

尋找在剪裁位置的 RNA 編輯事件

Using a Bioinformatics Approach to Search Splice Sites Regulated by RNA editing

許芳榮

逢甲大學資訊工程系

逢甲大學生醫資訊暨生醫工程碩士學位學程

frhsu@fcu.edu.tw

黃于展

逢甲大學資訊工程系

M9620667@fcu.edu.tw

摘要 — RNA 編輯經過被稱作 ADAR (adenosine deaminases that act on RNA) 的反應酶作用之後而產生核苷酸的轉換。RNA 編輯是在後轉錄的過程中發生的，但 RNA 編輯不一定會發生，發生的條件是要形成 RNA 二級結構，然後被 ADAR 包覆而產生的，產生的核苷酸轉換是從 adonsine (A) 變成 inosine (I)。最近的研究有指出因為 RNA 編輯的調控而使得一些基因產生了選擇性剪裁，而我們使用了生物資訊的方法找出大量位於剪裁位置的 RNA 編輯事件。我們對四種物種，人類、大鼠、小鼠的 mRNA 來做尋找和分析，全部的物種總共找到了 23 條 mRNA 有發生 RNA 編輯事件，總共找到 21 個不同位置的 RNA 編輯位置，並且我們也對所有找到的位置做選擇性剪裁、序列分析和跨物種的比較。

關鍵字：剪裁位置、選擇性剪裁、RNA 編輯、ADAR

Abstract — RNA editing is nucleotide conversion which occurred by the enzyme which the adenosine deaminases that act on RNA (ADAR). RNA editing is a post-transcription processing, the occurred conditions are the

sequences formed the double-stranded RNA, and the nucleotide is converted by ADAR enzyme, the nucleotide conversion is adonsine (A) to inosine (I). In recently studies, some genes have alternative splicing events regulated by RNA editing. We used a bioinformatics approach to search splice sites regulated by RNA editing and searched and analyzed for three species: human, rat, and mouse. Total 23 mRNA sequences that occurred RNA editing were found out, and 21 editing positions are different. Then also search alternative splicing regulated by RNA editing, trinucleotide analysis, and cross-species comparison.

Keyword: splice site, alternative splicing, RNA editing, ADAR.

一、緒論

RNA 編輯是一個普遍的生物分子現象。它發生在原核生物、植物和動物身上。RNA 編輯是後轉錄修改現象，RNA 編輯經過被稱作 ADAR 的反應酶作用之後而產生核苷酸的轉換，轉換的型態有兩種一種是 cytidine(C)轉換成 uridine(U) 和 adonsine(A)轉換成 inosine(I)，因為 I 的化學

夜蛾、家蠶、紅色麵粉甲蟲、義大利蜂進行跨物種分析，也發現發生在部分昆蟲發生 RNA 編輯。部分昆蟲在同源基因上的相同位置反而遺傳成核苷酸 G，這被認為 RNA 編輯經過萬年後核苷酸 A 可能遺傳成 G，不再發生 RNA 編輯。所以 RNA 編輯在演化上可能扮演重要的關係。而且現在已發現的 RNA 編輯數量並不，而且在剪裁位置上的 RNA 編輯數量更是少。我們是希望藉由生物資訊的方法可以分別幫助尋找出在 3'端和 5'端剪裁位置上的可能的 RNA 編輯位置，圖 1.3.所示。

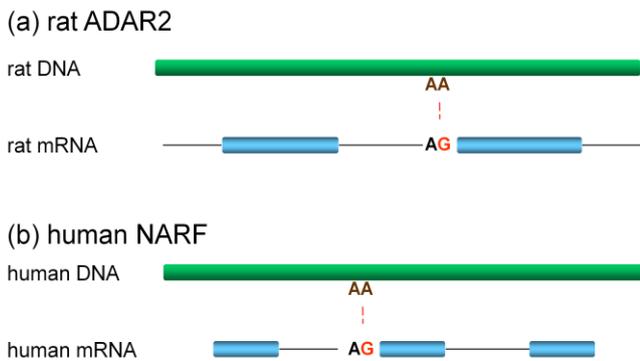


圖 1.2. (a)大鼠基因 ADAR2，因為 RNA 編輯，所以產生 3'選擇性剪裁事件，多出 47 的鹼基。(b)人類基因 NARF，也是因為 RNA 編輯而發生外顯子躍過。紅色核苷酸是經過 RNA 編輯的核苷酸。

