台灣農業研究 (J. Taiwan Agric. Res.) 71(3):267–279 (2022)
DOI:10.6156/JTAR.202209_71(3).0006

研究報告

土壤微生物擴增子定序物種分類指派策略之研究

陳涵葳[1] 林美君[2] 林素禎[3] 曾清山[4] 杜元凱[1,*]

摘要

陳涵葳、林美君、林素禎、曾清山、杜元凱。2022。土壤微生物擴增子定序物種分類指派策略之研究。台灣農業研究 71(3):267–279。

應用 16S rRNA 基因擴增子定序進行土壤微生物物種 DNA 條碼鑑定，是近年微生物群落的研究趨勢，是一種高通量、標準化的方法學。DADA2 是適合 Illumina 定序平台使用的新一代分裂擴增子降噪演算法，可提供高解析的擴增子序列變體 (amplicon sequence variants; ASVs) 資料，如何連結微生物二名法與高解析資料，對土壤微生物群落後續分析顯得更加重要。本研究使用 DADA2 套件處理土壤樣品定序資料，比較 3 種不同的物種分類指派 (taxonomic assignment) 流程，結果顯示 DADA2 套件內件之 assignTaxonomy 指令搭配包含菌種名的 SILVA 138 參考序列訓練集 (training set)，有最好的物種分類指派效能。另以二元分類法評估 DADA2 套件適用的 SILVA 138、SILVA 138.1、GTDB 與 RefSeq + RDP 參考序列訓練集，對土壤微生物物種分類指派之效能，研究顯示 GTDB 訓練集敏感度最高，SILVA 138 與 SILVA 138.1 訓練集具有最佳特異性，而 RefSeq + RDP 訓練集物種分類指派結果之正確率、正確覆蓋率、馬修斯相關係數、陽性預測率指標均高於其他訓練集。微生物多樣性分析結果則顯示，GTDB 訓練集之物種分類指派結果最貼近原始 ASVs 資料，最能反應真實土壤微生物群落狀況。本研究揭示物種分類指派流程與參考序列訓練集的選擇，對微生物物種鑑別有很大的影響，隨著 16S rRNA 基因參考序列資料庫不斷地更新，更應該謹慎選擇與反覆評估，才能更準確的描述微生物間的多樣性關係。

關鍵詞：土壤微生物、16S rRNA 基因擴增子定序、DNA 條碼、DADA2。

前言

微生物群落之功能、交互作用與其變動性對生態平衡至關重要 (Zhou *et al*. 2015)，土壤是一個複雜且動態的生態系統，每克土壤中約有 10^6–10^7 種細菌、真菌或古生菌，但由於培養基和培養條件的限制，目前可培養分離的微生物僅占估計數量的 1–5% (Amann & Ludwig 2000)。隨著次世代定序技術的發展，原核生物核糖體小亞基 (ribosomal small subunit) 之 16S rRNA 基因序列分析已經是公認的細菌分類鑑定方法 (Pace 1997)，16S rRNA 基因全長約 1.6 kb，可分為 9 個高度變異區 (V1–V9)，已知 V3 與 V4 變異速度快，適合進行菌種層級之判定 (Mizrahi-Man *et al*. 2013)，利用引子增幅 16S rRNA 基因特定區域後進行擴增子定序 (amplicon sequencing)，是近年研究微生物群落系統發生和分類的趨勢。根據特定片段 DNA 序列進行物種分類的方法，也稱 DNA 條碼 (DNA barcoding) 鑑定，對於物種豐度較高的環境，如土壤細菌、海洋浮游生物……等研究，提供一個高通量、標準化的方法學，可落實大尺度或全球性的物種普查，如參考土壤宏基因體計畫 (reference soil metagenome pro-

投稿日期：2021 年 11 月 9 日；接受日期：2022 年 5 月 30 日。
* 通訊作者：yktu@tari.gov.tw
[1] 農委會農業試驗所生物技術組助理研究員。台灣 台中市。
[2] 農委會農業試驗所生物技術組計畫助理。台灣 台中市。
[3] 農委會農業試驗所農業化學組副研究員。台灣 台中市。
[4] 農委會農業試驗所生物技術組副研究員。台灣 台中市。

ject) (Vogel *et al*. 2009)、地球微生物體計畫 (the earth microbiome project; EMP) (Gilbert *et al*. 2010)、中國土壤微生物體倡議 (China soil microbiome initiative; CSMI) (Wang *et al*. 2021) 等。土壤微生物體學的技術的發展，從傳統型態辨識走向高通量定序、生物資訊與大數據分析，要連結 DNA 序列與環境生態，必須提高微生物物種分類解析度 (taxonomic resolution)，可從擴增子序列分類方法以及 16S rRNA 基因參考資料庫之選擇兩方面著手。

擴增子序列分類方法包括：操作分類單元 (operating taxonomic units; OTU) 與擴增子序列變體 (amplicon sequence variant; ASV) 兩種。OTU 分析主要是利用聚類法 (clustering)，如 97% 相似性聚類法、無參考序列 OTU 聚類法 (reference-free OTU clustering)、封閉參考 OTU 聚類法 (closed-reference OTU clustering) ……等；缺點為運算資源過大、虛假 OTU 產生、無法辨識參考序列外之物種等 (Callahan *et al*. 2017)。ASV 分析法與 OTU 聚類的概念相反，以分裂分配演算法 (divisive partitioning algorithm) 進行讀序分析，如分裂擴增子降噪演算法 (divisive amplicon denoising algorithm; DADA)，結合讀序豐度校正錯誤率，有效降低偽陽性誤判 (Rosen *et al*. 2012)，而 DADA2 是適用於 Illumina 定序平台的降噪演算法 (Callahan *et al*. 2016)，其他 ASV 演算法套件如 Deblur (Amir *et al*. 2017)、UNOISE3 (Edgar 2016) 等。DADA2 較 Deblur、UNOISE3 有更高的靈敏度和分辨率，是目前擴增子讀序資料分析套件中最適合進行菌種層級分辨的 (Callahan *et al*. 2019; Prodan *et al*. 2020)。常用的 16S rRNA 基因資料庫包括核糖體資料庫計畫 (Ribosomal database project; RDP) (Cole *et al*. 2014)、SILVA 資料庫 (Quast *et al*. 2012)、GreenGenes 資料庫 (DeSantis *et al*. 2006)、RefSeq 資料庫 (O'Leary *et al*. 2016)，以及 GTDB 資料庫 (Parks *et al*. 2020)。由於 DADA2 演算法能產生極高分辨率的 ASVs 資料，倘若無法正確連結微生物二名法，這樣高解析的資料便失去其意義 (Curry *et al*. 2018)。本研究使用 DADA2 套件處理土塊、旱田、水田 3 種土壤樣品定序資料，比較不同指令對土壤微生物物種分類指派的效能，評估 SILVA 138、SILVA 138.1、GTDB 與 RefSeq + RDP 等 4 個符合 DADA2 格式之參考序列訓練集，於 ASV 分析流程的適用性，期望能建立最適合的分析方法，提升土壤微生物群落分析之正確性。

材料與方法

土壤樣本收集

本研究收集行政院農業委員會農業試驗所之試驗田區土壤，包括長期種植水稻之水田 (paddy field; PF)、長期種植果樹之旱田 (dry field; DF)，以及旱田周邊沒有植物根系穿越之土塊 (bulk soil; BS)，每種性質的土壤有 6 個土壤樣品 (n = 6)，共計 18 份土壤樣品，而每份土壤樣品是由 3 個採土點土面下 15 cm 之 100 g 土壤混合而成，於室溫下風乾後以 2 mm 篩網過濾去除植物殘體與碎石，之後立即進行土壤 DNA 萃取。

