

出國報告（出國類別：研習）

# 新一代 DNA 序列分析策略在基因體學 研究之應用技術研習

服務機關：行政院農委會高雄區農業改良場

姓名職稱：蔡奇助副研究員

派赴國家：美國

出國期間：99 年 9 月 11~20 日

報告日期：99 年 12 月 6 日

## 摘要

新一代 DNA 讀序(Next Generation Sequencing)是近年來快速發展的一項新技術，此技術顛覆以往傳統 DNA 讀序的方法，藉由 DNA 聚合酵素的反應過程中，即完程序列的讀取，而且利用螢光原理搭配高感度的螢光偵測系統，可以同時讀取數萬筆的序列資料，因此讀序的效能遠超過傳統的 DNA 讀序。傳統 DNA 讀序一個反應僅可讀取數百個鹼基，新一代 DNA 讀序可以讀取數百萬個，使得分子生物的研究從早期單一基因或幾個基因的研究，擴展至整個基因體的研究，讓原本需要花費相當龐大的人力與物力的基因體或轉錄體的研究變的相當簡單。雖然目前新一代 DNA 讀序能讀取的 DNA 長度遠較傳統法短，不過隨著反應條件的改進，此技術能讀取的序列長度持續增加，未來應能趨近於傳統的 DNA 讀序法。由於新一代 DNA 讀序法能讀取的序列較短，且同時獲得的 DNA 訊息相當大量，其中還有部分 DNA 訊息是重複的，所以需藉由生物資訊學的各種程式加以輔助，藉此進行大量 DNA 資料的比對、拼接。若所獲取的巨量 DNA 資料不能有效的整理與分析，將會是一堆無用的鹼基資料。此外，新一代 DNA 讀序技術可以應於物種基因體的解序、轉錄體的讀序、差異性基因表現、不同基因體的 SNP 發現、轉錄因子結合區的發現、microRNA 的選殖等。

## 目錄

一、目的-----	4
二、行程表-----	5
三、研習課程-----	6
四、研習內容摘述-----	11
五、心得與建議-----	21
六、附件-----	23

## 一、目的

由於目前新一代 DNA 讀序(Next Generation Sequencing)系統在台灣尚屬起步階段，已經有些民間廠商投入本技術的服務，雖然每次反應可以獲取數百萬個鹼基，但每次反應所需的費用也需要數十萬元，因此有必要瞭解此技術的優點、缺點及限制因子，否則將會浪費研究經費去獲取一些沒有用的巨量 DNA 序列資料，無法獲取有用的 DNA 或 RNA 訊息。此外，也可以藉由本次研習瞭解目前新一代 DNA 讀序的發展方向，以及本技術的應用潛力。也希望瞭解應用此技術所獲取的巨量 DNA 資料需要哪些生物資訊學的軟體來輔助，使之能將原本雜亂無章的 DNA 資料，變成有用的 DNA 訊息，如此即能將此嶄新、有效能的現代基因科技的效能加以發揮。

## 二、行程表

### 新一代 DNA 序列分析策略在基因體學研究之應用技術研習行程表

(計畫名稱：育種、栽培設施及生物技術農業科技人才培育計畫)

(計畫編號：99 農科-4.1.1-高-K1)

11<sup>th</sup> -20<sup>th</sup> Sep 2010

日期	行程	住宿地點
9/11 (六)	屏東→21:10 高雄小港機場搭乘中華航空 CI308→22:05 於桃園國際機場轉搭中華航空 CI004→舊金山	飛機
9/12 (日)	美國時間晚上 7:50 到舊金山機場	舊金山
9/13 (一)	舊金山機場→搭 BART 捷運至 Richmond→轉搭火車至加州大學戴維斯分校(UC Davis)報到	沙加緬度
9/14 (二)	UC Davis 基因體中心研習	沙加緬度
9/15 (三)	UC Davis 基因體中心研習	沙加緬度
9/16 (四)	UC Davis 基因體中心研習	沙加緬度
9/17 (五)	UC Davis 基因體中心研習	沙加緬度
9/18 (六)	UC Davis 基因體中心研習 沙加緬度→Richmond→舊金山	舊金山
9/19 (日) -9/20(一)	舊金山機場候機，搭乘中華航空 CI003 至桃園國際機場→轉搭中華航空 CI301 至高雄小港機場→屏東	飛機、屏東

### 三、研習內容

#### 9月14日研習內容

Time	Topic	Speaker	Room
8:30am - 9:15am	Welcome & Course Overview	Dr. Dawei Lin, UC Davis Bioinformatics Core Manager	GBSF 1005
9:15am - 10:15am	Next Generation Biology Through Next Generation Sequencing	Dr. Dawei Lin	GBSF 1005
10:30am - 11:30am	Sequence Assembly 1	Dr. Dawei Lin	GBSF 1005
11:30am - 12:30pm	Sequence Assembly 2	Dr. Joe Fass, UCD Bioinformatics Core Lead Scientific Programmer	GBSF 1005
1:30pm - 3:00pm	Overview of the Roche 454 Genome Sequencer	Dr. Teri Mueller, Roche, Field Applications Consultant - Bioinformatics	SLB 2060
3:15pm - 5:30pm	Hands On Exercises: Linux Boot Camp; Neubler assembler for 454 sequence data	The UCD Bioinformatics Core Team	SLB 2060

9月15日研習內容

Time	Topic	Speaker	Room
8:30am - 9:30am	Gene Prediction	Dr. Ian Korf, UCD Genome Center and Bioinformatics Program, Assistant Professor Molecular & Cellular Biology	GBSF 1005
9:30am - 10:15am	Comparative Genomics	Dr. Aaron Darling, UC Davis Genome Center and Bioinformatics Program, Fellow	GBSF 1005
10:30am - 11:30am	SNP Discovery 1	Dr. Dawei Lin	GBSF 1005
11:30am - 12:30pm	SNP Discovery 2	Dr. Joe Fass	GBSF 1005
1:30pm - 2:45pm	Hands On Exercise: Velvet Assembler for Illumina Sequence Data SNP Discovery	The UCD Bioinformatics Core Team	SLB 2060
3:00pm - 5:30pm	Hands On Exercise: Velvet Assembler for Illumina Sequence Data SNP Discovery	The UCD Bioinformatics Core Team	SLB 2060

