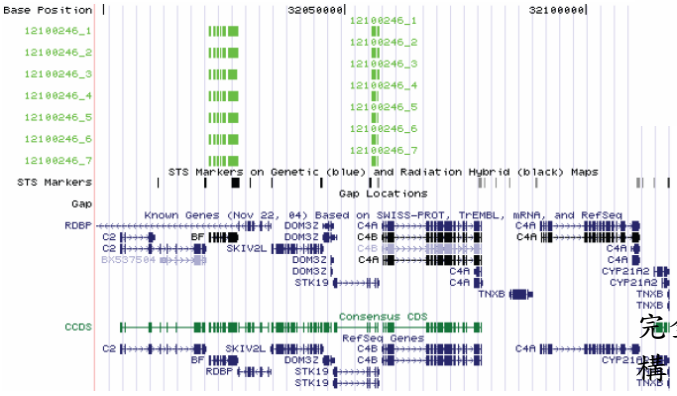

圖 13：SKIV2L 基因圖

第二種情況：在重組基因的下遊，如下圖 14 所示。出現此種現象主要因為 COX4I2 的表現序列標籤量太少、無重疊或重組結構的分數太低導致無被選取。

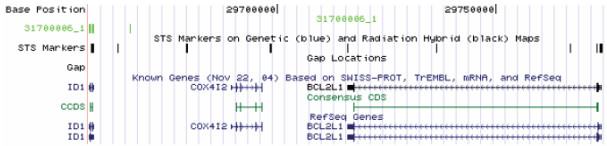

圖 14：COX4I2 基因圖

第三種情況：在重組基因的上遊，如下圖 15 所示。出現此種現象主要因為 CYP27B1 的表現序列標籤量太少、無重疊或重組結構的分數太低導致無被選取。

圖 15：CYP27B1 基因圖

部分無對應：重組的基因結構並沒有完全對到 UCSC 基因結構、Mugup 基因結構，重組的基因結構缺少部分的外顯子，如下圖 16 所示，會有此種現象應該為表現序列標籤量太少或重組結構的分數太低導

致無被選取。

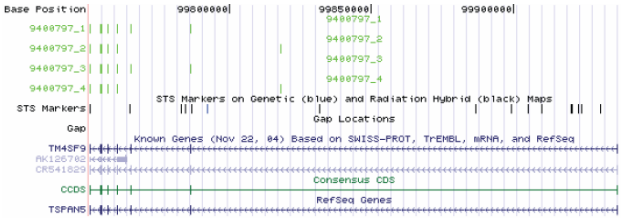

圖 16：TSPAN5 基因圖

內含子正確：重組的基因結構完全可建立出 UCSC 基因結構、Mugup 基因結構的內含子，在此有四種情形。在下面將一一說明和解釋。

第一種：重組基因結構和 UCSC 基因結構或 Mugup 基因結構的外顯子數量一樣，但重組基因結構在兩端的邊界與之不同，如圖 17 所示，在此邊界的精確建立可以在經由聚腺嘌呤尾端、啟動子的精確預測軟體來更進一步的確認。

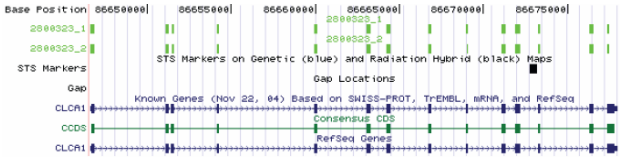

圖 17：CLCA1 基因圖

第二種：重組基因結構和 UCSC 基因結構或 Mugup 基因結構的外顯子數量不一樣，但重組基因結構在兩端的邊界與之不同並在上遊尚有外顯子，如圖 18 所示，在此情形在下遊可由聚腺嘌呤尾端的精確預測軟體來更進一步的確認，至於是在上遊且信任切位對齊程式，所以應該是存在的。

圖 18：CIB1 基因圖

第三種：重組基因結構和 UCSC 基因結構或 Mugup 基因結構的外顯子數量不一樣，但重組基因結構在兩端的邊界與之不同並在下遊尚有外顯子，如圖 19 所示。在此情形在上遊可由啟動子的精確預測軟體來更進一步的確認，至於是在下遊且信任切位對齊程式，所以應該是存在的。

圖 19：TAGLN 基因圖

第四種：重組基因結構和 UCSC 基因結構或 Mugup 基因結構的外顯子數量不一樣，但重組基因結構在兩端的邊界與之不同並在上遊和下遊尚有外顯子，如圖 20 所示，在此信任切位對齊程式，所以是存在的。

