细菌高通量数据荟萃分析的可行性研究

肖海钰，李 蕾，王 琳，秦明森，唐 赟*

（西华师范大学 生命科学学院，四川 南充 637000）

摘 要 细菌作为重要的分解者，对于生态系统功能至关重要，其群落多样性一直是生态研究的热点。通过荟萃分析，整理已发表的细菌高通量数据，已经取得一些重要的全球性研究结果。这些荟萃分析的核心是将不同引物来源序列提取统一区段，进行细菌群落分析，但该方法的可行性和准确性还一直未被探究。本研究使用了含有三种引物扩增的同一细菌 16S rRNA 高通量数据，分析了不同引物和提取统一区段后细菌群落情况。结果显示，不同引物以及提取统一区段后，得到细菌群落组成和多样性显著不同，且提取同一区段后的群落主要受原始引物的影响较大。该研究表明，细菌高通量数据荟萃分析中，引物差异会影响其统一分析的群落结果，荟萃方法需要慎重使用。

关键词 高通量测序；荟萃分析；扩增引物；16S rRNA；截取引物

中图分类号 Q939.9 **文献标识码** A **文章编号** 1005－7021(2023)02－0088－08

doi：10.3969/j.issn.1005－7021.2023.02.010

Feasibility of Meta-Analysis on Bacterial High-Throughput Data

XIAO Hai-yu，LI Lei，WANG Lin，QIN Ming-sen，TANG Yun*

（*Coll. of Life Sci.*，*West China Normal Uni*，*Nanchong* 637000）

Abstract Bacteria，as important decomposers，are crucial to ecosystem functions，and their community diversity has always been a hotspot in ecological research. The global research results has achieved some important findings about microbial ecology by sorting out published bacterial high-throughput data adopting meta-analysis. The core of these meta-analyses is to implement bacterial community analyses of unified sections extracted from different primer resource sequences，however the feasibility and accuracy of the approached has not been researched. In this study，an amplified identical bacterial 16S rRNA high-throughput data，containing three different primers，to analyze bacterial communities extracted from different primers and uniform sections. The results showed that different primers as well as after the extraction of the uniform sections，the obtained composition of bacterial population and diversity was significantly different，and the population after the extraction of uniform sections was affected mainly and greatly by the original primers. This study suggested that meta-analysis of bacterial high throughput data，the primer difference will affect the population results，and the assemble method needs to be used prudently.

Keywords high-throughput sequencing；meta-analysis；amplification primers；16S rRNA；interceptive primers

细菌（Bacteria）作为最重要的微生物类群[1]，其分布涵盖地球主要生境，包括各类极端环境，参与各生境中的物质和能量循环[2]，对于维持生态系统的功能和稳定具有极其重要的作用[3-4]。据估算，目前仅约 2% 的细菌被详细调查研究[5]。测序技术的兴起为细菌多样性研究带来了重要变革，第二、三代高通量测序的普及和应用，导致更多的细菌种类被发现，其更多的生

基金项目：四川省科技厅应用基础重点项目（2020YJ0346）；西华师范大学博士科研启动项目（19E048）；四川省自然科学基金项目（2023NSFSC1223）

