次世代定序在轉錄體學之應用：兼談白條海葵魚體色形成轉錄體學之研究

劉恩良[1]、陳定濰[2]、劉孜孜[3]、Frank Stahl[4]、何源興[1]

[1]水產試驗所東部海洋生物研究中心、[2]長庚紀念醫院生物醫學轉譯研究中心
[3]國立陽明大學榮陽基因體研究中心
[4]Leibniz Universität Hannover, Institute for Technical Chemistry, Hannover, Germany

前言

人類基因體解碼之後，加上微陣列 (microarray) 技術的發展，使得轉錄體學 (transcriptomics) 分析成為可行的研究題材。過去基因 (mRNA) 的表現，可藉由微陣列技術與即時定量 PCR (quantitative real-time PCR) 加以分析。前者比較不靈敏，而後者雖具有高靈敏度，但價格昂貴且無法適用全基因體層次的基因表現分析。DNA 定序一直是分子生物學研究中最主要工作平台之一。自從次世代定序 (next generation sequencing, NGS) 技術建立以來，已經提供高通量的基因表現研究、基因體的註解與非編碼 RNA (non-coding RNA) 的發現。這種快速大量的定序技術，使得過去費時 15 年，耗費 27 億美元的人類基因體計畫，在今日只需 8 天和 10 萬美元的花費便可完成。在 NGS 技術中，重要的核心原理是使用合成定序的概念 SBS (sequencing-by- synthesis)，利用次世代定序技術進行 RNA 的定序，都可稱作 RNA-Seq，近年來會特別只對 mRNA 的定序為 RNA-Seq。最近，RNA-seq 已經發展成轉錄體定量的重要系統，不同的生技大廠也以不同的定序原理，設計出各具特色的 NGS 定序機台 (圖 1)。

NGS

英國科學家 Friedrich Sanger，在 1970 年所發明的雙去氧核醣核酸鏈終止法 (dideoxynucleotide termination method)，是過去 DNA 定序的主要方法，而 pyrosequencing 主要是利用動態 DNA 合成，以冷光模擬合成過程中，每一個核苷酸嵌入時所釋出的焦磷酸鹽基 (pyrophosphate; PPi) 來偵測每一個對應的訊號，以螢光素酶為基礎的複數酵素套組，在每一個核苷酸嵌入後，會產生一個光訊號，四種不同核苷酸依序嵌入時，只有嵌入的核苷酸，會產生一個訊號，其中複數酵素套組，包含有 DNA Polymerase，ATP sulfurylase，luciferase 以及 apyrase，主要是放大合成反應。在一次序列操作中，NGS 可以分析百萬單位短片段 DNA。依據不同的 NGS 作業平台，片段長度在 25－450 鹼基對 (bp) 之間 (除 Roche 454 的定序長度可達

