淺談次世代定序技術（Next Generation Sequencing, NGS）發展與其應用

邱燕欣[1]、林詩舜[2]

一、前言

農業領域的發展與生物技術研發步伐息息相關，過去傳統育種技術需靠大量園藝性狀觀察進行子代篩選評估、抗病育種則需進行特定病原接種與篩選、病原菌檢測則必須透過病原分離培養與鑑別等過程才能進行確認。在生物技術導入應用後，可藉由分子標的篩選輔助，進行親代或子代的遺傳物質抽取後以特定標幟試驗判別，縮短育種試驗之時程、病原的檢測，可藉由分子或血清檢測快速得到診斷結果，加快栽培者對於作物管理之作為。因此農業栽培者、相關研究單位除了專注於栽培研究之外，應多關注醫學領域、動物研究領域與各項生物技術之研究發展，為植物研究領域注入新思維。次世代定序技術（Next Generation Sequencing, NGS）為近 10 年來快速發展應用的生物技術，本文即針對該技術之發展、原理、與差異比較，以及該項技術植物科學研究之應用作相關介紹。

二、定序技術之發展

在 50 年代前生物研究的發展，仍停留在生物體的遺傳性狀、表現特徵進行統計分析，但自 1953 年 James Watson 和 Francis Crick 兩位發現去氧核醣核酸的構造後，進而開啓了分子生物學的領域。1977 年 Sanger 與 Coulson 兩位發展鏈終止法（chain termination method）定序技術，定序技術成為分子生物技術之基礎資訊之一，Sanger 亦因此項技術贏得 1980 年諾貝爾化學獎。1990 年由美國啓動國際人體基因組計畫，其參加國家包括美國、英國、德國、日本、中國以及印度。最後提早於 2003 年 4 月公告完成，共計完成 3.3 Bbp （百萬鹼基 Billion bp），歷時 13 年定序共計 23 萬個基因。而生物化學材料的研究加速了定序技術的演進，2000 年開始陸續有生技公司投入次世代定序技術（Next Generation Sequencing, NGS）之研究。NGS 開發的原理，主要可分為以下三大部分：

1. 合成性定序（sequencing by synthesis）：包括 Roche 公司開發的 454 系列；Illumina 開發的 Illumina HiSeq 平台系列，包括 Hiseq® 2500、Hiseq 2000、Hiseq 1500、Hiseq 1000 等；以及 Life Technologies 公司旗下的 Ion Torrent 系統。
2. 接合性定序（sequencing by ligation）：包括 Life Technologies 公司旗下的 SOLiD 系統與哈佛大學試驗室合作開發的 Polonator G.007 系統。
3. 單一分子定序系統（Single Molecular Real Time Technology, SMRT）又稱第三代定序技術：PacBio （Pacific Biosciences）開發之 Pacific Biosciences；Helicos Biosciences 開發的 Helicos；Oxford Technologies 開發之 Nanopore 等。

NGS 之間的差異迥異，包括 1. 輸出、定序時間；2. 可定序之最大長度；3. 定序模組；4. 定序後續分析平台的選擇。舉例來說 Illumina 開發的 Hiseq® 2500 最長定

1 種苗改良繁殖場繁殖技術課　助理研究員
2 國立臺灣大學生物科技所　副教授

序長度爲 200 bp，可在 24 小時定序一組人類的基因組，約可產生 120 Bbp的定序資料，資料輸出量更可達 600 Gbp （十億鹼基Giga bp）。

三、次世代定序技術之比較

可以從定序的標的物及已開發次世代定序技術的機型及差異比較來介紹NGS。依照定序的對象（表一），可區分爲去氧核醣核酸（DNA）階段、核醣核酸（RNA）階段。依照定序生物是否爲模式生物又可區分爲全基因體定序（whole genome sequencing, WGS）、基因組重定序（Resequencing）以及非模式生物基因體定序（de novo sequencing）。依單一定序物種的複合性可分作單一物種或是多源性基因體學（Metagenomics）。依照DNA定序區間的特徵，可爲特殊區間的增幅子（Amplicon），再細分作爲外顯子定序（Exon）與標的區間（Target Region）定序。如人類白血球抗原基因定序（human leukocyte antigen (HLA) gene typing），染色質免疫沉澱核酸定序（Chromatin Immuno-precipitation sequencing, ChIP-seq），基因組甲基化與否的分析（Methylation analysis）等。經由全定序後的基因體組，可藉由後續生物資訊的比較，搜索特定的序列或設定搜索條件，進行多種基因偵測，包括單一核酸多型性 （single nucleotide polymorphism, SNP）偵測與驗證、核酸構造變異（Structure variants）、Indels偵測 DNA層次是否有序列插入（insertion）或缺失（deletion）、單倍體基因組及演變分析（Haplotypes and phasing）與鹼基修飾偵測（Base modification detection）等。而 RNA 則可依照定序標的是否爲模式生物區分爲轉錄體重定序或非模式生物轉錄體定序 （Transcriptome de novo）及族群標的核酸定序（Sequencing Capture metagenomics）。

