我們究竟知道了甚麼？

——「人類基因組序列」最新研究綜述

● 史季鍾（編譯）

「人類基因組序列」的破譯是科學史上的里程碑，其意義之重大可以與人類登上月球、分裂原子或是發明輪子相提並論。這項研究實際上是在兩個陣營的競爭中展開，一個是公共基金支持的「人類基因組計劃」(Human Genome Project，HGP)，另一個是美國塞萊拉遺傳信息公司 (Celera Genomics)，兩者各自獨立，方法不同，但獲得了大致相當的結果①。

2001年2月15日英國《自然》(*Nature*) 雜誌發表了HGP的草案報告，並於12日在互聯網上 (http://www.nature.com/nature/) 公布了所有研究結果。2月16日，美國《科學》(*Science*) 雜誌發表了塞萊拉公司的研究報告，也在網上 (http://www.sciencemag.org/) 公布了主要成果。

通過媒體連篇累牘的報導，整個世界都為「生命天書」的破譯而驚歎，但我們大多數人並不完全清楚自己在興奮甚麼或緊張甚麼：我們究竟知道了多少人類生命的秘密？目前的發現對人類社會的未來到底會產生怎樣的影響？本文根

據《自然》和《科學》雜誌公布的研究報告，以及相關的專家評論和分析文章進行編譯綜述，向中文世界的讀者介紹這項研究的進展和意義。

一　甚麼是「人類基因組序列」？

我們知道，生物是由細胞組織構成的，細胞的各種新陳代謝作用顯現為生物機體的發育、成長、病變、衰老和死亡。每一個高等生物的細胞都有細胞核，其中的染色體就儲存了整個生物的「生命藍圖」，包括其構造和一切自然變化。人類有23對46條分別來自父母的染色體。染色體結構主要由蛋白質和DNA分子組成。基因則是DNA分子的片斷，是含有特定遺傳信息的一段序列結構，也是遺傳物質的最小功能單位。基因具有編碼指令的功能，用來指示控制蛋白質的合成與複製過程，也決定了人類的生物種性特徵。簡單地說，就是基因決定蛋白質的合成，蛋白質決定代謝作用，代謝作用決定各種性狀。

那麼，DNA與基因又是甚麼關係呢？DNA分子的單元成份並不複雜，它是由四種核苷酸分子A、T、C、G構成（它們也被稱為遺傳編碼的字母），而在DNA分子的雙螺旋鏈上面，A和T、C和G總是兩兩配對的，這稱為鹼基對（base pair）。複雜而重要的是DNA分子中鹼基對的排列順序。這些巨長的、排列規則多變的鹼基對序列中，只有一些片斷才是基因。按照功能的不同，基因分為三類：第一類是編碼蛋白質的基因，它具有轉錄和翻譯功能；第二類只有轉錄功能而沒有翻譯功能，包括tRNA基因和rRNA基因；第三類是不轉錄的基因，它對基因表達起調節控制作用，包括啟動基因和操縱基因，有時被統稱為控制基因。

所謂破譯人類基因組序列，就是分辨和確定人類全部23對染色體中DNA分子的所有（大約32億對）鹼基對的完整序列結構，以及在這序列中的基因及其類別。設想一下，如果說人類基因組是一部「生命的天書」，那麼這本書是如此巨大而深奧，它由四種32億對字母（鹼基對）構成，其中的每個句子都極其冗長，而句子當中沒有任何標點符號分割。基因是這些長句子中的一個個獨立片段，其長度從大約一千到十萬對字母不等。可想而知，要確定辨識這本天書中的所有字母的排列順序是何等艱巨的工作。

二　我們發現了甚麼？

目前發表的基因組圖譜覆蓋了人類基因組的95%，平均測序精確程度為99.96%。而基因組序列中每個染色體的圖譜都是如此巨大，以至於我們無法在此直接顯示，有興趣的讀者可以到《自然》和《科學》雜誌的網頁中下載查看所有染色體（chromosomes）圖譜。