圖 20：GSTM3 基因圖

最後，有個外顯子跳躍的特殊例子，此例子發生是在兩邊的邊界一個相同一個不同，但相同的那一邊會把 UCSC 基因結構或 Mugup 基因結構的外顯子分除數個外顯子，如下圖 21 所示。

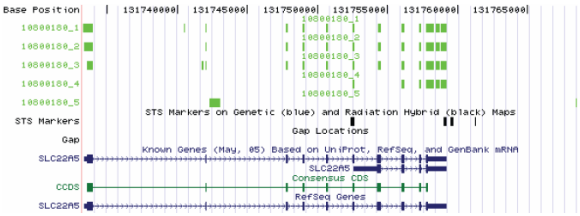

圖 21：SLC22A5 基因圖

第三節：尋找可能的基因

我們利用未對應到基因的多個外顯子群，選出包含 EST 數量的前 100 名當作測試集合。其中有 65 個群所建立的結構有一致的內含子切位，剩下的 35 的群所建立的結構最少有一個內含子切位與其他的內含子切位不一致。群之所建立的結構最長 CDS 在同一股的有 68 個，在另一股的有 32 個，發現另一股的 16 個群所建立的結構有一致的內含子切位。

深入研究群所建立之結構有一致的內含子切位且最長之 CDS 在相同股，而這樣情形有 49 個。在此我們又根據 CDS 長度佔全長的百分之 50 為一個標準來篩選，通過篩選的有 23 個。而這 23 個有以下幾種情況：

第一種情況：有對應之基因。此例子的群包含 94 條 EST，且為反股、內含子切位為 CT/AC 和位於 Contig：37550092。基因結構為如下圖 22 所示。

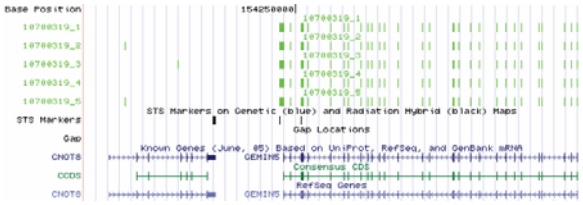

圖 22：對應 GEMIN5 基因

第二種情況：有對應之 UniProt 或 mRNA。此例子的群包含 99 條 EST，且為反股、內含子切位為 CT/AC 和位於 Contig：37549622。基因結構為如下圖 23 所示。

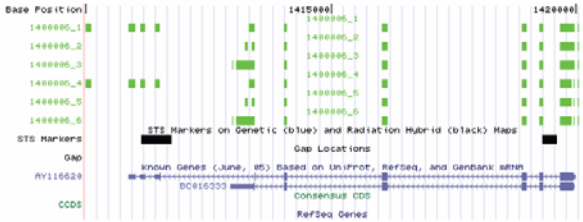

圖 23：對應之 UniProt 或 mRNA

第三種情形：另一股有基因。此例子的群包含 196 條 EST，且為正股、內含子切位為 GT/AG 和位於 Contig：51471365。基因結構為如下圖 24 所示。

圖 24：另一股有基因

第四種情況：無任何對應。此例子的群包含 148 條 EST，且為反股、內含子切位為 CT/AC 和位於 Contig：51471030。基因結構為如下圖 25 所示。

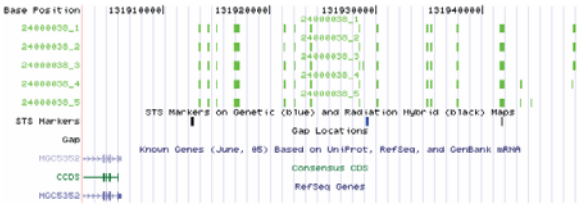

圖 25：無任何對應

第五種情況:有少部份重疊之 UniProt 或 mRNA。此例子的群包含 74 條 EST，且為正股、內含子切位為 GT/AG 和位於 Contig：27482319。基因結構為如下圖 26 所示。

圖 26：有少部份重疊之 UniProt 或 mRNA

第六種情況：有另一股之 UniProt 或 mRNA 和基因。此例子的群包含 52 條 EST，且為正股、內含子切位為 GT/AG 和位於 Contig：51460714。基因結構為如下圖 27 所示。