作者简介：肖海钰 女，硕士研究生。研究方向为微生物生态学。E-mail：2723179260@qq.com

* 通讯作者。男，教授，博士。研究方向为微生物生态学和微生物资源学。E-mail：tangyun_502@cwnu.edu.cn

收稿日期：2022-11-11

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

态功能正不断被揭示[6-7]。随着 NCBI、EBI 等高通量数据库的开放,以及 Muther、Qiime1、Qiime2 等高通量数据分析软件的完善,通过收集已发表的细菌高通量数据,进行整理、再分析,可以打通不同研究的壁垒,扩大了其研究尺度,取得了一些全球性的研究成果[8]。一部分荟萃分析(Meta-analysis)对细菌高通量数据采取了类似的处理策略,例如,Tian 等[9]使用 515F/806R 引物剪切不同引物来源的细菌序列,得到目标 V4 区段,用以研究全球污水处理厂中细菌群落和代谢功能; Xiong 等[10]利用 38 个全球数据集,通过对目标序列进行剪切,提取得到了细菌 16S rRNA 中 V4 区段序列,最终确定了全球不同生境中细菌群落情况; 而 Gwak 等[11]针对主流的细菌高通量注释数据库,使用特定序列将其 16S rRNA 中的 V1_V2 和 V3_V4 区段提取出,通过分析,将不同数据库中细菌种类进行统一注释,为人类肠道细菌群落的探究提供了重要参考。这些荟萃分析的核心是将不同来源的原始序列先进行标准化,即用同一序列(通常为细菌特定引物,可以理解为伪"引物")对序列进行剪切,提取得到统一区段序列之后进行再分析[12-13]。然而,这种剪切得到的目标区段(可称"伪"区段)与"真"引物扩增得到的区段是否存在差异,这一核心问题在上述研究中均未被考虑过。避免讨论该问题,将忽略不同引物所造成的测序影响,从而影响以上研究结果的准确性。为此,本研究使用了 78 个包含 V3_V4、V4 以及 V4_V5 区段的 16S rRNA 高通量测序数据[14],并使用 Qiime2 平台比较了 V3_V4 和 V4_V5 串联区截取得到的"伪"V4 区段与"真"V4 区段(即原生 V4 扩增片段)所得到的细菌群落,来探究提取细菌统一扩增区段的荟萃分析方法的可行性和准确性,该研究能为今后类似分析方法提供重要的参考。

1　材料与方法

1.1　材料

使用 Wear 等[14]测定的 83 个海水样品中的细菌 16S rRNA 数据,该数据已上传至 NCBI,接收号为 PRJNA412105。该数据集包含四对不同引物的测序结果,选取了其中 V3_V4、V4 以及 V4_V5 区段的高通量数据进行分析,对应的测序引物信息如表 1 所示。为排除较低测序对后续结果的影响,将包含任意测序深度低于 5 000 reads 的样品剔除,共得到 234 份高通量数据。

表 1　扩增 16S rRNA 不同区段的引物信息

Table 1　Primers used to amply different region of 16S rRNA

目标区段	引物名称	引物序列(5′-3′)	扩增长度/bp
V3_V4	341F	CCTACGGGNGGCWGCAG	460
	785R	GACTACHVGGGTATCTAATCC	
V4	515F	GTGYCAGCMGCCGCGGTAA	292
	806R	GGACTACNVGGGTWTCTAAT	
V4_V5	515F	GTGYCAGCMGCCGCGGTAA	411
	926R	CCGYCAATTYMTTTRAGTTT	

1.2　方法

所有高通量分析使用 Qiime2(2021.04)平台进行[15],主要步骤: 第一,训练分类集(train feature classifier)。使用 Silva 138(2020.11)16S rRNA全长数据库作为参考数据库[16]。从全长数据库中截取目标区段,提高物种分类精度,主要使用 RESCRIPt[17]剔除原本重复序列和截取引物后目标序列的重复序列; 使用 feature-classifier extract-reads 命令分别截取对应引物序列,设置 chunk-size = 1 000 的运算通量来训练细菌分类数据集。第二,合并双端测序结果。使用 Vsearch 2.17.1[18],按照-fastq_maxdiffs 30-fastq_minovlen 20-fastq_maxee 0.5 的参数进行合并。先合并序列能避免在下一步引物剪切后,各双端剩余序列过短,影响序列合并。第三,剪切前段引物,以保留 V4 区段,该步骤主要对 V3_V4 区段使用 515F 引物进行剪切,保留后端序列,即为"伪"V4 区段。使用 cutadapt 剪切对应引物。第四,细菌群落分析。该分析主要分为三部分: ①单独分析各扩增区段,即对 V3_V4 区段、V4 区段以及 V4_V5 分别截取至 420 bp、260 bp 以及 360 bp 长度后进行单独分析。②单独分析 V3_V4 以及V4_V5区段截取后的 V4 区段,包括 V3_V4 区段剪切后的"伪"V4 区段(剪切至 260 bp),以及截取至 260 bp 的 V4_V5 区段(剩余序列即为 V4 区段)。③将所有 V4 区段整合至一起进行分析,即将所有 V4 区段序列合并后,截取至 260 bp 进行分析。序列使用 DADA2 进行嵌合体筛选和划分扩增子序列变异(Amplicon Sequence Variants,

