應用次世代定序技術分析微生物相

郭裔培、劉清碩
水產試驗所淡水繁養殖研究中心

前言

自然環境中的微生物相具有高複雜性，傳統方法透過培養基分離出不同的菌落後，再進行計數與菌種鑑定，然而傳統培養方法除有菌種增殖速率不同的問題外，對於無法培養的菌種也有分析上的盲點。高通量的次世代定序 (next generation sequencing, NGS) 不須透過培養分離，即能於短時間完成總基因體分析 (metagenomic sequencing)，幫助研究人員從宏觀角度探討微生物相的變化。

定序方法簡介

一、一代定序

1977 年由英國學者 Frederick Sanger 提出的雙脫氧鏈終止法，利用聚合酶連鎖反應會以目標 DNA 為模板、去氧核醣核苷酸 (dNTP) 為原料進行複製合成的原理，於定序時加入少量標記的雙去氧核醣核苷酸 (ddNTP)。ddNTP 和 dNTP 相比，缺少一個 3’ 羥基，複製時若序列接上 ddNTP 而非 dNTP，複製反應即會終止，無法繼續合成。每個聚合酶連鎖反應都會形成隨機停留在不同 DNA 位置的片段，且末端均為具有標記的 ddNTP，透過電泳分離由小至大的片段，讀取 ddNTP 上的標記訊號，即可推得目標 DNA 序列 (圖 1)。自動化的定序儀透過毛細管電泳，可同時進行 8－96 個樣本的定序。

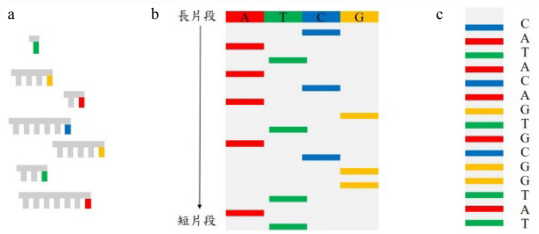

圖 1　Sanger 定序法 (Gauthier, 2007)
a：ddNTP 隨機停留在不同 DNA 位置的片段；b：電泳分離由小至大的片段；c：讀取 ddNTP 上的標記訊號

為了加速長序列的定序速度，霰彈槍定序法 (shotgun sequencing) 將長 DNA 片段逢機切割為大量互相有序列重疊的短片段，經由質體轉殖進大腸桿菌複製為序列庫，以便同時進行多個 Sanger 定序反應，所得的大量短片段序列，以演算法組裝重疊序列，接合出原本的 DNA 序列。

二、次世代定序

第一代定序的低通量限制，不適合用於分析環境中複雜的微生物相組成，次世代定序根據霰彈槍定序法的原理，先將長 DNA 片段逢機切割為大量的短片段，透過高通量定序，達到同時進行數百萬甚至數十億的定序反應，最後接合出原本的 DNA 序列。不同定序平台的差異主要在於定序時的檢測方法，大致可分為下列 4 種定序方法 (表 1)：

(一) 焦磷酸定序 (pyrosequencing)

核苷酸結合到序列上時，釋出的焦磷酸鹽 (pyrophosphate) 經由三磷酸腺苷硫酸酶

(ATP sulfurylase) 轉化為三磷酸腺苷，三磷酸腺苷會提供冷光酶 (luciferase) 能量，將冷光素 (luciferin) 氧化產生冷光訊號。由於冷光訊號只能分辨核苷酸是否成功接上，無法分辨攜帶的鹼基種類，因此定序時依序加入不同的核苷酸，每次只加入一種，若未成功接上，則分解後加入下一種核苷酸，反覆直至偵測到冷光訊號，即能確定鹼基的種類。

(二) 合成反應定序 (sequencing by synthesis)

四種核苷酸帶有可被切除的終止基團，並分別以不同的可逆螢光分子標記。定序時同時加入 4 種核苷酸，當序列與其中一種核苷酸結合後，終止基團會避免下一個核苷酸接合，清除未接合的核苷酸，透過螢光波長檢測接上序列的鹼基種類，而後終止基團和螢光分子會被切開，進行一下個鹼基的定序。

(三) 序列連接定序 (sequencing by ligation)