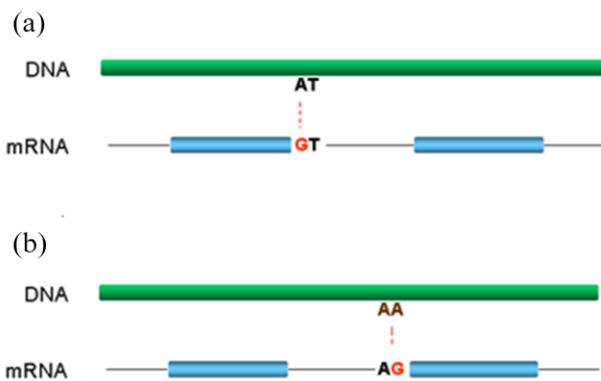


圖 1.3. (a)尋找在 5'端剪裁位置的 RNA 編輯事件 (b)尋找在 3'端剪裁位置的 RNA 編輯事件。

二、資料來源和方法

(一) 資料來源

本研究所使用的資料，基因體序列和 mRNA 序列是從 NCBI[9]的 FTP 下載下來的，如表 2.1 是我們從 NCBI 得到的各種物種基因體序列的版本和 mRNA 序列個數。

表 2.1. 三個物種的版本與 mRNA 的序列個數。

物種	版本	mRNA 序列
人類	36.3	45,044
小鼠	37.1	40,009
大鼠	RGSC v3.4	40,672

(二) 方法與步驟

本研究所使用的剪裁比對工具，一個 MUGUP [3]，一個是 sim4 [5]，兩者都是剪裁比對的工具。之所以選擇這兩種來當作本研究的工具，是因為發現所發現典型的剪裁型態 GT-AG 和 GC-AG 兩種，如圖 2.1 有 MUGUP 只會尋找 GT-AG 剪裁型態，當 MUGUP 找不到 GT-AG 剪裁型態就會在真正的剪裁位置附近去尋找適當的 GT-AG，因此部分剪裁位置附近會有些許的缺口(gap)。相反地，如果 sim4 比對不到上述這兩種剪裁型態，sim4 就不會強迫要比對 GT-AG 和 GC-AG。而是會比對到其他適當的位置。但是 sim4 和 MUGUP 的比對方法不同。另外 sim4 是使用動態規劃(dynamic programming)，MUGUP 是使用多層次的獨特標記(multi-layer unique marker)。因此 MUGUP 比對的速度比 sim4 還要快，所以我們才先用 MUGUP 比對，然後再使用 sim4 比對抓取下來的序列。

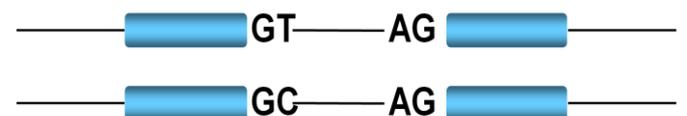


圖 2.2. 三種典型的剪裁型態 GT-AG、GC-AG 和 AT-AC。

第一步:

首先我們從 MUGUP 搜尋 mRNA 序列的條件是比對分數非 100，因為 RNA 編輯發生在剪裁位置，則 MUGUP 無法正確比對在基因體上，而會在真正剪裁位附近尋找似 GT-AG 的剪裁型態。這會讓部分外顯子的長度不正確，使得無法與基因體完全吻合，導致序列分數下降。但分數太低的 mRNA 也不需要，因為我們只尋找在剪裁位置的 RNA 編輯，分數不會下降太低，因此才抓取 94 到 99 之間的分數。然後我們抓取不是 100 分的目標外顯子，再以目標外顯子為基準抓取上游的外顯子、下游的外顯子等序列和相對的基因體序列，我們所抓取的外顯子序列稱"RNAseq"。如圖 2.2 所示。

第二步:

接下來使用 sim4 把相對的基因體序列和 RNAseq 做剪裁序列比對，並且尋找在剪裁型態為 GT-AA 和 AT-AG，如圖 2.4 所示。我們尋找的不是只有目標外顯子的剪裁位置，還包括上下游的剪裁位置，這是當初可能因為 RNA 編輯的發生而導致比對會左右位移。

第三步:

最後過濾不是 RNA 編輯的位置。我們先利用 NCBI 所提供的資訊，NCBI 會在基因上提供 mRNA 資料，其中一個資料是 GenBank。GenBank 有提供外顯子序列位置的註解。因此用 GenBank 的註解來核對是否與 sim4 的比對結果一致，接著再過濾所尋找的 RNA 編輯是否為 SNP 的位置，最後留下的就可能是 RNA 編輯的位點。

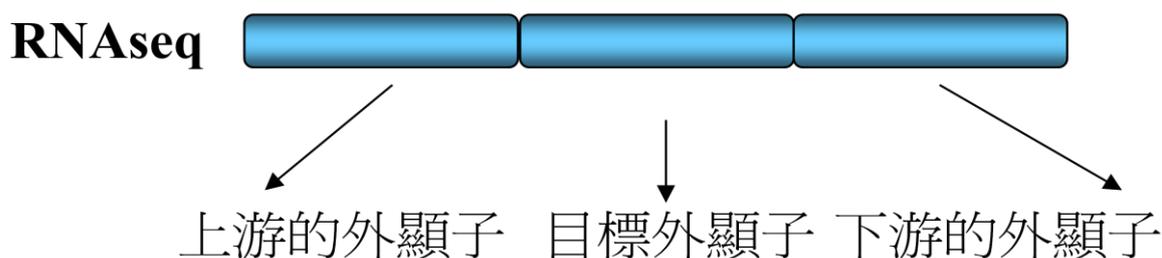


圖 2.4. RNAseq 包含了上游的外顯子、外顯子，下游的外顯子，目標外顯子是所搜尋 94 到 99 的外顯子。

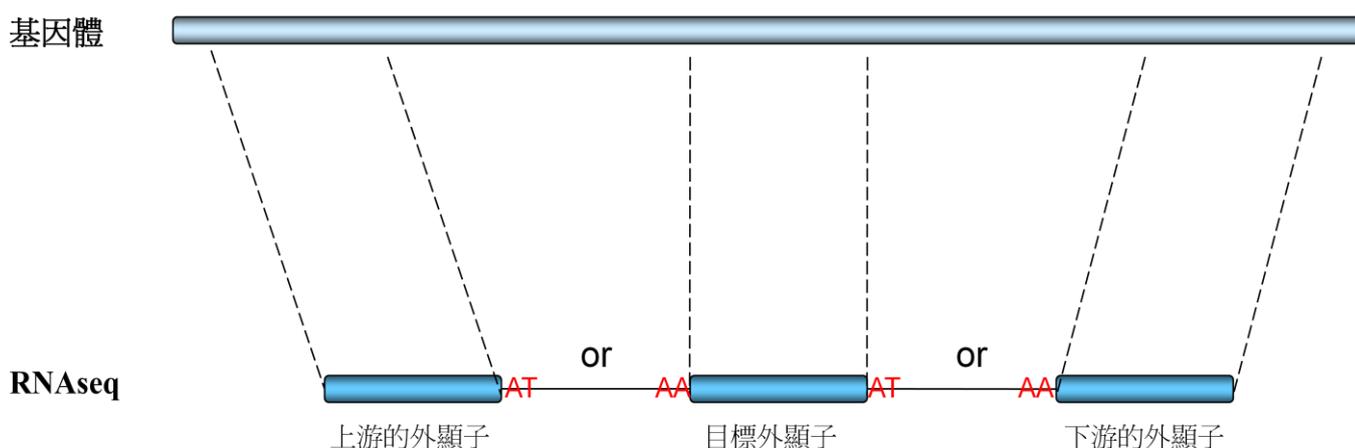


圖 2.5. 使用 sim4 對基因體序列和 RNAseq 做剪裁序列比對，尋找 GT-AA 和 AT-AG 的剪裁型態。

三、實驗結果

首先經過第一個步驟找出 94 到 99 分的 mRNA，我們分別找出人類 17774 條 mRNA、小鼠 15169 條 mRNA 和大鼠 9232 條 mRNA。再經過第二個步驟找出 sim4 所比對的剪裁位置分別是 AA 或 AT，我們過濾掉分別剩下人類 50 條 mRNA、小鼠 62 條 mRNA 和大鼠 77 條 mRNA 最後再經過 NCBI 和 SNP 的過濾之後，分別剩下人類 3 條 mRNA、小鼠 6 條 mRNA 和大鼠 14

