宏基因组学技术与微生物群落多样性分析方法

彭玺[1,2],冯凯[1],厉舒祯[3],邓晔[1,2*]

1. 中国科学院生态环境研究中心,中国科学院环境生物技术重点实验室,北京 100085
2. 中国科学院大学资源与环境学院,北京 100049
3. 大连理工大学环境学院,工业生态与环境工程教育部重点实验室,大连 116024

摘要　从宏基因组学的2个主要技术——扩增子测序与宏基因组测序出发,对其在微生物群落检测中的基本分析流程进行了介绍。概述了微生物群落多样性的概念、相关的统计分析原理以及相关分析结果的解析方法等。提出利用正处于蓬勃发展时期的大数据分析技术与手段来克服宏基因组学数据解析,并将分析结果用更易理解的形式展现出来是未来研究的重点和难点,也是从事环境微生物学、生物信息和统计学研究人员共同的挑战。

关键词　环境微生物;宏基因组学;扩增子;生物信息学

微生物在地球上分布广泛,在种类、数量和功能上千差万别。自然界中大量的微生物无法利用纯培养手段在实验室条件下进行培养,传统微生物学的技术手段限制了环境微生物的研究[1-2]。免培养的组学技术的出现实现了从分子水平(DNA、RNA、蛋白质和代谢产物)对环境微生物及其功能进行检测与分析,为研究者了解完整的环境微生物全貌提供了有效的途径。而微生物生态学作为近年来发展最快的交叉学科之一,其研究理论和方法广泛服务于环境科学、环境工程、土壤学、湖沼学、农学、医学、公共卫生学等多个领域,在环境污染的调查与治理、农业生产的管理与指导、食品工业的检测与加工、生物能源的生产与应用、人类疾病的发生与预防、动物的共生与进化、生态系统功能的

收稿日期:2021-06-21;修回日期:2021-12-17
基金项目:国家自然科学基金山东联合基金重点项目(U1906223)
作者简介:彭玺,博士研究生,研究方向为生物信息学、微生物生态学,电子信箱:xipeng_st@rcees.ac.cn;邓晔(通信作者),研究员,研究方向为环境微生物生态,电子信箱:yedeng@rcees.ac.cn
引用格式:彭玺, 冯凯, 厉舒祯, 等. 宏基因组学技术与微生物群落多样性分析方法[J]. 科技导报, 2022, 40(3): 99-111; doi: 10.3981/j.issn.1000-7857.2022.03.009

发掘与保护等多个方面发挥着越来越重要的作用。

1 宏基因组学技术简介

高通量组学技术的迅速发展使得人类对各类生态系统中的复杂微生物群落有了前所未有的认知。微生物组学通常是宏基因组、宏转录组、宏蛋白组、宏代谢组等各类系统生物学技术和方法的总称,它们不同于以往仅研究少数几个基因、蛋白质或生化通路的分子生物学方法,而是注重研究生物系统组成及群落中物种之间的相互关系、系统结构和功能的关联、以及群落结构与生态系统的关联等整体上的科学问题。其中,以高通量测序技术为基础的宏基因组学是目前最为关键和成熟的组学方法,它也为其他组学的研究提供了研究基础。目前关于宏转录组学、宏蛋白组学和宏代谢组学的研究虽然仍处于起步阶段,但它们却显示出了巨大的发展前景[3]。

宏基因组(metagenome)概念由Handelsman等[4]于1998年首次提出,其定义为某特定环境中所有微生物的基因组的总和。作为环境微生物学的重要研究手段,宏基因组学研究无需对环境微生物进行分离培养,而是直接分析环境中微生物的DNA来获知微生物群落的遗传、功能与生态特性。目前的宏基因组研究紧密依赖高通量测序技术,包括扩增子测序(amplicon sequencing)与宏基因组测序(metagenome sequencing)。扩增子测序主要针对核糖体RNA基因(ribosomal RNA gene, rDNA)和功能基因,前者对细菌或古菌16S rDNA及真菌18S rDNA与内部转录间隔区(internal transcribed spacer, ITS)序列等分子标记进行扩增,后者对于微生物某些特定功能基因(如参与碳、氮循环的功能基因)进行扩增。宏基因组测序则是对环境中所有DNA进行测序。基因组测序成本较高,且对于后续数据分析的计算资源要求也相对较高。相比之下,扩增子技术凭借测序与分析成本较低的优点目前成为了环境微生物组学研究的主要手段。

近年来,随着三代测序技术的兴起,宏基因组的研究手段被进一步强化。在三代测序技术流行以前,二代测序技术是宏基因组学研究的主要测序方法。以典型的二代测序平台Illumina Hiseq为例,其测序原理都是基于桥式聚合酶链式反应(PCR)的边合成边测序[5]。它的特点是读长短(通常在100 bp至300 bp之间)、准确性高,和一代测序(Frederick Sanger的DNA双脱氧链终止法)相比具有测序速度快、准确性高以及成本低的优良特性。三代测序技术均采用了单分子测序原理,最大的特点是DNA样本无需经过PCR扩增,避免了PCR反应对于DNA样本本身的偏好性影响。目前最典型的2个三代测序平台分别是Pacific Biosciences公司推出的单分子实时测序(Single Molecule Real Time Sequencing, SMRT)[6]和Oxford Nanopore Technologies公司开发的纳米孔单分子测序技术[7],前者基于光信号判断碱基类型,后者则根据电信号来判断碱基类型。三代测序在对二代测序的一些缺点进行改进的同时,也有测序错误率相对较高、通量较低且成本贵等问题。除了测序本身的特性外,二代与三代测序技术对于宏基因组数据后续的组装也会带来影响。

高通量组学技术的出现掀起了一场环境微生物领域的革命,同时能够帮助我们进一步了解微生物群落的遗传潜力和功能活动规律。而高度复杂的微生物群落组成和庞大的数据量,使信息分析从组学技术诞生的那一刻起就成为了它在应用上的瓶颈[3]。针对这一问题,目前已有一些环境微生物宏基因组分析工具及群落多样性的统计分析方法被应用到相关研究中,本文将对其进行重点介绍。

2 宏基因组学测序分析流程

宏基因组学测序得到的原始数据常通过碱基形式存储在形如FASTQ格式文件中,需要通过一系列的生物信息学算法与软件的处理才能将其转化为可以被直接进行分析的形式。目前宏基因组学测序的2种主要测序技术——扩增子测序与宏基因组测序由于其得到的数据类型、分析目的常常不同,因此对这2种数据的分析也采用了不同的方法。

