**16S扩增子分析中常用软件及数据库应用现状**

**The Application Status of Commonly Used Software and Database in 16S Amplicon Analysis**

杨潇瀛，张浩林，韩莹莹，翁强，袁峥嵘\*

1生物科学与技术学院，北京林业大学，北京

\*通讯作者邮箱: [zryuan@bjfu.edu.cn](mailto:zryuan@bjfu.edu.cn)

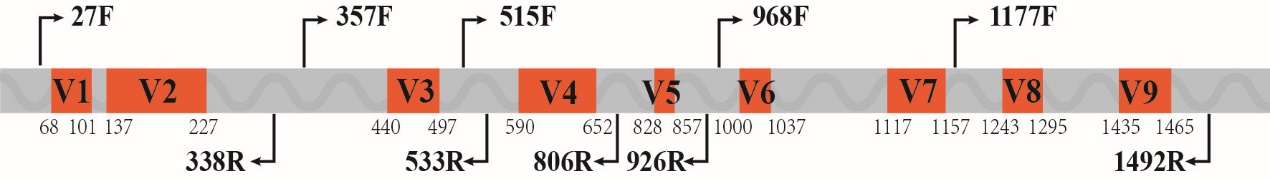
**摘要**：随着高通量测序技术 (High-throughput sequencing, HTS) 的发展，微生物组领域的研究日益广泛。其中扩增子测序技术因其操作便捷、成本较低等特点逐渐成为研究者的首选。但随之产生的问题是面对测序产生的大量数据，非专业分析人员较难通过生物信息学分析从数据中挖掘出有用信息。本文全面介绍了用于16S扩增子测序数据分析的几种常用软件及参考数据库，以及近年来推荐的基于去噪分析生成扩增子序列变体 (amplicon sequence variants, ASVs) 的几种算法。目的在于为初学者在选取分析软件及数据库方面提供参考，使其能高效进行扩增子数据分析，挖掘其中蕴含的生物学意义。

**关键词**：扩增子，微生物组，分析软件，数据库，功能预测

**研究背景：**高通量测序技术 (High-throughput sequencing, HTS) 又称“下一代”测序技术 ("Next-generation" sequencing technology, NGS)，可以并行的对数百万到数十亿个小片段DNA进行测序。与Sanger测序相比，NGS以其高数据输出、低成本、高时间效益，应用多样等特点改变了基因组的研究 (Behjati and Tarpey 2013; Kumar*等，* 2019)。Illumina为当下主流平台之一，采用“桥式扩增”技术，包括iSeq、MiniSeq、MiSeq、NextSeq、HiSeq和NovaSeq等多种测序系统，具有广泛的应用空间。在临床医学 (Yohe and Thyagarajan 2017)、法医学 (Yang*等，* 2014)、环境科学 (Mahnert*等，* 2019)、农业 (Berkman*等，* 2012) 等多领域中，NGS都有着广泛深入的应用。虽然NGS有强大的技术支持和广泛的应用前景，但是限制它发展的主要因素为相对较短的读取长度 (35~700 bp)。当基因组中包含的大量重复序列超过NGS可测量长度便会导致错配和缺口，增加测序错误率 (0.1~15%) (Goodwin*等，* 2016; van Dijk*等，* 2018)。因此以Pacific Biosciences (PacBio) 公司研发的单分子实时测序 (single-molecule real-time sequencing, SMRT) (Eid*等，* 2009) 和Oxford Nanopore Technologies (ONT) 公司的新型纳米孔测序法 (nanopore sequencing) 为首的第三代测序技术 (Third-Generation Sequencing, TGS) (Jain*等，* 2015) 应运而生。与前两代测序技术相比，TGS在保证一定准确性的同时可以在更短的时间获取更多的读长，从而更好的进行从头组装，并能够直接检测单倍型，甚至整个染色体定相 (Schadt*等，* 2010)。除了基因组测序，TGS还有更广泛的用途，包括转录组的综合表征、甲基化模式的鉴定、表观遗传修饰的检测等(van Dijk*等，* 2018)。