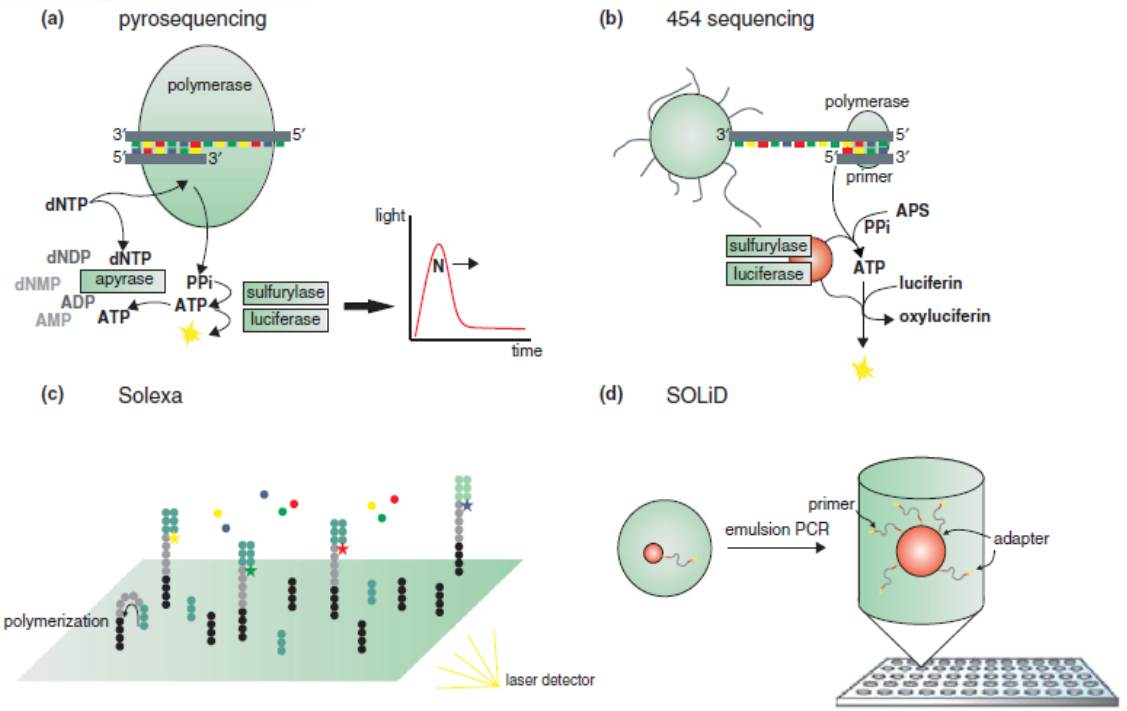

圖 1　NGS 技術基本原理

(a)焦磷酸測序法：每一個新核苷酸嵌入時，會產生一個光訊號；(b)Roche 454 機型：隨著焦磷酸基釋放會產生一個對應可測的光訊號；(c)Solexa 機型：DNA 片段構築成雙股分子橋，被螢光所標誌的分子加入後，測序循環反應隨即啟動；(d)SOLiD 機型：DNA 片段與連接分子結合後，進行乳化 PCR 反應，產生待測的分子 (Mutz et al., 2013)

700－1,000 bp)，雖然沒有達 Sanger 法的長度，但 NGS 一次操作可達 50 GB (gigabases) 的高通量，提供了有效分析大量樣品的基礎。RNA-seq 分析流程主要由四個基本分析步驟組成：(1)原始影像數據轉換成短片段可讀之序列；(2)依序排列對應到已知基因體或轉錄體標準參考資料庫；(3)計算對應排列的讀數量和演算基因表現程度；(4)統計分析決定不同基因表現 (圖 2)。

原始 NGS 數據是由螢光訊號組成，之後藉由操作平台專一性螢光訊號與核苷酸的轉換，成為一段核苷酸序列，當輸出成標準 FASTQ 格式之後，序列誤差之計算，依廠家

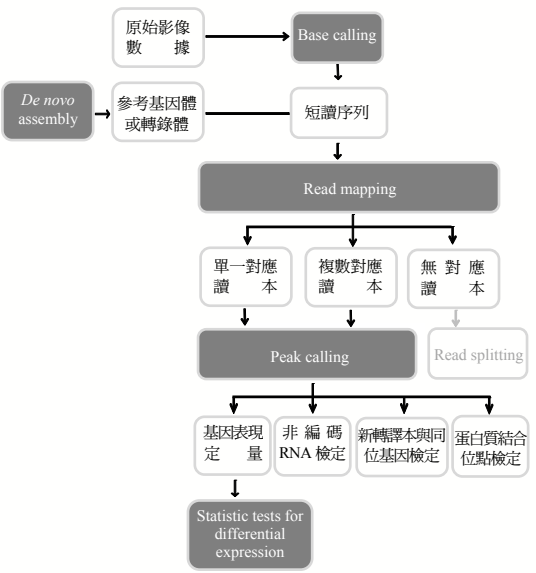

圖 2　RNA-seq 分析流程 (Mutz et al., 2013)

內建軟體而有所不同，因此，對結果分析應該使用適當對應方式來解讀。此外，讀數排列時需要有參考序列，目前基因體序列可以從線上資料庫 genome online database (GOLD) 取得，目前已有超過 3,000 種生物的基因體序列可用，而新註冊的物種持續增加中。

與傳統對應排列不同，NGS 使用索引式 (indexing) 策略，使數百萬個短片段序列，在一定允許時間內完成對應排列的動作。目前最常用的運算方式是散列表運算 (harsh table lookup) 和塊排序壓縮 (burrow-wheeler transformation)。前者具有較高的靈敏度，但後者演算時間較短。大多數的序列排列方法允許少數錯誤對應，原因可能來自定序本身的錯誤，單核苷酸多型態 (SNPs) 與突變；但不允許過大錯誤的缺口，而且可允許錯誤比對核苷酸的數目與所選取序列長度有關。對真核生物 RNA-seq 分析而言，cDNA 讀序結果，最好與可參考的轉錄體比對，然而現今缺乏轉錄體完整資料庫的情況下，多數研究人員還是以基因體序列參考。為了以 RNA-seq 來估算基因表現高低時，對應到特定基因的讀次必須被定量。這些定量後的讀次，經回歸標準化後，可用來比較不同基因和不同實驗組別中基因表現的程度。其中，回歸後基因表現的分數與統計分析有兩種方法，分別為參數與非參數運算法。參數運算法使用一般機率分布，如二名式分布 (Binomial) 或帕松分布 (Poisson)。非參數運算法則是根據實際數據模擬雜訊分布 (noise distribution)，對定序深度依賴性較小，而且呈現較穩定的結果。