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

ASV，可简单理解为 100% 相似度进行聚类)[19]，ASV 使用 feature-classifier classify-sklearn 命令，结合对应区段的训练分类集进行细菌 ASV 物种鉴定，其中置信度最低为 0.8。最后，将所有细菌矩阵 reads 数抽平至 1 849，以避免不同测序深度对结果的影响[20]。

1.3 数据分析

细菌群落多样性、群落组成等均在 R 语言平台分析完成[21]，主要使用“vegan”程序包[22]。多样性差异使用配对的 t 检验进行，采用最小二乘法(OLS)拟合多样性和相对丰度之间的线性关系，使用 Mantel test 进行群落相似性分析。

2 结果与分析

2.1 不同扩增片段下细菌群落差异

同样测序深度下，不同扩增区段得到的细菌各纲相对丰度存在明显差异(图 1A)，得到的多样性也有显著不同(图 1B)。其中，扩增 V4 区段得到的细菌丰富度(Richness)最高，V3_V4 区段次之。群落均匀度(Evenness)也有显著差异，扩增 V3_V4 区段的均匀度最高，V4 区段次之(图 1C)。三种扩增区段下，细菌丰富度相关性显著(表 2)，其中 V3_V4 与 V4_V5 扩增下，细菌丰富度相关程度最高(表 2)。就均匀度来看，V4 与 V4_V5 扩增下细菌群落均匀度相关性最高(表 2)。通过对群落相似性距离的相关性分析发现，三种扩增区段所构建的群落存在着一定程度的相似，其中扩增 V4 与 V4_V5 之间相似程度最高(表 3)。进一步对丰度最高的三种细菌纲之间的相关性分析发现，V4 和 V4_V5 扩增得到的三种细菌纲相对丰度的相关性最高(表 4)。

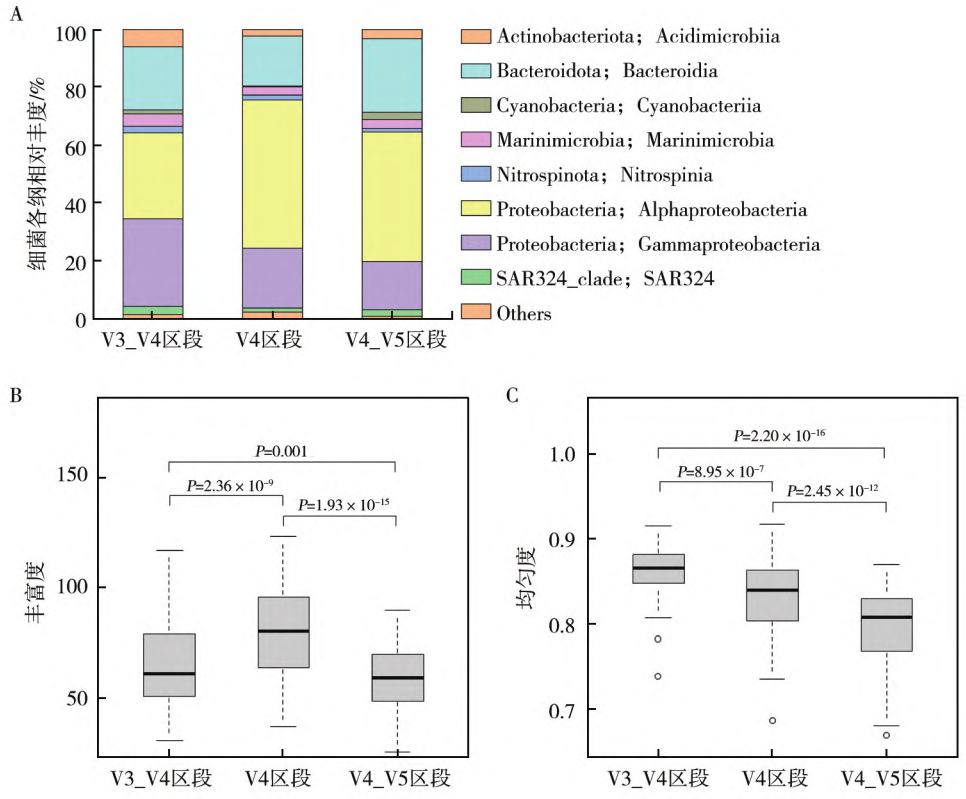


图 1 不同引物下细菌组成和多样性的差异

Fig. 1 The composition and diversity of bacterial communities of different primers