土壤樣本 DNA 萃取與建庫

秤取 0.3 g 土壤樣品萃取土壤 DNA，使用 Power Soil 土壤 DNA 純化試劑組 (Qiagen, Hilden, Germany)，依照使用說明進行 DNA 純化。所得 DNA 經微量分光光度計 (ND-1000, NanoDrop Technologies, Wilmington, DE, USA) 測定濃度與純度，再以 1% 洋菜膠體電泳確認 DNA 完整性後，以無菌水將 DNA 稀釋成 1 ng μL^{-1} 後保存 -20℃。利用細菌 16S rRNA 基因 V3–V4 區域通用引子對 341F (CCTACGGGAGGCAGCAG)/805R (GACTACHVGGGTATCTAATCC) 進行增幅反應，每個 polymerase chain reaction (PCR) 反應總體積為 25 μL，其中包含 1 ng 土壤 DNA、0.5 μM 正向引子與反向引子、0.5 μL High-Fidelity PCR Master Mix (KAPA Biosystems, Wilmington, MA, USA)。PCR 反應條件為 95℃ 3 min；95℃ 30 s、57℃ 30 s、72 ℃ 30 s， 循環 30 次；72 ℃ 5 min。PCR 產物經 2% 洋菜膠體電泳確認後，以 QIAquick 膠體回收試劑組 (Qiagen, Hilden, Germany) 回收 450–500 bp 片段，再以 Qubit 2.0 螢光定量系統進行 DNA 片段定量 (Thermo

Fisher Scientific, Waltham, MA, USA)。DNA 文庫使用 Truseq nano DNA Library Prep Kit (Illumina, San Diego, CA, USA) 依原廠建議步驟進行構築。DNA 文庫於 Illumina MiSeq 平台上進行定序，生成 2 × 300 bp 雙端讀序資料。

讀序分析流程

所有 MiSeq 讀序資料都在 R (版本 4.1.0) 環境下，以 DADA2 (版本 1.16) 套件生成 ASVs 資料。讀序處理流程參照 Callahan *et al*. (2016) 之建議，匯入 *.fastq.gz 格式之讀序資料，以 removePrimers 指令去除讀序內引子序列；使用 filterAndTrime 指令進行序列過濾，設定 truncLen = c (240, 160)，對正向讀序與反向讀序分別裁剪至 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，設定 maxEE 參數調整序列之錯誤期望值極大值；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；以 derepFastq 指令，將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；使用 dada 指令將樣品內的單一序列進行分裂分配，在此更改 pool 參數為 TRUE，將所有樣品的單一序列進行分裂分配，可以提高極低豐度序列的敏感度；輸入 mergePairs 指令，比對正向與反向互補序列，預設 minOverlap = 12 而 maxMismatch = 0，接著以 makeSequenceTable 指令生成高解析度的 ASVs 資料，再以 removeBimeraDenovo 指令進行比對，移除雙源嵌合體 (bimera)。

物種分類指派流程 (pipeline)

流程 1 (pipeline_1)：下載並載入 DECIPHER (版本 2.6.0) 套件適用之 SILVA 138 參考序列訓練集 (http://www2.decipher.codes/Classification/TrainingSets/SILVA_SSU_r138_2019.RData)，輸入 IdTaxa 指令並確認參數 strand = “both”，即可進行物種比對。流程 2 (pipeline_2)：下載並載入 DADA2 套件適用的 SILVA 138 參考序列訓練集 (包含菌種種名) (https://zenodo.org/record/3986799/files/silva_nr99_v138_wSpecies_train_set.fa.gz?download=1)，輸入 assignTaxonomy 指令，將 tryRC 參數之設定改為 TRUE，進行物種比對。流程 3 (pipeline_3)：下載 SILVA 138 參考序列訓練集 (https://zenodo.org/record/3986799/files/silva_nr99_v138_train_set.fa.gz?download=1)，將訓練集載入 DADA2 套件後，輸入 assignTaxonomy 指令並更改 tryRC 參數為 TRUE，進行物種比對；下載 SILVA 138 二名法參考訓練集 (https://zenodo.org/record/3986799/files/silva_species_assignment_v138.fa.gz?download=1)，載入 DADA2 並以 addSpecies 指令添加菌種種名。

參考序列訓練集效能評估

為評估參考序列訓練集於高解析 ASVs 資料的適用性，於 DADA2 網站 (https://benjjneb.github.io/dada2/training.html) 下載了 4 個 16S rRNA 基因參考序列訓練集 (皆包含菌種種名)，分別為：SILVA 138 (2020 年 8 月 15 日發布)、SILVA 138.1 (2021 年 3 月 7 日發布)、GTDB (2020 年 4 月 28 日發布)、RefSeq + RDP (2020 年 6 月 11 日發布)。使用前述流程 2 步驟進行物種分類指派。參考 Escobar-Zepeda *et al*. (2018) 的策略，以二元分類測試法評估不同參考序列訓練集的分類效能；為取得菌種分類之「參考值」，參照 McClenaghan *et al*. (2020) 的分析流程，選取土壤樣品豐度排序前 250 名之 ASVs 資料，將序列資料轉為 *.fast 格式，使用 Basic Local Alignment Search Tool (BLAST) 之 blastn 比對工具 (於 2021 年 8 月 6 日進行比對)，設定 e-value 最低閥值為 0.01，若比對後有 2 個或以上的物種分類指派，則以最低共同分類為分類結果，輸出比對結果作為菌種分類指派的參考值。4 個參考序列訓練集比對結果再與參考值進行比較，若物種分類指派與參考值一致，標註為真陽性 (true positive; *TP*)；若物種分類指派與參考值不一致，標註為偽陽性 (false positive; *FP*)；若參考序列訓練集沒有辦法給予物種分類指派，則標註為偽陰性 (false negative; *FN*)；另加入 100 條隨機序列於 ASVs 資料中，標註為真陰性 (true negative; *TN*)。本研究所使用的效能評估指標計算公式如下：

$$\text{正確率 (Accuracy; } ACC) = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$\text{正確覆蓋率 (Coverage; } COV) = \frac{TP}{TP + FP + FN + TN} \quad (2)$$

$$\text{馬修斯相關係數 (Matthews correlation coefficient; } MCC) = \frac{(TP \times TN) - (FP \times FN)}{[(TP + FP)(TP + FN)(TN + FP)(TN + FN)]^{1/2}} \quad (3)$$

$$\text{陽性預測率 (Positive predictive value; } PPV) = \frac{TP}{TP + FP} \quad (4)$$

$$\text{真陰性率 (True negative rate; } TNR) = \frac{TN}{FN + TN} \quad (5)$$

$$\text{真陽性率 (True positive rate; } TPR) = \frac{TP}{TP + FN} \quad (6)$$

數據處理與統計分析

土壤樣品之 ASVs 資料使用 vegan 套件 (版本 2.5.7)、phyloseq 套件 (版本 1.36.0)、ROCit 套件 (版本 2.1.1) 進行分析，各項數據以 SAS 統計分析軟體 (SAS Enterprise Guede 7.1) 進行變方分析 (analysis of variance; ANOVA) 後，再以最小顯著差異性測驗 (least significant difference test; LSD test)，比較 5% 顯著水準下各處理間的差異。