2.1　扩增子测序分析流程

基于核糖体rDNA高度可变区序列的高通量扩增子测序对微生物多样性以及群落组成差异的研究方法被广泛应用后,而一系列生物信息学分析流程也应运而生[8]。环境微生物群落研究的随机取样导致其扩增子分析可重复性(reproducibility)低,但通过增加生物学重复、删除单一样品仅出现一次的序列等手段可以较好地加以弥补[9]。为了从定量和定性2个角度表征生物多样性,研究者定义了多项指数对生物多样性进行量化,同时也为不同生物多样性的比较提供了方法。针对群落结构分析,相当多的数理统计学,尤其是多元统计分析方法被广泛应用于生态学分析。在微生物生态学兴起后,不少宏观生态学的研究方法和手段也被逐渐应用于微生物生态学的研究中,为其提供新的研究思路。

微生物扩增子测序的分析方法多样,分析流程也不尽相同,目前常用的扩增子分析主要步骤与方法如图1所示。随着测序技术的更迭,Illumina测序平台逐渐占据着微生物扩增子测序的大部分市场,目前以双端250 bp测序策略居多[10]。该方案建库测序前,对不同样品分别使用含不同条形码(barcode)标签的引物对目的片段进行PCR扩增,原始下机数据需要去除条形码以及引物序列。针对低质量序列和模糊碱基,对数据进行质量控制。之后,通过FLASH[11]等软件对双端序列进行拼接。拼接完成的序列可以作为挑选代表序列的起始文件。目前OTU(operational taxonomic unit,可操作分类单元)与ASV(amplicon sequence variant,扩增子序列变异)是2种主要的代表序列形式。目前主流的挑选代表序列的工具有UPARSE、DADA2、UNOISE3与Deblur等,其主要算法与原理及特点如表1所示。

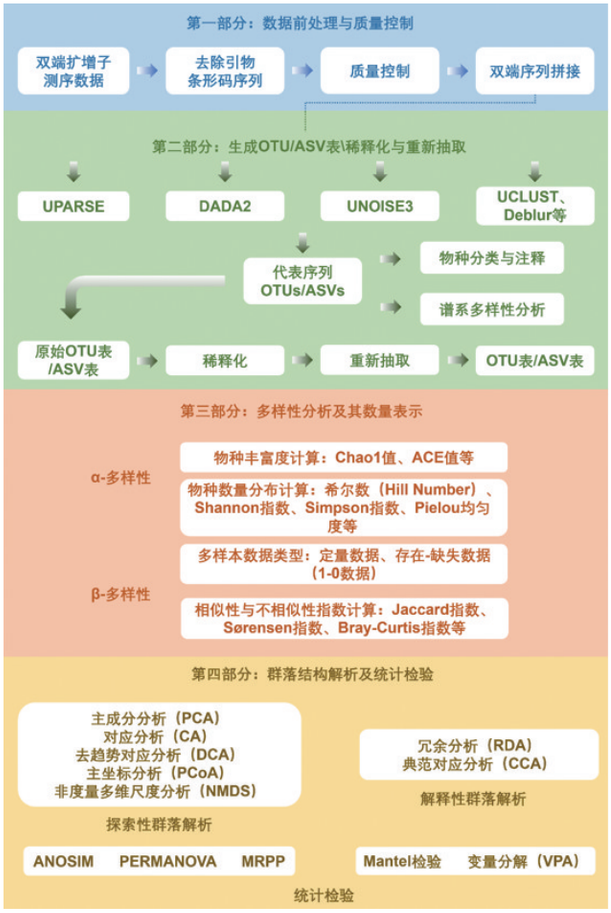

图1　以16S rDNA测序为例,扩增子测序分析的主要方法和流程

表1　主流的挑选代表序列的工具

工具名称	主要算法与原理	文献
UPARSE	使用贪心算法(Greedy Algorithm)对去嵌合体后序列进行聚类,常使用97%相似度,按照序列质心与丰度规则划分OTU	[12]
DADA2	通过机器学习方法对序列是真实变异或是测序错误进行判断并进行相应的校正。其得到的代表序列称为ASV(Amplicon Sequence Variant)	[13]
UNOISE3	改进的降噪纠错算法。其得到的代表序列称为zOTU(zero-radius OTU)	[14]
Deblur	基于由汉明距离(Hamming Distance)计算概率在单核苷酸精度水平上对序列进行降噪纠错	[15]

2.2 宏基因组测序分析流程

近年来,随着测序技术在通量和读长方面的持续提高,其成本也不断降低,因此针对微生物群落的全部基因组DNA的鸟枪测序也在不断增加。除此之外,三代测序技术也正逐渐在微生物生态领域普及。然而,三代测序在对二代测序的一些缺点进行改进的同时,也有测序错误率相对较高、通量较低且成本高等问题。三代测序技术得到的原始数据往往是光信号或电信号信息,常需要碱基判定(base-calling)将其转化为我们所需要的FASTQ格式文件。

虽然二代与三代测序原理不同,但其后续的数据处理策略却很相似。宏基因组测序数据的数据量大,相比于扩增子分析来说需要更加专门的算法与软件来处理与分析。宏基因组学的分析常有一套通用的流程(图2)。宏基因组数据产出后,首先需要进行质量控制与去除宿主污染,通过去除低质量序列、引物、adaptor和宿主序列,输出高质量序列。至此,对于宏基因组的数据分析可分为2个水平,即基于reads的分析与基于组装的分析。前者直接利用reads挖掘其中的信息,其中部分分析方法与扩增子分析类似,如从宏基因组测序reads中提取mOTUs(phylogenetic marker gene-based operational taxonomic units)的分析[16];而后者则需要通过后续步骤将reads拼接为更长的结构进行分析,其中,组装是其中的关键一环。序列组装是将短的高质量序列拼接为更长的序列,即重叠群(contig)或基因组支架(scaffold)。根据重叠群信息,在去除冗余基因后,可以用于基因预测生成基因丰度表。再后,对于宏基因组数据分析,分箱(binning)是重要的工具之一,分箱根据不同序列的四核苷酸频率模式或丰度频率规律将不同序列进行分类,其得到的是高质量的基因组草图。最后,根据不同处理的宏基因组数据可以进行组成分析、差异分析或根据研究目的设计的个性化分析。

