知識天地

人為什麼是人？簡介人和黑猩猩基因體序列之間的插入和刪除事件

莊樹諄副研究員（基因體研究中心）

引言

人科動物（Hominidae）包括人類和大猿（great ape）。其中大猿包括：黑猩猩（chimpanzee；或者稱為普通黑猩猩），侏儒黑猩猩（bonobo），大猩猩（gorilla）和紅毛猩猩（orangutan）。**圖 1** 顯示這些人科動物從共同祖先分開演化的大約時間表，這也說明了黑猩猩是地球上和人類親緣關係最接近的物種。在這篇文章裡，我們所謂的黑猩猩指的是普通黑猩猩，因為人類對普通黑猩猩的研究遠比對侏儒黑猩猩多得多。人和黑猩猩大約在 4~6 百萬年前從共同

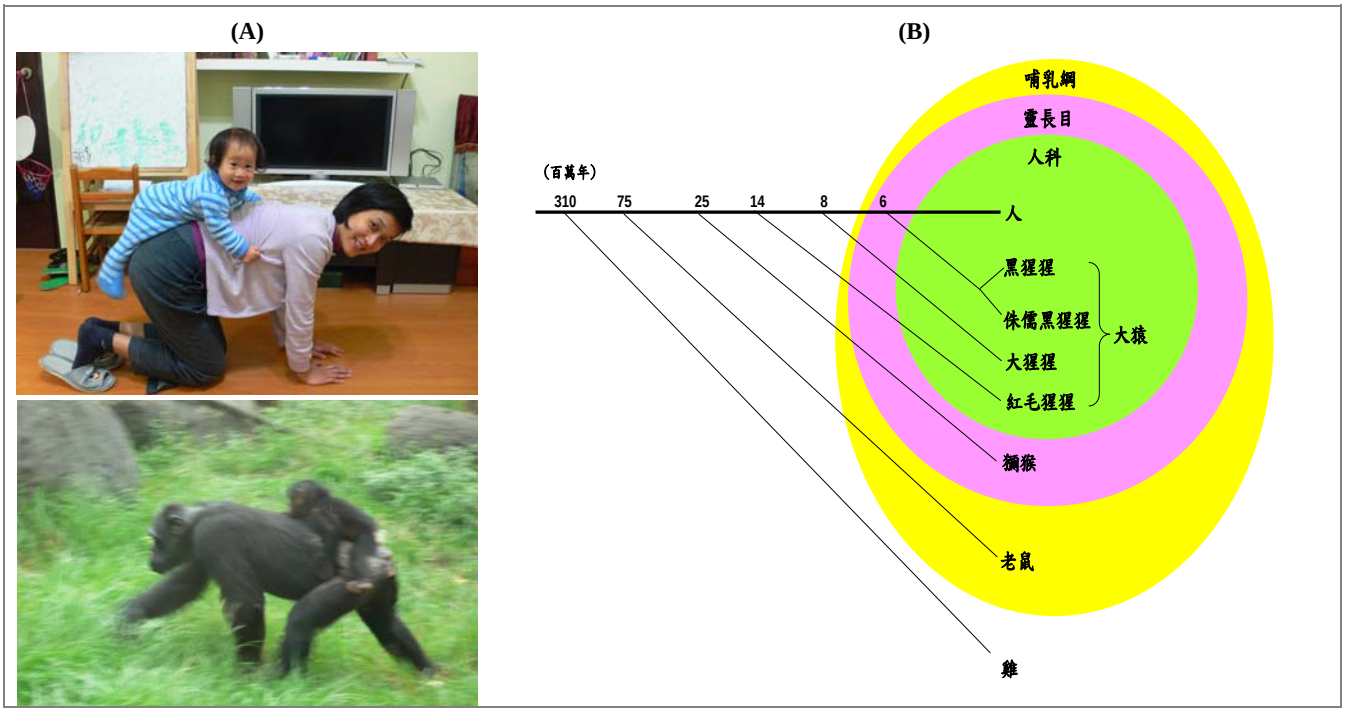

圖 1　(A)人與黑猩猩（圖片皆由作者自攝）；(B)人類、大猿和其他動物之間的親緣關係，以及從共同祖先大約分開演化的時間（此為示意圖，線條的長度非按比例）。

祖先彼此分開演化。這麼短的演化時間，在人和黑猩猩之間產生些微的遺傳的距離，但這已經造成兩個物種在型態上及各種基因表現上有很大差別。研究人和黑猩猩間遺傳差別的工作是非常令人著迷的，因為這能夠，至少可以提供一些線索，幫助我們回答一個很根本的問題：什麼原因使我們成為人？第一版的黑猩猩基因體序列草稿，在 2005 年由 The Chimpanzee Genome Sequencing and Analysis Consortium（簡稱 TCGSAC）完成並發表，這更加速了探討人和黑猩猩之間的形態、行為上等等差別的研究。TCGSAC 的報告指出兩個物種可能共有的 DNA 序列高達 98~99%，換句話說，人和黑猩猩的差異，若單從 DNA 序列中核苷酸變異的比率來看是很少的。然而，人與黑猩猩間基因體的演化差異可能遠比簡單的核苷酸變異來得複雜。人和黑猩猩基因體在各自演化過程中，除了核苷酸變異之外，還經歷廣泛的基因體架構上的重新安排、基因與大片段序列複製、以及序列的插入與刪除(insertion/deletion 簡稱 indel) 事件等等。而且，除了在 DNA 序列上的差異之外，物種間的差別也很可能在其他階段發生，包括例如在 RNA、蛋白質體、後轉譯修飾、代謝等等，這些事件都對兩個物種間的差異，扮演了極為關鍵的角色。在這篇文章裡，主要是介紹我們新近在關於人類與黑猩猩序列的插入與刪除事件上的研究成果。