完整的定序平台主要包括樣品的取得與製備、定序及定序資料分析三大步驟（如圖 1）。根據不同的定序技術所得的序列資訊，必須經過連續性的生物資訊處理平台進行，包括 1. 定序資料回收進行品質管控篩選獲得正確定序資料；2. 定序資料相互比對（alignment）、組合（assembly）與定序次數分析 （read count）；3. 資料庫比對，確認重組序列之完整性及正確性；4. 基因功能性註解（annotation）；5. 定序資料間比對等。

因定序目的有不同的定序需求，必須事

表一、以定序標的區分次世代定序技術

核酸層次	次世代定序技術	細項應用
去氧核醣核酸	全基因體定序	基因組重定序
	非模式生物基因體定序	單一核酸多型性偵測與驗證 Indels偵測 核酸構造變異偵測 單倍體基因組及演變分析 鹼基修飾偵測
	表觀遺傳物質定序	染色質免疫沉澱DNA定序 DNA 甲基化定序
	外顯子定序與標的區間定序	
	多源性基因體學定序	
核醣核酸	非模式生物之轉錄體定序	
	小分子RNA定序	
	特定訊息去氧核糖核酸定序	

先評估包括單次定序長度需求（Read length）、單端定序（single end）或雙端定序（paired end）及定序覆蓋度（dept of coverage）等。而最大產出數據可套用以下公式得出：最大產出數據=單次定序長度 × 覆蓋度（dept of coverage）×1 單端定序（single end）或×2雙端定序（paired end）。

2014年*Barba*等人針對7家次世代定序平台進行比較（如表二），可以得知定序長度、準確度、最大產出數據等會因定序技術不同而有所差異。覆蓋率則決定於最大產出數據與基因體的大小的倍數。在相同的定序平台下，較大基因體的覆蓋率則會下降。表三爲該團隊整理，應用研究領域根據不同次世代定序平台的特色，選擇不同定序平台。

四、NGS技術在植物科學研究之應用

基因體的定序技術的開發，加速了各項生物科學的發展。在植物科學領域亦然，研究主題可以藉由NGS的導入，增加領域的深度與廣度，不同定序平台的結合更能爲研究帶來更多的想法與驗證。NGS 技術應用於植物科學研究之領域區分爲6大部分：

1. 演化與生態分析（Phylogenetic and ecological studies）：在NGS技術未廣泛利用以前，非模式植物之基因庫資料相對模式植物來說薄弱。在野外進行生態或演化調查而言，物種的廣泛度與歧異性更是大，鮮少針對生態群進行相關分析。NGS 的發展可以利用物種間的定序資訊比對演化親源性，進行族群定序，了解區塊生態間的生物分類。例如土壤材料間的生物族群分析、種子族群的潔淨度。
2. 基因庫建立（Applications for large gene bank collection）：保留生物材料之際，同步進行多種植物基因體組的定序，保留分子資訊材料的多樣性。如果在該屬作物發現可利用之特殊基因，可藉由生物資訊的快速搜尋分類，之後接續挑選利用。
3. 作物育種與分子標記輔助（Breeding molecular marker development）：利用育種特殊分子標記，在基因體組定序拼裝後，在生物資訊電腦運算中，加入搜尋序列特徵，分析 SNP 之位點，作後續分析之應用。或利用植物分子標記之分析，加入資訊演算之特徵法，找出可利用於作物品種鑑定之簡單序列重複（simple sequence repeat, SSR）。2014年Satya等人歸納了利用NGS協助

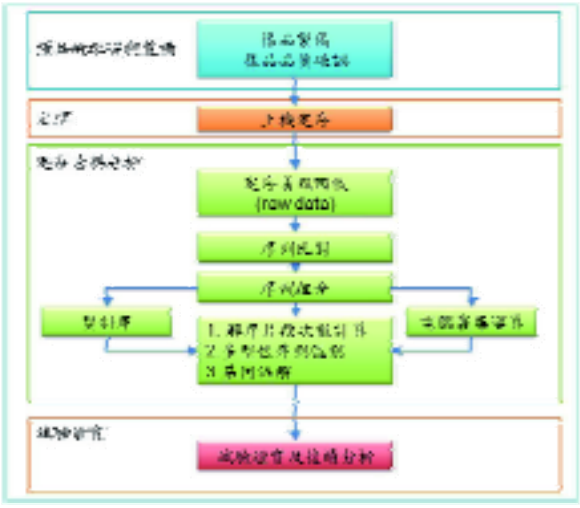

圖1｜定序平台執行流程

表二、次世代定序技術比較 (*Barba* et al. 2014)

