# 微陣列資料分析(Microarray Data Analysis)

蔡政安副教授

## 前言

在人類基因體定序計劃的重要里程碑陸續完成之後，生命科學邁入了一個前所未有的新時代，在人類染色體總長度約三十億個鹼基對中，約含有四萬個基因，這是生物學家首次以這麼宏觀的視野來檢視生命現象，而醫藥上的研究方針亦從此改觀，科學研究從此正式進入後基因體時代。微陣列實驗 (Microarray) 及其它高產能檢測 (high-throughput screen) 技術的興起，無疑將成為本世紀的主流；微陣列實驗主要的優勢再於能同時大量地、全面性地偵測上萬個基因表現量，透過基因晶片，可在短時間內找出可能受疾病影響基因，作為早期診斷的生物指標(biomarker)。然而，由於這一類技術的高度自動化、規模化及微型化的特性，使得他們所生成的資料量非常龐大且資料型態比一般實驗數據更加複雜，因此，傳統統計分析方法已經不敷使用。在此同時，統計學家並未在此重要時刻缺席，提出非常多新的統計理論和方法來分析微陣列實驗資料，也廣受生物學家所使用。由於微陣列資料分析所牽涉的統計問題層面相當廣且深入，本文僅針對整個實驗中所衍生的統計問題加以介紹，並介紹其中一些新的圖形工具用以呈現分析結果。

## 基因晶片的原理

微陣列晶片即一般所謂的基因晶片，也是基因體計畫完成後衍生出來的產品，花費成本雖高，但效用無限，是目前所有生物晶片中應用最廣的，由於近年來不斷改進，也是最有成效的生物技術。一般而言，基因晶片是利用微處理技術，先把人類所有的基因分別固著在一小範圍的玻璃片(glass slide)、薄膜(membrane)或者矽晶片上；然後，可以平行地、大量地、全面性地偵測基因體中 mRNA 的量，也就是偵測基因的調控及相互作用表現。目前微陣列晶片大致分為以下兩種平台(如圖一)：cDNA 晶片及高密度寡核甘酸晶片(high-density oligonucleotide)，兩種系統無論在晶片的製程及樣本處理上皆有相當的差異，因此在分析上也略有不同，以下便就晶片的特性約略介紹。

1. cDNA 晶片: 基本上晶片上的探針(probes)及準備進行雜合反應(hybridization)的樣本(Targets)皆來自於 cDNA。正常及癌組織中萃取的 mRNA 經反轉錄後，分別標上綠色(Cy3)和紅色(Cy5)螢光標記，並同時和晶片進行雜合反應，反

應後經過雷射掃描器顯像，綠色螢光點表示正常組織的基因表現高於癌組織；紅色螢光點表示癌組織的基因表現高於正常組織；當基因表現不變時，即呈黃色螢光。經影像分析軟體可將影像強度轉換成數據資料，用以分析有顯著差異表現之基因。

2. 高密度寡核甘酸晶片: 高密度寡核甘酸晶片主要由 25 個鹼基所構成的探針對(probe pair)所組成，而每一個基因由 16-20 個探針對來代表，每組探針對包括 perfect-match (PM) 和 miss-match (MM) 探針，MM 探針除了中間鹼基不同於 PM 探針外，兩者有相同的 DNA 序列，主要爲內部對照之用。不同於 cDNA 晶片，正常及癌組織中萃取的 mRNA 分別和不同的晶片進行雜合反應，所以只使用單色螢光標記。經影像分析軟體可將螢光強度轉換成數據資料，再利用不同的統計模型將每個基因所對應的探針對整合來顯示基因的表現程度。

**<u>微陣列資料統計分析</u>**

雖然微陣列實驗能快速有效地偵測表現差異的基因，也已廣泛應用在生物研究上，然而由於實驗的複雜性和特異性也使得分析上的困難度增加；近年來，由於各學術領域研究學者的加入探索並針對實驗中各步驟提出各式改進分析的方法，使得整個微陣列實驗的精確性及可靠度增加至一定的水準，從早期僅用表現差異(fold-change)的大小來篩選有差異表現基因到現在許多複雜計算的統計或數學模型。本文將微陣列資料分析分成五大部份(如圖二)，並介紹其中所牽涉相關的統計問題，這五大分析要素關係整體分析的品質及準確性，分別為:

(一)實驗設計: 透過詳細完整的實驗設計可以使得資料的品質和效度達到最佳化。實驗設計包括樣本數估計，其中樣本數可分為生物性(biological replicates)及技術性樣本(technical replicates)；在晶片上品質管制的設計；根據不同微陣列平台及研究因子設計最佳實驗配置等。

(二)資料的前置處理: 由於微陣列實驗的雜訊、系統及非系統上變異等干擾因子，因此在進行統計推論之前，需要對資料先行處理。前置處理包括影像分析及正規化用以移除系統性變異；資料轉換及篩選；缺失值插補等。資料的前置處理相當繁複，且不同微陣列平台各有不同處理程序，但是此步驟卻非常關鍵，關係著往後分析的精

確性，不可輕忽。在雙色 cDNA 微陣列中常用的正規化方法如 LOWESS 平滑曲線調整(如圖三(b) )。

(三)顯著性分析: 以統計方法檢定有顯著差異的基因，這也是微陣列實驗主要目的之一。近年來有非常多學者提出不同統計方法來偵測有顯著差異的基因，但由於在微陣列實驗中需要同時檢定上萬個基因，其中有一個非常重要的統計議題，是關於多重檢定(multiple testing)的問題，有別於傳統控制 family-wise error rate(FWER) 的方法太過保守以至於檢定力過低，另外控制 false discovery rate(FDR) 的方法可提供有效解決方案。常用的統計方法有 SAM(如圖三(c))及混合模型(Mixture model) 等可控制挑選基因中犯錯的比率(FDR)至研究者設定的標準，此外可同時利用兩種以上檢定法則來挑選有顯著差異的基因，如圖三(d)所示之 Volcano plot 利用表現平均差異質(fold-change)和統計檢定的 P 值(p-values)來挑選有顯著差異的基因。

(四)群集分析和預測分析: 群集分析(Clustering analysis)可由兩個方向來討論，基因和受測組織(如圖三(a))，基因的群集分析主要想找出具有相似表現型態的基因群集，並配合生物上代謝及傳導功能來輔助解釋；而受測組織的群集分析可用來評估受測樣本的變異程度(variation)及實驗的再現性(reproducibility)，同時也可藉由群集分析中發現疾病的次型態。預測分析(Prediction)或分類法則(Classification)主要目的想利用基因表現資料建構分類法則(如圖三(e))，用以預測疾病的發生，其中包括如何從眾多基因中挑選重要的預測因子(feature selection)，以及預測模型的建構等，此分析的目標是希望從微陣列實驗中找出可能受疾病影響基因，作為早期診斷的生物指標(biomarker)，並成功建立診斷模型。