NGS 的運用

一、人類疾病的分子診斷

最常見 NGS 技術平台運用在基因體定序。NGS 技術革命使得大量定序工作顯著減少時間與金錢的需求，另外 NGS 也被運用在基因體或特定目標再次定序對應到已知參考序列的片段，可藉由再次定序確認，諸如 SNPs、小片段 DNA 嵌入或缺失、重複片段數差異，以及其他基因結構上的變異。因此，提升了對遺傳疾病基因與外表型特徵的分析靈敏度。最近，NGS 平台發展出偵測基因體突變的模式，用以檢測不正常 DNA 序列改變與新型疾病基因，這項技術運用在產前與產後遺傳病的診斷上，例如唐氏症 (Down syndrome) 與帕金森氏症 (Parkinson's disease)，目前許多個人化癌症基因體也被成功的定序，並發現許多與癌症初期發生有關的 DNA 突變。

二、表觀遺傳學 (Epigenetics) 研究

NGS 技術也被利用在基因體層次上表觀遺傳學修飾 (epigenetic modification) 的研究，諸如 DNA 甲基化、染色絲結構變異、組織蛋白後轉譯修飾，這些與基因體調控、疾病發生與癌症形成有關細胞生物學重要的過程。另外，DNA 與蛋白質的相互作用，轉錄因子結合分析與核小體定位都藉由 NGS 技術導入，促成相關領域資料庫大幅累積。

三、總體基因體學 (Metagenomics) 研究

總體基因體學是近年發展出來，使用 NGS 技術分析環境與生態的學門，從水、土壤、沉積物、腸道內容物，分析細菌 (16S) 真菌 (18S) 在樣品中的組成與數量。經過 NGS

測序後，利用相似度將序列分群，再與相關參考資料庫比對後，就可以得知樣品中菌相組成與各菌種數量，進一步評估這些微生物對環境、人類或其他生物的影響。

四、非編碼 RNA (ncRNA) 研究

為了強化對已定序基因體的註解，NGS 也被運用到小片段非編碼 RNA (ncRNA) 的發現與側錄分析上，ncRNA 包括 tRNA、rRNA、核內分子 RNA、核仁小分子 RNA、微型 RNA (miRNA)、短小干擾 RNA (siRNA)。這些 RNA 雖不被轉譯成蛋白質，但對基因表現卻有重要的影響。NGS 技術已經促成許多全新小型 RNA 的發現。

五、海水觀賞魚轉錄體學之研究

近十年來，東部海洋生物研究中心在海水觀賞魚繁殖研究，累積了相當豐碩的成果；為了深化研究的內涵與促進知識的永續發展，與長庚大學、陽明大學合作，並獲得德國漢諾威大學技術之協助，以臺灣東部常見的白條海葵魚 (*Amphiprion frenatus*) 為模式魚種，藉 Illumina HiSeq 2500 技術平台，分析海葵魚孵化初期到體色條紋形成之基因表現 (圖 3)。我們以孵化當天與孵化後 25 天為採樣點，每個採樣點各建立 2 個待定序的

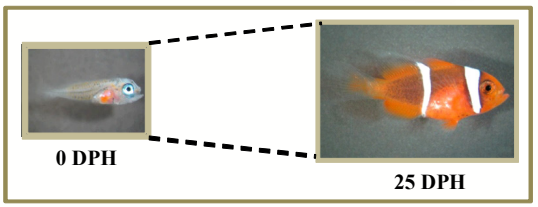

圖 3　海葵魚孵化後到體色條紋形成之形態學與體表色素之變化 (DPH：孵化後天數)。孵化當日體表除黑色素之外，不具有其他體色與體斑

基因庫。所建立基因庫片段分布在 250－900 鹼基對之間；平均長度約 440 鹼基對；基因片段讀數 74－95 百萬讀次；總計 352.90 百萬讀次；平均讀長為 80 鹼基對。與斑馬魚 (*Danio rerio*) 基因體比對後，發現 68.4 百萬讀次 (19.4%) 有對應關係 (如表)。基因彼此間並非單獨運作，也會受到生理或外在環境的影響。要人工手動分析上千個至數千個基因數據點的相關性，是生物科學領域的一大挑戰，因此對 OMIC 差異表達的數據進行〝生物學解析〞是當前不斷增長的重要需求。我們交互運用兩種分析策略來進行生物學解析的註解工作：