随着NGS的发展加上生物信息学的进步，微生物组 (Microbiome) 迎来了快速发展时期。微生物组指整个栖息地，包括微生物 (细菌、古菌、低等真核生物、高等真核生物和病毒)、它们的基因组和周围的环境条件(Marchesi and Ravel 2015)。其中存在于特定环境中的微生物的集合称为微生物群落 (Microbiota)，主要依赖于对16S rRNA、18S rRNA、内转录间隔区 (Internal transcribed spacer，ITS) 基因或其他标记基因和基因组区域的分析，从给定的生物样本中扩增特定片段并进行测序 (Marchesi and Ravel 2015)。

16S rRNA扩增子测序 (16S rRNA amplicon sequencing) 是在微生物群落研究中的代表性方法，可以测得样本中的细菌及古菌。16S中的“S”代表非国际单位制下的沉降系数 (Sedimentation coefficient)。16S rRNA长度适中，全长约1,540 nt，包括9个高变区 (Variable region，V1~V9) 和10个保守区 (Conserved region，C1~C10) (图1)，存在于所有的细菌中，因在功能结构上具有高度的保守性，常被用于微生物分类研究的标志物(吴悦妮*等，* 2020)。但目前的二代测序技术不能覆盖16S全长，需要对一个或多个高变区进行测序。一般来说，V4区特异性较好，可识别多数序列到属水平，为了增加测序的准确性通常增加V3区域以增加测序长度，故而V3~V4区 (约465bp) 是常用目标区域(Zhang*等，* 2018)。但对于V区的选择没有统一的标准，研究者可根据研究目的、生境条件、可变性、保守性、连续性、可比性等因素综合选择合适的V区 (张军毅*等，* 2015)。



**图1. 16S rRNA结构示意图**

第二代测序会得到海量的原始下机数据，须通过合适的生物信息学软件加以处理分析才可得到具有可读性的数据或图表。如何从海量的数据集中得到有助于科研发现的线索也是当下的一个挑战 (Leonelli 2019)。对于大数据下的生物信息学分析一方面需要高性能的计算资源，另一方面依赖于高效的分析软件和强大数据库支持。在过去十年中，扩增子分析的流程框架逐渐完善并在不断更新。在众多软件之中，选取适合自身高效的分析软件成为科研中不可或缺的一步。因而本文综述了在16S扩增子分析中常用的软件以及数据库，总结了其优缺点以供初学人员参考选择。

**软件和数据库**

1. **16S扩增子分析中常用软件**
2. 主流分析软件 (表1)

(1) mothur：2009年2月发布了mothur的第一个版本，目前被引用13,658余次，是由美国密歇根大学Patrick D. Schloss教授团队基于C++编写的开源单一的软件平台，第一次实现了较完整的扩增子分析流程(Schloss*等，* 2009)。mothur可从项目网站获取 (<http://www.mothur.org>)，作为与Windows兼容的可执行文件，或作为Unix/Linux或Mac OS X环境中编译的源代码。其整合了DOTUR (Schloss and Handelsman 2005)、SONS (Schloss and Handelsman 2006)、UniFrac (Lozupone and Knight 2005)等多款软件，可实现OTU表生成、多样性比较、差异分析等功能。mothur没有图形界面或数据可视化工具，后续可使用R语言进行绘图，但其有良好的可重复性，并提供了易于理解的教程，使用户掌握命令行界面和其他计算技能。在过去的10年中，mothur也在不断完善来提高其可用性和实用性，并将继续开发完善这款软件(Schloss 2020)。

(2) QIIME：Quantitative Insights Into Microbial Ecology (QIIME，读作’chime’)发布于2010年，是由加州大学圣地亚哥分校微生物学家Rob Knight教授团队使用Python语言编写的一款扩增子测序分析流程，包括了200余个常用软件包和150余个脚本 (Caporaso*等，* 2010)。该软件实现了从原始数据到发表级图表的全部分析流程，已被引用21,300余次，是目前使用最多的扩增子分析软件，但其依赖关系众多，安装较为复杂。随着测序技术和生物信息学的发展，2018年QIIME2发布，正式替代QIIME。QIIME2使用Python3重新设计编写，解决了QIIME1中的许多局限性，不仅可以分析扩增子数据，而且可以作为一个多维、强大的数据科学平台，为分析代谢组学和shotgun宏基因组学数据提供初步支持(Bolyen*等，* 2019)。QIIME2的优势在于实现了分析流程的可追溯，确保了分析的可重复，并且提供了许多交互式可视化工具，还提供了QIIME2 Studio图形用户界面和QIIME2 View供零基础人员使用。此外QIIME2增加了许多新算法，还具备了插件系统，可以调用composition (Mandal*等，* 2015)、cutadapt (Kechin*等，* 2017)、DADA2、Deblur、Vsearch等一系列插件，并允许第三方提供相应功能，使用conda安装也简化了安装流程 (Bolyen*等，* 2019)。