結果與討論

讀序處理結果

BS、DF 與 PF 土壤樣品經 Illumina MiSeq 高通量定序平台取得了 651,173、604,447 及 616,843 條原始讀序 (raw reads)，通過 DADA2 套件 removePrimers 指令去除 16S rRNA 基因擴增片段引子序列，約保存 99.4% 之讀序資料 (表 1)，扣除引子後的平均長度為 409 bp。以 filterAndTrime 指令移除品質不佳的讀序，可保留 70% 過濾讀續 (filtered reads)，BS、DF 與 PF 土壤樣品之組裝讀序 (merged reads) 依序為 454,292、397,095 與 418,949 條，經分裂分配演算法得到之 ASVs 資料量如表 1 所列，BS 土壤樣品之 ASVs 資料量為 2,279，低於 DF 與 PF 土壤樣品之 ASVs 資料量 (表 1)。植物根部分泌多種物質，包括碳水化合物、胺基酸、有機酸、二級代謝物……等，根圈附近的微生物較容易獲得養分 (Walker *et al*. 2003)，可以解釋為何沒有植物根系穿過的 BS 土壤樣品 ASVs 資料量較低，微生物多樣性少。稀釋曲線 (rarefaction curve) 或豐富度曲線可以反應定序資料量是否足以代表定序樣品的物種多樣性，根據圖 1 結果顯示，隨著抽樣數量的增加，BS、DF 與 PF 土壤樣品之 ASVs 數量趨於平緩。

物種分類指派流程之效能評估

比較DECIPHER/IdTaxa、DADA2/assignTaxomomy、DADA2/assignTaxonomy/addSpecies 三種分類指派流程進行運算，並以物種分類指派率 (assignment to taxon) 來評估不同流程之物種分類指派效能，流程 1 的物種分類指派率在門、綱、目、科與屬的層級均較流程 2、3 為低，顯示 DADA2 套件的物種分類指派運算效能較 DECIPHER 套件佳，在門、綱的層級均可達到 97% 的物種分類指派率 (表 2)。流程 2 與流程 3 的差別在於是否使用 addSpecies 步驟添加菌種種名，在門、綱、目、科與屬的層級，兩者沒有差異，但流

表 1. 土塊、旱田、水田土壤樣品之原始讀序、過濾讀序、組裝讀序與 16S rRNA 基因擴增子序列變體數量。

Table 1. The number of raw reads, filtered reads, merged reads and amplicon sequence variants (ASVs) in bulk soils (BS), dry field soils (DF) and paddy field soils (PF).

Method	BS	DF	PF
Raw reads	651,173	604,447	616,843
Filtered reads	479,360	421,824	443,899
Merged reads	454,292	397,095	418,949
ASVs	2,279	4,496	5,665

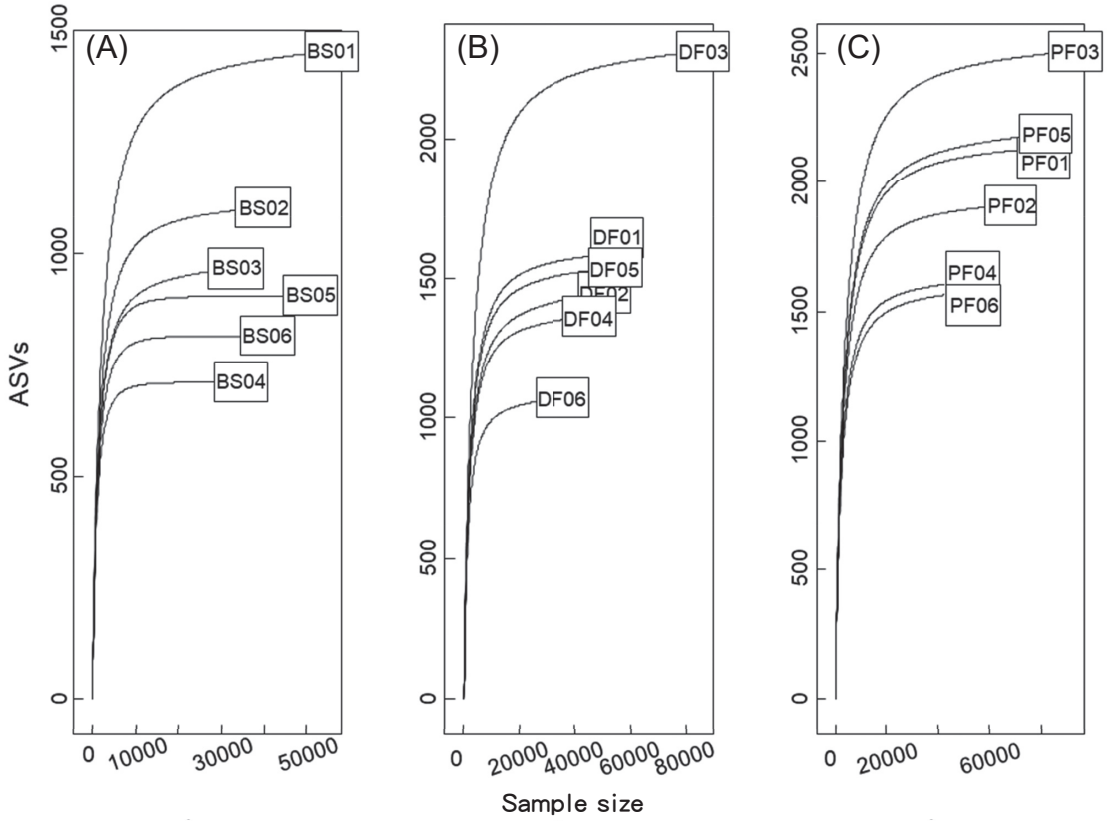

圖 1. 稀釋曲線。(A) 6 個土塊 (bulk soil; BS) 樣品 16S rRNA 基因擴增子序列變體稀釋曲線；(B) 6 個旱田 (dry field; DF) 樣品 16S rRNA 基因擴增子序列變體稀釋曲線；(C) 6 個水田 (paddy field; PF) 樣品 16S rRNA 基因擴增子序列變體稀釋曲線。

Fig. 1. Rarefication curves. (A) Rarefication curves for amplicon sequence variants (ASVs) in 6 bulk soil (BS) samples; (B) rarefication curves for ASVs in 6 dry field (DF) samples; and (C) rarefication curves for ASVs in 6 paddy field (PF) samples.

程 2 在種的層級具有最佳的比對效能，顯示二名法參考序列訓練集與 addSpecies 步驟並不是必要，直接選擇包含菌種種名的參考序列訓練集即可 (表 2)。DADA2 套件使用 RDP 分類器，是建構於單純貝氏分類模型與 Bergey 分類法的機器學習分類 (Callahan *et al*. 2016)，RDP 分類器為了加快比對速度，將序列拆分成 8 bp 一組的 k-mer，適合高通量 rRNA 基因序列分析，但對於較偏離參考序列訓練集的資料，通常會給予可信度較低的分類結果 (Wang *et al*. 2007)，這也是 RDP 分類器常發生過度分類 (over classification) 的原因。流程 1 使用 DECIPHER 套件的 IdTaxa 分類器，能整合系統發育、機器學習和距離特徵等分類法，與 RDP 分類器相比，具有較低的過度分類錯誤率，但在人類腸道微生物定序資料比對結果也顯示，在屬的分類層級上，IdTaxa 能提供的物種分類指派率低於 RDP 分類器 (Murali *et al*. 2018)，這與本研究結果一致 (表 2)。此外，適合 DECIPHER 套件使用的 SILVA 138 參考序列訓練集並不包含種的分類層級，雖然 IdTaxa 分類器整合多種分類方法，具有運算速度快、避免過度分類等優點，但目前沒有完備的參考序列訓練集可供下載使用，是美中不足的地方。本研究建議 ASVs 資料之物種分類指派，仍以 DADA2 套件內建的 assignTaxonomy 指令，搭配使用包含菌種種名的參考序列訓練集 (流程 2)，會有最佳的物種分類指派率。

表 2. 不同流程之物種分類指派效能比較。

Table 2. Comparison with the percentage of taxonomic assignments at different taxonomic ranks using different pipelines.