9月16日研習內容

Time	Topic	Speaker	Room
8:30am - 9:15am	Overview of ChIP-Seq Analysis and Demo	Dr. Dawei Lin and Nikhil Joshi, Scientific Programmer, UCD Bioinformatics Core Facility	GBSF 1005
9:15am - 10:15am	SOLiD	Dr. Asim Siddiqui, Applied Biosystems, Director of Bioinformatics	GBSF 1005
10:30am - 11:30am	Using ChIP-chip and ChIP-Seq For Genome-Wide Analyses of Transcription Factor Binding Sites	Dr. Peggy Farnham, UCD Pharmacology Professor, Assoc. Director of Genomics	GBSF 1005
11:30am - 12:30pm	High-Throughput, Short-Read Resequencing of Samples of Genomes From Natural Populations of <i>Drosophila melanogaster</i>	Dr. Charles Langley, UCD Center for Population Biology, Dept. of Evolution and Ecology, Distinguished Professor Genetics	GBSF 1005
1:30pm - 3:00pm	Hands On Exercise: UCSC Browser Tutorial ChIP-Seq	The UCD Bioinformatics Core Team	SLB 2060
3:15pm - 5:30pm	Hands On Exercise: UCSC Browser Tutorial ChIP-Seq	The UCD Bioinformatics Core Team	SLB 2060



9月17日研習內容

Time	Topic	Speaker	Room
8:30am - 9:15am	Hands On Review	Dr. Dawei Lin and Nikhil Joshi	GBSF 1005
9:15am - 10:15am	Transcriptional Profiling	Dr. Dawei Lin	GBSF 1005
10:30am - 12:30am	The Illumina Genome Analyzer II: Technology, Applications and Data Analysis	Dr. Gary Shroth, Illumina, Sr. Director, Gene Expression Applications R&D	GBSF 1005
1:30pm - 3:00pm	Hands On Exercise: Genome Annotation (Gene Finding) Blast2GO (Functional Annotation)	The UCD Bioinformatics Core Team	SLB 2060
3:15pm - 5:30pm	Hands On Exercise: Genome Annotation (Gene Finding) Blast2GO (Functional Annotation)	The UCD Bioinformatics Core Team	SLB 2060

9月18日研習內容

Time	Topic	Speaker	Room
8:30am - 9:15am	Informatics Issues	Dr. Dawei Lin	GBSF 1005
9:15am - 10:15am	Interesting Topics From Students	Dr. Dawei Lin, Dr. Joe Fass, Nikhil Joshi, Dr. Monica Britton, Analyst, UC Davis Bioinformatics Core	GBSF 1005
10:30am - 11:30am	TILLING by Next Generation Sequencing	Dr. Luca Comai, UC Davis Plant Biology, Professor	GBSF 1005
11:30am - 12:30pm	Lettuce Transcriptome Sequencing and Assemblies	Dr. Marta Matvienko, UCD Genome Center and Bioinformatics Program, Assoc. Project Scientist	GBSF 1005
1:30pm - 2:30pm	Sequencing Services at the DNA Technologies and Expression Analysis Cores	Dr. Charles Nicolet, UC Davis, Manager DNA Tech & Exp. Core	GBSF 1005
2:30pm - 3:30pm	Experimental Design and Cost Estimation	Dr. Dawei Lin	GBSF 1005
3:15pm - 4:00pm	Question & Answer Session	Dr. Dawei Lin	GBSF 1005
4:00pm - 5:30pm	Final Remarks & Certificates	Dr. Judy Kjelstrom, Dr. Dawei Lin	GBSF 1005

## 四、研習內容摘述

### (一)研習簡介

美國 UC Davis 基因體中心第三次舉辦該類型研習，研習內容包含上午的課程原理及應用之介紹，下午都進行電腦教室實習。本研習分成兩階段，第一階段是新一代 DNA 讀序技術原理及實作，第二階段為雲端運算。其中第二階段是本年度新加入的研習內容，需要有電算、數學背景的學者方能參與，筆者僅參與第一階段研習課程。參與第一階段研習人數約 55 位，主要是美國各地的教師、研究人員、博士後研究人員，另外有八位來自不同國家的研究人員。第一階段研習內容彙整成以下幾個主題，分別加以摘要敘述。

### (二)加州大學戴維斯分校基因體中心介紹

加州大學戴維斯分校基因體中心(UC Davis Genome Center)是該大學最重要的基因體學的技術支援單位。該中心提供相關、有用的基因體學技術，協助學校各領域之研究的技術支援。然而，由於基因體學技術發展快速，必須不斷地精進與發展方能跟上腳步。基因體中心以提供校內或校外無法提供之基因體技術支援的為目標，目前我們提供五個核心服務重點：分別為基因體學 (Genomics)、轉錄體學 (Transcriptomics)、蛋白質體學 (Proteomics)、代謝體學 (Metabolomics) 及生物資訊學 (Bioinformatics)等，分述如下：

#### 1. 基因體學：

自從人類基因體完成解序，鼓舞基因體學的發展，目前已有各式各樣微生物、動物及植物完成基因體的解序。傳統 DNA 讀序技術需要花費相當多的人力與物力進行物種的基因體解序，新一代 DNA 讀序技術的導入，使得物種基因體的解序突飛猛進。

## 2. 轉錄體學：

在特定組織或發展時期時，其各種基因所轉錄出的整體 RNA 分子稱為轉錄體(transcriptome)，研究轉錄體的科學稱為轉錄體學。這也是目前最多學者參與研究的一個領域，目前也有發展很多以 PCR 為基礎來進行轉錄體差異性表現分析的技術，目前轉錄體學的研究也可以利用新一代 DNA 讀序技術進行快速且大規模的研究，讓轉錄體學的研究進展更為快速，且能進行全面系統性的研究。

