應用DADA2套件進行環境微生物擴增子分析

農試所生技組 陳涵葳 杜元凱 林彥君

一、前言

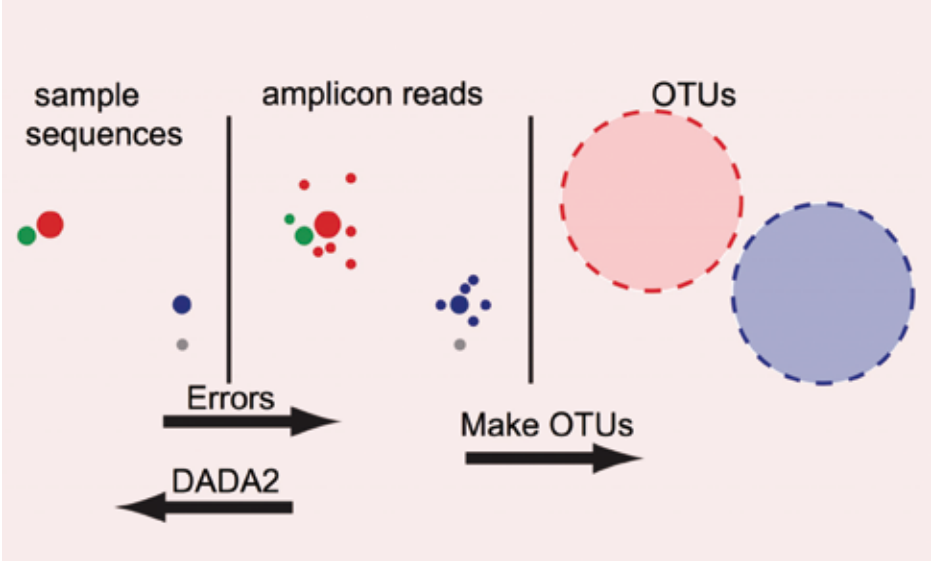
環境微生物相 (microbiota) 包括了細菌、真菌、病毒及原生生物，與土壤養分轉換、物質水解、植物殘體分解、土壤團粒結構的維持都極其相關；雖然我們認知微生物參與自然界物質的循環，但對微生物群落的了解卻相當有限。目前僅有不到1%的微生物菌種可在實驗室中進行培養，且需耗費大量時間與資源，但其結果之判讀仍高度仰賴經驗累積，這也是為什麼微生物群落組成與生態功能關聯性研究進展緩慢的主因。隨著次世代定序的普及，目前最具性價比的環境微生物物群落分析方法，是利用PCR增幅微生物特定片段後進行定序分析，藉由樣本擴增子 (amplicon) 組成與分群 (cluster)，初步了解環境樣品內微生物的豐富度與種類，更進一步則可推斷環境參數與微生物間的相關性。目前較常作為微生物物種鑑別的擴增子片段包括細菌16S rRNA、真菌ITS以及原核生物18S rRNA，但要區別擴增子序列間的差異是來自生物變異 (biological variation) 還是擴增子定序誤差 (amplicon sequencing error)，僅依靠Illumina定序平台的資料分析在實務上仍是很大的挑戰。而QIIME-uclust、MOTHUR和USEARCH-UPARSE等套件的演算方法，雖然能利用序列品質過濾、操作單元 (operational taxonomic units, OTU) 聚類演算法等方式來排除定序誤差，但也常因此錯失了微生物基因組間的細微變化，無法釐清複雜的微生物相與環境間的相關訊息。對此，本文將簡介R語言環境適用的DADA2套件，除了安裝簡便、功能強大外，還能有效判斷定序誤差與物種生物變異，提高環境微生物分析比對的準確性，對於環境內微生物群落多樣性分析、微生物體研究或系統微生物學的研究大有助益。