定序平台	增幅技術	定序原理	定序長度 (bp)	定序速度 (bp/小時)	最大產出數據	準確度	*MID 偵測率
454 (Roche)	乳化PCR	焦磷酸定序（Pyrosequencing）	400-700	13 Mbp	700 Mbp	99.9	0.10, 0.3, 0.02
Illumina (Illumina)	橋式PCR	Reversible terminators	100-300	25 Mbp	600 Gbp	99.9	0.12, 0.004, 0.006

定序平台	增幅技術	定序原理	定序長度 (bp)	定序速度 (bp/小時)	最大產出數據	準確度	*MID 偵測率
SOLiD (Life Technologies)	乳化PCR	Ligation	75-85	21-28 Mbp	80-360 Gbp	99.9	誤差率高於Illumina
PacBio (Pacific Biosciences)	不經增幅單一分子解讀技術 (No amplification Single molecular real-time, SMRT)	Fluorescently labeled nucleotides	4,000-5,000	50-115 Mbp	200 Mbp-1 Gbp	95	1, 2, 12
Helicos (Helicos Biosciences)	不經增幅單一分子解讀技術	Reversible terminators	25-55	83 Mbp	35 Gbp	97	誤差值率略高於454及Illimina與InDels有關
Ion Torrent (Life Technologies)	乳化PCR	Detection of released H	100-400	24 Mbp-16 Gbp	100 Mbp-64 Gbp	99	M, 0.06, I+D 1.38
Nanopore (Oxford Technologies)	不經增幅單一分子解讀技術		可達50Kbp	150 Mbp	>10 Gbp (Tens of Gbp)	96	

*MID偵測率：M爲鹼基配對錯誤（mismatch）、I爲鹼基插入（Insertion）、D爲鹼基缺失（Deletion）

表三、現今定序平台與相關應用領域 （Barba et al. 2014）

定序平台	技術應用
454 GS FLX + System （GS FLX Titanium XL+/ FLX Titanium XLR70） *GS junior System （bench top）	DNA定序：基因全定序、非模式生物定序、大片段（1Kbp）基因組重定序。增幅子定序；RNA定序：轉錄組定序、捕捉性族群定序（sequencing capture meta-genomics）。 時間：10-23 小時 *定序片段大小爲400bp
HiSeq System （2500/2500/1500/1000） Genome analyzer IIx, HiScan SQ, * MiSeq （bench top）	DNA定序：表基因組定序（epigenetic sequencing）：染色質免疫沉澱定序（ChIp-Seq）、DNA甲基化定序分析。 RNA定序：轉錄組定序、小分子RNA定序、訊息RNA（mRNA）定序、表現基因組定序分析 時間：8-14 天 *在 25M 的定序次數及每次 2x300bp 的資料輸出量可達 15 Gbp（= giga base pairs = 1,000,000,000 bp），可應用於表現子定序（Exome）、族群定序（meta-genomics）、人類白血球抗原基因定序（human leukocyte antigen（HLA） gene typing）、訊息RNA定序、標的基因表現（包含蛋白質表現與非表現區間如rRNA、tRNA、smRNA等） 定序時間：20-35 小時
SOLidD 5500 W Series Genetic Analysis Systems （5500W, 5500xlw）	DNA 定序：基因全定序、表現子定序（Exome）、表觀遺傳基因定序（epigenetic sequencing）；RNA定序 時間：7-12 天
PacBio PACBIO RSII	DNA定序：應用SMRT技術，爲現階段最長之定序技術。 基因變異性、甲基化（methylation）、標的基因定序及偵測差異：SNP偵測與驗證、基因插入與去除（indels）、構型變異、多套體、單體型定相（Haplotype phasing）、鹼基修飾偵測（確認基因表現）、病原寄主交互作用偵測、DNA損壞與修復 時間：30 分鐘
Helicos Genetic Analyzer System	DNA定序與RNA定序 時間：8 天 使用半導體定序原理，最長定序長度爲400bp，可適用於基因組小的生物
Ion Torrent Ion PGM System （bench top） *Ion Proton System（bench top）	DNA定序：適用於微生物基因、基因組、增幅子、表現子、標的基因、病原型及其他病原相關基因定序。 RNA定序時間：4.5 小時 *使用半導體定序原理，可適用於基因組小的生物如微生物基因組、表現子以及轉錄體。 *時間：4.5 小時
Nanopore GridION System （bench top） *MinION System （a miniaturizes disposable device for single use）	DNA定序：表觀遺傳基因組、基因組驗證 RNA定序：適用於直接分析RNA，非反轉錄成cDNA後再行定序 時間：可小於60分鐘 *僅適用於DNA定序，如血液DNA定序。