條 mRNA。總共找出了 23 條 mRNA 在剪裁位置 RNA 編輯，21 個不同的 RNA 編輯的位點，其中所找到的 NM_001111056.1 和 NM_001111057.1 兩條 mRNA 是之前 Susan M. Rueter[11] 等學者所發現的 RNA 編輯。NM_031968.2 是 Galit Lev-Maor[6] 等學者所發現的 RNA 編輯，此外我們觀察到我們其中在 NM_018658.1 和 NM_170741.1 上的 RNA 編輯位置 SNP，此 SNP 的編號 rs7226242。表 3.1 是找到的 RNA 編輯的各種資訊。

表 3.1. 人類、小鼠和大鼠等三個物種的 RNA 編輯資訊，MRNA GI: mRNA 的 gi 編號，MRNA REF: mRNA 的 ref 編號，EXON NO: 第幾個外顯子，CONTIG GI: contig 的 gi 編號，GENE NAME: 基因的名稱，EDIT SITE: RNA 編輯的位置。

MRNA GI	MRNA REF	EXON NO	CONTIG GI	GENE NAME	EDIT SITE
在人類 3'端剪裁位置的 RNA 編輯事件					
134284353	NM_031968.2	8	37542676	NARF	655149
在人類 5'端剪裁位置的 RNA 編輯事件					
74027268	NM_001014443.2	11	88943682	USP21	11624758
74027266	NM_012475.4	10	88943682	USP21	11624758
在小鼠 3'端剪裁位置的 RNA 編輯事件					
6680422	NM_008365.1	12	149233934	I18r1	18083945
149272752	XM_887152.3	6	94409721	LOC622706	446239
在小鼠 5'端剪裁位置的 RNA 編輯事件					
38348569	NM_198668.1	8	149255466	4930590J08Rik	44182277
94367098	XM_892197.2	4	149250280	LOC627525	31059790
118130743	NM_019408.2	1	149270586	Nfkb2	39492202
124257962	NM_001080925.1	2	149262584	Rapefl1	10242728
在大鼠 3'端剪裁位置的 RNA 編輯事件					
161377410	NM_001111056.1	4	34852431	Adarb1	3829297
161377412	NM_001111057.1	4	34852431	Adarb1	3829297
109464345	XR_006173.1	10	34853764	LOC502504	1303987
109489168	XM_221021.4	16	34874015	LOC686084	2761905
12018317	NM_022628.1	27	34855852	Nphs1	1668685
109474552	XM_232595.4	2	34866468	Rbpsuh_predicted	2886601
18158440	NM_080775.1	4	62646202	Smgb	34608407
109468766	XM_238280.4	7	62646202	Stard7_predicted	4970717

在大鼠 5'端剪裁位置的 RNA 編輯事件

32996706	NM_019363.2	2	62655325	Aox1	137382
51948451	NM_001004240.1	8	34867094	Gpaal	8983754
109478416	XR_005855.1	1	34866353	LOC680736	4963486
109488687	XM_220804.4	5	62656853	RGD1310166_predicted	14449166
109475488	XM_575907.2	3	62649682	RGD1560436_predicted	5437703
109458491	XM_001067115.1	6	34855852	Zfp569	1255774

四、分析與討論

接下來我們針對所尋找出來的 RNA 編輯位置做分析，我們分成兩個重點來分析選擇性剪裁的分析和跨物種分析。

(一)RNA 編輯和選擇性剪裁事件的分析

接下來我們對四個物種尋找是否有發生選擇性剪裁事件，結果發現除了之前 Susan M. Rueter 和 Galit Lev-Maor 等學者有選擇性剪裁事件外，也有其他的選擇性剪裁事件。我們發現了人類和斑馬魚在 EST 上有選擇性剪裁事件。如圖 4.2 示，人類在基因 USP21 上有 EST 跟 mRNA 發生了選擇性剪裁事件，分別是 5' 選擇性剪裁和內含子保留選擇性剪裁事件，除此之外沒發現其他的選擇性剪裁事件。