作　　者：陳涵葳助理研究員
連絡電話：04-23317322

二、DADA2套件之核心概念與分析流程

史丹福大學統計學系Holmes實驗室所開發的分裂擴增子降噪演算法(Divisive Amplicon Denoising Algorithm, DADA)，以誤差校正模型取代傳統OTU聚類演算法，對Roche 454平台擴增子序列進行降噪，減少偽陽性聚類單元的產生 (Rosen *et al*. 2012)，而DADA2則是適用於Illumina定序平台的新一代降噪演算法 (Callahan *et al*. 2016a)。我們假設環境樣本中具有已知的紅、綠、藍、灰四個微生物種 (圖一左)，面積大小代表了該物種序列豐度，距離則代表物種序列間的差異，經擴增子定序後衍生多個誤差序列 (圖一中)，OTU聚類演算法將97%相似性的序列視為同一群，無法正確區分四物種之差異 (圖一右)，DADA2演算法則根據誤差模型判斷擴增子定序誤差，可還原高解析的物種分析結果 (圖一左)。DADA2的核心概念是利用無監督學習法建立錯誤模型 (error model)，核苷酸轉移機率由原始核苷酸、置換核苷酸、定序品質參數 (quality score) 共同決定，由錯誤模型和豐富度資訊判定差異是來自於定序誤差或是物種變異，以修正Illumina定序平台定序資料的誤判，此外，DADA2套件不需要參考序列，適用於任何基因座，且套件本身具備完整的擴增子處理流程，如去除引子、序列過濾、序列組裝與去除嵌合…等，並能減少套件間資料格式轉換的問題，這些都是選用DADA2套件的優點。DADA2是R語言套件，安裝最新版本的R (4.0) 後，可從Bioconductor軟體平台 (https://www.bioconductor.org/) 搜尋DADA2安裝，也可直接在R中輸入BiocManager::install("dada2")指令進行安裝，DADA2具備完整的擴增子處理流程 (Callahan *et al*. 2016b)，簡述分析流程如下 (圖二)：