對於一般讀者而言，基因組序列圖譜幾乎是無法讀解的「天書」。那麼，科學家從中發現了甚麼呢？

(1) 人類基因的數量稀少

一個重要的發現是確定了人類基因組的大約32億個碱基對中，包含了三萬個左右蛋白編碼基因 (HGP估計為31,780個，塞萊拉公司的報告認為總數不會超過3.5萬個) 。這比原來預計的六萬至十萬個要少得多。與低等動物相比，人類基因的數量只相當於蚯蚓和果蠅體內基因數目的兩倍，人類蛋白質有61%與果蠅同源，43%與線蟲同源，46%與酵母同源。將人類與老鼠的基因相比，人類只不過多了大約300個基因。人類17號染色體上的全部基因幾乎都可以在小鼠11號染色體上找到。

由此看來，人類與低等動物的差別不在於基因數量的多寡，而主要在於人類某些基因的功能和控制蛋白質產生的機制更為高級複雜。人類基因中至少有35%可以有不同的組合方式，因此，它們所能結合成的蛋白質種類比果蠅和蛔蟲要多四倍。換句話說，人類在使用基因方面很「節約」，與其他物種相比更高效。科學家原來認為一個基因只負責合成一種蛋白質，現在看來每個基因平均可以製造三種蛋白質。HGP的首席科學家柯林斯 (Francis Collins) 說，人類不是靠「自我開發」新基因來獲取新功能，而是通過重新編排或擴充已有可靠資源來達到「創新」目的。塞萊拉公司研究計劃的首席科學家樊特 (Craig Venter) 指出，發現人類基因數量稀少具有不同尋常的意義，它改變了科學家原有的「一種基因與一種疾病對應相關」的觀念，今後用於診斷疾病的基因檢測可能將被蛋白質檢測所代替，後者將更精確。

(2) 人類基因的散漫分布、碎片化和重複

人類基因不僅數量稀少，而且在染色體中的分布密度很低、非常分散，整個基因組中只有5%的碱基對序列含有真正 (對蛋白質生成予以指令) 的基因。據估算，人類每100萬個DNA基對中只有12個真正的基因，這個分布密度與果蠅 (117個) 、蛔蟲 (197個) 和水芥草 (Arabidopsis，221個) 相比都是很低的。基因組中存在着大片「荒漠」，大約1/4的區域是長長的、沒有基因的片段。在這上百萬個DNA中尋找12個真正的基因，其難度可想而知，這對於研究所使用的計算機軟件也是嚴峻的考驗。從分布來說，在第17、第19和第22號染色體上基因密度較高，而在X染色體、第4、第18號和Y染色體上相對貧瘠。

另外，人類基因的「碎片化」程度也比其他生物更高。要理解碎片化的問題，需要了解「編碼段」(exon) 和「內區段」(intron) 這兩個概念。就功能機理而言，基因並不是直接合成蛋白質，而是首先要形成 (包含編碼信息的) mRNA分子，這些mRNA才成為蛋白質合成的模塊。基因序列中實際上只有一些片斷才能

形成mRNA分子，這就是(有用的)「編碼段」，而編碼段之間的區域是(無用的)「內區段」。

研究發現，人類基因序列中無用的內區段既多又長，而有用的編碼段既少又短，這就是所謂的「碎片化」。《自然》的高級編輯吉(Henry Gee)用了一個形象的比喻形容基因的「碎片化」：基因就像是插播了太多廣告的電視節目，包含遺傳信息的真正節目(編碼段)通常比「插播」的那些無用的廣告(內區段)小得多。

比較而言，果蠅和蛔蟲基因的內區段長度一般只有幾十個到上百個碱基對，人類基因的內區段則長短懸殊，大多數都只有87個碱基長，但還有許多長得驚人，甚至長達一萬個碱基。平均下來，人類基因的內區段長度約為3,300多個。而編碼段序列要短得多，已知的編碼段只有19個碱基對大小。