## 3. 蛋白質體學：

基因經過轉錄產生 RNA，RNA 經過轉譯產生蛋白質，蛋白質是真正參與生理功能的分子。是屬於巨分子，需要靠先進的蛋白質分離技術進行純化，然後依賴質譜儀(LC-MS-MS)進行蛋白質序列分析，此外，蛋白質的結構往往影響該蛋白質能否扮演其特定功能重要因子，因此蛋白質的結構模擬在蛋白質體學上也是相當重要的一環。

## 4. 代謝體學：

基因經過轉錄產生 RNA，RNA 經過轉譯產生蛋白質，蛋白質可以參與各式各樣的生化反應產生各種小分子代謝物，因此代謝物體學就應運而生。代謝體學是在研究特定的生化反應其獨特的系統性化學指紋 (chemical fingerprints)，亦即研究該反應所參與的系統性小分子代謝物(small-molecule metabolite)的輪廓。

## 5. 生物資訊學：

生物資訊學是新一代 DNA 讀序技術能否發揮其效能的最關鍵的一環，而且也常常是此技術的瓶頸。依照所需解決的生物問題與實驗目的，會使用各式各樣的計算和分析方法去解讀新一代 DNA 讀序後的資料，這些分析程式可用來進行如，序列組裝和分析，大分子結構預測，推理和分析的途徑，調控網絡、數據庫設計、開發和彙整，以及統計分析。本部門可以將新一代 DNA 讀序技術所獲取

的巨量資料，將之轉換成有用的資訊。此外，爲了推展生物資訊學在新一代 DNA 讀序技術之應用，此部門每年也會進行核心技術的培訓，以及教授生物資訊學之工具的使用方法，以及分享一些大規模計算資源。

### (三) 新一代 DNA 讀序技術介紹

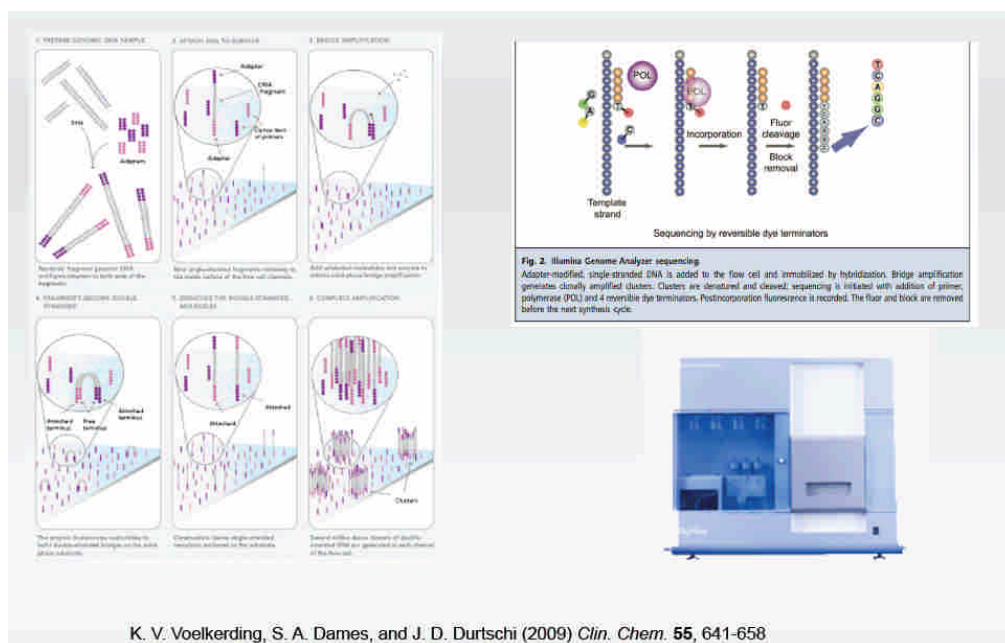
新一代 DNA 讀序技術(Next-Generation Sequencing Technology)，主要是應用高效能的螢光劑及靈敏的偵測器，在同一時間，進行數百萬個讀序反應，此類的讀序反應有別於傳統的 Sanger 氏的讀序法，因而稱之爲新一代 DNA 讀序技術。此種技術是近年來才發展出的技術，由於不同的發展基礎與原理，造就此類技術並非單一技術，而是有幾種不同的策略都可以達到相同的目的。不同的新一代基因測序技術，其原理有很大的差別。目前比較成熟且已市場化的技術的發展廠商有三家，分別爲 Illumina 公司的 Solexa 技術、Roche 公司的 454 技術及 ABI 公司的 SOLiD 技術。以下分別介紹各技術的原理。

#### 1. Solexa 測序原理

Solexa 讀序技術乃利用 DNA 聚合酶的延伸反應過程中，從 DNA 的 5' 端往 3' 端複製的延伸過程中，同時也進行讀序，由於所使用的 dNTP，dATP, dTTP, dGTP 及 dCTP 分別標定不同的螢光基團，同時又在所有的 dNTP 中加入了 3' 末端保護基團，即封閉基團，以使每一次的延伸反應只能延伸一個鹼基，螢光信號被偵測器收集後，會將每一個不延伸讀序分子的螢光基團去除，最後將封閉基團去掉，清洗後進行下一步反應，而每次反應所收集到的螢光信號，都對應了所要檢測的序列。

在實驗過程中，會在每個測序列之核酸片段的兩端分別含有特定序列的連接子(linker)，然後將上述核酸片段進行變性展佈於特製的晶片上，由於晶片表面佈滿一層分別與待核酸測片段兩個末端的特定序列對應互補的兩個單鏈引子，可