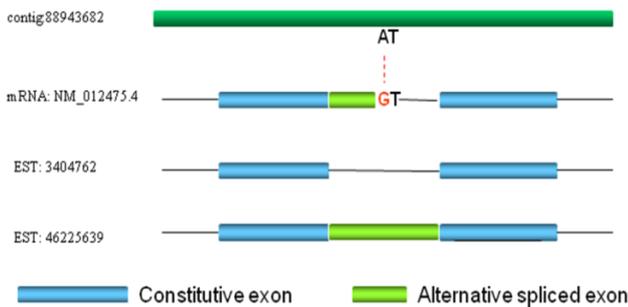


圖 4.1. 在人類基因 USP21 上發現有 5' 端和 intron retin 的選擇性剪裁事件。

(二)RNA 編輯在物種演化上的分析

最後我們針對所找出來的 RNA 編輯位置進行跨物種比較，我們所比較的物種有人類、黑

猩猩、恆河猴、小鼠和大鼠等五種物種。我們使用 ClustalX[12] 來做跨物種的工具。經過跨物種比對之後發現某些 RNA 編輯位置在同源基因上的保留區域位置已經被保留成核苷酸 G 了。

比較小鼠和大鼠 RNA 編輯的同源基因位置，我們發現在大鼠基因 Nphs1 和 Aox1 上有發生 RNA 編輯事件，但是在小鼠同源基因上的保留區域位置已經被遺傳編碼成核苷酸 G 了，而且這個位置也是一個剪裁位置，如圖 4.3 所示。接下來針對 nfkb2 這個基因做跨物種比較，不過是全部五種，這個基因是小鼠有發生 RNA 編輯。經過比對之後發現靈長類，也就是人類、黑猩猩和恆河猴在保留區域的位置上已經被遺傳編碼成 G，而大鼠和小鼠則是核苷酸 A。很有趣地並不是這個位置上並不是所有的物種都是剪裁位置，人類、黑猩猩和小鼠是剪裁位置，但是恆河猴和大鼠上的位置則是在外顯子裡，如圖 4.4。依照核苷酸 A 和 G、在剪裁位置 and 在外顯子裡畫出了演化樹，如圖 4.5。從這演化樹來看，只從靈長類的演化樹來看，這子演化樹跟 Rhesus Macaque Genome Sequencing and Analysis Consortium 團隊[9]裡的學者所提的演化樹相同。最後我們針對基因 USP21 來做分析，這個基因是人類發生了 RNA 編輯，但是我們發現黑猩猩和恆河猴在保留區域的位置上都是核苷酸 A，不過黑猩猩的兩條 mRNA：XM_00173091.1、XM_00173063.1 和恆河猴的兩條 XM_001117772.1、XM_001117782.1 在這個位置上有發生內含子保留選擇性剪裁事件，所