(一) 檔案匯入：原始數據 (raw data) 需去除轉接子 (adaptor trimming)，再進行樣品拆分 (demultiplexing) 與品

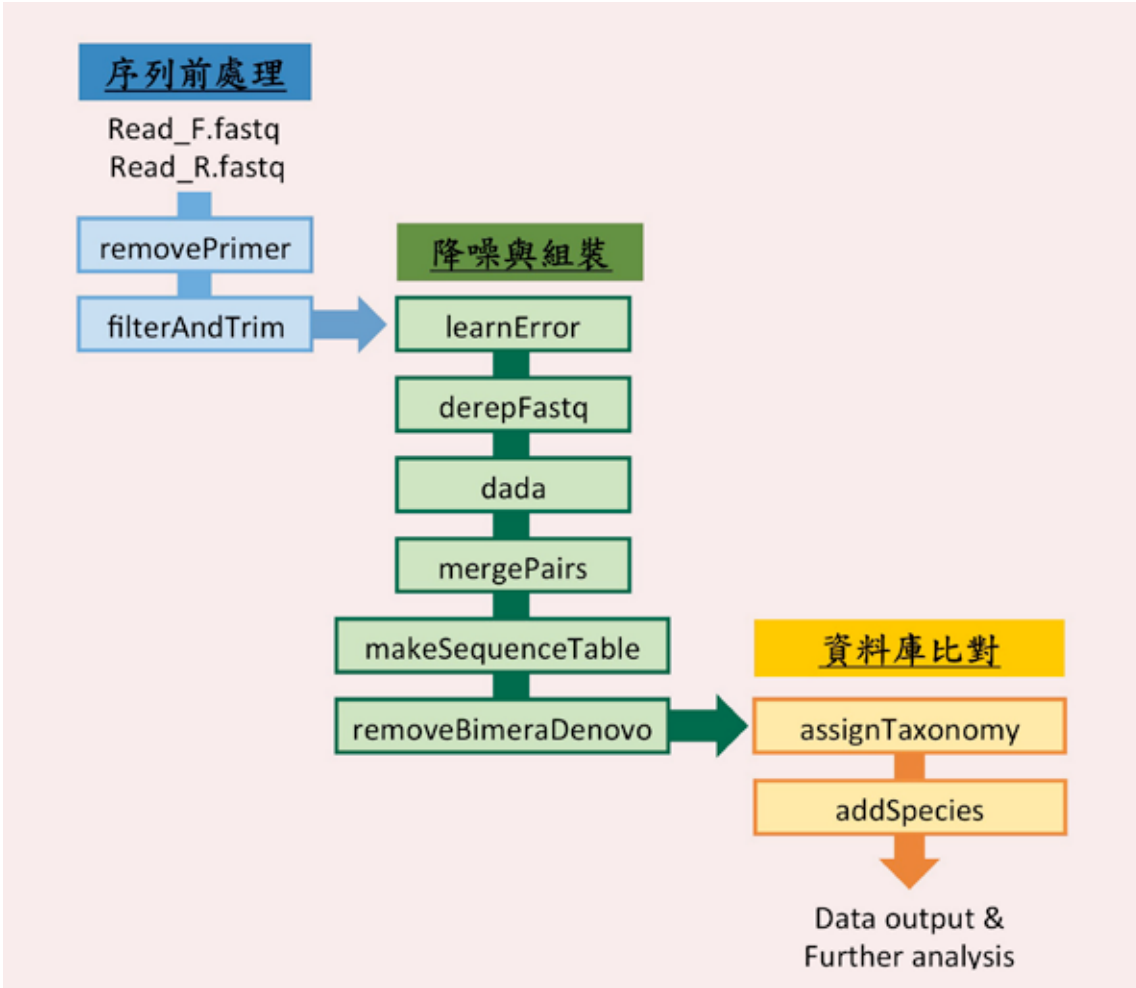


圖一、擴增子定序誤差示意圖。

質過濾 (quality filtering) 等步驟，才會得到可信序列 (clean reads)，DADA2可讀取*.fastq與*.fastq.gz兩種格式。

(二) 去除引子：細菌的16S rRNA基因全長約1.6 kb，具有9個高度變異區 (high variable regions)，V3－V4的變異區擴增子約428 bp，非常適合Illumina平台之短片段雙端 (paired-end) 定序，內建的removePrimers指令可去除指定之引子序列。

(三) 序列過濾：此步驟為DADA2的關鍵步驟，Illumina平台定序時末端品質較低，保留過多低品質的序列會影響錯誤模型之建立，但剪裁過多低品質序列則無法正確配對組裝。DADA2套件下使用filterAndTrime指令進行序列過濾，序列剪裁設定truncLen參數，確保序列之高品質與20個核苷酸重疊，以細菌16S V3－V4擴增子經MiSeq平台使用250 bp雙端定序為例，設定truncLen



圖二、DADA2套件分析流程。

= c (240,160)，對正向序列與反向序列分別裁剪至240 bp和160 bp，因為Illumina平台中反向序列品質明顯較差，故而剪裁至160 bp以保留較高品質的序列，但真菌ITS序列長度變化較大，建議改設minLen = 50。設定maxEE參數調整序列之錯誤期望值極大值，一般預設為2，設定愈大保留資料愈多，但也需要更多的處理效能。

(四) 誤差模型：DADA2使用selfConsist無監督學習模型建構誤差模型，假設最豐富的序列是正確的，其餘均為錯誤，使用learnErrors指令隨機擷取參考樣本交替估算錯誤率，推斷樣品組成直到收斂。

(五) 降低重複：使用derepFastq指令，具有相同序列的擴增子讀序重新分配形成單一序列，提高處理效能。

(六) 分裂分配：使用dada指令將樣品內的單一序列進行分裂分配，也可以更改pool參數為“TRUE”，將所有樣品的單一序列進行分裂分配，可以提高極低頻率序列偵測的敏感度，適合鑑定樣品中的低濃度混雜或汙染，但相對需要更多的處理效能，若更改pool參數為“pseudo”，可提高敏感度但仍保有很好的處理效能。