虽然宏基因组测序分析的流程类似,但由于其测序数据量大,目前缺乏标准的分析工具进行统一化处理,不同的分析工具和方法在性能和速度方面差异较大,尤其是不同类型的微生物组数据往往还需要个性化的调整。随着三代测序技术的普及,国内外现有针对宏基因组数据分析各个步骤设计的软件也都处于飞速发展中,常用软件如表2所示。

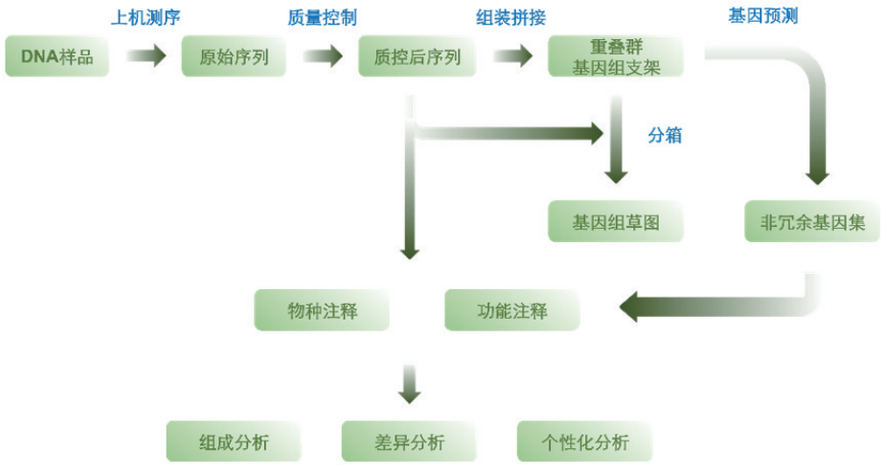

图2 宏基因组测序分析的常用分析流程

表2　宏基因组分析流程常用软件

软件名称		最新版本	是否支持二代测序	是否支持三代测序	文献
数据质控	FastQC	0.11.9(2019.1.8)	是	支持PacBio	—
	MultiQC	1.10.1(2021.4.1)	是	支持模块处理ONT	[17]
	fastp	0.20.1(2020.4.8)	是	—	[18]
	Trimmomatic	0.39	是	—	[19]
	Nanoplot	1.33.0(2020.9.12)	—	支持长读长	[20]
序列组装	MEGAHIT	1.2.9(2019.10.15)	二代短读长		[21]、[22]
	metaSPAdes (SPAdes)	3.15.2(2021.3.11)	二代短读长,支持二三代混合组装		[23]、[24]
	IDBA-UD	1.1.3(2016.7.12)	二代短读长		[25]
	SOAPdenovo2	r242(2020.10.23)	二代短读长		[26]、[27]
	Canu	2.1.1(2020.10.19)	单分子测序(PacBio RS II/Sequel or Oxford Nanopore MinION)		[28]
	OPERA-MS	0.8.3(2020.6.30)	二三代混合组装		[29]
	metaFlye (Flye)	2.8.3(2021.2.11)	三代长读长		[30]
分箱	metaBAT2	2.12.1(2017.9.1)	基于四核苷酸频率与丰度得分分箱		[31]
	MaxBin2	2.2.7(2020.6.12)	—		[32]、[33]
	CONCOCT	1.1.0(2019.8.2)	基于序列组成与丰度分箱		[34]
	VAMB	3.0.2(2020.10.27)	基于深度变分自编码器改进的分箱		[35]
	metaWRAP	1.3(2020.8.5)	综合了质控、组装与多种分箱工具与模块的宏基因组数据分析流程		[36]
预测与注释	DIAMOND	2.0.9(2021.4.12)	蛋白序列比对软件		[37]
	Kraken2	2.1.2(2021.5.10)	宏基因组物种注释工具		[38]
	Kaiju	1.7.4(2020.11.4)	基于Read水平的物种注释软件		[39]
	MetaPhlan	3.0.9(2021.5.17)	基于Read水平的物种注释软件		[40]
	Prodigal	2.6.3(2016.2.12)	按照基因编码区结构的基因预测软件		[41]

注:二代短读长与三代长读长具体指二代测序技术生成的长度约为200 bp至300 bp的短读与三代测序技术生成的长度能达到几千甚至上万bp的长读。

3　微生物群落多样性分析方法

无论是使用扩增子测序还是宏基因组测序,后续都需要大量的群落生态学方法,针对环境微生物群落生物多样性和群落组成与结构,进行深入分析和统计。

生物多样性(biodiversity)被提出后,其定义层出不穷。其中,国际生物多样性公约(The International Convention on Biological Diversity)对其定义为“来源于包含陆地、海洋与其他水生生态系统以及它们组成的复合生态系统中的生物的可变性,其中包括物种内、物种间以及生态系统的多样性”[42]。微生物生态学中的微生物多样性按照描述物种的尺度进行层级划分,通常主要有分类多样性(taxonomic diversity)、谱系多样性(phylogenetic diversity)、遗传多样性(genetic diversity)和功能多样性(functional diversity)[43]。其中分类多样性与功能多样性常通过分析分类单元、功能基因或通路在不同环境下的分布情况进行衡量,谱系多样性通过计算不同分类单元在系统发育水平上的接近程度衡量,而遗传多样性相对来说较难研究,因为其涉及到了同一物种不同菌株水平,需要通过更精细水平的组学研究技术进行相应的描述。在微生物群落多样性的描述中,常使用分类多样性与谱系多样性。

3.1 微生物多样性的数量描述

按宏观生态学描述习惯,多样性常根据空间尺度分为3个种类:α-多样性、β-多样性与γ-多样性[44-45]。α-多样性主要描述局部群落或斑块中的多样性,β-多样性主要描述不同群落间(或整个景观的)物种差异,γ-多样性则关注更大区域性尺度的多样性[45]。而对于微生物生态学的研究,多样性分析常常聚焦于α-多样性与β-多样性。