	Assignment to taxon (%)		
Taxonomic ranks	Pipeline_1[z]	Pipeline_2[y]	Pipeline_3[x]
Kingdom	100.00 ± 0.00 a	100.00 ± 0.00 a	100.00 ± 0.00 a
Phylum	83.23 ± 2.92 b	99.57 ± 0.21 a	99.68 ± 0.19 a
Class	81.24 ± 3.11 b	97.19 ± 0.95 a	97.25 ± 0.88 a
Order	71.72 ± 4.20 b	90.84 ± 1.84 a	90.84 ± 1.91 a
Family	61.19 ± 4.74 b	77.19 ± 3.62 a	76.84 ± 3.75 a
Genus	38.18 ± 4.69 b	51.07 ± 4.75 a	50.37 ± 4.88 a
Species	0.00 ± 0.00 c	11.96 ± 2.81 a	5.92 ± 1.95 b

Data represent mean ± standard error of three replicates. Values in the same row with different letters are significantly different ($P < 0.05$) based on Fisher's protected least significant difference test (LSD) test.
[z] Pipeline_1: IdTaxa (DECIPHER) + SILVA 138 training set.
[y] Pipeline_2: assignTaxonomy (DADA2) + SILVA 138 training set (with Species).
[x] Pipeline_3: assignTaxonomy (DADA2) + SILVA 138 training set (without Species) → addSpecies (DADA2) + SILVA 138 training set (Species assignment).

參考序列訓練集物種分類指派效能評估

根據序列進行物種指派的策略包括了序列相似性搜尋、序列組合方法與系統發育方法；序列相似性搜尋是基於序列同源性或比對的方式，如 BLAST，而系統發育方法則是將進化模型應用於資料庫中，尋找目標序列在系統發育中最適合的位置。本研究所評估之 SILVA 138、SILVA 138.1、GTDB 和 RefSeq + RDP 等 4 個符合 DADA2 格式之參考序列訓練集，皆是擷取微生物公共資料庫序列資料的優化版本，以 BLAST 結果作為「參考值」，配合二元分類測試法評估 4 個參考序列訓練集對於土壤微生物物種分類指派之效能。針對屬與種分類階層進行 *ACC*、*COV*、*MCC*、*PPV*、*TNR* 與 *TPR* 等效能評估指標計算，RefSeq + RDP 訓練集在屬的分類層級上，*ACC*、*COV*、*MCC* 與 *PPV* 的表現均顯著優於其他訓練集，在種的分類階層中 RefSeq + RDP 訓練集 *COV* 指標可達 50%，顯示該訓練集具有最大的正確覆蓋率，*ACC* 與 *MCC* 指標表現亦較 SILVA 138、SILVA 138.1 與 GTDB 訓練集佳 (圖 2)。GTDB 訓練集 *TPR* 指標顯著高於其他訓練集，由於 *TPR* 指標也稱敏感性 (sensitivity) 指標，即 GTDB 訓練集在屬與種的分類階層，對絕大部分的 ASVs 都可給予對應的物種指派 (圖 2)。雖然 SILVA 138 與 SILVA 138.1 訓練集在 *ACC*、*COV*、*MCC*、*TPR* 指標雖遜於 RefSeq + RDP 訓練集，但 *TNR* 指標表現最佳，*TNR* 指標也稱特異性 (specificity) 指標，顯示 SILVA 訓練集判定的物種雖少，但所判定的物種大部分與「參考值」一致 (圖 2)。四個參考序列訓練集物種分類指派之特徵曲線 (receiver operating characteristic curve; ROC curve) 分析結果也顯示，RefSeq + RDP 資料集之曲線下面積 (area under curve; AUC) 達 0.85，優於 GTDB、SILVA 138 與 SILVA 138.1 之 0.81、0.78 與 0.71 (圖 3)。綜合二元分類 *ACC*、*COV*、*MCC* 指標與 AUC 檢定結果，本研究之土壤 ASV 數據集使用 DADA2 套件內建的 assignTaxonomy 指令，搭配 RefSeq + RDP 參考序列訓練集，會有最好的物種分類指派效能。GTDB 與 RefSeq + RDP 訓練集的資料節點項數，在屬與種的分類階層遠較 SILVA 訓練集為多 (圖 4)，在物種分類指派效能評估上較占優勢。參考序列資料庫或訓練集的物種命名方法，對於物種分類指派結果也深具影響；GTDB 訓練集在 *PPV* 與 *TNR* 評估指標上表現不佳，主要是因為 GTDB 資料庫是一個完全基於基因體序列建構的分類資料庫，並連接蛋白質系統發育作為細菌分類學的基礎架構，這種新的分類方法可擴大 30% 以上細菌及古生菌多樣性，但有 58% 的序列與

現有物種分配不一致 (Parks *et al*. 2018)，將 GTDB 資料庫與 National Center for Biotechnology Information (NCBI) 資料庫共有序列的物種進行比較，約 35.1% 的物種具有不同的命名，主要是因為 GTDB 資料庫採用相對進化分歧 (relative evolutionary divergence) 的策略進行屬分類階層的歸一化，並對不能培養的菌株給與占位名稱 (placeholder names)，變化最多的分類為 *Pseudomonas* 屬、*Bacillus* 屬與 *Lactobacillus* 屬 (Parks *et al*. 2020)，在 Bacillus 屬的分類研究中，歸納了 6 個分類進化枝 (phyletic clades)，對應形成 6 個新的屬，包括 *Peribacillus*、*Cytobacillus*、*Mesobacillus*、*Neobacillus*、*Metabacillus* 與 *Alkalihalobacillus*，與 GTDB 資料庫的分組情形一致 (Patel & Gupta 2020)，顯示 GTDB 分類命名基礎上與 NCBI 資料庫雖有較大的差異，但對於不能培養的菌株或是系統發育不連貫的群體 (phylogenetically incoherent group)，提供了新穎且具代表性的物種分類指派。