研究還發現，基因組的字母序列中有1/3以上重複，19號染色體中有57%的重複。半數以上的DNA包含有不同類型的重複序列。因為以前發現，有些脊椎動物的基因組中重複序列很少，卻也具有完好的編碼功能，人們由此認為，重複的DNA片斷是無用的、寄生性的，即所謂的「垃圾DNA」(junk DNA)。但目前科學家認為，這些重複的序列到底有沒有功能是值得進一步研究的。

同時，人類寄生性DNA的重複序列中，很少有基因突變所要求的「重新插入」(reinserting)的迹象，多半是垂老不變的，這可能是漫長進化所留下的遺迹，為研究人類的基因進化提供了線索。

(3) 人類基因與種族和性別

塞萊拉公司的研究計劃採用了三位女性和兩位男性貢獻的DNA樣品，其中有一位是非洲裔美國人、一位是亞洲華人、一位是西班牙裔墨西哥人、還有兩位是白種人。研究發現，在整個基因組序列中，人與人之間的變異僅為萬分之一，只有1,250個序列片斷是不同。也就是說，地球上人的個體與個體之間有99.99%的基因密碼是相同的，差異變化僅僅為0.01%。實際上，來自不同人種的兩個人完全可能比來自同一人種的兩個人在基因上更為接近。因此並不能通過單獨基因或組合基因來確定不同的人種。也就是說，在基因生物學的尺度上，「種族」並不是一個科學的概念。

研究還表明，男性染色體Y的序列是高度重複的，其中的大多數序列不具有功能，被稱為「遺傳垃圾場」(genetic junkyard)。只要想想人類有一半人口沒有Y染色體，就不會對此感到意外。雖然Y染色體的基因含量相對較低，卻有很高的研究價值。Y染色體表現出一種「固步自封」的品性，其中95%的DNA不與X交換，這包括了那個界定了男性性別基因的區域，這使得X與Y染色體互相間離，產生穩定的男女性別差異。X染色體仍然與其「姐妹」在雌性細胞中交換DNA，而Y染色體因為與X隔離，相比而言就變得退化、失去了許多遺傳材料，累積了突變的可能。

通過分析Y染色體上300萬個重複單元，檢測重複散布的形式，研究者估計了X和Y染色體的相對突變率，發現男性和女性的突變比例為2：1。這表明，人類進化的主要因素是定義男性性別的染色體Y，即男性在人類進化過程中扮演了至關重要的角色。

⑷ 基因與疾病

人類以及其他物種的基因組序列的完成將大大增進我們對疾病的認知。疾病的一個重要成因是單個或多個基因的突變，主要發生在基因組的所謂「多態性區域」(polymorphism，即人類個體之間可能發生差異的DNA序列區域)。人體中的基因發生突變，就可能讓病變細胞以無法控制的方式增長和分裂。通過對「單核苷多態」(SNP，由單個核苷的變化導致的多態性)圖譜的研究，我們將會對許多已知和未知的疾病的基因成因加深了解。

HGP的研究發現了1,778個疾病基因，相當於未知疾病基因的40%，其中至少有30種疾病基因已經通過克隆定位法被完全定位。在23對染色體中，至少有三組基因和遺傳性疾病高度相關，第1號染色體中有與阿茲海默症相關的基因，第6號染色體的基因與人類智能有關，X染色體更是帶有許多疾病基因，和數十種疾病有聯繫。研究者認為，癌症實際上也是一種基因疾病。

⑸ 基因與蛋白質：複雜的機理

人類基因組中的基因數量不多，但功能強大而複雜。某些基因能夠不斷地進行轉錄和翻譯，給出編碼指令，合成各種蛋白質，這被稱為「基因表達」(gene expression)。每個細胞都有一套完整的基因表達調控系統，使各種蛋白質只有在人體需要時才被合成，這使得生物得以適應多變的環境，避免生命活動中的浪費現象和有害後果，保持人體正常的代謝過程。但基因表達的機理是如此複雜，對此我們還遠遠沒有獲得完全準確的知識。