由于取样和测序的随机性,分析结果并不能完全反应群落的真实状态。对于这类数据的物种累积曲线,随着样本大小的增加,序列(read)数量以恒定速度线性增加,同时观测到的物种数目以递减的对数速率累积。Sanders[46]于1968年提出稀释化(rarefaction)方法,使得不同样本大小的物种累积曲线可以进行比较。该方法对于物种累积曲线的改进在于它强调了物种累积曲线本身的形状,而非样本的绝对大小数值。使用该方法绘制的累积曲线称为稀释曲线(rarefaction curve),其绘制方式为保持样本中OTU百分比组成不变,构建具有相同OTU组成但具不同样本大小的样本物种累积曲线。一般认为,当某一样本的稀释曲线末端趋于平缓时,即认为更多的采样量或更深的测序深度基本上无法再测到更多的新的OTU了,即认为该样本的采样和测序已经近似完全。稀释化方法的缺点在于,由于根据百分比组成缩放了物种丰富度信息,构建稀释曲线对于稀有种等信息会造成失真,故一般认为样本内物种符合随机分布或均匀分布时,稀释曲线才能有效工作[46-47]。

在获得扩增子数据并根据这些数据计算α-多样性后,一般需要构建稀释曲线,并对OTU表进行重新抽取(resample)。重新抽取是指全体样本在该操作后仍能满足稀释曲线趋于平缓的条件下,按照所有样本中序列数的最小值R_{min}对所有其余样本重新随机抽取R_{min}条序列进行多样性分析。该方法可以降低样本大小对于多样性指标间的比较的影响。

3.1.1 α-多样性的数量描述

α-多样性的数量描述对象主要是物种丰富度(richness)与物种数量分布(distribution)。以下以OTU为例进行说明,对于ASV的分析同理。OTU表中OTU观察值(observed OTU number, S_{obs})可作为物种丰富度的观测指标。常用α-多样性的量化参数如表3所示。

除表3列出的指数外,希尔数(Hill Number)也是一类重要的描述群落α-多样性的指数。希尔数是一类多样性指数组成的指数家族,其整合了相对

表3 常用作α-多样性数量表示的参数

指数名称	计算公式	指数意义
Chao1值(S_{chao1})*	$S_{chao1}=S_{obs}+\frac{n_1^2}{2n_2}$ $S_{chao1}=S_{obs}+\frac{n_1(n_1-1)}{2(n_2+1)}$	估计物种丰富度[48]
Shannon指数[49-50](H')**	$H'=-\sum_{i=1}^{n}p_i\ln p_i$	用于度量序列落在OTU中的均匀程度,与α-多样性呈正相关关系[51-52]
Simpson指数[53](λ)***	$\lambda=\sum_{i=1}^{n}p_i^2$(大样本) $\lambda=\sum_{i=1}^{n}\left[\frac{n_i(n_i-1)}{N(N-1)}\right]$(小样本或有限样本)	用于度量序列落在OTU中的均匀程度,与α-多样性呈负相关关系,常用其倒数($1/\lambda$)作分析[51]

注:*表示S_{obs}为样本内OTU数,n_1为仅含1条序列的OTU(singleton)数目,n_2为仅含2条序列的OTU(doubleton)数目;**表示n为样本内全体OTU包含序列数,p_i为OTU_i包含序列数占全体OTU包含序列数目的比例;***表示n为样本内全体OTU包含序列数,p_i为OTU_i包含序列数占全体OTU包含序列数目的比例,n_i为OTU_i包含序列数,N为全体序列数。

丰度、物种丰富度并消除了它们所具有的一些缺陷。希尔数实际上是上述几个指数的外推形式,也是这些指数的统一形式。该指数最先由Hill于1973年提出[54]。希尔数是符合复制原则(replication principle)的,即两完全相异群落的希尔数之和等于两群落混合后的希尔数,很容易计算得知,上述指数中,除了物种丰富度符合复制原则外,其他的指数显然都是不符合这一原则的。希尔数的0阶形式(Hill Number ^{0}D)为物种丰富度,而1阶(^{1}D)与2阶(^{2}D)形式则分别表示Shannon熵指数的指数形式和Simpson指数的倒数形式[55]。hilldiv为常用的用于希尔数计算的R包[56]。

3.1.2　β-多样性的数量描述

相对于α-多样性所关注某一个微生物群落或样本内的多样性情况,β-多样性所关注的是多个微生物群落或样本间的相似性(similarity)或不相似性(dissimilarity)。在对β-多样性的数量描述中,互补性(complementarity)是一个重要的描述角度,它指的是2个样本之间包含对方所不包含物种的数量,因此,2个样本的互补性越强,可以认为它们的β-多样性越高[52,57]。由于对于样本间互补性的描述与数学集合之间元素的互补性有相似之处,因此对于互补性的计算、描述以及延伸多仿用了集合中的相应规则。相应地,互补性可以用维恩图(Venn Diagram)来进行可视化表示,同时通过样本间的共享物种(shared species)与特有物种(unique species)可以计算样本间的相似性或不相似性。样本间不相似性可以用距离指数(distance index)来衡量,对于OTU表来说,全体样本的成对距离形成的矩阵称为距离矩阵(distance matrix)或不相似性矩阵(dissimilarity matrix)。

常见的OTU表中的数据代表了各样本中各OTU下的序列数,即每个样本不仅有OTU种类信息,同时还含有每个OTU的丰度信息,这一类数据常被称为定量数据(quantitative data)。应用中,还有另外一类数据,这类数据只包含有每个样本具有哪些OTU,但不包含每个OTU的丰度信息,这类数据常称为存在-缺失数据(presence-absence data),由于这类数据在矩阵中常用1和0表示存在和缺失2种状态,故也被称为1-0数据。这2类数据微生物生态学的应用主要取决于研究是否需要考虑每个OTU的丰度信息。