A: 主要细菌纲; B: 丰富度; C: 均匀度。下同

A: The main bacterial class; B: Richness; C: Evenness. The same for below

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 2　不同引物或剪切区段细菌多样性相关性

Table 2　The linear relationships of bacterial diversity among different primers and the cutoff regions

比较区段	目标区段	丰富度		均匀度	
		r	*P*	*r*	*P*
原始区段	V3_V4 vs V4	0.428	<0.001	0.217	0.032
	V3_V4 vs V4_V5	0.531	<0.001	0.000	0.482
	V4 vs V4_V5	0.395	<0.001	0.720	<0.001
原始区段与 V4 区段	V3_V4 vs V4(V3_V4)	0.874	<0.001	0.303	0.004
	V4_V5 vs V4(V4_V5)	0.933	<0.001	0.939	<0.001
所有 V4 区段	V4(V3_V4) vs V4	0.362	0.001	0.000	0.371
	V4(V3_V4) vs V4(V4_V5)	0.626	<0.001	0.486	<0.001
	V4 vs V4(V4_V5)	0.421	<0.001	0.713	<0.001

注: V4 后面括号代表原引物扩增区段,下同

表 3　不同引物和剪切区段细菌群落组成的相关性

Table 3　Mantel tests for relationship of the bacterial community among different primers and cutoff regions

比较区段	目标区段	群落相似性	
		Mantel *r*	*P*
原始区段	V3_V4 vs V4	0.873	<0.001
	V3_V4 vs V4_V5	0.907	<0.001
	V4 vs V4_V5	0.965	<0.001
原始区段与 V4 区段	V3_V4 vs V4(V3_V4)	0.957	<0.001
	V4_V5 vs V4(V4_V5)	0.984	<0.001
	V4(V3_V4) vs V4	0.929	<0.001
所有 V4 区段	V4(V3_V4) vs V4(V4_V5)	0.962	<0.001
	V4 vs V4(V4_V5)	0.973	<0.001

表 4　不同引物和剪切区段主要细菌纲丰度的相关性

Table 4　The linear relationships of the main bacterial class abundance among different primers and cutoff regions

比较内容	目标区段	Alphaproteobacteria		Bacteroidia		Gammaproteobacteria	
		r	*P*	*r*	*P*	*r*	*P*
原始区段	V3_V4 vs V4	0.781	<0.001	0.946	<0.001	0.841	<0.001
	V3_V4 vs V4_V5	0.820	<0.001	0.943	<0.001	0.856	<0.001
	V4 vs V4_V5	0.947	<0.001	0.964	<0.001	0.918	<0.001
原始与 V4 区段	V3_V4 vs V4 (V3_V4)	0.958	<0.001	0.991	<0.001	0.980	<0.001
	V4_V5 vs V4 (V4_V5)	0.995	<0.001	0.995	<0.001	0.992	<0.001
所有 V4 区段	V4 (V3_V4) vs V4	0.703	<0.001	0.959	<0.001	0.877	<0.001
	V4 (V3_V4) vs V4 (V4_V5)	0.772	<0.001	0.955	<0.001	0.872	<0.001
	V4 vs V4 (V4_V5)	0.942	<0.001	0.969	<0.001	0.928	<0.001

2.2　同种引物下不同区段的细菌群落差异

通过对 V3_V4 以及 V4_V5 区段进行剪切,得到“伪”V4 区段,分析发现,“伪”V4 区段主要细菌纲的组成与原始区段差别不大(图 2A) ,但物种多样性差异明显(图 2B、2C) 。剪切后的“伪”V4 区段丰富度高于剪切前区段(图 2B) ,所得的均匀度低于剪切前区段(图 2C) 。剪切前后的区段所得的细菌多样性之间相关性显著,但 V3_V4区段剪切前后均匀度相关程度相对较低(表 2) 。剪切前后细菌群落之间具有极为显著的相关性,其中 V4_V5 区段剪切前后群落相似性更高(表 3) ,且主要的三种细菌纲相对丰度之间相关性也高于 V3_V4 剪切前后(表 4) 。