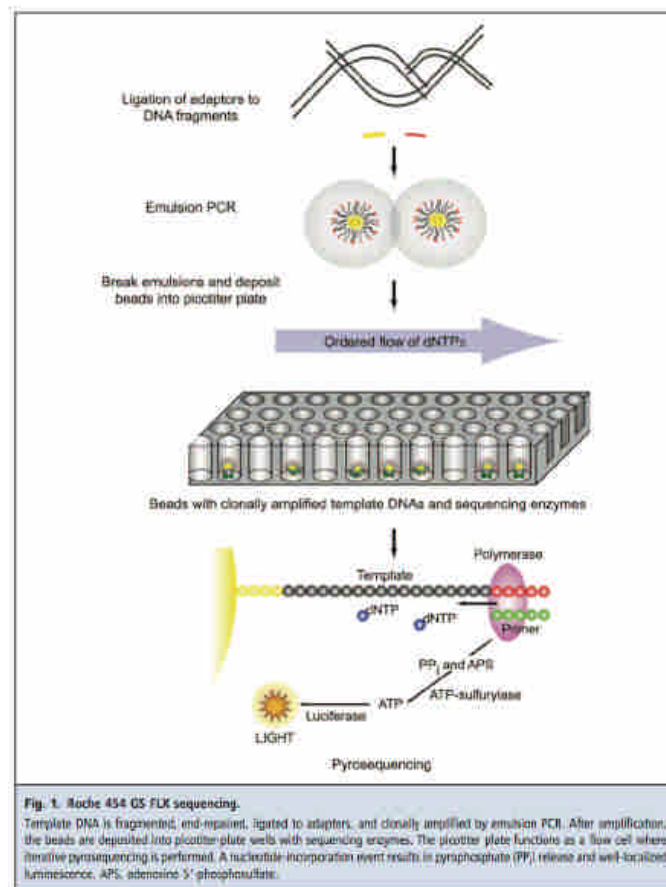
以通過互補煉合，捕獲被變性成單鏈的待測片段，以進行待測序列在晶片上的原位擴增，引子(primer)延伸使得單股 DNA 成爲雙股，再將雙股變性後成爲單股，其中一端煉合在晶片上，另外一端(5'或 3')隨機與附近的另外一個引子互補煉合，形成一個“橋”。上述反應在晶片上的數以萬計的 DNA 分子同時進行，形成的單鏈橋再以引子進行延伸，於晶片表面的所有核酸片段進行延伸，形成雙股 DNA。雙股經變性成單股，再次形成橋，成爲下一輪擴增的模板繼續擴增，經過 30 循環複製，產生數以萬計的不同“DNA 簇群”。最後所有的“DNA 簇群”在 Genome Analyzer 分析儀進行讀列反應。本技術早期能夠讀取的序列長度約 30 個核苷酸左右，不過最新的技術已經可以達到 150 核苷酸 (如下圖)。



## 2. 454 測序原理

454 讀序技術是利用 pyrosequencing 的策略，先以特異性的測序引子 (specific primer)和單股 DNA 模板結合，加入 DNA 聚合酶(DNA polymerase)、ATP 硫酸化酶(ATP sulfurylase)、螢光素酶(luciferase)和雙磷酸酶(apyrase)以及底物 APS 和 luciferin 等，數以萬計的核酸分子同時反應，若引子能夠延伸一個 dNTP(註：每次循環僅加一種，dATP, dTTP, dGTP 或 dCTP 依序反應)，則會釋放出一個分子的

焦磷酸(PPi)。在有 ATP 硫酸化酶的催化反應下，生成的 PPi 可以和 APS 結合形成 ATP；在有螢光素酶的催化下，產生之 ATP 又可以與螢光素結合形成氧化螢光素，同時產生可見光。通過 CCD 光學偵測系統即可獲得一個特異的訊號。上述反應中剩餘的 dNTP 和殘留的少量 ATP 在 Apyrase 的作用下發生降解，這樣，就可以在反應體系中，加入另一種 dNTP，使以上反應重複進行，即可讀取 DNA 序列，目前 454 讀序技術已經可以達 800 個核苷酸（如下圖）。