距离指数的形式多种多样,每种形式的距离指数对于定量数据与存在-缺失数据的计算方法也不同。对于常用的Jaccard距离与Bray-Curtis距离,前者为典型的存在-缺失数据距离指数,而后者则为定量数据距离指数。Jaccard指数是以相似系数形式提出的[58],它的经典计算公式为:

$$S_{\text{Jaccard}} = \frac{a}{a + b + c} \tag{1}$$

其中,a为2个样本共享OTU个数,b与c分别为2个样本各自特有OTU个数。

根据定义,Jaccard指数的值域为[0,1]。相应地,定义Jaccard距离或Jaccard不相似性为:

$$D_{\text{Jaccard}} = 1 - S_{\text{Jaccard}} \tag{2}$$

之后,有研究者提出了Sørensen指数,又称Sørensen-Dice指数[59-60],该指数的计算形式与Jaccard指数相似:

$$S_{\text{Sørensen}} = \frac{2a}{2a + b + c} \tag{3}$$

相似地,a为2个样本共享OTU个数,b与c分别为2个样本各自特有OTU个数。

相比于Jaccard指数,Sørensen指数加大了2个样本共有OTU的权重,认为样本间共有的OTU是比较样本间多样性的关键。实际上,Jaccard指数与Sørensen指数在计算上考虑2个样本共享OTU的情况下,同时忽略了2个样本中均没有出现的其他OTU。即假设某环境中可被检测的OTU总数为s,a、b、c的定义仍与上述公式中保持一致,定义d为在取得的2个样本中均未出现的OTU个数,则有:

$$s = a + b + c + d \tag{4}$$

在上述2种指数的定义与计算中,d是没有被考虑的。这就是生态数据分析中的"双零问题(double-zero problem)"[52],即当某些OTU同时在2个样本中均表现为缺失,那么这一类数据对于生态学结论的贡献应该是怎样的。针对双零问题,有时双零数据不参与到数据计算中,如Jaccard指数与Sørensen指数计算,这类指数称为非对称指数,因为它们对待"双存在(double-presence)数据"与"双缺失(double-absence)数据"的方式不同,其他在计

算中考虑双零数据的指数则称为对称指数。

在上述2种指数中,OTU丰度信息没有参与到计算当中。而Bray-Curtis不相似度在计算中考虑了丰度信息,该指数的计算方法在1950年首次被提出[61],随后在1957年被Bray与Curtis在论文中使用[62]。假设A与B为2个样本,OTU总数为n,样本A、B中OTU_i(i=1, 2, …, n)分别为X_{1i}与X_{2i},则Bray-Curtis不相似度的计算公式为:

$$D_{\text{Bray-Curtis}}=\frac{\sum_{i=1}^{n}\left|X_{1i}+X_{2i}\right|}{\sum_{i=1}^{n}\left(X_{1i}+X_{2i}\right)}=1-2\frac{C}{A+B} \quad (5)$$

其中,A与B分别是2个样本OTU总数,C为2个样本共有OTU均取较小值的总和。

同样地,Bray-Curtis不相似度也属于非对称指数。值得注意的是,许多程序包或文献中将Bray-Curtis不相似度(Bray-Curtis dissimilarity)称为Bray-Curtis距离(Bray-Curtis distance)的这一提法可能是有误的,Legendre等在《Numerical Ecology》中[63]证明,Bray-Curtis不相似度与Sørensen指数无法满足三角不等式定理并无法在欧几里得空间中合理排序,它们在严格意义上不应被称为距离指数,该指数为半度量的(semimetrics)。

除上述指数外,常用的相似性-不相似性指数还有很多,在先前的一些文献综述中已有详细总结[64],R语言[65]中vegan包[66]的vegdist函数提供了大多数指数的计算方法。

对于类别多样性来说,不同类别对于多样性的贡献是平等的,而Vane-Wright认为进化关系在类别多样性的评估中也应该起到作用[57]。于是,Faith于1992年首先提出谱系多样性(phylogenetic diversity, PD)这一概念并定义其为待观察物种在分支树上的最短进化分支长度之和[67]。谱系多样性考虑了物种之间在进化水平上的差异,其中包含了物种表型性状与生态位等信息[68]。谱系α-多样性的计算方式基于Faith对于PD的基本定义[67,69]。在计算微生物群落间谱系多样性,即谱系β-多样性时,UniFrac指数是常用的计算指标[70],该指数根据不同群落包含的共有与特有谱系结构计算群落间的不相似性。UniFrac距离是度量性(metric)指数,符合三角不等式定理。UniFrac距离可以通过蒙特卡洛方法的随机化过程用于查看两群落是否在谱系多样性上具有显著性差异,也可以用来产生距离矩阵,存储多样本间成对的系统发育距离。

3.2 群落结构的解析方法

多数微生物生态学研究主要关注于不同生境或不同环境梯度下微生物群落的变化,一组高通量扩增子数据是对于微生物在多组时空样本下的观测。这类研究所产生的数据集,现多用多元统计方法进行分析。多元统计方法作为一类独立的统计学方法,其中的方法分类与应用场景根据研究目的与数据类型会有变化。常用的多元统计分析方法及其在R语言中应用程序包与函数列在表4中[71]。

主成分分析(principal component analysis, PCA)是最常见、应用最广泛的多元统计方法之一[76]。在数学上,PCA实际是降维的过程[77]。一般地,认为选取前2~3个主成分变量能够解释原矩阵中大部分方差。PCA采用欧几里得距离来度量样本之间的差异性,样本覆盖梯度太长时(即多样本中有很多相同OTU)该分析方法会出现马蹄效应等问题。

对应分析(correspondence analysis, CA)通常用于衡量由样本-OTU数据反映的样本群落间的差异。CA排序用列联表中的卡方距离(chi-square distance)找到样本与OTU之间的对应关系并将其在降维空间中展示出来[78-79],因此CA规避了PCA中会出现的马蹄效应。然而,CA排序常会伴有弓形效应,即CA1轴上的样品分布信息会部分体现在CA2轴上,这是因为CA排序选取的排序轴之间并不一定是线性不相关的。因此,去趋势对应分析(detrended correspondence analysis, DCA)被提出用于尽量减小弓形效应。

主坐标分析(principal coordinates analysis, PCoA)在概念上由PCA衍生而来。与PCA类似地,PCoA同样遵循了降维的基本思路,将样本空间压缩并投射到低维空间。PCoA使用样本间两两成对不相似性矩阵,其中不相似性指数可以使用Jaccard指数、Bray-Curtis不相似性指数与UniFrac指数等。由于PCoA排序的依据是不相似性矩阵,PCoA的排序轴与原始变量间并不存在直接关系,但其方