2.3　同一区段下不同引物细菌群落差异

将来自三种不同引物的 V4 区段合并分析发现,细菌纲的相对丰度存在明显差异(图 3A) ,群

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

落多样性也有显著不同，其中 V3_V4 截取后的 V4 区段细菌群落丰富度最高（图 3B），扩增 V4 区段所得的均匀度最高（图 3C）。V3_V4 截取后的 V4 区段与 V4_V5 截取后的 V4 区段细菌群落丰富度相关性最高；V4 区段与 V4_V5 截取后的 V4 区段均匀度相关性最高；而 V3_V4 截取后的 V4 区段与 V4 区段均匀度并无明显相关关系（表 2）。不同扩增序列 V4 区段下细菌群落间具有极为显著的相关性，其中来自 V4 与 V4_V5 的 V4 区段群落相似程度最高（表 3），该结果与群落排序结果一致（图 4），且来自 V4 与 V4_V5 的 V4 区段细菌纲相对丰度相关性最高（表 4）。

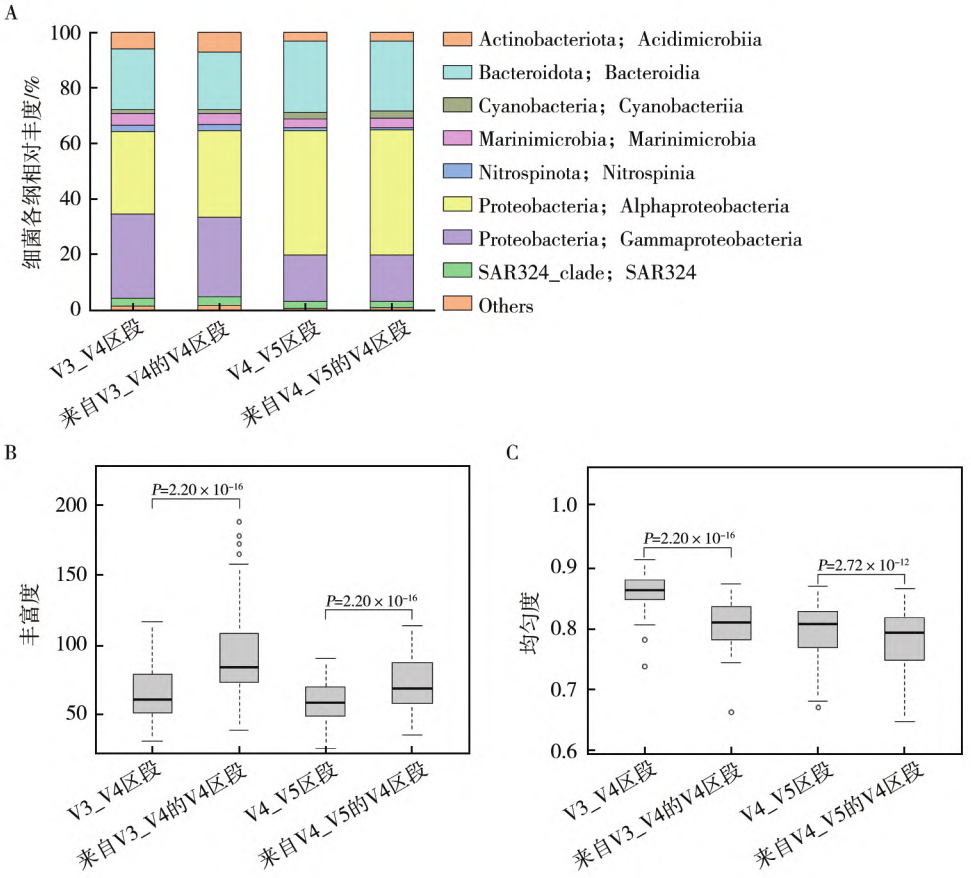
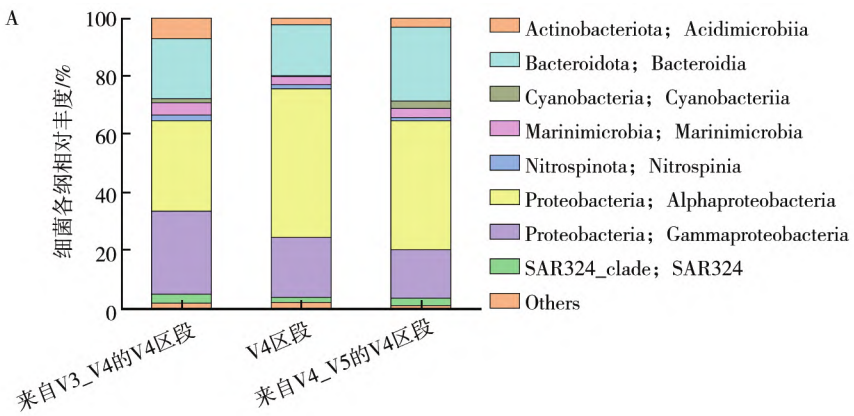