參考序列訓練集對微生物群落多樣性分析之影響

已知 16S rRNA 基因參考序列資料庫之組成、品質和完整性，決定了 DNA 條碼對物種的鑑別能力，對物種分類指派影響極大 (Park

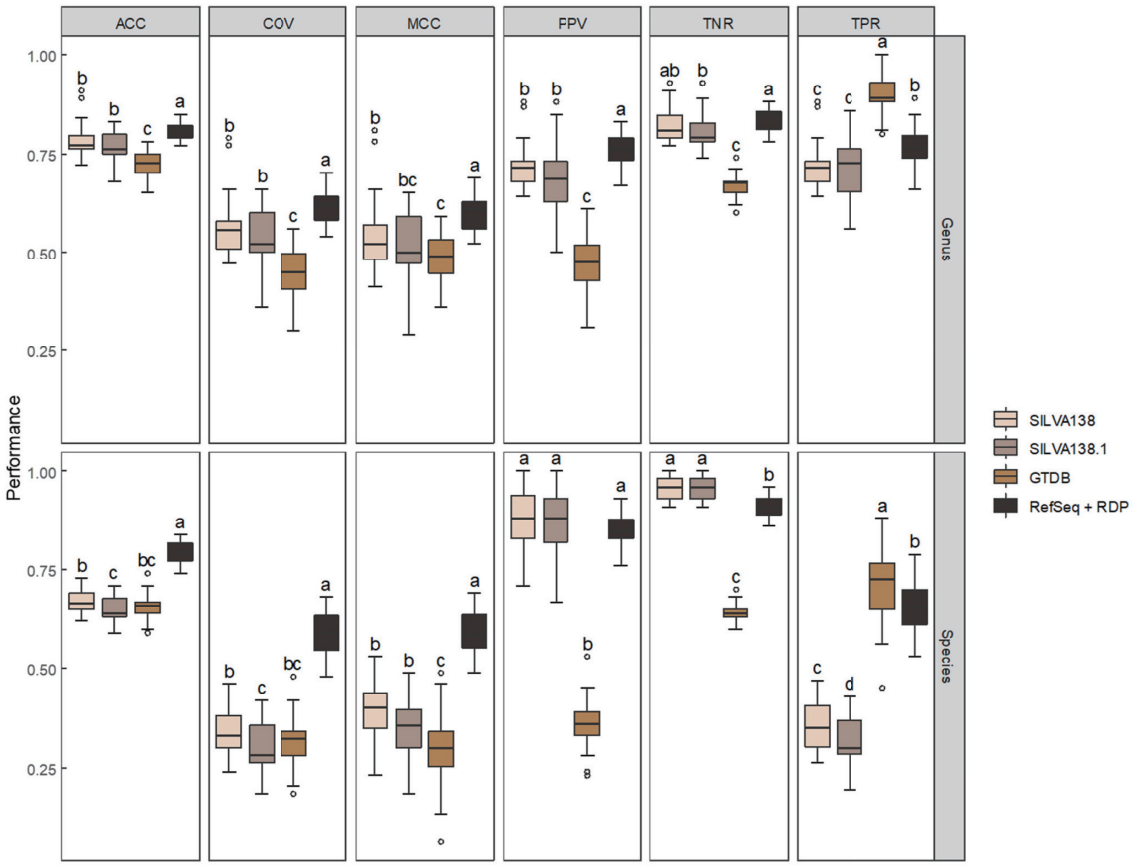

圖 2. 不同參考序列訓練集之效能指標評估。

Fig. 2. Performance descriptors for taxonomic assignments at Genus or Species ranks based on different training sets. Six indicators in terms of accuracy (ACC), coverage (COV), Matthews Correlation Coefficient (MCC), positive predictive value (PPV), true negative rate (TNR), and true positive rate (TPR) were used to describe the performance of taxonomic assignments. Different letters above bars indicate a significant difference based on analysis of variance (ANOVA) test followed by Fisher's protected least significant difference test (LSD) test ($P < 0.05$).

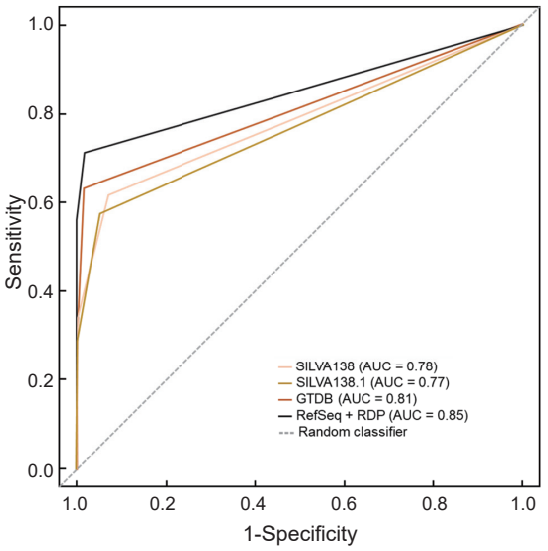

圖 3. 不同參考序列訓練集之特徵曲線評估。

Fig. 3. Receiver operating characteristic (ROC) curves for taxonomic assignments of different training sets.

& Won 2018; Henderson *et al*. 2019)。由於 SILVA 138.1 是 SILVA 138 訓練集修正版本，僅修正少部分不確定分類位置 (*incertae sedis*) 與未分類缺失，為維持資料的一致性與可比較性，後續僅保留 SILVA 138、GTDB 和 RefSeq + RDP 三個參考序列訓練集之物種分類指派結果，利用 phyloseq 套件整併相同物種分類指派結果之 ASVs，之後再進行微生物群落多樣性分析，土壤樣品在種的分類階層之 α 多樣性指數箱型圖如圖 5，原始 ASVs 資料中，PF 土壤樣品多樣性最高，其次為 DF 土壤樣品，BS 土壤樣品多樣性最低，但經物種分類指派與整併後，α 多樣性指數降低，SILVA 訓練集之物種分類指派結果在 Chao1、ACE、Shannon、InvSimpson 與 Fisher 等多樣性指數上已無法反應原始情況，GTDB 訓練集之物種分類指派結果較接近原始 ASVs 分析結果。從主座標分析 (principal co-ordinates analysis; PCoA) 雙

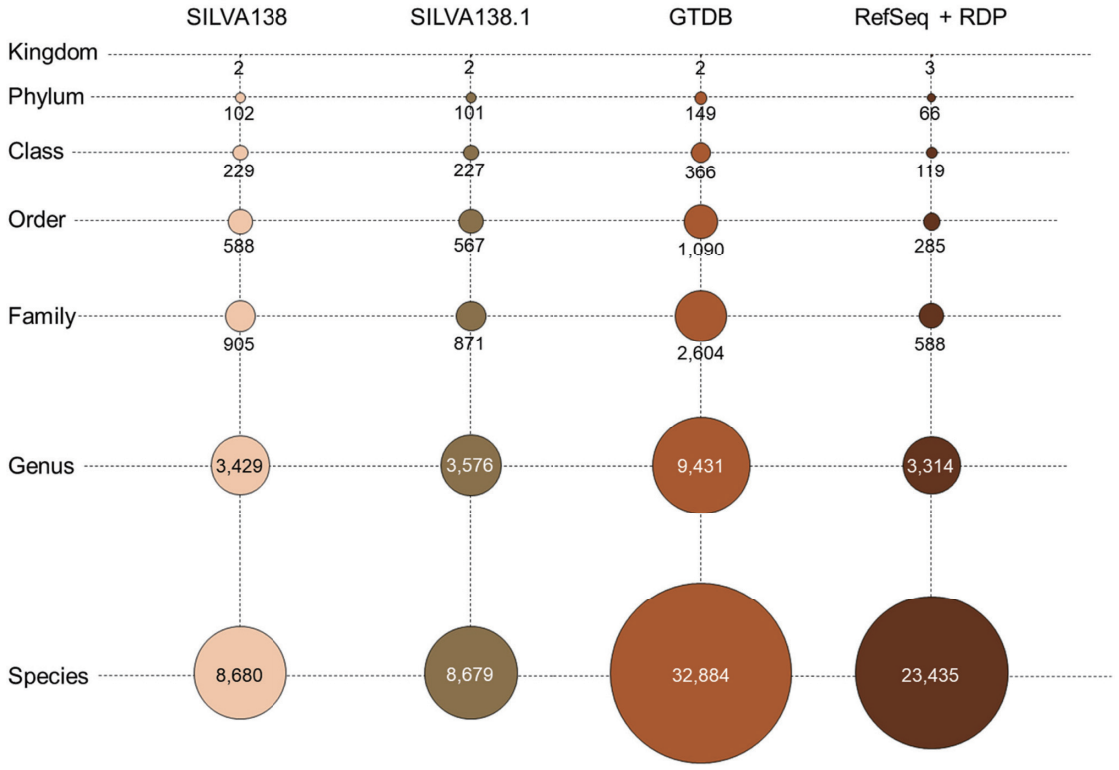

圖 4. 參考序列訓練集資料節點之比較。

Fig. 4. Composition with the number of nodes at each rank. Circle area correspond to the number of nodes at each rank in SILVA 138, SILVA 138.1, GTDB, and RefSeq + RDP 16S rRNA training set.