與其它定序方法不同，序列連接定序不使用聚合酶，改以 16 個八聚體寡核苷酸探針。每個八聚體從 3' 端開始，由 2 個鹼基 (ATCG × ATCG，共有 16 種，AA、AT、AC、AG、TA、TT…) 接上 6 個變性鹼基 (圖 2a)，探針以排列組合方式，在 5' 端接上 4 種螢光分子中的其中一種 (圖 2b)。反應開始時先以引子接合轉接子，探針上和序列互補的 2 個鹼基會引導 DNA 連接酶雜交，清除未接合的探針並記錄螢光訊號，而後探針上的螢光分子和 5' 端最後 3 個變性鹼基會被切除，並開始下一次的雜交，每輪會重複大約 7 次的雜交循環。第一輪雜交結束後，DNA 變性為單股，比先前引子位移一個鹼基的新引子進行第二輪的雜交接合，共計完成 5 輪的雜交接合後 (圖 2c)，樣品序列上的每個鹼基位置皆有 2 個螢光訊號，分別為探針的第一和第二鹼基，每個探針的第一個鹼基種類確定後，透過螢光訊號對應探針的排列組合，可定序出探針第二鹼基位置的種類，由於引子接合區的序列已知，透過各鹼基位置的螢光訊號，即可向下依序推得目標序列 (圖 2d)。

(四) 離子半導體定序 (ion semiconductor sequencing)

核苷酸結合到序列上時，會釋放氫離子，造成 pH 下降，透過半導體的離子選擇性場效電晶體 (ion sensitive field effect transistor, ISFET) 轉為電子訊號，與焦磷酸定序相似，離子半導體定序法無法直接分辨接上序列的鹼基種類，定序時依序加入一種核苷酸，若未結合則替換下一種核苷酸，直

表 1　不同次世代定序方法比較

	單一定序長度 (bp)	優　　點	缺　　點
焦磷酸定序	600-1000 (Roche 454)	定序片段長	1. 試劑成本高 2. 6 個以上的重複序列容易有插入或缺失的錯誤
合成反應定序	36-250 (Illumina HiSeq)	解決焦磷酸定序在重複序列錯誤率高的問題	隨定序反應進行，背景雜訊增加，造成序列後端的錯誤率提高
序列連接定序	50-75 (ABI SOLiD)	雙鹼基定序法的準確度高	單一定序長度短，定序長片段序列耗費的時間長
離子半導體定序	200-400 (Thermo Fisher Ion Torrent)	與焦磷酸定序步驟相似，但不需要光學判讀設備，成本較低且設備體積小	單一定序長度較焦磷酸定序低

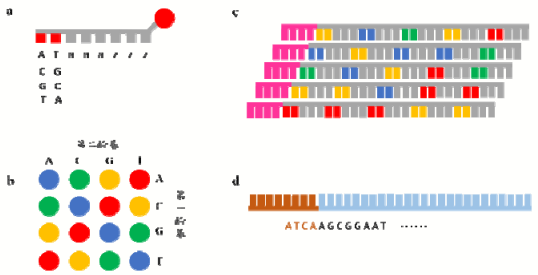

圖 2　序列連接定序原理
a：探針八聚體；b：探針螢光標記對照表；c：雜交接合，粉紅色代表引子；d：分析螢光訊號，咖啡色為已知序列的引子接合區

至偵測到電子訊號。

三、第三代定序

次世代定序大幅提高了定序通量，但相較於一代定序，單一定序長度較短，長序列必須經由切割、定序和組裝，對於沒有參考基因體 (reference sequence) 輔助組裝的從頭定序 (*de novo* sequencing)，次世代定序的應用性受到限制。第三代定序採單分子定序，樣品不須先進行聚合酶連鎖反應放大即可直接定序，目前提供第三代定序平台的公司如 Pacific Biosciences 和 Oxford Nanopore Technologies。