(七) 序列組裝：使用mergePairs指令，比對正向與反向互補序列，預設minOverlap = 12而maxMismatch = 0，無法順利配對的序列資料將被刪除減少雜訊，大部分的序列皆能順利配對，但若丟失太多的序列則建議檢查序列過濾步驟。

(八) 生成檔案：DADA2套件不再使用97%相似度的OTU聚類演算法，而是以makeSequenceTable指令生成高解析度的擴增序列變體 (amplicon sequence variants, ASV)。

表一、土壤樣品微生物擴增子分析統計資料

Name	Input reads	Filtered reads	Merged reads	Nom-chimera	Mapped reads	ASVs
Sample-1	126419	95197	90017	76181	76181	536
Sample-2	122660	88623	85176	73624	73624	209
Sample-3	138848	97804	91653	73803	73803	584
Sample-4	93610	72074	69058	55463	55463	268
Sample-5	87235	58339	53644	46910	46910	348
Sample-6	116097	87224	82296	65261	65261	463
Sample-7	105152	77903	74101	55471	55471	404
Sample-8	75909	49557	46189	41597	41597	335
Sample-9	74643	49598	46332	40280	40280	371
Sample-10	78171	51323	47788	41154	41154	391

(九) 去除嵌合：雙源嵌合體 (bimera又稱two-parent chimera) 常出現在雙端定序組裝後，這些嵌合體的正向與反向股分別來自高豐度的不同序列，使用removeBimeraDenovo指令，可比對並移除之。

(十) 資料庫比對：DADA2套件使用單純貝氏分類演算法 (naive Bayesian classifier method) 進行物種資料庫比對，細菌資料庫Silva、RDP、Greengene與真菌UNITE資料庫，皆持續更新並支持DADA2套件使用；下載資料庫檔案後匯入R，即可使用assignTaxonomy指令與addSpecies指令進行比對。

三、土壤樣品微生物擴增子分析

本研究室已初步建立DADA2套件擴增子定序資料分析流程，利用前述十個步驟進行土壤樣品擴增子定序資料分析，結果顯示DADA2套件在序列過濾這步驟保留70%高品質讀序建構誤差模型、計算豐富度p值、去除重複，經序列組裝、去除嵌合等步驟又移除約10%讀序，最終保留近60%的有效讀序進行物種比對 (表一)。取得土壤樣品之物種 (ASVs) 與豐富度 (每個ASVs對應的讀序數)，便可接續進行樣品內 (alpha) 與樣品間 (beta) 之多樣性分析。

四、結語

在環境微生物的研究方法中，利用選擇性培養基篩選、菌落形態辨識、計數，或利用生化指標測定，均十分仰賴經驗且耗時費工。隨著分子生物學的發展，次世代定序技術提供一個快速、簡便、高靈敏度的方法，幫助我們了解環境微生物群落結構和功能多樣性方面的資訊。DADA2套件具備完整的擴增子處理流程，可減少套件間資料格式轉換的問題，且能有效排除Illumina平台之定序誤差，提高辨識靈敏度，在微生物群落多樣性之研究上頗具優勢。若能配合採樣環境之pH值、有機質含量、微量元素含量等參數，可將微生物群落組成與環境間的關係初步連結，或更進一步推斷特定微生物之生態功能，對於了解複雜的環境微生物組成有著極大的幫助。

五、參考文獻

Callahan, B. J., P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. Johnson, and S. P. Holmes. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nat. Methods 13(7): 581–583.

Callahan, B. J., K. Sankaran, J. A. Fukuyama, P. J. McMurdie, and S. P. Holmes. 2016. Bioconductor workflow for microbiome data analysis: from raw reads to community analyses. F1000Research 5:1492

Rosen, M. J., B. J. Callahan, D.S. Fisher, and S. P. Holmes, 2012. Denoising PCR-amplified metagenome data. BMC Bioinformatics 13: 283.