以我們認為在黑猩猩和恆河猴有發生 RNA 編輯， 在演化上有一定程度的關係。如圖 4.6 所示。由這些分析得知 RNA 編輯可能

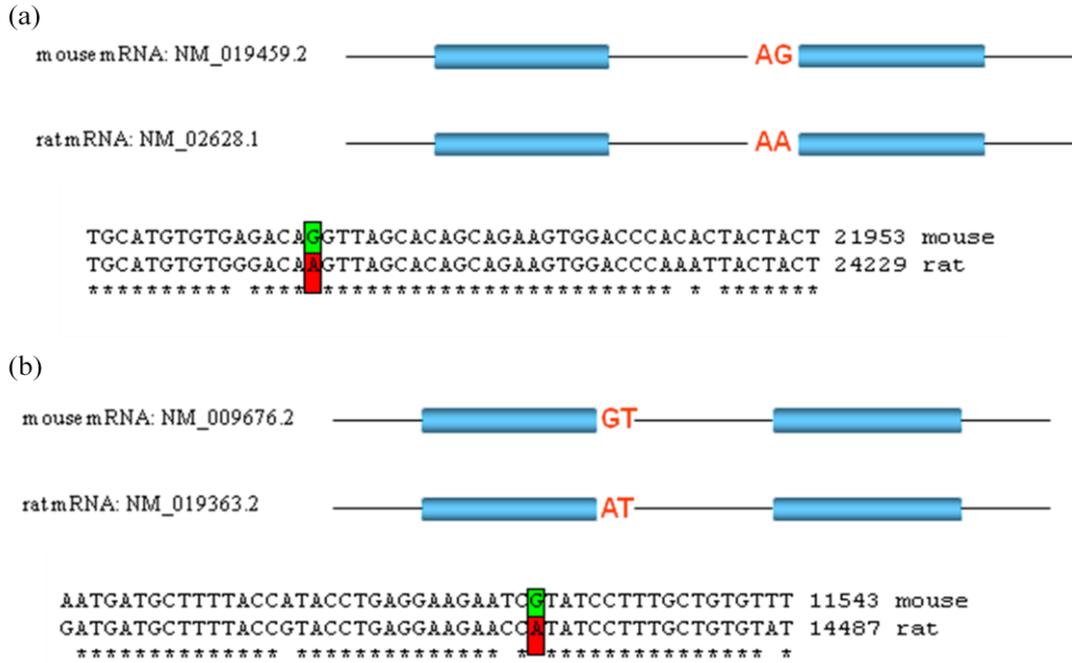


圖 4.2.(a)基因 Nphs1 的跨物種比對，發現在小鼠的保留區域位置上發生了 3'端剪裁事件，但沒有發生 RNA 編輯。(b)基因 Aox1 的跨物種比對，發現在小鼠的保留區域位置上發生了 5'端剪裁事件，但是也沒發生 RNA 編輯事件。紅色的核苷酸是我們發現的可能 RNA 編輯，綠色的核苷酸是被遺傳編碼成 G。