Pacific Biosciences 定序晶片上有許多奈米級的零模波導孔 (zero-mode waveguides)，每個孔直徑約 70 nm、深 100 nm，由於孔洞比光波長小，從底部發射的螢光激發光，在經過孔洞通道內時能量會快速衰減，避免背景訊號干擾 (圖 3)。每個零模波導孔的底部固定有一個 DNA 聚合酶，透過 4 種不同螢光標記的 dNTP，當與樣品序列互補的 dNTP 被聚合酶抓住時，零模波導孔底部發出激發光偵測螢光訊號，且 dNTP 的螢光基團設計與磷酸鏈相接而非鹼基，因此當核苷酸添加到 DNA 序列上時，螢光基團會隨磷酸基團一起脫落，達到與 DNA 聚合酶同步的定序，定序速率約 1－3 bp/s。Pacific Biosciences 定序平台容易發生隨機的插入錯誤，定序準確率僅約 87.5%，相比次世代定序 99% 左右的準確率，有相當程度的差距。為提高定序準確率，以特殊環形轉接子連接樣品序列的正反股進行循環共識定序 (circular consensus sequencing) (圖 4)，重複定序同一序列以校正錯誤，大約重複 6 次後，準確率可達 99% 以上。

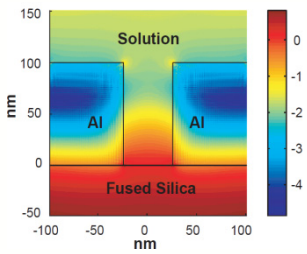

圖 3　零模波導孔內能量分布 (Levene et al., 2003)

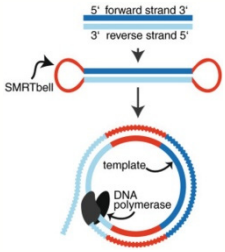

圖 4　循環共識定序 (Fichot and Norman, 2013)

Oxford Nanopore 奈米孔定序平台在薄膜上插有跨膜蛋白，以解旋酶將 DNA 解開為單股，當薄膜給予電位差時，單股 DNA 會穿透跨膜蛋白，造成電流改變，因為不同鹼基的電流訊號不同，能快速的分析序列，定序速率約 250－450 bp/s。由於結構的相似性，腺嘌呤 (A) 容易和鳥嘌呤 (G) 誤判，胞嘧啶 (C) 容易和胸腺嘧啶 (T) 誤判，且複雜度低的序列 (如連續相同的鹼基)，錯誤率也較高，整體來說定序準確率約 85%。為了提高準確率，Oxford Nanopore 開發 $1D^2$ 轉接子技術，當樣品雙股序列中的第一條被定序後，可以引導互補序列接著被定序，達到相互校正的目的，定序準確率可提高至 90－95%。此外，針對錯誤率較高的簡單序列，

亦可搭配次世代定序進行輔助校正，達到理想的準確性。

以次世代定序分析微生物相之流程

以目前市佔率最高的 Illumina HiSeq 次世代定序平台，進行 16S 擴增子定序 (16S amplicon sequencing) 為例：

一、樣品前處理

16S 核醣體基因 (16S rDNA) 由 10 個保留區 (conserved regions) 和 9 個高度變異區 (hypervariable regions) 構成，高度變異區在不同菌種間有明顯差異，常被用來作為微生物之分子鑑定標的。定序前會針對樣本來源 (土壤、皮膚、糞便或水體等)、單一定序長度和比對用的資料庫數據，決定適合的定序區域。

採集樣品後進行 DNA 的萃取和純化，萃取的宏基因組 (metagenome) DNA 濃度以 Qubit 螢光定量，再經由聚合酶連鎖反應擴增 16S rDNA 的目標變異區，確認核酸電泳後有單一條帶。

確認 DNA 符合定序要求的品質後，將 PCR 產物 3’ 接上一個 A 鹼基 (若欲分析的片段經切割處理，則須先進行修補，避免末端不平整)，3’ 和 5’ 端透過連接酶 (ligase) 分別接上 P7 和 P5 轉接子，完成樣品的建庫 (library) (圖 5a)。

二、橋式聚合酶連鎖反應

單一序列的螢光訊號強度不足，定序前會先在定序晶片 (flow cell) 上進行橋式聚合酶連鎖反應，每條序列經多次的複製，放大形成一個叢聚 (cluster)，再以叢聚為單位進行定序的螢光訊號檢測。

橋式聚合酶連鎖反應的原理是定序晶片上，具有多個和樣品兩端轉接子互補的寡核苷酸，樣品序列庫先變性為單股 DNA，其中一端的轉接子和晶片上的互補寡核苷酸雜交，第一次複製後，去除作為模板的樣品序列，僅留下固定在晶片上的新合成序列，完成序列的固定。第二次的聚合酶連鎖反應開始時，第一次新合成的序列彎曲成拱橋狀，讓另一端的轉接子與晶片上對應的寡核苷酸雜交，進行第二次的複製，變性後分離為兩條固定在晶片上的直線單股 DNA (圖 5b)。