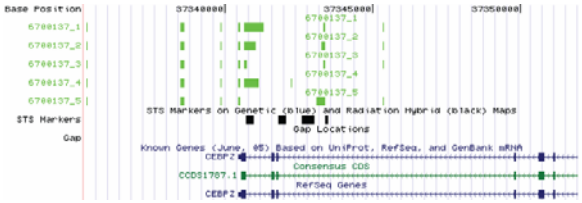

圖 27：有另一股之 UniProt 或 mRNA 和基因

以上六種情形的數量如下表 5 所示。根據我們的資料顯示，除了已對應之基因的部分外，其餘的可能需要經由設計實驗方式來驗證是存在此一基因。

型態：	數量	比例
有對應之基因	6	23%
有對應之 UniProt 或 mRNA	9	34%
另一股有基因	1	3%
無任何對應	1	3%
有少部份重疊之 UniProt 或 mRNA	1	3%
有另一股之 UniProt 或 mRNA 和基因	3	11%
其他	2	7%

表 5 : 23 個群之結構與 UCSC

第四節：基因融合

基因融合（Gene Fusion）在此為相同股的基因由 EST 重組建立外顯子內含子結構所連接起來。而非不同股基因有部分的重疊。我們發現有以下幾種基因融合現象。

第一種情況：相鄰基因融合。此例子的群包含 992 條 EST，且為正股、內含子切位為 GT/AG、位於 Contig：51468814 和融合 C11orf2、TM7SF2 基因。基因結構為如下圖 28 所示。

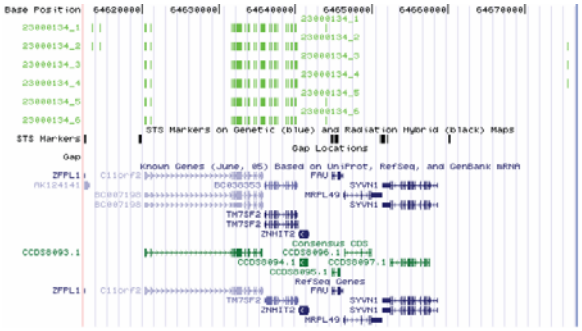

圖 28：相鄰基因融合

第二種情況：相鄰三個基因融合。此例子的群包含 404 條 EST，且為正股、內含子切位為 GT/AG、位於 Contig：51459255 和融合 CTRC、CLA2A、ELA2B 基因。基因結構為如下圖 29 所示。

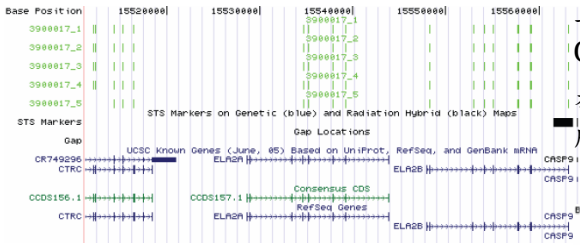

圖 29：相鄰三個基因融合

第三種情況：相鄰基因融合且頭尾重疊。此例子的群包含 419 條 EST，且為正股、內含子切位為 GT/AG、位於 Contig：51458073 和融合 RNPC3、AMY2B 基因且頭尾重疊。基因結構為如下圖 30 所示。

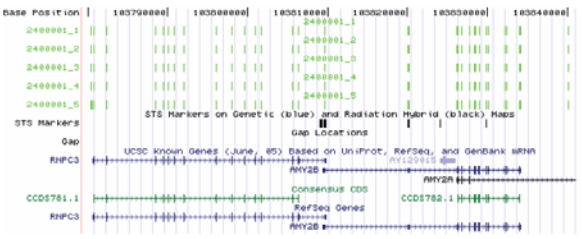

圖 30：相鄰基因融合且頭尾重疊

第四種情況：相鄰基因融合且部分重疊。此例子的群包含 467 條 EST，且為正股、內含子切位為 GT/AG、位於 Contig：37550092 和融合 ANKHD1、MASK-BP3 基因且部分重疊。基因結構為如下圖 31 所示。

圖 31：相鄰基因部分重疊

第五種情況：跨數個基因之相鄰基因融合。此例子的群包含 495 條 EST，且為正股、內含子切位為 GT/AG、位於 Contig:29800594 和融合 SNRP70、PPFIA3 基因跨過 LIN7B。基因結構為如下圖 32 所示。