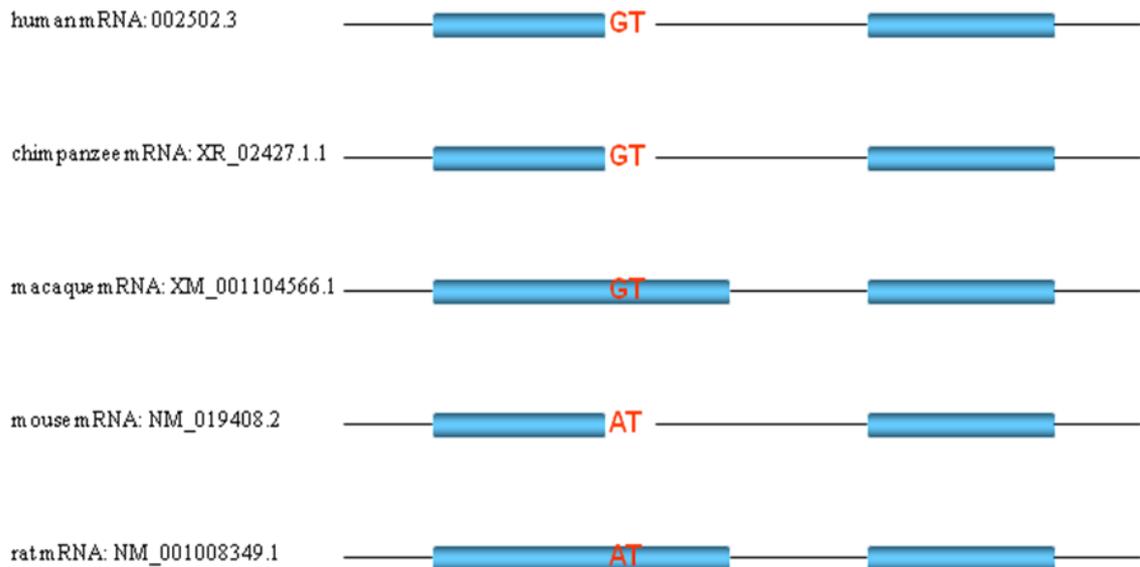


圖 4.4 五種跨物種比對後對他們剪裁比對，發現並不是有的物種都在剪裁位置，有些在外顯子裡。

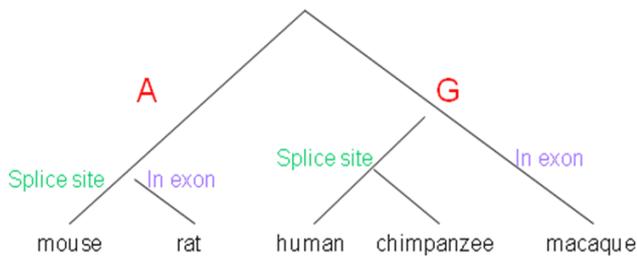


圖 4.5. 依照核苷酸 A 和 G、在剪裁位置和在外顯子裡畫出的演化樹。

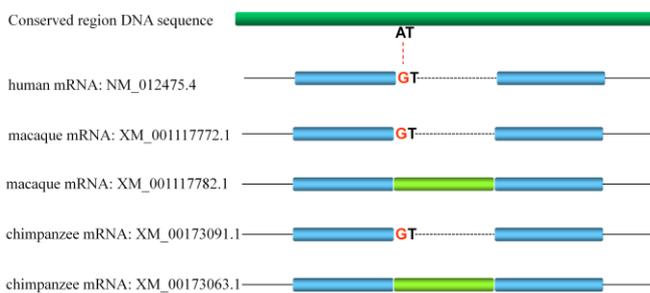


圖 4.6. 經過跨物種比對之後，再做剪裁序列比對分析。發現黑猩猩和恆河猴有 intron retain 的選擇性剪裁事件發生。第一條序列是人類 mRNA，第二、三是恆河猴的 mRNA，最後第四、第五是黑猩猩的 mRNA。紅色的核苷酸假設是經過從 A 轉換成 I 的 RNA 編輯。

五、結論

我們在所有人類，小鼠和大鼠的 mRNA 裡總共發現 23 條 mRNA 有發生 RNA 編輯和 21 個不同位置的 RNA 編輯，此外過濾的資料裡有一個 SNP 位置是位在剪裁位置。針對選擇性剪裁方面來分析從 EST 發現人類和斑馬魚個 RNA 編輯可能調控選擇性剪裁的發生。另外物種演化的分析，RNA 編輯可能會在遺傳上核苷酸 A 被編碼核苷酸 G，而而且部分被遺傳編碼在演化上有一定程度的關係。最後我們希望能提供這些資訊給有興趣研究 RNA 編輯的生物學家來做研究。

六、參考文獻

[1] Anant S; Davidson NO, “Molecular

mechanisms of apolipoprotein B mRNA editing”, *Current opinion in lipidology*, Vol. 12, pp. 159-65, 2001.

- [2] Dennis D.Y. Kim, Thomas T.Y. Kim, Thomas Walsh, Yoshifumi Kobayashi, Tara C. Matise, Steven Buyske and Abram Gabriel, “Widespread RNA Editing of Embedded Alu Elements in the Human Transcriptome,” *Genome Research*, Vol. 14, pp. 1719-1725, 2004.
- [3] F.R. Hsu and J. F. Chen, “Aligning ESTs to Genome Using Multi-Layer Unique Markers,” *Proc. Of the IEEE Computational Systems Bioinformatics Conference*, pp. 564-566, 2003.
- [4] Fang Rong Hsu¹, Hwan-You Chang, Yaw-Lin Lin, Yin-Te Tsai, Hui-Ling Peng, Ying Tsong Chen, Chia YangCheng, Min Yao Shih, Chia-Hung Liu, Chin-Feng Chen, “AVATAR: A database for genome-wide alternative splicing event detection using large scale ESTs and mRNAs,” *Bioinformatics*, Vol. 1, pp. 16-18, 2005.
- [5] Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W, “A computer program for aligning a cDNA sequence with a genomic DNA sequence”, *Genome Res.*, Vol. 8, pp. 967-74, 1998.
- [6] Galit Lev-Maor, Rotem Sorek, Erez Y Levanon, Nurit Paz, Eli Eisenberg and Gil Ast, “RNA-editing-mediated exon evolution”, *Genome Biology*, Vol. 8 pp.R29, 2007.
- [7] Johan Ohlson, Jakob Skou Pedersen, David Haussler, “Editing modifies the GABAA receptor subunit α 3”, *RNA*, Vol. 13, pp. 698-703 2007.
- [8] Nan Tian, Xiaojie Wu, Yaozhou Zhang, “A-to-I editing sites are a genomically

encoded G: Implications for the evolutionary significance and identification of novel editing sites”, RNA, Vol. 17, pp. 211-216 2008.

[9] NCBI URL:

<http://www.ncbi.nlm.nih.gov/>

[10] Rhesus Macaque Genome Sequencing and Analysis Consortium, et al. “Evolutionary and Biomedical Insights from the Rhesus Macaque Genome”, Science, Vol. 316, pp.

222-234, 2007.

[11] Rueter SM, Dawson TR, Emeson RB, “Regulation of alternative splicing by RNA editing”, Nature, Vol. 399, pp. 75-80, 1999.

[12] Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G., “The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools”, Nucleic Acids Research, Vol.24, pp. 4876-4882, 1997.