作物育種分析之流程（如圖 2）。

4. 雜交技術與基因滲透（Hybridization and introgression）：可利用長度較長、覆蓋率較高的轉錄體組定序，獲得物種的轉錄體體組後，可以計算物種基因在同義變異核酸的比率（Ks 值=同義核酸鹼基置換數／總同義核酸鹼基位點）。可藉由比較旁系同源基因（paralogous genes/paralogs）Ks 值，了解比較物種間相較於共同始祖之演化關係。
5. 轉錄體研究（Transcriptome investigations）：過去常以 microarray 來進行特定基因的定量熱點分析（hot spot analysis），可以利用轉錄體定序進行全面性的比對研究，針對同一生物在不同生理階段或不同組織取樣，觀察時期或是組織在基因的表現差別性。必須注意取樣時期的代表性、差異處理與試驗重複數的適宜性、樣品間比較分析前須進行常態化的步驟（normalization）等。
6. 植物病理研究(Plant pathology)：可分作病理研究、病原偵測、流行病學族群分析等。可藉由 RNA-sequencing 分析罹病過程中寄主與病原生物的交替作用，寄主植物病程各階段的基因表現與病徵相關性。病原偵測上，可藉由單一物種複合材料的 RNA 全定序、DNA 定序與病原資料庫比對，瞭解該作物病原的族群相，在多樣品的集合下，瞭解特定作物的病原相與區域的演化差異。

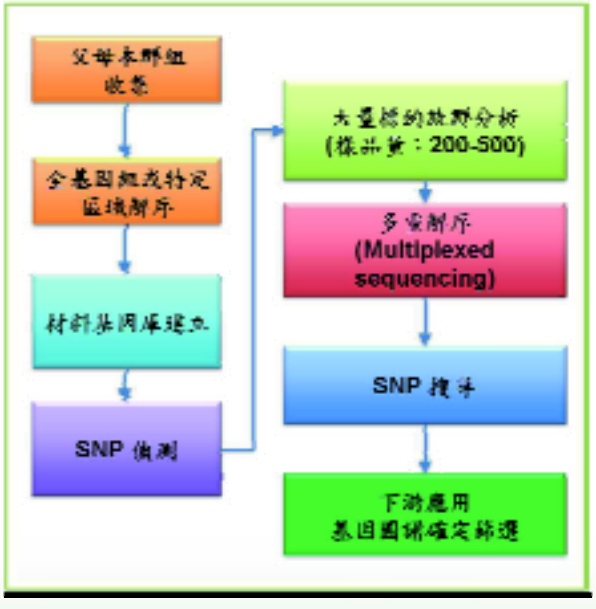

圖 2｜2014 年 Satya 等人歸納了利用 NGS 協助作物育種分析之流程

五、結語

在次世代定序技術發展前期，因爲高價的定序機台，單樣品分析成本高。且研究人員必須仰賴專業資訊人員，將定序材料之生物特性、分析之序列特性等篩選條件與資訊人員溝通，將生物描述特性轉爲篩選參數。專業資訊人員在高規格的電腦資訊設備及應用程式語言分析巨大數據庫下才能進行相關資訊運算。正因如此，導致許多生物研究人員不敢踏入次世代定序技術應用領域。近年來因NGS機型日益增加，多家廠商如Roche、Illumina、Life Technologies以及Oxford Technologies Nanopore著力開發桌上型定序機器，期能成爲試驗室常設設備，而期刊：*Bioinformatic* 甚至在 2009 年推出NGS 專刊對相關分析軟體與平台進行介紹，使得更多生物研究人員願意投入資訊研發。資訊分析可經由工作流的架設（work flow）設定指令，利用雲端操控資料存取等工作，可藉由計畫之技術合作與大專院校租借高規格的電腦資訊空間，進行資訊運算。再加上相關試驗耗材成本下降，依照 2014 年 Satya 等人的推估，在第 3 代定序技術加入定序能量之後，定序技術的容量將持續擴大、速度也將增加、定序成本下降。單一生物之全定序基因體組應可下降至 1,000 元美金，定序時間也大幅減少爲 15 分鐘。成本的降低也意謂著未來的試驗研究，可考量納入定序技術來增加相關研究的深度，在不同的生物研究領域提供跨世代的藍圖。