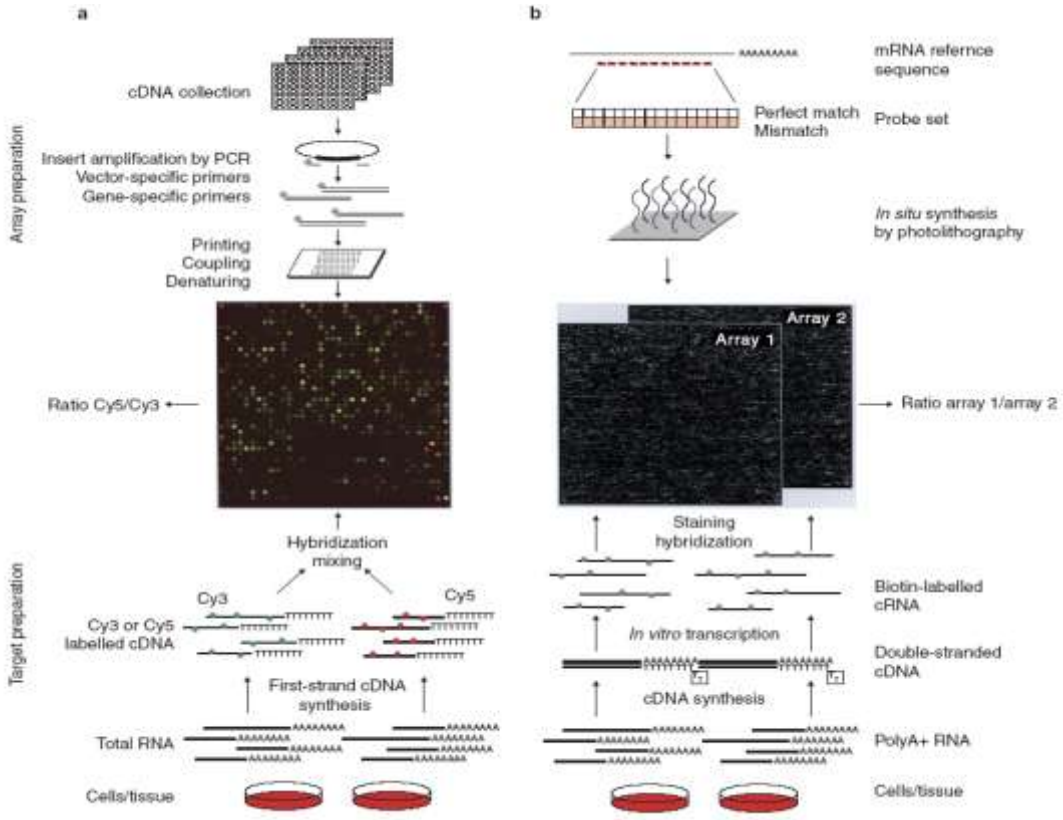
(五)相關分析及實驗確認: 經過以上分析，我們可找出具有表現差異或疾病診斷的基因，但是還是要和生物現象做緊密結合，可以經由對照大型公用生物資料庫，如 GO、KEGG 和 BioCarta Pathways 等，來描述及觀察基因在生物功能註解及動態圖解模型互動關係。此外，使用較精確的實驗(如 RT-PCR)來作進一步分析確認也是不可獲缺的步驟。

## 結論

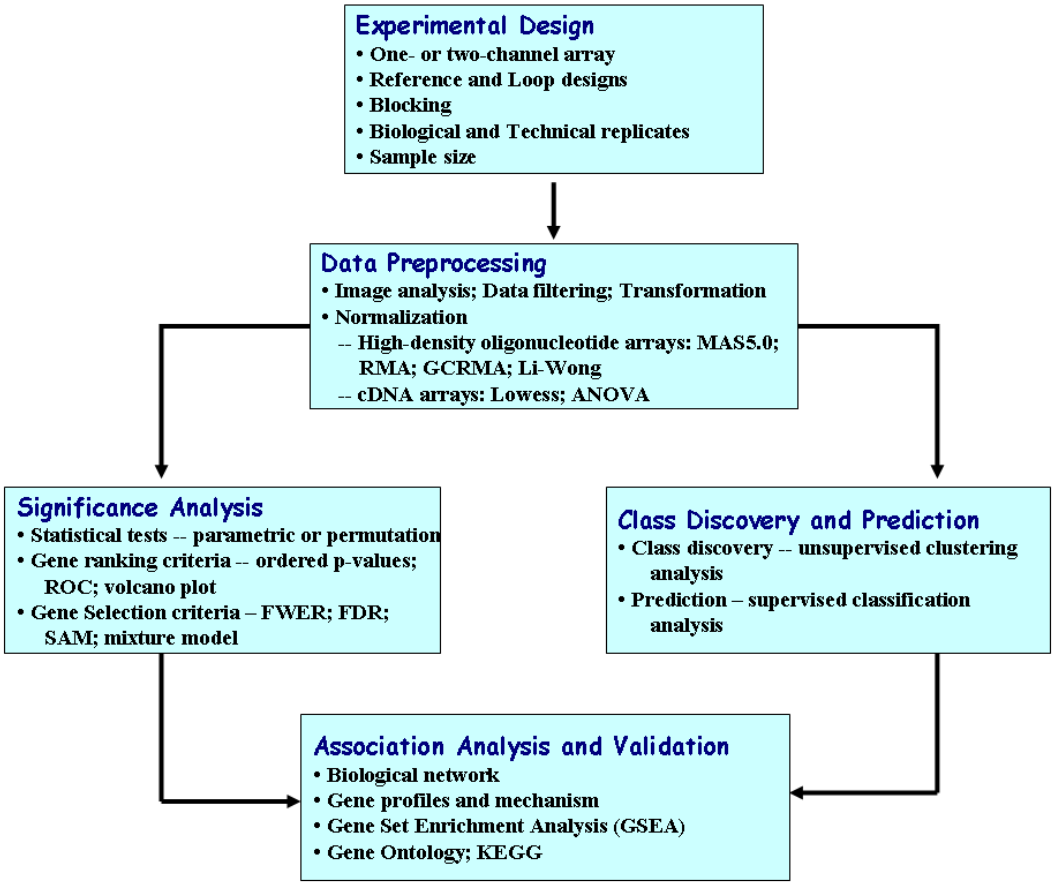
DNA 雙股螺旋結構模型發表至今 50 年，在全世界科學家不斷地探索下已了解七千多個基因的功能。在四萬個基因中，目前尚有三萬多個基因的功能，或可能有的致病因子及生物醫學用途，我們仍一無所知。透過基因體定序計畫及基因晶片的應用，可快速探測這些基因在各類疾病或生物體變動中的功能，加速我們對各生物體所有基因的了解。