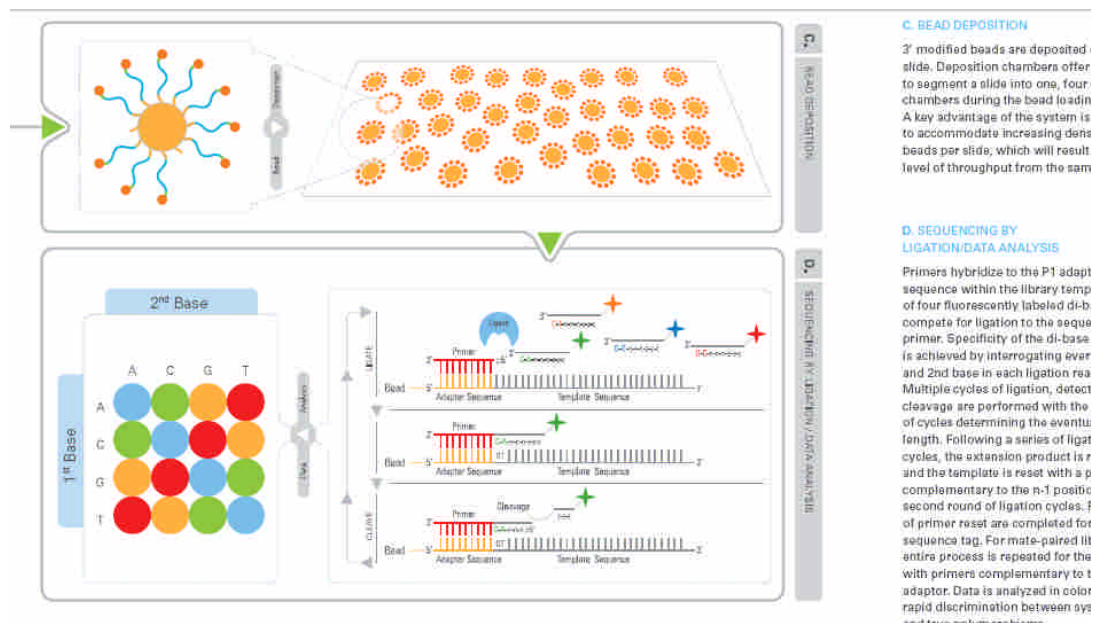


### 3. SOLiD 測序原理

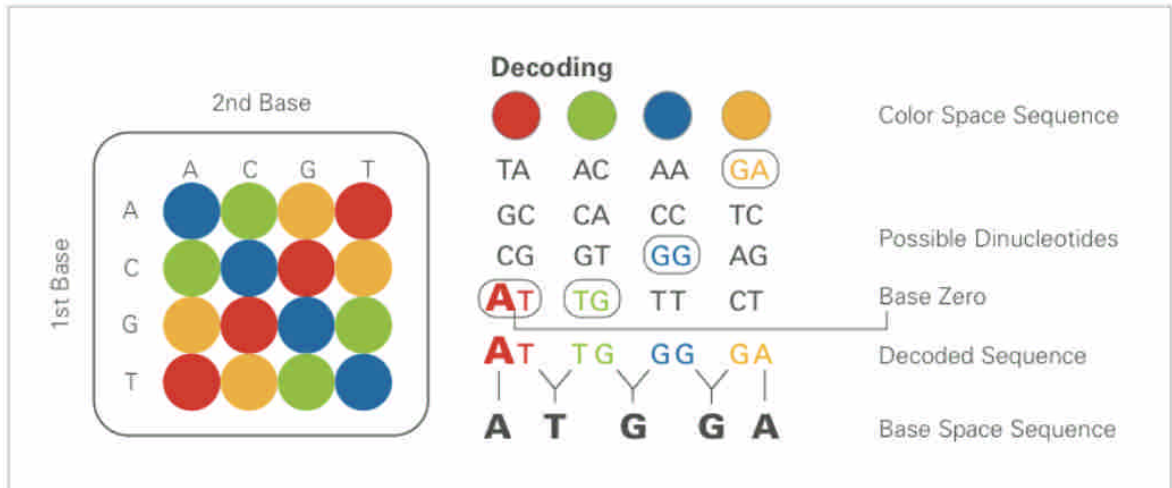
SOLiD 應用連接法進行讀序，利用了獨特的雙鹼基編碼原理進行讀序。在等待讀序之短的 DNA 片段的兩側連上 SOLiD 接頭，分別是 P1 接頭和 P2 接頭，然後對加上 SOLiD 接頭的待測 DNA 片段，在特定的磁珠表面進行增殖，通過油包水成一隔離空間進行 PCR 反應，其中和 P1 接頭對應的 P1 引子，被固定在 P1 磁珠的表面，在 PCR 反應前，將含有 PCR 反應所有成分，包括 P1 磁珠和對應於

P2 接頭的 P2 引子等之水溶液，注入高速旋轉的礦物油表面，水溶液就可以被分離成無數被礦物油包圍的小液滴，並各自成獨立的反應空間，理想狀況下，每個小液滴僅含一個 DNA 模板和一個 P1 磁珠，隨著 PCR 反應的進行，磁珠上就形成了若干具有相同來源的擴增產物，以利後續之讀序反應。

如下圖所示，測序反應是由上述製備的磁珠的基礎上進行；先利用一個讀序引子，讓該讀序引子與磁珠上的 PCR 增殖產物的 P1 接頭進行互補雜合，在其鄰近的位置上若存在另一個互補鏈，那麼讀序引子和鄰近的互補股之間可以進行連接反應。SOLiD 系統使用了特有的 8 個鹼基長的寡核苷酸鏈，其 3'端第 1、2 位構成的鹼基對是特定探針染劑類型的編碼區，5'末端標有螢光染劑，因此不同的序列組合就被標記上不同的螢光基團。寡核苷酸鏈競爭性與讀序引子鄰近的序列雜合連接，透過顏色判斷序列組成，當其標記顏色被讀取後，即將連接上的寡核苷酸在第 5 及 6 位點間切斷，以移除螢光基團標記，然後進行下一次反應，依此重複進行反應。第一次反應中，可以獲取的鹼基資訊之位點為第 1、2、6、7、11 及 12 位鹼基等。重複上述反應過程中，往左偏移一個鹼基，使用較第一次反應少一個鹼基的引子進行反覆反應，直至整個序列讀序完成。此技術目前的讀長為 30-35 個鹼基 (如下圖)。







#### 4. 結語：

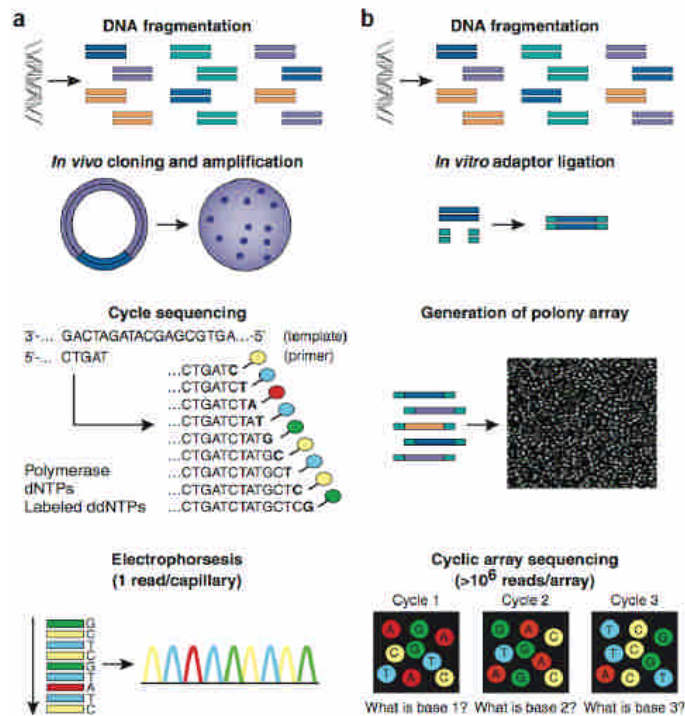
由以上可知，雖然上述測序技術在原理上有些不同，但都有一個共同之處，即都在測序反應進行的同時，就將反應信號收集起來，使得讀序成本可以降低。下表是新一代 DNA 讀序系統的比較表如下表。在台灣目前有從事 454 及 Solexa 法新一代 DNA 讀序的服務，價格分別為 70-80 萬與 20-30 萬，遠高於美國的價格，不過相信在不久的將來，隨著機器的普遍與技術的成熟，會愈來愈縮短與國外的價差。至於到底要選擇哪一種新一代 DNA 讀序技術來進行讀序，需依照不同的研究需求而異。由於 454 能夠讀取較長的 DNA 序列，因此對於沒有基因體相關訊息之物種(novel species)，若要進行基因體的解序，適合利用 454 來分析。不過每次 454 反應能涵蓋的基因體百分比較低，若要探討之物種的基因體已經有背景資料庫可比對與分析，因此可以選擇 Solexa 法 (或 SOLiD 法)，理由是，雖然 solexa 法所能讀取的序列較短，但能同時讀取的 DNA 片段較多，能夠涵蓋的基因體百分比較高。需要重複反應的次數較低就能將整個基因體涵蓋。