表4　常用多元统计分析方法及R语言函数

统计分析方法		基本说明	R语言函数(程序包)
探索性方法	主成分分析(PCA)	利用正交主成分将数据在矩阵空间中按照欧氏距离排序	rda (vegan)
	对应分析(CA) 去趋势对应分析(DCA)	利用对应列联表将数据在矩阵空间中按照卡方距离排序	decorana (vegan)
	主坐标分析(PCoA)	利用不相似性矩阵将数据压缩至低维空间排序	pcoa (ape)
	非度量多维尺度分析(NMDS)	利用不相似性数值秩序不变规则将数据在低维空间排序	metaMDS (vegan)
解释性方法	冗余分析(RDA)	PCA分析加入约束性解释变量的拓展	rda (vegan)
	典范对应分析(CCA)	CA分析加入约束性解释变量后的典范形式	cca (vegan)
统计检验方法	ANOSIM	基于不相似性度量不同样本组间的显著差异[72]	anosim (vegan)
	PERMANOVA	利用非参数方法对不同样本组间差异进行置换检验[73]	adonis (vegan)
	MRPP	检验每个分组下随机置换检验的不相似性变化[74]	mrpp (vegan)
	MANTEL	对两个矩阵间的相关性显著性的检验[75]	mantel (vegan)

差解释度还是能够通过校正后的不相似性矩阵特征值给出。

非度量多维尺度分析(non-metric multidimensional scaling, NMDS)是一种特殊的排序方法。在该方法中,先利用原始数据计算样品成对不相似性矩阵并根据不相似性数值大小对样品赋予秩序(rank order),然后规定降维空间的维数,如二维或三维,将数据按照秩序不变的原则投射在降维空间上,在这个过程中原有的样品间的不相似性会被改变,这种改变的程度的量化指标为胁迫值(stress),在投射的过程中不相似性相较原始矩阵变化越小,其胁迫值越小。在进行NMDS分析时,一般会进行多次迭代排序,以求取得尽量小的胁迫值。一般认为胁迫值小于0.15是可以接受的[72]。NMDS分析中排序距离已经和样本之间的原始不相似性无关,因此排序轴不具有解释样本不相似性方差的作用,故NMDS排序图的排序轴上无法给出合理的解释度。

上述的4种多元统计分析方法均属于探索性方法(exploratory methods),这一类分析方法提供了样本变化的主要梯度以及样本的相似程度,而事实上,即使样本经过上述方法分析后体现出了某些规律,这些规律是否为偶然出现仍然需要验证。很多情况下,研究者希望能够得到对于观察到的差异进行统计分析得到其显著性。常见的对样本间差异进行统计检验的方法有:ANOSIM(analysis of group similarities,分组相似性分析)、PERMANOVA(permutational multivariate analysis of variance,置换方差分析)与MRPP(multi-response permutation procedures,多响应置换过程)等。

在分析不同样本组间的微生物群落差异时,研究人员往往还关注造成这种差异的环境因子是哪些,即将微生物群落差异看作响应变量(因变量),而将环境因子看作解释变量(自变量)。于是,另一类不同于探索性方法的排序方法被用于这方面的研究分析,被称为解释性方法(interpretive methods),这一类方法在前者的基础上,增加了一组解释变量,解释变量在每个排序轴上的分量表示该变量对于样品沿该轴分布的贡献。冗余分析(redundancy analysis, RDA)与典范对应分析(canonical correspondence analysis, CCA)是2种典型的解释性排序方法。RDA可以看作是PCA排序的一种拓展,其中加入解释变量,使得排序轴(主成分)被约束为解释变量的线性组合[80]。同时,RDA也具有PCA的缺点,即并不适合于处理样本覆盖梯度长的数据集。类似于PCA无法使用时可以选用CA一样,CCA是当RDA不适用时更好的选择。CCA是

利用解释变量约束响应变量后进行对应分析的典范形式[81]。RDA与CCA的可视化展示在PCA排序图基础上增加了代表解释变量的向量(数量变量)或点(类别变量)。

在进行环境因子与样本群落间差异的相关性统计性检验时,传统的相关系数检验往往不能很好地实施,因为它计算的是两列向量的相关性,而群落间差异常以二维矩阵形式表示。Mantel于1967年提出的Mantel检验方法可以对两个矩阵之间的相关性显著性进行检验[75]。当要对多变量矩阵或控制变量矩阵进行相关性统计检验时,还可以使用偏Mantel检验(partial Mantel test),即选取一个解释变量,其余解释变量作为协变量。变量分解分析(variation partitioning analysis, VPA)利用了偏分析的思想,将响应变量数据集中的总方差划分为单个解释变量的独立解释贡献以及联合解释贡献。VPA分析常用于在确定了哪些环境因子对于微生物群落间差异有显著影响之后进一步说明不同环境因子对于不同群落的差异的贡献度。

4 结论

目前,多组学技术的联合应用已经成为认识环境微生物群落及其功能的重要手段[82],而通过组学技术的应用,研究者逐渐意识到生活在土壤、淡水、海水、空气,甚至人体等环境中的微生物,其系统发育的多样性和功能的复杂度远远超过以往的认识。目前,如何利用正处于蓬勃发展时期的大数据分析技术与手段来克服宏基因组学数据解析这一难关,并将分析结果用更易理解与操作的形式展现出来,这是从事环境微生物学、生物信息和统计学研究人员共同的挑战。

参考文献(References)

[1] Borneman J, Skroch P W, O'Sullivan K M, et al. Molecular microbial diversity of an agricultural soil in Wisconsin [J]. Applied and Environmental Microbiology, 1996, 62 (6): 1935-1943.

[2] Trevors J T. Bacterial biodiversity in soil with an emphasis on chemically-contaminated soils[J]. Water, Air, and Soil Pollution, 1998, 101(1): 45-67.

[3] 魏子艳, 金德才, 邓晔. 环境微生物宏基因组学研究中的生物信息学方法[J]. 微生物学通报, 2015, 42(5): 890-901.

[4] Handelsman J, Rondon M R, Brady S F, et al. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products[J]. Chemistry & Biology, 1998, 5(10): R245-R249.

[5] Mardis E R. Next-generation DNA sequencing methods [J]. Annual Review of Genomics and Human Genetics, 2008, 9: 387-402.

[6] Eid J, Fehr A, Gray J, et al. Real-time DNA sequencing from single polymerase molecules[J]. Science, 2009, 323 (5910): 133-138.

[7] Niedringhaus T P, Milanova D, Kerby M B, et al. Landscape of next-generation sequencing technologies[J]. Analytical Chemistry, 2011, 83(12): 4327-4341.

[8] Tremblay J, Yergeau E. Systematic processing of ribosomal RNA gene amplicon sequencing data[J]. GigaScience, 2019, 8(12): giz146.