橋式聚合酶連鎖反應約需要進行 35 次循環，才能形成一個可用於定序的叢聚，此時晶片上每一個叢聚含有正股 (forward) 與反股 (reverse) 兩種序列，為避免定序時的訊號互相干擾，變性為直線單股 DNA 後，會去除與 P5 寡核苷酸相連的反股序列，並透過 ddNTP 阻斷晶片上以及正股序列上的 P5 轉接子，避免錯誤的位點被定序，即完成可用於定序的叢聚。

三、定序與序列組裝

以 16S rDNA 高度變異區 V3－V4 定序為例，目標片段約 460 bp，受限於單一定序長度約僅有 300 bp，採雙端定序 (paired-end sequencing)，首先從正股序列 5’ 端開始第一次合成反應定序 (read 1)，第一次定序結束後，去除定序產物與 P5 轉接子的 ddNTP，再進行一次橋式聚合酶連鎖反應合成反股序列，去除作為模板的正股序列，僅留下新合成的反股序列，再次從反股序列 5’ 端定序 (read 2) (圖 5c)。

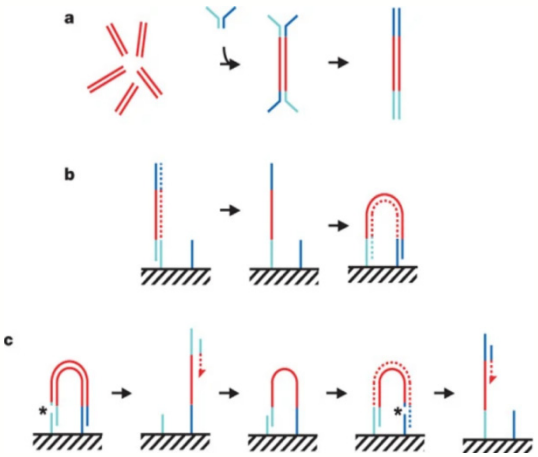

圖 5　Illumina 定序原理 (Bentley et al., 2008)
a：DNA 序列切割後接上轉接子；b：橋式聚合酶連鎖反應；c：雙端定序，虛線為新合成的序列，星號為切割位點

獲得的原始數據 (raw reads)，透過序列重疊部分拼接為原始序列 (raw tags)，將低品質 (錯誤率高) 的鹼基位點截斷，並去除過短的序列得到乾淨序列 (clean tags)，最後以演算法和資料庫比對，過濾嵌合體序列 (片段錯誤接合的序列)，得到可用於數據分析的有效序列 (effective tags)。

四、物種注釋

若要對所有的有效序列比對資料庫，會耗費大量的時間，因此先將相似度高 (一般設定為 97%) 的序列分類在同一個操作單元 (operational taxonomic unit)，將每個操作單元中，出現頻率最高的作為代表序列，並與 16S rDNA 資料庫比對，取得分類資訊。

五、多樣本混和定序

隨定序技術進步，定序晶片能提供的讀數，已超過單一樣品微生物相分析所需，當有多個不同來源的樣品需要分析時，可透過帶有不同條碼 (index) 的轉接子，分別針對不同樣品進行建庫，混和後在同一晶片上進行定序，透過讀取條碼上的特殊序列，即可分辨樣品來源，縮短定序時間與晶片成本。

數據分析

一、操作單元分析

(一) 物種階層分析

為了解定序片段的長度是否滿足研究所需，可依據定序結果比對到的分類階層 (界、門、綱、目、科、屬、種)，繪製成物種階層序列豐度累積柱狀圖，若分類階層過高，則可評估使用不同的變異區片段，或是增加片段長度。以目前文獻最常使用的 16S rDNA 高度變異區 V3－V4 為例，一般約有 80% 可達屬分類層級。

(二) 相對豐度

依據不同分類階層，選取豐度最高的前 10 種物種，繪製相對豐度柱狀圖，以便快速尋找不同分類階層中，各菌種的豐度比例 (圖 6)。另外將兩兩組別的物種平均豐度相減，繪製相對豐度均互差長條圖，可比較組間的菌種增減。