**Table 1. Comparison of massively parallel sequencing technologies**

	<b>FLX (Roche)</b>	<b>Solexa (Illumina)</b>	<b>SOLiD (Applied Biosystems)</b>
Website addresses	<a href="http://www.roche-applied-science.com">http://www.roche-applied-science.com</a>	<a href="http://www.illumina.com">http://www.illumina.com</a>	<a href="http://www.appliedbiosystems.com/index.cfm">http://www.appliedbiosystems.com/index.cfm</a>
Chemistry	Emulsion PCR [67] of bead-anchored oligos [68]. Clonal plate amplification. Pyrosequencing using light emission and detection [69,70].	Solid-phase-anchored oligo bridge amplification [71]. Cluster sequencing using reversible fluorescent dNTP terminators [72].	Paired-end oligo cloning. Emulsion PCR of bead-anchored oligos. Fluorescent oligo ligation and detection [73,74].
Read length	~250 bases	~35 bases	~25 bases
Machine costs	US\$500 000	US\$520 000	US\$600 000
Costs/run (reagents)	Not currently available	US\$3000	US\$3000
Availability	Immediate	Immediate	Immediate
Capacity/run	0.1 Gb	1 Gb	1 Gb
Run duration	7.5 h	67-91 h	4 days for fragment library, 8 days for paired library
Advantages	Relatively long read length	Possible to sequence through single-repeat DNA nucleotide stretches	High base-calling accuracy (bases are read twice). Homopolymer sequencing.

### (三)傳統讀序與新一代 DNA 讀序之比較

傳統 DNA 讀序法針對整個基因組之讀序流程，需先將基因組打成小片段 DNA，然後將各片段 DNA 與載體接合，然後轉型進入大腸桿菌進行載體大量複製，然後才進行選殖系的讀序，每一個選殖系都要經過一次的讀序反應，然後所有的 DNA 序列在進行拼接(流程如下圖 a)。新一代 DNA 讀序也是將基因組 DNA 打斷成小片段，然後才上轉接頭進行複製，沒有經過轉型作用，就直接進行讀序，而且是在 DNA 聚合酵素反應的過程中就同時獲取序列，而且可以同時進行數萬個選殖系的讀序。此外，新一代 DNA 讀序也沒有經過接合反應與轉型作用，直接用 PCR 產物進行讀序，因此可以快速且便宜的完成基因體的解序(流程如下圖 b)。



## Parallel Sequencing (100s vs. billions)

### (五) 新一代 DNA 讀序技術生物資訊相關程式操作實習

本研習除了介紹新一代 DNA 讀序技術的原理及應用外，有很多時間在電腦教室進行研習，藉由 step by step 的指導原則下讓學員可以實際進行巨量 DNA 資訊的操作，使原本巨量且雜亂無章的 DNA 資料變成有用的資訊，電腦實習總共有十個主題，以下分別介紹之(註：黑體字表程式的名稱)，

#### 1. Linux system: Basic Unix Commands :

本主題在介紹 Linux 系統的原理及操作所需具備的基本認知，以及相關的指令，由於 Linux 作業系統有別於一般 Windows 或 Mac 的作業系統，主要是其操作過程需要下達一些特定的指令，因此需要熟記許多相關的指令。

#### 2. Short Read Quality Control :

本主題在介紹如何判定新一代 DNA 讀序系統所讀取之 DNA 序列的品質好壞，以決定是否進行後續的相關程式的分析。由於新一代 DNA 讀序所獲取的資料相當的多，後續的分析技術層次高且費時，因此進行分析前需要對其所

獲取之資料是否適合需要事先確認。

### 3. The **Velvet** Assembler :

確認新一代 DNA 讀序所獲得之資料品質無虞後，接下來就是要先將所獲取之巨量的資料先做一事先的整併與拼接，因為有很多資料會是重複的，完成拼接後才進行後續的程式的分析。因此 Velvet 程式在新一代 DNA 讀序資料中扮演很重要的角色。尤其是 Solexa 與 SOLiD 系統所獲取的資料是屬於較短片段的 DNA 序列，因此能否有效的拼接就顯的相當重要。

### 4. Introduction to UCSC Genome Browser (<http://gb.genomecenter.ucdavis.edu>) :

本主題在介紹 UC Davis 基因體中心的 Browser，經由 Linux 系統，以及新一代 DNA 讀序相關的應用程式編輯分析後，可以進入 UC Davis 基因體中心的 Browser 進行分析結果的呈現，會獲得較容易判讀的資訊，這些資訊是可以當作最後的分析結果。

### 5. Paired and single End Assembly Comparision Using Illumina Reads :

本主題是針對 Illumina 公司的 Solexa 讀序系統所讀取的資料進行不同方式拼接之比較，由於 Solexa 讀序系統所讀出的 DNA 序列短於 Roche 公司的 454 讀序系統，因此拼接技術就顯的相當重要，Solexa 讀序系統的資料可以藉由 paired ends 的方式，可以獲得較佳的分析結果。