[9] Zhou J, Wu L, Deng Y, et al. Reproducibility and quantitation of amplicon sequencing-based detection[J]. The ISME Journal, 2011, 5(8): 1303-1313.

[10] Liu Y X, Qin Y, Chen T, et al. A practical guide to amplicon and metagenomic analysis of microbiome data[J]. Protein Cell, 2020, 12: 315-330.

[11] Magoc T, Salzberg S L. FLASH: Fast length adjustment of short reads to improve genome assemblies[J]. Bioinformatics, 2011, 27(21): 2957-2963.

[12] Edgar R C. UPARSE: Highly accurate OTU sequences from microbial amplicon reads[J]. Nature Methods, 2013, 10: 996-998.

[13] Callahan B J, McMurdie P J, Rosen M J, et al. DADA2: High-resolution sample inference from Illumina amplicon data[J]. Nature Methods, 2016, 13: 581-583.

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

[15] Amir A, McDonald D, Navas-Molina J A, et al. Deblur rapidly resolves single-nucleotide community sequence patterns[J]. mSystems, 2017, 2(2): e00191-16.

[16] Milanese A, Mende D R, Paoli L, et al. Microbial abundance, activity and population genomic profiling with

mOTUs2[J]. Nature Communications, 2019, 10: 1014.

[17] Ewels P, Magnusson M, Lundin S, et al. MultiQC: Summarize analysis results for multiple tools and samples in a single report[J]. Bioinformatics, 2016, 32(19): 3047-3048.

[18] Chen S, Zhou Y, Chen Y, et al. fastp: An ultra-fast all-in-one FASTQ preprocessor[J]. Bioinformatics, 2018, 34 (17): i884-i890.

[19] Bolger A M, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data[J]. Bioinformatics, 2014, 30(15): 2114-2120.

[20] De Coster W, D'Hert S, Schultz D T, et al. Nanopack: Visualizing and processing long-read sequencing data [J]. Bioinformatics, 2018, 34(15): 2666-2669.

[21] Li D, Liu C M, Luo R, et al. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph[J]. Bioinformatics, 2015, 31(10): 1674-1676.

[22] Li D, Luo R, Liu C M, et al. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices[J]. Methods, 2016, 102: 3-11.

[23] Bankevich A, Nurk S, Antipov D, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing[J]. Journal of Computational Biology, 2012, 19(5): 455-477.

[24] Nurk S, Meleshko D, Korobeynikov A, et al. metaSPAdes: A new versatile metagenomic assembler[J]. Genome Research, 2017, 27(5): 824-834.

[25] Peng Y, Leung H C, Yiu S M, et al. IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth[J]. Bioinformatics, 2012, 28(11): 1420-1428.

[26] Luo R, Liu B, Xie Y, et al. SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler[J]. GigaScience, 2012, 1(1): 2047-217X-1-18.

[27] Luo R, Liu B, Xie Y, et al. Erratum: SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler[J]. GigaScience, 2015, 4(1): s13742-015-0069-2.

[28] Koren S, Walenz B P, Berlin K, et al. Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation[J]. Genome Research, 2017, 27(5): 722-736.

[29] Bertrand D, Shaw J, Kalathiyappan M, et al. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes[J]. Nature Biotechnology, 2019, 37(8): 937-944.

[30] Kolmogorov M, Bickhart D M, Behsaz B, et al. metaFlye: Scalable long-read metagenome assembly using repeat graphs[J]. Nature Methods, 2020, 17(11): 1103-1110.

[31] Kang D D, Li F, Kirton E, et al. MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies[J]. PeerJ, 2019, 7: e7359.

[32] Wu Y W, Tang Y H, Tringe S G, et al. MaxBin: An automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm[J]. Microbiome, 2014, 2: 26.

[33] Wu Y W, Simmons B A, Singer S W. MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets[J]. Bioinformatics, 2016, 32(4): 605-607.

[34] Alneberg J, Bjarnason B S, de Bruijn I, et al. Binning metagenomic contigs by coverage and composition[J]. Nature Methods, 2014, 11(11): 1144-1146.

[35] Nissen J N, Johansen J, Allesøe R L, et al. Improved metagenome binning and assembly using deep variational autoencoders[J]. Nature Biotechnology, 2021, 39(5): 555-560.

[36] Uritskiy G V, DiRuggiero J, Taylor J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis[J]. Microbiome, 2018, 6(1): 158.

[37] Buchfink B, Reuter K, Drost H G. Sensitive protein alignments at tree-of-life scale using diamond[J]. Nature Methods, 2021, 18(4): 366-368.

[38] Wood D E, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2[J]. Genome Biology, 2019, 20: 257.

[39] Menzel P, Ng K L, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju[J]. Nature Communications, 2016, 7: 11257.

[40] Beghini F, McIver L J, Blanco-Miguez A, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3[J]. eLife, 2021, 10: e65088.

[41] Hyatt D, LoCascio P F, Hauser L J, et al. Gene and translation initiation site prediction in metagenomic sequences[J]. Bioinformatics, 2012, 28(17): 2223-2230.

[42] Hamilton A J. Species diversity or biodiversity?[J]. Journal of Environmental Management, 2005, 75(1): 89-92.

[43] Naeem S, Duffy J E, Zavaleta E. The functions of biological diversity in an age of extinction[J]. Science, 2012, 336(6087): 1401-1406.

[44] Whittaker R H. Vegetation of the Siskiyou Mountains, Oregon and California[J]. Ecological Monographs, 1960, 30: 279-338.

[45] Whittaker R J, Willis K J, Field R. Scale and species richness: Towards a general, hierarchical theory of species diversity[J]. Journal of Biogeography, 2001, 28(4): 453-470.

[46] Sanders H L. Marine benthic diversity: A comparative study[J]. The American Naturalist, 1968, 102(925): 243-282.

[47] McMurdie P J, Holmes S. Waste not, want not: Why rarefying microbiome data is inadmissible[J]. PLoS Computational Biology, 2014, 10(4): e1003531.

[48] Chao A. Nonparametric estimation of the number of classes in a population[J]. Scandinavian Journal of Statistics, 1984, 11(4): 265-270.

[49] Shannon C E. A mathematical theory of communication [M]. New York: Bell System Technical Journal.