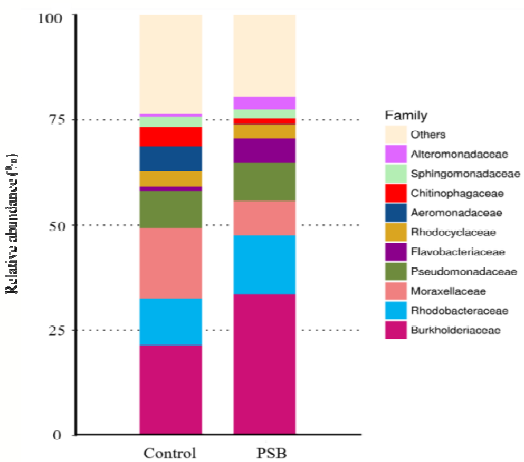

圖 6　以科分類階層繪製的菌種相對豐度

二、多樣性指數

(一) α 多樣性指數

評估同一群落（樣品）內的物種歧異度，包含群落豐富度 (community richness) 和群落多樣性 (community diversity)，豐富度能反映物種數目，常用如 Chao 1 和 Ace 指數，估算群落母體的操作單元數，兩者指數越高，代表群落豐富度越高；群落多樣性則評估物種分配上的均勻性，常用如 Shannon 和 Simpson 指數。Shannon 指數越高、Simpson 指數越低，代表群落多樣性越高。

稀釋曲線依據 α 多樣性指數，反映從該樣品中隨機抽取不同的定序量，含有多少對應的物種，可用於評估定序的數據量是否足夠，當曲線漸趨向平緩時，代表更多的數據量只會鑑定出少量的新物種，已具備足夠的代表性 (圖 7)。

(二) β 多樣性指數

反映群落和群落間的物種組成差異，常用的分析方法如主成分分析 (principal component analysis, PCA)、主座標分析 (principal co-ordinates analysis, PCoA) 和非度量多維度分析 (non-metric multidimensional scaling, NMDS) 等，兩點之間的距離越長，物種組成差異越大。

上述幾種方法均是透過降維處理，將群落間複雜的關係，轉換為平面或 3D 圖。以主成分分析為例，樣品中每個物種的相對豐度以 1 個座標表示，1 個物種 (x)、2 個物種 (x, y)、3 個物種 (x, y, z)，n 個物種即會形成 n 維向量，由於超過 3 維向量即無法直接繪製為圖表，因此透過降維演算，將多維度數據投射為 2 維平面或 3 維 3D 圖 (圖 8)。

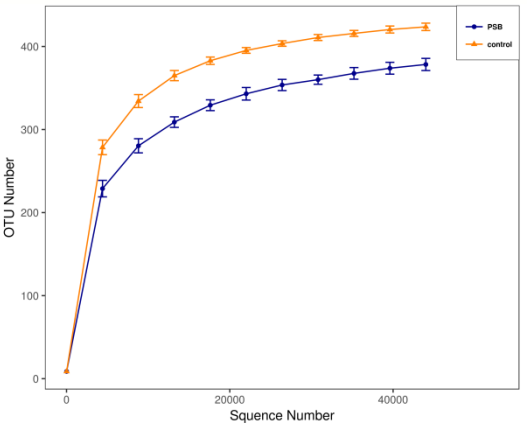

圖 7　稀釋曲線

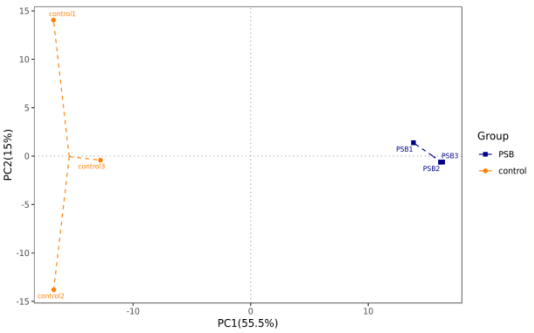

圖 8　主座標分析

三、統計分析

(一) 群落結構差異顯著性檢驗

β 多樣性指數經降維繪圖後，可看出群落組成的分類趨勢，但缺少是否達統計顯著性的資訊，為比較組間的差異是否顯著大於組內差異，常用的統計方法如相似性分析 (analysis of similarity, ANOSIM) 和多響應置換過程分析 (multiple response permutation procedure, MRPP)。