**參考文獻**

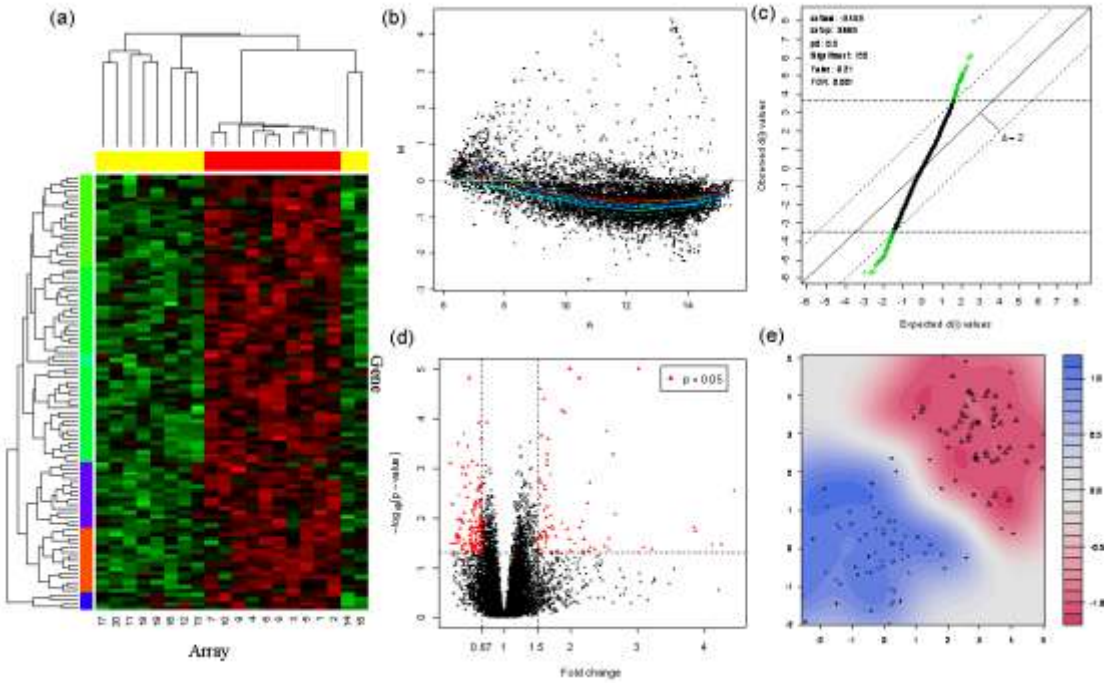
David B. Allison, Xiangqin Cui, Grier P. Page, Mahyar Sabripour, (2006). Microarray data analysis: from disarray to consolidation and consensus. NATURE REVIEWS GENETICS, 7(1), 55-65.




圖一: Principles of two major microarray platforms: cDNA array and high-density oligonucleotide array.




圖二: Guidelines for the statistical analysis of microarray experiments.




圖三: Visualization tools for microarray analysis