作為HGP相關研究的一部分，圖普勒(Rossella Tupler)、佩里尼(Giovanni Perini)和格林(Michael R. Green)在本期《自然》上發表了他們關於基因表達的研究結果。他們着重研究了控制基因表達的三個階段：從基因到mRNA的轉錄；mRNA的初步接合，以及多聚A尾的加入。他們發現，雖然人類的許多基因序列與酵母和果蠅的序列十分相似，但人類有更多的特殊轉錄因子相關序列TFIID，這意味着人類TFIID具有更豐富的潛在多樣性。他們還發現，在轉錄激活劑的過程中，有2,000多個基因編碼了這些蛋白。在尋找與mRNA接合有關的基因時也發現了相同結果。在研究多聚A尾加入的有關基因時，他們意外地發現了新的基因，這表明人類的基因表達機制極其複雜，許多新基因可能表達了大蛋白複合體的亞單位，需要進行進一步的研究來理解其功能。由此，未來的生命科學研究將以人體的蛋白質形成為主要目標。

三　人類基因組圖譜的意義

設想有一群語言考古學家，在岩洞中發現了一部遠古時代的巨大書稿，記錄着一個逝去文明的全部知識，他們打開了這本書稿，甚至辨識了書稿中刻寫的所有符號字母，這無疑是令人興奮的重大發現。但一切仍在費解之中，因為他們不知道這些符號字母所構成的每個單詞的確切意思，也不知道這種語言的句法和結構，幾乎無法破譯其中的秘密。

《自然》雜誌的顧問編輯鮑爾 (Philip Ball) 以這樣一個情景來類比目前人類基因組序列 (基本) 完成的意義。我們「生命的天書」的狀況很像是那部古代的神秘書稿，即使你可以辨識每一個字母並確定所有字母的排列順序，你仍然還無法真正讀懂它。對生物學的真正革命而言，目前的研究成果僅僅是一個開始，更為重要的或許不是對基因組序列的確定，而是發現在這些序列中基因活動和相互作用的模型。用鮑爾的話來說，「基因組沒有敍事而只是一部詞典，所有的故事是在詞與詞的相互動態關聯中才能展開，正像細胞所展開的故事一樣」。

HGP的首席科學家柯林斯說，真正的工作現在才開始。研究人員將把注意力放在基因表達的分析上，他們會面臨數以千計極端複雜的蛋白質功能探測工作，而這個工作可能需要花上100年的時間才能完成。

顯然，面臨的挑戰是巨大的。因為細胞中有上千個基因在即時活動，以回應不同的人體狀況，其中涉及到複雜的活化與復原模式。最為關鍵的是，生物學目前還沒有任何理論框架來對此進行描述，甚至不清楚應該如何建立這樣的理論構架。

但科學家已經開始了閱讀「細胞故事」的艱巨探索。其中一個方法是運用所謂「DNA微列」(DNA microarrays) 或稱「基因芯片」(gene chips) 技術。以一種特殊的裝置和標記技術，可以同時監測到細胞中哪些基因在活動 (即使還不知道它們的功能)，從而促進我們探索基因表達的機理。雖然這項技術是初步的，但開啟了「後基因生物學」的一個路徑。

無論如何，人類基因組序列的基本確定是一個重要的開端，它將極大地推動生命科學領域的一系列基礎研究的發展，闡明基因的結構與功能關係，細胞的發育、生長、分化的分子機理，疾病發生的機理等等。如果在這些研究領域獲得重大的進展，那將會對人類生活產生更為直接的衝擊和影響。

註釋

① 塞萊拉公司的核心分析方法被稱為「霰彈法」，HGP則採用了「克隆法」，兩個研究組將數據進行的對比以及人類基因組工程的科學家、《科學》和《自然》雜誌高級指導編輯的評估表明，塞萊拉公司的基因組分析與人類基因組計劃的分析結果雖然存在一些差異，但大部分地方都極為吻合。