插入和刪除事件

TCGSAC 的研究已經在人和黑猩猩基因之間找到約 500 萬個 indel 事件，這些 indels 的長度總共約 90 Mb(Mb = Mega bases；bp = base pair；1 Mb = 1 百萬 bp)。人與黑猩猩的基因體長度都大約有 30 億個鹼基對，所以這些 indels 約佔整個基因體長度的 3%，這個比率顯然遠大於 DNA 序列中核苷酸變異 (~1.23%)，因此研究 indels 對這兩個物種間的差異的重要性不言而喻。我們若進一步來探討這些 indels 的長度，可以發現絕大部分的 indels 長度都很短，大片段的 indels 其實不多，在 500 萬個 indels 中，高達 96%的 indels 都小於 20 bp (其中 1 bp 的 indels 就超過 225 萬個)，只有 1.4%的 indels 是大於 80 bp 的。雖然這些大於 80 bp 的 indels 的數量只佔所有 indel 的 1.4%，但他們的總長度卻佔 indel 總長度的 73%，這顯示這些大片段 indels 在兩物種間的序列差異佔有顯著的比重。許多科學家對 indel 的形成原因做了很多研究，他們發現有一大部分 indel 的成因是由基因體序列中的反覆元素(repetitive element) 所造成的。反覆元素是一段 DNA 序列，它們會反覆插入基因體序列中不同的地方，因這些片斷插入的個數與位置的差異，當兩物種的序列一拿來比對，就會找到很多 indels。主要的反覆元素包括四大類 : short interspersed repetitive elements(SINEs), long interspersed repetitive elements（LINEs), SVE elements , 以及 endogenous retroviruses(ERVs)。其中前三類是造成人和黑猩猩之間 indels 的大宗，在反覆元素所引起的 indels 中，有超過 95%的 indels 便是由這三大類反覆元素所造成。反覆元素在靈長類動物的基因體序列中出現相當頻繁，有研究指出人類的基因體中有 45%包含反覆序列。所以，indels 的形成和反覆元素息息相關。