图 2 剪切至 V4 区段后细菌组成和多样性的差异

Fig. 2 The composition and diversity of bacterial communities of V4 region and corresponding amplified region



(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

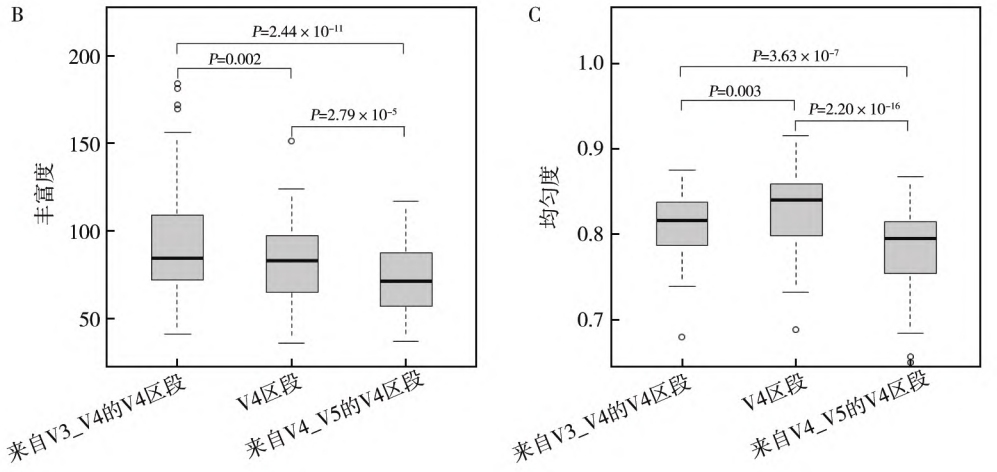


图 3　不同引物的 V4 区段细菌组成和多样性的差异

Fig. 3　The composition and diversity of bacterial communities of V4 region from different primers

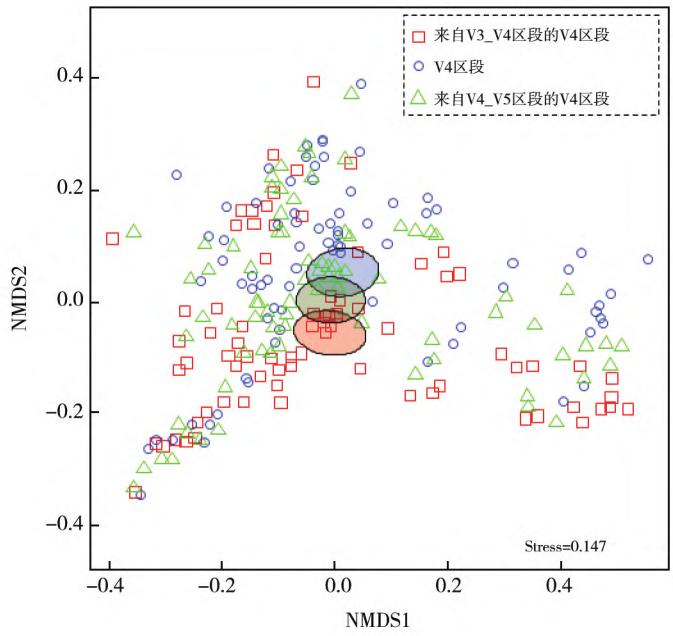


图 4　不同引物的 V4 区段细菌群落 NMDS 排序

Fig. 4　Nonmetric multidimensional scaling analysis (NMDS) of bacterial communities of V4 regions from different primers

椭圆圈代表对应颜色细菌群落的 95% 置信区间

Ellipses with different colors indicate 95% confidence of corresponding bacterial community