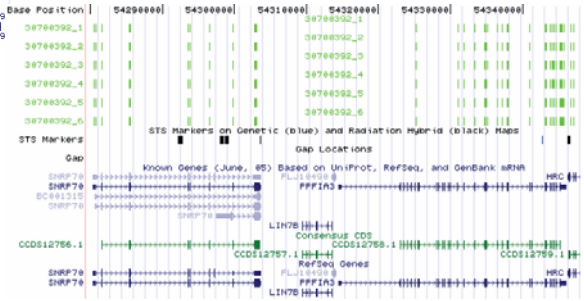

圖 32：跨數個基因之相鄰基因融合

群包括兩個基因以上，有 3227 個群。而這些群是有可能發生基因融合。經由我們所建立之結構且基因有包括兩個以上，我們以 EST 數量支持的前 100 名，以人工的方式於 UCSC Human Genome Browser 檢測基因融合且有外顯子共享。五種情形的數量如下表 6 所示。而這些情況，已有少部分已在討論，且根據我們的資料結果發現，可能有多種融合的方式，可能是 SNP、Promoter 等，或其他不明原因之影響所造成的，需要進一步經由設計實驗方式來驗證有此一融合情況。

型態：	數量	比例
相鄰基因融合	22	22%
相鄰三個基因融合	2	2%
相鄰基因融合且頭尾重疊	1	1%
相鄰基因融合且部分重疊	3	3%
跨數個基因之相鄰基因融合	4	4%

表 6：100 個群之結構與 UCSC 基因融合

四、 結論

本論文所建立之分群，雖然在分群的結果和 UniGene 群不盡相同，但根據在一個對應的區域中包含一個片段和另一個對應區域的片段相同，可能是基因中包含另一個相同的基因，且根據 mRNA 測試資料顯示，本論文提出分群之方法對有發生選擇性裁切發生的群是有意義的。

本論文所建立之完整基因結構是可信任的，其中由 EST 所建立的外顯子和內含子的結構與 mRNA 所建立的外顯子和內含子的結構之分析，內含子完全一樣的有 41，完全不一樣的 12。

提供生物資訊學者和生物學者如想知道位於某物種的某個基因的完整結構則可以方便使用，讓生物資訊學者和生物學者不必為了知道完整的基因結構或相關訊息，而浪費不必要的時間，可以繼續進一步的相關研究。

至於未來研究可分為兩部分，第一部分：針對相似基因改進分群之方法，第二部分：利用群之基因結構，進行未知基因的驗證。

五、 參考文獻

[1] About the Human Genome Project : http://www.genome.gov/10001772

[2] database for "expressed sequence tags": http://www.ncbi.nlm.nih.gov/dbEST/

[3] E. Eyras, M. Caccamo, V. Curwen and M. Clamp, "ESTGenes: alternative splicing from ESTs in Ensembl", Genome Res., Vol. 14, pp. 976-987, 2004.

[4] F. R. Hsu, H. Y. Chang, Y. L. Lin, Y. T. Tsai, H. L. Peng, Y. T. Chen, C. F. Chen,, C. Y. Cheng, C. H. Liu, M. Y. Shih, "Genome-wide alternative splicing events detection through analysis of large scale ESTs", Proceedings of the IEEE 4th symposium on bioinformatics and bioengineering, pp. 310-318, 2004.

[5] F. R. Hsu and J. F. Chen., “Aligning ESTs to Genome Using Multi- Layer Unique Markers”, Proc. Of the IEEE Computational Systems Bioinformatics Conference (CSB2003), pp. 564-566, 2003.

[6] L. Florea, G. Hartzell, Z. Zhang, G. M. Rubin,and W. Miller, “A Computer Program for Aligning a cDNA Sequence with a Genomic DNA Sequence”, GENOME RESEARCH, pp. 967-974, 1998.

[7] N. Kim, S. Shin, and S. Lee, "ECgene: Genome-based EST clustering and gene modeling for alternative splicing", Genome Res., Vol. 15, pp. 566-576, 2005.

[8] R. Sorek and H. M. Safer., “A novel algorithm for computational identification of contaminated EST libraries”, Nucleic Acids Res., Vol. 31, pp. 1067-1074, 2003.

[9] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool”, J. Mol. Biol.,

Vol. 215, pp. 403–410, 1990.

[10] S. Rogic, A. Mackworth and F. Ouellette, “Evaluation of gene conding programs”, Genome Research, Vol. 11, 817-832, 2001.

[11] UniGene Build Procedure ： http://www.ncbi.nlm.nih.gov/UniGene/build.shtml

[12] W.J. Kent, C.W. Sugnet, T.S. Furey, J.M. Roskin,T.H. Pringle,A.M. Zahler and A.D. Haussler, “The Human Genome Browser at UCSC”, Genome Research, Vol. 12, pp. 996-1006, 2002.

[13] Z. Kan, E. Rouchka, W. Gish, and D. States., “Gene Structure Prediction and Alternative Splicing Analysis Using Genomically Aligned ESTs”, Genome Res., Vol. 11, pp. 875-888, 2001.