具人類專一性的插入和刪除序列

前面，我們已經大略說明人和黑猩猩基因體之間發生的 indels 的形成因素。大多數尋找這些 indels 的方法，都是基於兩個物種間的序列比對，然而，這樣找出來的 indel 並不能釐清到底是哪一個物種的插入還是刪除。如**圖 2 A** 所示的 case 1，我們無法判斷這個 indel 的形成，在演化過程中是因為人多得到一個片段還是黑猩猩失去一個片段。此外，TCGSAC 在 2005 出版的黑猩猩基因體序列只是一個草稿版本，其精確度可能遠比不上人類的。在我們的研究中估計，利用人和黑猩猩基因體序列兩兩比對的方式所找到的 indels，可能有 15~19%是高估的 (最近 TCGSAC 又發表黑猩猩第二版的基因體序列，這樣高估的情況可能會改善一些)。為了回答上述兩個問題，我們利用具高精確度的第三種物種，如老鼠 (mouse 和 rat) 等，當作參考外群來找尋具物種專一性的 indels。在**圖 2** 的 B 所示，我們把人、黑猩猩、小鼠、大鼠、狗五種物種的基因體序列一起拿來比對，如此便可以找到具人類專一性的插入與缺失片段。如**圖 2B** 的 Case 1，我們便可以假設這個片段是人類在演化過程中獲得的，因為現存的其他四個物種並沒有這個片段。利用這樣的方法，我們找到超過 840,000 個具人類專一性的小 indels (即 indels<100 bp)，這些 indels 總共影響人和黑猩猩基因體間差異度的 0.21%。由於有其他物種當做參考外群，我們所找到的這些 indels 是非常可信的。我們觀察到 indel 所造成的序列差異度，在會轉譯產生功能的區域 (coding sequences) 與假基因 (絕大部分的假基因沒有功能) 的區域的影響力是差別很大的 (indel rate 是 0.03% vs. 1.4%)，這顯示如果一段序列在功能上是重要的，則在這個區域所產生的 indels 便會在演化過程中因為自然的篩選而被剔除，換句話說具有這個 indel 的個體，便會被自然所淘汰無法存活而留下後代。反之，如果某個區段並不是功能上所必須的，那麼在此所發生的任何形式的突變，並不會對個體的生存造成致命的影響，這些突變便可以被保留下來。因為會轉譯產生功能的區域在演化過程中受到很強的演化壓力，所以這些區域不容許太多突變。根據不同型態的基因體序列我們做了統計，在 84 萬個具人類專一性的 indels 中，有超過 51 萬個落在兩個基因間的區域，將近 32 萬個落在介入子 (intron) 區域，約 1 萬兩千個落在表現子(exon)的非轉譯區域(UTR)，只有 1 千 7 百個落在表現子會轉譯的區域。這樣的分布趨勢，符合序列在功能上的重要性，也再次顯示表現子會轉譯的區域受到特別強大的演化壓力，比較不容許發生突變。在此我們特別強調，我們所找到的非轉譯的區域的 indel 個數與比率極可能是低估，因為這些區域受到的演化壓力較低，很容易發生突變，通常比較不能在多個物種間共同保留。我們進一步分析落在表現子的 indels，令人驚訝的，這些 indels 影響超過七千個人類基因，這佔人類的總基因數約三分之一。我們進一步發現那些落在表現子會轉譯的區域的 indels 的長度高達 55%的不是 3 的倍數，因為不是三的倍數會造成轉譯時的嚴重差異，所以可見我們所找到

的這些具人類專一性的 indels 對人類為何是人類，可能具有相當程度的重要性。不過，因為黑猩猩基因體序列的準確度相對較低，indel 的長度很可能有一定程度的誤差，這不是三的倍數的 indel 比率可能是高估。在基因的功能分析中揭示這些落在表現子會轉譯的區域 indels 與轉錄/轉譯調控、病毒的生命週期、以及催化與運輸等活動有關。特別值得注意的，人和黑猩猩對病毒感染的的敏感度差異很大，例如廣受注目的愛滋病和Ｂ/C 型肝炎等，對人都是具致命性的疾病，但對黑猩猩則可能完全不具致命性或病癥相對輕微許多。研究這些具人類專一性的 indels，很可能可以提供治療這些對人類非常重要的疾病的線索。