3　讨　论

16S rRNA 基因的可变区具有属或种的特异性[23]，通过单一可变区或可变区串联测序可以确定微生物物种组成。然而，研究发现不同可变区或串联组合测序所得到的结果均相同[24]。本研究中，不同引物扩增不同区段下，在纲水平上其群落组成具有较大差异，一方面有可能来自于不同可变区的鉴定差异，另一方面，不同细菌可变区的引物有不同的扩增偏差[25]。本研究通过对不同扩增引物进行剪切，得到统一的 V4 区段数据分析结果来看，同种引物下，剪切前后的群落主要细菌纲相对丰度几乎相同。该结果表明不同区段引物扩增导致的群落差异并非由不同可变区对细菌

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

种鉴定差异引起，而是由引物特异性导致。另外，Marcus 等[26]研究表明 V3_V4、V4_V5 这两个可变区串联组合对细菌的鉴定准确性较高，而本研究中 V4 区段引物扩增后细菌丰富度最高，表明即使 V4 区段引物 515F/806R 受到地球微生物组计划(EMP)的推荐[27]，但该引物有可能会高估细菌多样性，并损失细菌物种鉴定的准确性。

本研究使用了相似度阈值在 100% 的 ASV 进行代表序列划分，该分析中单个碱基的差异将导致划分出新的 ASV。基于该分析方法，理论上序列长度越长，变异的 ASV 也会越多。本研究 V3_V4和 V4_V5 剪切至 V4 区段后，部分前段(V3_V4)或后段(V4_V5)的可变区序列被移除，理论上，序列变短将导致检测到的 ASV 数量下降，而实际结果与此相反，这与理论结果不相符。这是因为 DADA2 会将丰度极低的物种当做测序错误而剔除[19]，部分 V3_V4 前段和 V4_V5 后段涵盖不同碱基的低丰度序列被剪切后，剩余 V4 区段的低丰度序列将被被识别为“较高”丰度序列，从而纳入分析，最终提高了群落的丰度。剪切后，群落更低的均匀度也能佐证部分低丰度的序列被纳入最终群落分析。

合并“伪”V4 和“真”V4 区段序列统一分析发现，不同引物的同段序列群落组成和多样性具有显著差异，该结果表明通过截取同一段序列进行荟萃分析的方法受原始引物的影响比较大，不能进行整合分析。但考虑到群落组成的显著相关性，该方法对于群落组成分析具有一定的可行性。本研究主要对细菌纲进行了分析，而纲的分类鉴定来自于更细的种属分类结果，细菌纲的差异能够体现其种属注释结果的差异巨大。总之，本研究发现不同引物对于群落结构和多样性的调查结果均会产生显著影响，采用序列截取的方法，得到同一区段的荟萃分析，也会引入引物差异，其结果值得重新分析考虑[9-11]，部分真菌[28-29]、原生生物[30]等其他微生物高通量数据的荟萃分析结果是否会引入引物误差，还需要更多的研究来探讨。

参考文献:

[1] Bar-On YM, Phillips R, Milo R. The biomass distribution on Earth [J]. Proceedings of the National Academy of Sciences of the United States of America, 2018, 115(25): 6506-6511.

[2] Flemming HC, Wuertz S. Bacteria and archaea on Earth and their abundance in biofilms [J]. Nature Reviews Micobiology, 2019, 17(4): 247-260.

[3] Bardgett RD, Putten WH, Van Der. Belowground biodiversity and ecosystem functioning [J]. Nature, 2014, 515(515): 505-511.

[4] Crowther TW, van den Hoogen J, Wan J, et al. The global soil community and its influence on biogeochemistry [J]. Science, 2019, 365(6455): eaav0550.

[5] Delgado-Baquerizo M, Oliverio AM, Brewer TE, et al. A global atlas of the dominant bacteria found in soil [J]. Science, 2018, 359(6373): 320-325.

[6] Bahram M, Espenberg M, Parn J, et al. Structure and function of the soil microbiome underlying N_2O emissions from global wetlands [J]. Nature Communications, 2022, 13(1): 1430.

[7] Ziegler M, Grupstra CGB, Barreto MM, et al. Coral bacterial community structure responds to environmental change in a host-specific manner [J]. Nature Communications, 2019, 10(1): 3092.