[50] Spellerberg I F, Fedor P J. A tribute to Claude Shannon (1916-2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon-Wiener' index[J]. Global Ecology and Biogeography, 2003, 12: 177-179.

[51] Lemos L N, Fulthorpe R R, Triplett E W, et al. Rethinking microbial diversity analysis in the high throughput sequencing era[J]. Journal of Microbiological Methods, 2011, 86(1): 42-51.

[52] Magurran A E. Measuring biological diversity[M]. Hoboken: Wiley-Blackwell, 2004.

[53] Simpson E H. Measurement of diversity[J]. Nature, 1949, 163(4148): 688.

[54] Hill M O. Diversity and evenness: A unifying notation and its consequences[J]. Ecology, 1973, 54: 427-432.

[55] Chao A, Gotelli N J, Hsieh T C, et al. Rarefaction and extrapolation with Hill numbers: A framework for sampling and estimation in species diversity studies[J]. Ecological Monographs, 2014, 84(1): 45-67.

[56] Alberdi A, Gilbert M T P. hilldiv: An R package for the integral analysis of diversity based on Hill numbers[J]. bioRxiv, 2019, doi: 10.1101/545665.

[57] Vane-Wright R I, Humphries C J, Williams P H. What to protect?—systematics and the agony of choice[J]. Biological Conservation, 1991, 55(3): 235-254.

[58] Jaccard P. The distribution of the flora in the alpine zone [J]. New Phytologist, 1912, 11(2): 37-50.

[59] Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analysis of the vegetation on Danish commons[J]. Biologiske Skrifter, 1948, 5 (4): 1-34.

[60] Dice L R. Measures of the amount of ecologic association between species[J]. Ecology, 1945, 26(3): 297-302.

[61] Odum E P. Bird populations of the highlands (North Carolina) plateau in relation to plant succession and avian invasion[J]. Ecology, 1950, 31(4): 587-605.

[62] Bray J R, Curtis J T. An ordination of the upland forest communities of Southern Wisconsin[J]. Ecological Monographs, 1957, 27(4): 325-349.

[63] Legendre P, Legendre L. Developments in Environmental Modelling[M]. Amsterdam: Elsevier, 1998.

[64] Legendre P, De Cáceres M. Beta diversity as the variance of community data: Dissimilarity coefficients and partitioning[J]. Ecology Letters, 2013, 16(8): 951-963.

[65] R Core Team. R: A language and environment for statistical computing[EB/OL]. [2021-01-09]. https://www. r-project.org/.

[66] Oksanen J, Blanchet F G, Friendly M, et al. vegan: Community Ecology Package[EB/OL]. [2021-06-14]. https://github.com/vegandevs/vegan.

[67] Faith D P. Conservation evaluation and phylogenetic diversity[J]. Biological Conservation, 1992, 61(1): 1-10.

[68] Webb C O, Ackerly D D, McPeek M A, et al. Phylogenies and community ecology[J]. Annual Review of Ecology and Systematics, 2002, 33(1): 475-505.

[69] Faith D P, Baker A M. Phylogenetic diversity (PD) and biodiversity conservation: Some bioinformatics challenges [J]. Evolutionary Bioinformatics Online, 2007, 2: 121-128.

[70] Lozupone C, Knight R. UniFrac: A new phylogenetic method for comparing microbial communities[J]. Applied and Environmental Microbiology, 2005, 71(12): 8228-8235.

[71] Paliy O, Shankar V. Application of multivariate statistical techniques in microbial ecology[J]. Molecular Ecology, 2016, 25(5): 1032-1057.

[72] Clarke K R. Non-parametric multivariate analyses of changes in community structure[J]. Australian Journal of Ecology, 1993, 18(1): 117-143.

[73] Anderson M J. A new method for non-parametric multi-

variate analysis of variance[J]. Austral Ecology, 2001, 26 (1): 32-46.
[74] Mielke P W, Berry K J. Permutation methods: A distance function approach[M]. New York: Springer, 2001.
[75] Mantel N. The detection of disease clustering and a generalized regression approach[J]. Cancer Research, 1967, 27(2): 209-220.
[76] Pearson K L. On lines and planes of closest fit to systems of points in space[J]. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1901, 2(11): 559-572.
[77] Hotelling H. Analysis of a complex of statistical variables into principal components[J]. Journal of Educational Psychology, 1933, 24(6): 417-441.
[78] Hill M O. Correspondence analysis: A neglected multivariate method[J]. Journal of the Royal Statistical Society. Series C (Applied Statistics), 1974, 23(3): 340-354.
[79] Hill M O. Reciprocal averaging: An eigenvector method of ordination[J]. Journal of Ecology, 1973, 61: 237-249.
[80] Rao C R. The use and interpretation of principal component analysis in applied research[J]. Sankhyā: The Indian Journal of Statistics, Series A (1961-2002), 1964, 26 (4): 329-358.
[81] Ter Braak C J F. Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis[J]. Ecology, 1986, 67(5): 1167-1179.
[82] 马海霞, 张丽丽, 孙晓萌, 等. 基于宏组学方法认识微生物群落及其功能[J]. 微生物学通报, 2015, 42(5): 902-912.

Analytical methods for metagenomic technology and microbial community diversity

PENG Xi[1,2], FENG Kai[1], LI Shuzhen[3], DENG Ye[1,2*]

1. CAS Key Laboratory of Environmental Biology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China
2. College of Resources and Environment, University of Chinese Academy Sciences, Beijing 100049, China
3. Key Laboratory of Industrial Ecology and Environmental Engineering, Ministry of Education; School of Environmental Science and Technology, Dalian University of Technology, Dalian 116024, China

Abstract The rapid development of omics technology represented by metagenomics technology has greatly promoted our understanding of microbial diversity, composition, structure, and function in the natural ecosystem. However, the big data generated by the technology can be a great challenge to researchers′ data analyzing and mining abilities. This review, based on two technical aspects, amplicon and whole-genome shotgun sequencing, summarizes the analysis workflow of metagenomics for discovering microbial community. Then, the concepts of microbial community diversity, the related principles of statistical analysis, and related interpretation of statistical tests are illustrated. Finally, the paper points out that to overcome the complexity of metagenomic data and huge amount information, using big data analysis so as to illuminate analytical results are the common challenge for environmental microbiologists, bioinformaticians, and statisticians.

Keywords environmental microbiology; metagenomics; amplicon; bioinformatics

（责任编辑　徐丽娇）