### 6. A SNP Discovery using **Pipeline** :

本主題是介紹 Pipeline 程式，此程式可以將新一代 DNA 讀序資料進行不同樣本，或實驗組與對照組之基因體進行比對，以獲取基因體間之單核苷酸多型性(single nucleotide polymorphism, SNP)。

#### 7. **Sissrs** and **MEME** :

本主題介紹包含 Sissrs 及 MEME 程式。此兩程式乃在搜尋新一代 DNA 讀序技術所獲得的巨量資料中具備的轉錄因子(transcription factor)結合位置(binding site)，亦即針對整個基因體進行轉錄因子的結合位置分析。

#### 8. Finding miRNAs Using **miRDeep** :

本主題是介紹 miRDeep 程式，此程式可以將以 cDNA 為樣本進行新一代 DNA 讀序之資料進行分析，主要是用來發掘 miRNAs 的程式。由於 miRNA 是屬於短片段的 RNA，因此利用傳統選殖 miRNA 技術甚為困難，因此若能利用新一代 DNA 讀序技術進行大規模分析，可以獲取大量的 miRNAs，在未來 miRNA 的應用上相當有用。

#### 9. Using **BWA**, **SAMTools**, and **R** to analyze Differential Expression of Genes in RNA-seq Data :

本主題是介紹 **BWA**, **SAMTools** 及 **R** 程式，這些程式可以探討轉錄體差異性表現之分析。可以取代傳統利用 DNA microarray、SAGE (Serial Analysis of Gene Expression)、cDNA-AFLP、SSH (Supression Subtractive Hybridization)、DD (Differential Display)、(cDNA-Amplified Fragment Length Polymorphic)進行差異性表現分析，而且所獲得的資料也較為完整。

#### 10. Transcriptome Assembly with **Velvet/Oases**

本主題是介紹 Velvet/Oases 程式，此程式進行搭配，可以進行轉錄體資料的拼接，以快速獲取某種生物特定組織，或特定條件處理下的 EST library。

## 五、心得與建議

1. 新一代 DNA 讀序技術系統是一門整合性的技術，不僅需要對生物有基本概念的人才，還需要對分子生物相關技術熟悉，最重要的是還需要生物資訊相關及電子計算相關人才，以及軟體設計等。尤其資訊相關人才在新一代 DNA 讀序中扮演相當吃重的角色，若不能將巨量的資料轉變成有用的資訊，那麼所有的努力將會白費功夫。
2. 台灣是資訊業發展不錯的國家，不過目前在生物資訊方面的人才還相當少，主要是目前台灣之資訊業所創造的產值遠高於生物科技，因此無法吸引優秀的資訊人才投入生物科技業，這是未來台灣在應用新一代 DNA 讀序技術時所需要面對的問題，亦即經由新一代 DNA 讀序所獲取的大量資料是否能轉變成有用的資訊的關鍵。
3. 目前新一代 DNA 讀序技術之發展已趨於穩定，已可應用於生物相關研究領域，不僅可在短時間內獲取大量的 DNA 訊息，而且也可以讓研究變的較為單純且簡單，對未來生物科技的發展影響至鉅，對於基因科技之研究有革命性的影響。
4. 爲了吸引更多年輕的資訊人才投入生物科技業，應該提供足夠的誘因，亦即在待遇上應該要比照資訊業，讓人才願意留在生物科技業發展。因此建議在中研院或重要的大專院校積極培育生物資訊相關人才，來面對新一代 DNA 讀序技術時代來臨各國的競爭，以協助台灣研究人員進行 DNA 資料整理與分析。因此基因科技的競爭是一項團體戰，擁有最完整資源的國家或團隊方能掌握先機，積極發展。