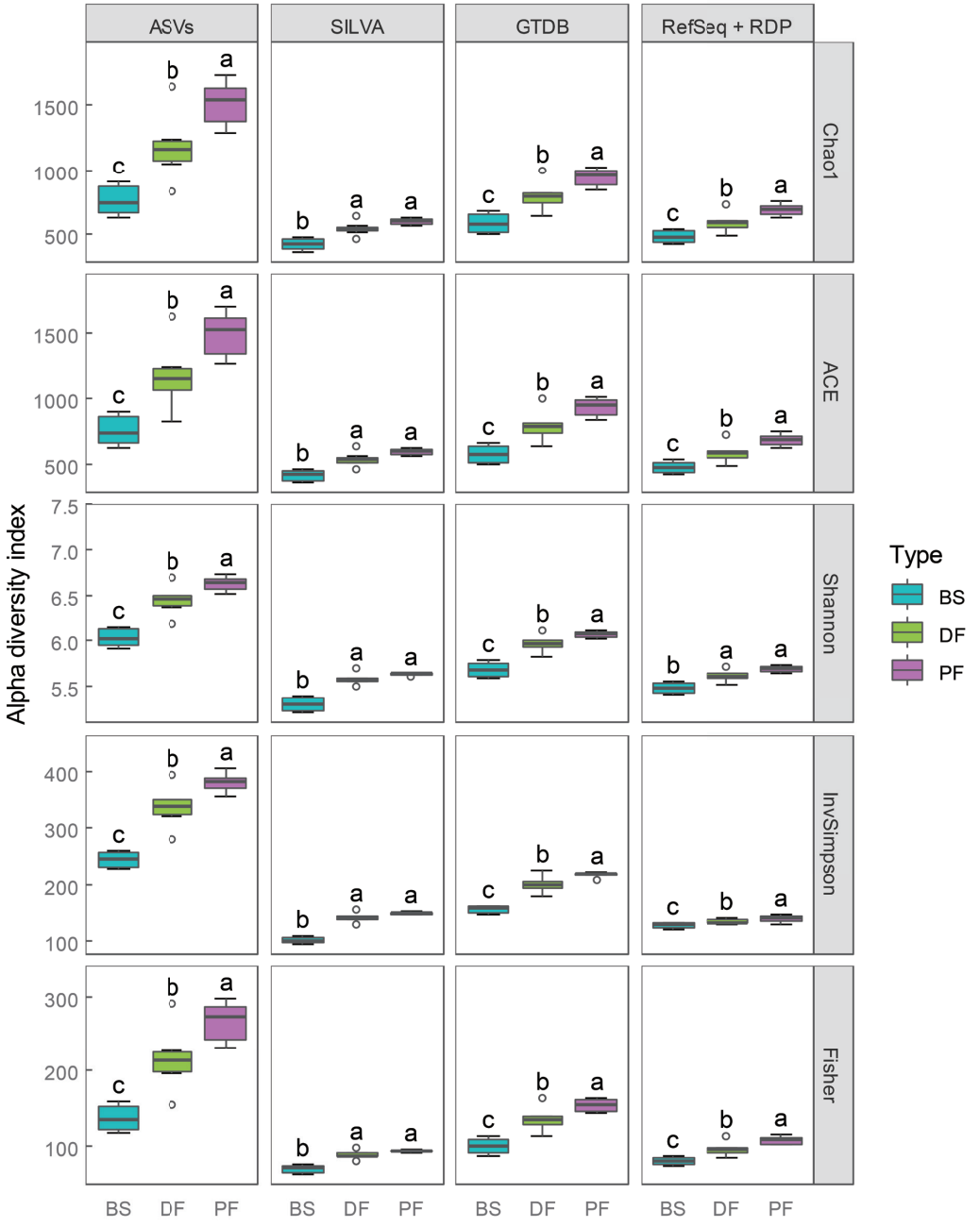

圖 5. 不同參考序列訓練集結果之 α 多樣性。

Fig. 5. α diversity for taxonomic assignments of different training sets. α diversity assessed by Chao1, ACE, Shannon, InvSimpson, and Fisher. Letters indicate significant differences ($P < 0.05$) among bulk soils (BS), dry field soils (DF), and paddy field soils (PF).

標圖 (biplot) 結果可以清楚看到不同訓練集對 β 多樣性的影響 (圖 6)，2,601 筆 ASVs 資料經 SILVA 訓練集進行物種分類指派與整併後，在種的分類階層上僅保留 238 筆資料，樣品分布位置與原始資料具有落差 (圖 6A、E)，而 GTDB 與 RefSeq + RDP 訓練集之物種分類

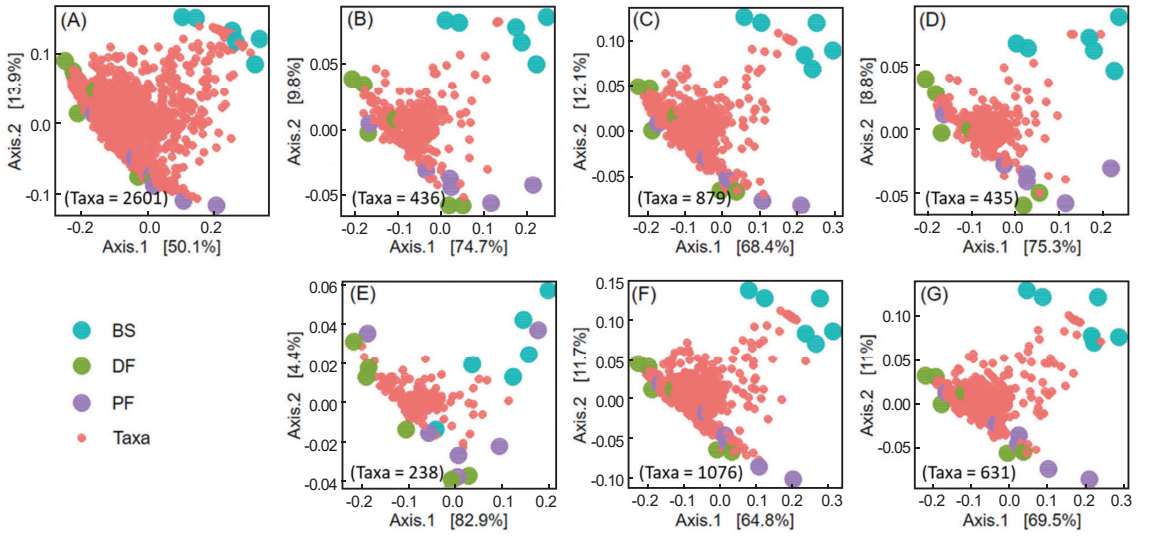

圖 6. 主座標分析雙標圖。土壤樣品依照土塊、旱田、水田給予不同顏色的分群標示。(A) 向量圖型繪製參照原始 ASVs 資料；(B、E) SILVA 138 訓練集物種分類指派結果；(C、F) GTDB 訓練集物種分類指派結果；(D、G) RefSeq + RDP 訓練集物種分類指派結果。(B–D) 為屬的分類階層；(E–G) 為種的分類階層。所有圖形皆依照 Bray-Curtis 相異度進行繪製。

Fig. 6. Principal co-ordinates analysis (PCoA) plot. The 18 soil samples grouped and colored according to bulk soil (BS), dry field (DF), and paddy field (PF). (A) Ordination biplot were generated by original ASVs data; (B, E) by taxonomic assignment of SILVA 138 training set; (C, F) by taxonomic assignment of GTDB training set; and (D, G) by taxonomic assignment of RedSeq + RDP training set. (B–D) Plots of rank scores at Genus level and (E–G) at Species level. All plots were generated with Bray-Curtis dissimilarity.