[8] Yuan J, Wen T, Zhang H, et al. Predicting disease occurrence with high accuracy based on soil macroecological patterns of Fusarium wilt [J]. The ISME Journal, 2020, 14(12): 2936-2950.

[9] Tian L, Wang L. A meta-analysis of microbial community structures and associated metabolic potential of municipal wastewater treatment plants in global scope [J]. Environmental Pollution, 2020, 263: 114598.

[10] Xiong W, Jousset A, Li R, et al. A global overview of the trophic structure within microbiomes across ecosystems [J]. Environment International, 2021, 151: 106438.

[11] Gwak H-J, Rho M. Data-driven modeling for species-level taxonomic assignment from 16S rRNA: Application to Human Microbiomes [J]. Frontiers in Microbiology, 2020, 11: 570825.

[12] Bickel S, Chen X, Papritz A, et al. A hierarchy of environmental covariates control the global biogeography of soil bacterial richness [J]. Scientific Reports, 2019, 9(1): 12129.

[13] Bickel S, Or D. The chosen few-variations in common and rare soil bacteria across biomes [J]. The ISME Journal, 2021, 15(11): 3315-3325.

[14] Wear EK, Wilbanks EG, Nelson CE, et al. Primer selection impacts specific population abundances but not community dynamics in a monthly time-series 16S rRNA gene amplicon analysis of coastal marine bacterioplankton [J]. Environmental Microbiology, 2018, 20(8): 2709-2726.

[15] Bolyen E, Rideout JR, Dillon MR, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2 [J]. Nature Biotechnology, 2019, 37(8): 852-857.

[16] Quast C, Pruesse E, Yilmaz P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools [J]. Nucleic Acids Research, 2012, 41(D1): 590-596.

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[17] Robeson MS, O´Rourke DR, Kaehler BD, et al. RESCRIPt: Reproducible sequence taxonomy reference database management [J]. PLoS Computational Biology, 2021, 17(11): e1009581.

[18] Rognes T, Flouri T, Nichols B, et al. VSEARCH: A versatile open source tool for metagenomics [J]. Peer J, 2016, 4: e2584.

[19] Callahan BJ, McMurdie PJ, Rosen MJ, et al. DADA2: high-resolution sample inference from Illumina amplicon data [J]. Nature Methods, 2016, 13(7): 581-583.

[20] Zhu D, Delgado-Baquerizo M, Ding J, et al. Trophic level drives the host microbiome of soil invertebrates at a continental scale [J]. Microbiome, 2021, 9(1): 189.

[21] R Core Team. R: A language and environment for statistical computing [M]. Vienna, Austria: R Foundation for Statistical Computing, 2020.

[22] 薛毅, 陈立萍. 统计建模与 R 软件 [M]. 北京: 清华大学出版社, 2007: 546-555.

[23] 朱诗应, 戚中田. 16S rDNA 扩增及测序在细菌鉴定与分类中的应用 [J]. 微生物与感染, 2013, 8(2): 6.

[24] 刘赛. 基于 16S rRNA 序列可变区的微生物物种鉴定及分析 [D]. 西安: 西安电子科技大学, 2018.

[25] 刘瑛, 张琼琼, 张蕾, 等. 不同 16S rDNA 高变区选择对细菌性阴道病菌群研究差异比较 [J]. 现代妇产科进展, 2019, 28(11): 4.

[26] Claesson M, Wang Q, O´Sullivan O, et al. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions [J]. Nucleic Acids Research, 2010, 38(22): e200.

[27] Thompson LR, Sanders JG, McDonald D, et al. A communal catalogue reveals Earth´s multiscale microbial diversity [J]. Nature, 2017, 551(7681): 457-463.

[28] Vetrovsky T, Kohout P, Kopecky M, et al. A meta-analysis of global fungal distribution reveals climate-driven patterns [J]. Nature Communications, 2019, 10(1): 5142.

[29] Meiser A, Bálint M, Schmitt I. Meta - analysis of deep-sequenced fungal communities indicates limited taxon sharing between studies and the presence of biogeographic patterns [J]. The New Phytologist, 2014, 201(2): 623-635.

[30] Singer D, Seppey CVW, Lentendu G, et al. Protist taxonomic and functional diversity in soil, freshwater and marine ecosystems [J]. Environment International, 2021, 146: 106262.

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net