(一) 路徑地圖 (pathway maps) 的建立

基於先前生物化學知識，所預先定義的調控關係。路徑地圖是具備明確定義的生物

白條海葵魚經 Illumina HiSeq 2500 定序平台之後所得之數據量

sample name	library size (bp)	library avg size (bp)	reads (PE read)	total base (Mbases)
AF25DPH-1	250-900	448	93,908,312	9,391
AF25DPH-2	250-900	449	95,326,327	9,533
AF0DPH-1	250-900	444	85,356,836	8,536
AF0DPH-2	250-900	447	74,282,884	7,428

學共識，路徑訊息可從特定公開的單一資料庫獲得 (如 KEGG, Biocarta, Gene Ontology, etc.)。

(二) 生物網絡 (biological networks) 的建立

生物網絡是由新測得的訊息類型、所採用的實驗和自然環境知識數據庫所建構。這些數據庫結合多個經由實驗驗證與預測的來源訊息，諸如轉錄因子調控網絡 (TF regulatory networks)、表基因體調控 (epigenetic regulation)、微型核醣核酸與信息核醣核酸調控網絡 (miRNA-mRNA networks)、基因代謝物網絡 (gene metabolites) 等。

依照路徑地圖的註解，將上千個差異表達的基因加以分群，找出可能顯著參與的基因叢集，再進一步運用生物網絡的註解資訊，將特定基因叢集裡與叢集間的調控關係建構起來。交互運用註解資訊的深入探索，有助於我們擷取特定生理現象背後的基因調控機制。這也是利用高通量的分子定序資料，探索物種發育過程中基因與染色體所產生複雜變化的有效方法。藉由上述的方法，我們從 31,967 個基因中發現 4,245 個基因，在孵化當天與孵化後 25 天，依照表現量可區分為兩大群組；其中 866 個基因 (20%) 在 Pathway Studio 程式分析中有顯著直接互動作用的關係。同時使用 Gene Ontology Enrichment 分析時發現過程表現差異的基因群中，顯著與代謝過程 (1,687 genes, p = 2.95E-41) 和分化過程 (437 genes, p = 1.40E-16) 有關。代謝過程相關的是具催化活性 (1,311 genes, p = 5.09E-19)，生物合成 (465 genes, p = 8.26E-20)，分解作用 (285 genes, p = 4.45E-10) 的基因。特別的是有 27 個跟醣解作用有關的基因，在孵化後 25 天，表現量全數下降。我們進一步發現：在孵化階段，與代謝有關的基因 (995 genes, p = 4.44E-37) 表現量有上升的現象；而與分化有關的基因 (258 genes, p = 2.91E-23) 表現量卻是下降的。進一步基因表現的位置，實驗正在進行中。

結語

近年來，功能性轉錄體學藉由微陣列技術與 RNA-seq 平台，已獲得突破性的進展。然而，微陣列技術已到達技術極限，而逐漸的被高通量 NGS 技術所取代。RNA-seq 技術可以計算轉錄層次所有轉錄本 (transcripts) 的數量，偵測出新型轉錄本與同位基因，以及確認先前註解 5’ 端與 3’ 端 cDNA 序列，對應列比外顯子 (exon) 與內含子 (intron) 交界序列，顯示序列變異，如 SNPs 和 RNA 剪接 (RNA-splicing) 變異。過去 7 年中，NGS 已經被發展成可適用於每一種含 DNA 分子研究領域所適用的技術平台 (all in one platform)，然而考量費用，經濟上仍誘因不足，目前有不同的研究團隊與生技大廠正著手進行所謂第三代定序技術的研發，達成人類基因體定序費用在 1,000 美元以下的目標，將是指日可待。東部海洋生物研究中心也將在海水觀賞魚體色形成轉錄體學的研究上，緊跟世界尖端科學技術的步伐，對臺灣水產養殖業的知識寶庫作出應有的貢獻。

註：本文主要參考 Current Opinion in Biotechnology 2013, 24: 22-30.