指派結果，分別保留 1,076 和 631 筆資料，樣品分布位置與原始資料相當 (圖 6F、G)；在屬的分類階層上，SILVA 138、GTDB 和 RefSeq + RDP 三個訓練集之物種分類指派結果依序為 436、879 與 435，樣品分布位置沒有明顯變化 (圖 6B、C、D)，上述結果顯示在種的分類階層上，以 GTDB 訓練集之物種分類指派結果最貼近原始 ASVs 資料，α 多樣性與 β 多樣性分析結果能反應原始的高解析 ASVs 資料分布；SILVA 訓練集在種的分類階層上，物種分類指派效能不佳，但在屬的分類階層上則與 RefSeq + RDP 訓練集結果相近。SILVA 138 參考序列訓練集選自 SILVA 138 SSU 資料庫，所包含的參考序列資料量為 452,522 筆，大於 GTDB 訓練集之 32,884 筆、RefSeq + RDP 訓練集之 23,589 筆，但 SILVA 138 訓練集包含種名的資料量為 111,122 筆，僅占訓練集資料量之 25%，而 GTDB 與 RefSeq + RDP 訓練集內所有的參考序列資料皆含種名，物種分類資料節點分析也顯示，SILVA 138 訓練集包含種名的資料節點為 8,680 項，遠小於 GTDB 與 RefSeq + RDP 訓練集之節點項數 (圖 4)，這也說明了為何 SILVA 138 訓練集在種的分類階層上無反應高解析 ASVs 資料的物種多樣性。最多人使用的 SILVA 資料庫自 2007 年上線，參考序列的資料量早已超過初始建模的 10 倍以上，Glöckner *et al*. (2017) 建議 SILVA 資料庫應導入不能培養的候選分類單元 (candidate taxonomic unit; CTU) 概念，並重新整合細菌與古生菌序列資料，建構新的系統發育進化樹，增加亞種或菌株的分類層級，解決分類錯誤或是不明確的歸類問題 (Glöckner *et al*. 2017)。隨著高解析 ASV 演算法的開發，通用資料庫已無法精準提供種或更低分類階層的環境微生物物種分類，建構棲地特異性 (habitat-specific) 或棲地專用 (habitat-dedicated) 參考序列訓練集，如瘤胃專用的 RIM-DP (Seedorf *et al*. 2014) 與 SILVA 19Rum (Henderson

et al. 2019)、人類腸道專用的 HITdb (Ritari *et al*. 2015)……等，可提升特定棲地微生物屬與種分類階層的分辨率。

結論

次世代定序的普及促進環境基因體研究的蓬勃發展，不斷精進的演算法可極大化的提高 16S rRNA 基因擴增片段之分辨率，但如何連結定序資料與正確的微生物物種分類指派，仍是我們需要努力的課題。本研究結果也揭示了物種分類指派流程與參考序列訓練集的選擇，對物種分類指派結果之影響，隨著 16S rRNA 基因參考序列資料庫不斷地更新，進行高解析的 ASVs 資料集分析時，更應該謹慎選擇反覆評估，才能更準確地描述環境與微生物多樣性間的關係。

引用文獻

Amann, R. and W. Ludwig. 2000. Ribosomal RNA-targeted nucleic acid probes for studies in microbial ecology. FEMS Microbiol. Rev. 24:555–565. doi:10.1111/j.1574-6976.2000.tb00557.x

Amir, A., D. McDonald, J. A. Navas-Molina, E. Kopylova, J. T. Morton, Z. Zech Xu, E. P. Kightley, L. R. Thompson, E. R. Hyde, A. Gonzalez, and R. Knight. 2017. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems 2:e00191-16. doi:10.1128/mSystems.00191-16

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 13:581–583. doi:10.1038/nmeth.3869

Callahan, B. J., P. J. McMurdie, and S. P. Holmes. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 11:2639–2643. doi:10.1038/ismej.2017.119

Callahan, B. J., J. Wong, C. Heiner, S. Oh, C. M. Theriot, A. S. Gulati, S. K McGill, and M. K. Dougherty. 2019. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. Nucleic Acids Res. 47:e103. doi:10.1093/nar/gkz569

Cole, J. R., Q. Wang, J. A. Fish, B. Chai, D. M. Mcgarrell, Y. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske, and J. M. Tiedje. 2014. Ribosomal database project: Data and tools for high throughput rRNA analysis. Nucleic Acids Res. 42:D633–D642. doi:10.1093/nar/gkt1244

Curry, C. J., J. F. Gibson, S. Shokralla, M. Hajibabaei, and D. J. Baird. 2018. Identifying North American freshwater invertebrates using DNA barcodes: Are existing COI sequence libraries fit for purpose? Freshw. Sci. 37:178–189. doi:10.1086/696613

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. 72:5069–5072. doi:10.1128/aem.03006-05

Edgar, R. C. 2016. UNOISE2: Improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv 081257. doi:10.1101/081257

Escobar-Zepeda, A., E. E. Godoy-Lozano, L. Raggi, L. Segovia, E. Merino, R. M. Gutiérrez-Rios, K. Juarez, A. F. Licea-Navarro, L. Pardo-Lopez, and A. Sanchez-Flores. 2018. Analysis of sequencing strategies and tools for taxonomic annotation: Defining standards for progressive metagenomics. Sci. Rep. 8:12034. doi:10.1038/s41598-018-30515-5

Gilbert, J. A., F. Meyer, D. Antonopoulos, P. Balaji, C. T. Brown, C. T. Brown, N. Desai, J. A. Eisen, D. Evers, D. Field, W. Feng, D. Huson, J. Jansson, R. Knight, J. Knight, E. Kolker, K. Konstantindis, J. Kostka, N. Kyrpides, R. Mackelprang, A. McHardy, C. Quince, J. Raes, A. Sczyrba, A. Shade, and R. Stevens. 2010. Meeting report: The terabase metagenomics workshop and the vision of an Earth microbiome project. Stand. Genom. Sci. 3:243–248.