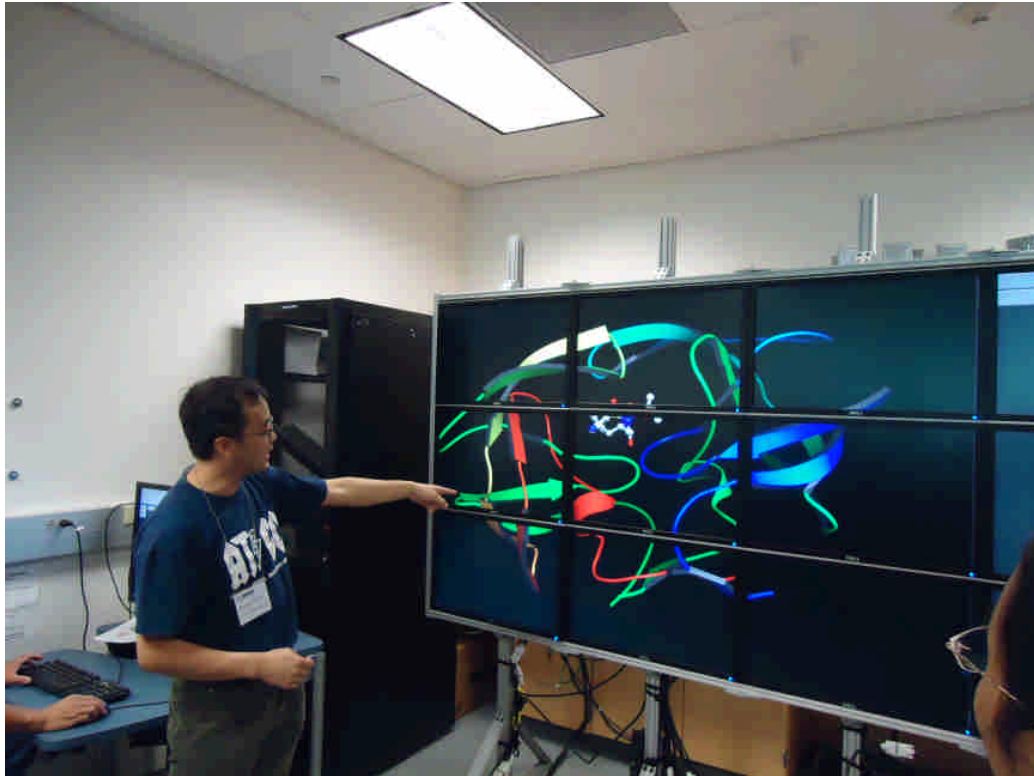
## 六、附件



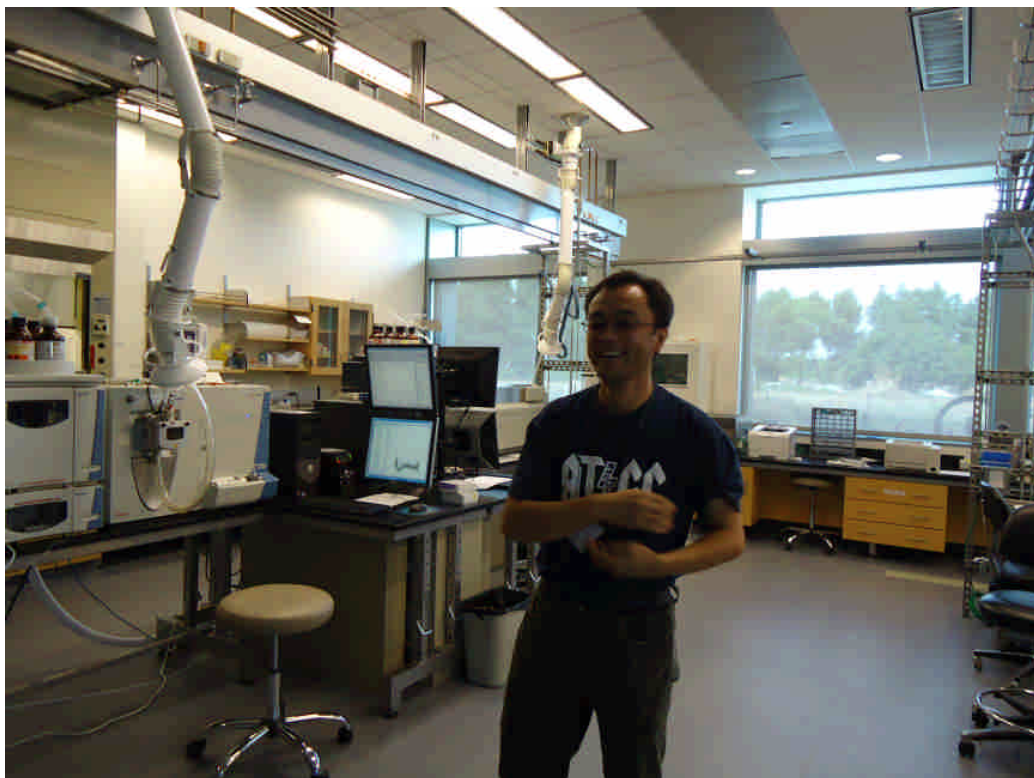
美國 UC Davis 基因體中心



UC Davis 基因體中心，實驗部門新一代 DNA 讀序 Solexa 反應時資料蒐集情形



UC Davis 基因體中心，蛋白質體部門蛋白質結構模擬情形



UC Davis 基因體中心，代謝體部門實驗室





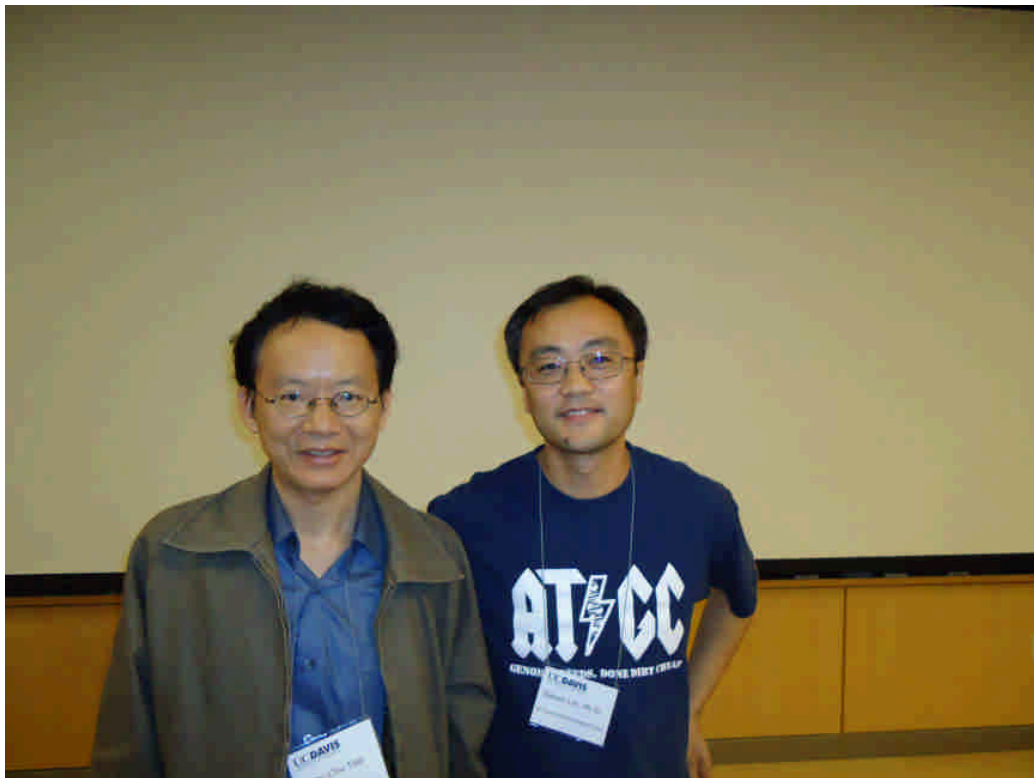
研習簡報室



程式實習教室



UC Davis 基因體中心生物資訊學部門成員



筆者與生物資訊部門主管合影