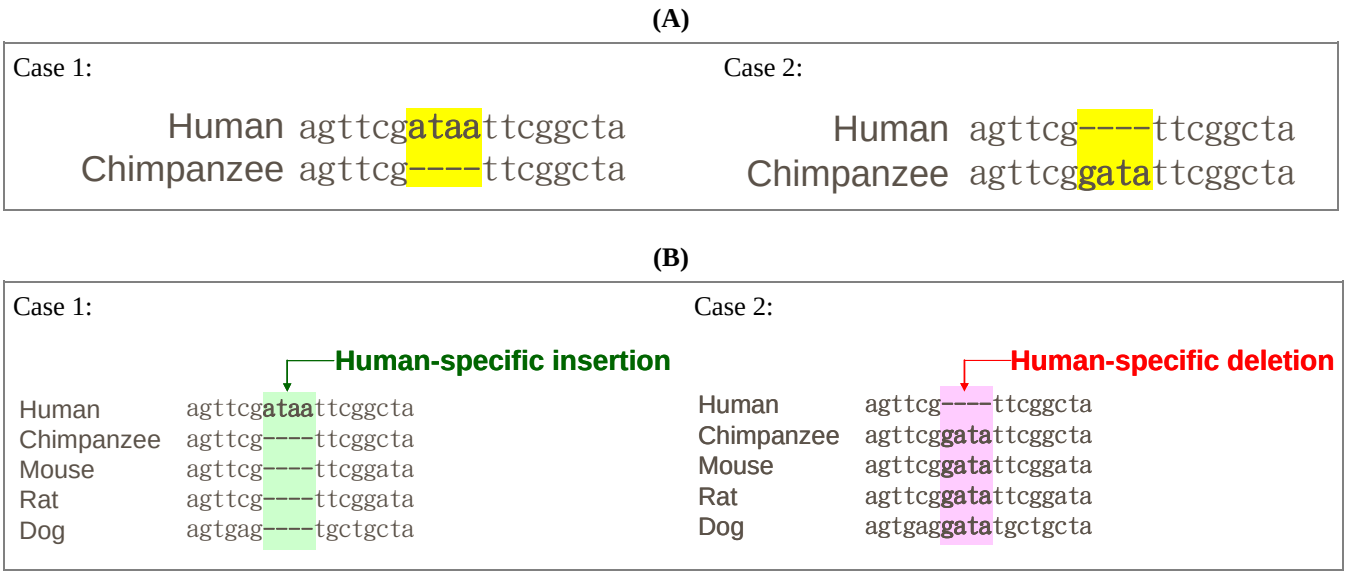

圖 2　(A)基於人和黑猩猩兩物種間基因體序列比對所找到的 indels；(B)利用五種物種多個序列多重比對所找到的 indels，因此我們可以找到具人類專一性的插入（如 case 1）與缺失（如 case 2）片段。

結語

人類和黑猩猩間的遺傳差異，實際上是不同的演化力量和分子機制混合所造成的。兩基因體之間的差別，比一般所謂 1%的單一核苷酸變異來得複雜得多。在這裡我們僅僅討論許多種遺傳變異中的插入與缺失現象而已，要進一步了解人和黑猩猩間的遺傳差異，可以探討的面向還非常多且複雜，如果能結合多種面向的研究，包括基因體學、蛋白質體學，系統生物學和基因環境相互作用等等，或許將使我們更接近「人為什麼是人？黑猩猩為什麼是黑猩猩？」答案。最後，竭誠歡迎**資科、資工本科系**對學術研究有熱忱的畢業生或研究生加入我們的行列，歡迎參觀我們的研究網頁 http://www.sinica.edu.tw/~trees/。欲進一步瞭解以上內容，詳見我們發表的相關文獻：

1. Feng-Chi Chen and Trees-Juen Chuang* (2008). Nucleotide Sequence Divergence between Humans and Chimpanzees. Encyclopedia of Life Sciences (ELS). Invited review article.
2. Feng-Chi Chen, Chueng-Jong Chen, and Trees-Juen Chuang* (2007). INDELSCAN: a web server for comparative identification of species-specific and non-species-specific insertion/deletion events, *Nucleic Acids Research*, 35 (Web Server issue): W633-8.
3. Feng-Chi Chen, Chueng-Jong Chen, Wen-Hsiung Li*, and Trees-Juen Chuang* (2007). Human-specific insertions and deletions inferred from mammalian genome sequences. *Genome Research*, 17(1), 16-22.