Glöckner, F. O., P. Yilmaz, C. Quast, J. Gerken, A. Beccati, A. Ciuprina, G. Bruns, P. Yarza, J. Peplies, R. Westram, and W. Ludwig. 2017. 25 years of serving the community with ribosomal RNA gene reference databases and tools. J. Biotechnol. 261:169–176. doi:10.1016/j.jbiotec.2017.06.1198

Henderson, G., P. Yilmaz, S. Kumar, R. J. Forster, W. J. Kelly, S. C. Leahy, L. L. Guan, and P. H. Janssen. 2019. Improved taxonomic assignment of rumen bacterial 16S rRNA sequences using a revised SILVA taxonomic framework. PeerJ 7:e6496. doi:10.7717/peerj.6496

McClenaghan, B., N. Fahner, D. Cote, J. Chawarski, A. McCarthy, H. Rajabi, G. Singer, and M. Hajibabaei. 2020. Harnessing the power of eDNA metabarcoding for the detection of deep-sea fishes. PLoS One 15:e0236540. doi:10.1371/journal.pone.0236540

Mizrahi-Man, O., E. R. Davenport, and Y. Gilad. 2013. Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: Evaluation of effective study designs. PLoS One 8:e53608.

doi:10.1371/journal.pone.0053608

Murali, A., A. Bhargava, and E. S. Wright. 2018. IDTAXA: A novel approach for accurate taxonomic classification of microbiome sequences. Microbiome 6:140. doi:10.1186/s40168-018-0521-5

O'Leary, N. A., M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt. 2016. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44:D733–D745. doi:10.1093/nar/gkv1189

Pace, N. R. 1997. A molecular view of microbial diversity and the biosphere. Science 276:734–740. doi:10.1126/science.276.5313.734

Park, S. C. and S. Won. 2018. Evaluation of 16S rRNA databases for taxonomic assignments using mock community. Genomics Inform. 16:e24. doi:10.5808/GI.2018.16.4.e24

Parks, D. H., M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P. A. Chaumeil, and P. Hugenholtz. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. 36:996–1004. doi:10.1038/nbt.4229

Parks, D. H., M. Chuvochina, P. A. Chaumeil, C. Rinke, A. J. Mussig, and P. Hugenholtz. 2020. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat. Biotechnol. 38:1079–1086. doi:10.1038/s41587-020-0501-8

Patel, S. and R. S. Gupta. 2020. A phylogenomic and comparative genomic framework for resolving the polyphyly of the genus *Bacillus*: Proposal for six new genera of *Bacillus* species, *Peribacillus* gen. nov., *Cytobacillus* gen. nov., *Mesobacillus* gen. nov., *Neobacillus* gen. nov., *Metabacillus* gen. nov. and *Alkalihalobacillus* gen. nov. Int. J. Syst. Evol. Microbiol. 70:406–438. doi:10.1099/ijsem.0.003775

Prodan, A., V. Tremaroli, H. Brolin, A. H. Zwinderman, M. Nieuwdorp, and E. Levin. 2020. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. PLoS One 15:e0227434. doi:10.1371/journal.pone.0227434

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner. 2012. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res. 41:D590–D596. doi:10.1093/nar/gks1219

Ritari J., J. Saloärvi, L. Lahti, and W. M. de Vos. 2015. Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. BMC Genom. 16:1056. doi:10.1186/s12864-015-2265-y

Rosen, M. J., B. J. Callahan, D. S. Fisher, and S. P. Holmes. 2012. Denoising PCR-amplified metagenome data. BMC Bioinform. 13:283. doi:10.1186/1471-2105-13-283

Seedorf, H., S. Kittelmann, G. Henderson, and P. H. Janssen. 2014. RIM-DB: A taxonomic framework for community structure analysis of methanogenic archaea from the rumen and other intestinal environments. PeerJ 2:e494. doi:10.7717/peerj.494

Vogel, T. M., P. Simonet, J. K. Jansson, P. R. Hirsch, J. M. Tiedje, J. D. van Elsas, M. J. Bailey, R. Nalin, and L. Philippot. 2009. TerraGenome: A consortium for the sequencing of a soil metagenome. Nat. Rev. Microbiol. 7:252. doi:10.1038/nrmicro2119

Walker, T. S., H. P. Bais, E. Grotewold, and J. M. Vivanco. 2003. Root exudation and rhizosphere biology. Plant Physiol. 132:44–51. doi:10.1104/pp.102.019661

Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. Naïve Bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73:5261–5267. doi:10.1128/AEM.00062-07

Wang, S., X. Bao, K. Feng, Y. Deng, W. Zhou, P. Shao, T. Zheng, F. Yao, S. Yang, S. Liu, R. Shi, Z. Bai, H. Xie, J. Yu, Y. Zhang, Y. Zhang, L. Sha, Q. Song, Y. Liu, J. Zhou, Y. Zhang, H. Li, Q. Wang, X. Han, Y. Zhu, and C. Liang. 2021. Warming-driven migration of core microbiota indicates soil property changes at continental scale. Sci. Bull. 66:2025–2035. doi:10.1016/j.scib.2021.01.021

Zhou, J., Z. He, Y. Yang, Y. Deng, S. G. Tringe, and L. Alvarez-Cohen. 2015. High-throughput metagenomic technologies for complex microbial community analysis: Open and closed formats. mBio. 6:e02288-14. doi:10.1128/mBio.02288-14

Analysis of Taxonomic Annotation Strategies for Soil Microbiota Amplicon Sequencing

Han-Wei Chen[1], Mei-Chun Lin[2], Suh-Jen Lin[3], Ching-Shan Tseng[4], and Yuan-Kai Tu[1,*]

Abstract

Chen, H. W., M. C. Lin, S. J. Lin, C. S. Tseng, and Y. K. Tu. 2022. Analysis of taxonomic annotation strategies for soil microbiota amplicon sequencing. J. Taiwan Agric. Res. 71(3):267–279.

The 16S rRNA gene amplicon sequencing is a high-throughput and gold-standard approach employed in DNA barcoding technique for soil microbial community study. DADA2 implements the divisive amplicon denoising algorithm and produces higher-resolution data sets of amplicon sequence variants (ASVs) for the Illumina sequencing platform. The importance is even greater to link microbial binomial nomenclature and high-resolution ASVs data for subsequent community diversity analysis. In this study, we performed a comparative study of three taxonomic assignment pipelines using DADA2 processed datasets. The efficiency of taxonomic annotation showed that DADA2's assign Taxonomy algorithm goes well with the SILVA 138 reference training set (with Species). Here we used a binary classification test to evaluate the ability of four DADA2-formatted reference training sets (SILVA 138, SILVA 138.1, GTDB, and RefSeq + RDP) in soil microbial classification. The results showed that the GTDB training set had the highest sensitivity, and both SILVA 138 and SILVA 138.1 training sets had the best specificity. While the RefSeq + RDP training set showed the best performing descriptors of accuracy, coverage, Matthews correlation coefficient, and positive predictive value than other training sets. However, the results of microbial diversity analysis showed that the taxonomic assignment of the GTDB training set was the closest to the original ASVs data, reflecting the best soil microbial community compositions. This study revealed that the selection of the taxonomic assignment pipelines and the 16S rDNA reference training set had a great impact on microbial identification. With the continuous updating of the 16S rDNA reference database, we should curate our taxonomic profiling results more carefully to obtain a better microbial diversity description.

Key words: Soil microbiota, Amplicon sequencing, DNA barcoding, DADA2.

Received: November 9, 2021; Accepted: May 30, 2022.
* Corresponding author, e-mail: yktu@tari.gov.tw
[1] Assistant Research Fellows, Biotechnology Division, Taiwan Agricultural Research Institute, Taichung City, Taiwan, ROC.
[2] Project Assistant, Biotechnology Division, Taiwan Agricultural Research Institute, Taichung City, Taiwan, ROC.
[3] Associate Research Fellow, Agricultural Chemistry Division, Taiwan Agricultural Research Institute, Taichung City, Taiwan, ROC.
[4] Associate Research Fellow, Biotechnology Division, Taiwan Agricultural Research Institute, Taichung City, Taiwan, ROC.