(二) 組間群落差異分析

為找出組別間豐度發生顯著變化的生物標誌 (biomarker)，線性判別分析效力量鑑別分析 (linear discriminant analysis (LDA) effect size, LEfSe) 先利用 Kruskal-Wallis test 找出

組間具有顯著豐度差異的物種，再以 Wilcoxon test 比較是否在子組中具有差異一致性，符合上述兩者的物種，最後以線性回歸分析 (LDA) 計算對組間群落差異的影響大小，LDA 距離大於設定值 (一般為 4.0)，判定該物種為具顯著差異的生物標誌 (圖 9)。

次世代定序應用

一、腸道

腸道是生物體中微生物相最複雜的器官，以人類腸道為例，約高達 40 屬、1,000 種微生物組成，這些微生物協助腸道發揮消化、對抗病原菌、參與免疫反應等功能。研究指出，過敏、心血管與腦慢性病、自體免疫疾病和肥胖等可能與腸道微生物相失衡有關。次世代定序可快速分析腸道微生物相組成，評估疾病風險，作為個人精準醫療的基礎。

水產養殖上，養殖生物的腸道微生物相可應用於益生菌、飼料開發、疾病治療等研究，如投餵某種益生菌後，是否能在魚類消化道定植增生，以及對腸道有害病原菌是否有抑制效果等。

二、環境

次世代定序技術可應用於環境污染程度評估，如四氯乙烯和三氯乙烯是工業常用的有機溶劑，進入環境後會滲漏至地下水層，造成污染擴散，脫氯菌群能降解氯烯類化合物，針對受氯烯類污染的場域，可透過脫氯菌群的相對豐度，作為污染鑑識及生物整治工法的成效評估依據。

此外，如污水處理廠、畜產排放水、養殖池水和底泥等生態系，仰賴複雜的微生物相分解有機物，並非少數菌種可獨立完成，次世代定序分析能協助探討不同菌種間的交互關係，開發高效能的生化處理複合菌。

三、食品

近年來次世代定序也開始應用於食品安全的檢驗上，例如檢測發酵加工過程是否有雜菌污染，運輸和倉儲過程是否有致病菌增殖等；當食物中毒案件發生時，更可快速釐清可能的致病原，並進一步透過全基因體定序確定菌株類型，除有助於追溯污染源頭外，更能提供醫療人員診療病患的重要資訊。

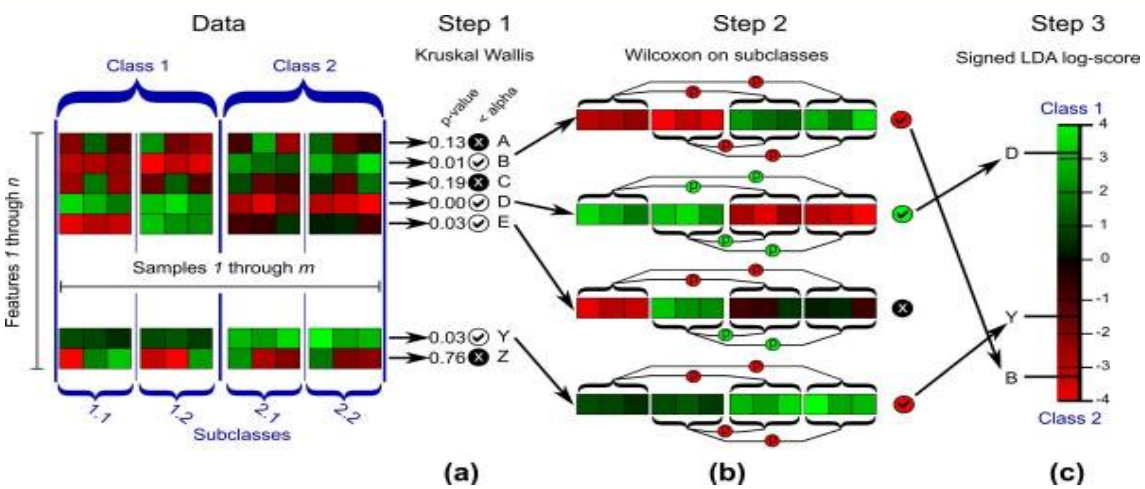

圖 9　線性判別分析效力量鑑別分析原理 (Segata et al., 2011)