(3) USEARCH：是继mothur、QIIME后的第三大流行扩增子分析流程，由独立研究员Robert Edgar开发 (Edgar 2010)。该软件体积小巧、安装便捷，32位版本可免费下载，可使用4G内存，能够满足大多数用户的需求；64位版本适用于大数据集，需购买后使用，目前已更新到v11版本。该软件内包括了Robert开发的鉴定操作分类单元 (operational taxonomic unit, OTU) 代表性序列的UPARSE算法。该算法将97%相似度的序列聚类为OTU，可以更大的提高准确性，使得聚类所得OTU数量更接近群落中预期的物种数量 (Edgar 2013)。此外Robert还开发了UCHIME算法，可以从头或者基于参考数据库对嵌合体进行检测，提高了检测嵌合体的灵敏度和速度，可独立使用也可在USEARCH中调用(Edgar*等，* 2011)。对扩增子数据分析准确度和效率的提高使得USEARCH软件在这一领域占有一席之地，2010年发表后已被引用12,952余次。

(4) VSEARCH：由于USEARCH软件代码不开源、算法无具体描述、内存受限，Torbjørn Rognes团队于2016年发表了开源免费多线程的VSEARCH软件，现已引用2,093余次 (Rognes*等，* 2016)。VSEARCH主要功能与参数与USEARCH相似，特点为体积小、易安装、可跨平台、采用64位设计，无内存限制，可以处理非常大的数据。VSEARCH v2.0.3 (64-bit) 在执行搜索、聚类、嵌合体检测和抽样时精确度高于USEARCH v7.0.1090和v8.1.1861；在合并双端序列时与USEARCH相当，成为USEARCH的替代软件。速度方面，VSEARCH在读取和双端序列合并时快于USEARCH，但在嵌合体检测时较慢(Westcott and Schloss 2015; Rognes*等，* 2016)。自开发以来已发布了106个版本，2021年1月更新到最新的2.15.2版本，也可在QIIME2中使用。

**表1. 16S扩增子分析主流软件**

Table1. Mainstream software for 16S amplification

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 软件  名称 | 是否开源 | 编程  语言 | 使用  系统 | 最新  版本 | 软件  首页 | 参考  文献 |
| mothur | 是 | C++ | Linux/MacOS/Windows | v1.44.3 | http://www.mothur.org | (Schloss*等，* 2009) |
| QIIME | 是 | Python2 | Linux/MacOS | 2018.1.1被QIIME2取代 | http://qiime.org | (Caporaso*等，* 2010) |
| QIIME2 | 是 | Python3 | Linux/MacOS/Windows | v2020.11 | https://library.qiime2.org | (Bolyen*等，* 2019) |
| USEARCH | 否 | C | Linux/MacOS/Windows | v11.0.667 | http://www.drive5.com/usearch | (Edgar 2010) |
| VSEARCH | 是 | C++ | Linux/MacOS/Windows | v2.15.2 | https://github.com/torognes/vsearch | (Rognes*等，* 2016) |

* 1. 基于ASVs分析算法

测序错误使得生物真实的核苷酸序列及测序错误的人工序列在分析中难以区分，降低了结果的准确性，为解决这一问题，通常以97%为特定阈值，将序列聚类到操作分类单元 (operational taxonomic unit, OTU) (Schloss*等，* 2009; Caporaso*等，* 2010)，这一阈值为1994年提出，随着测序技术的进步，Robert指出要得到更准确的结果，对于全长序列的最佳同一性阈值需在~99%，V4高变区的最佳同一性阈值为~100％(Stackebrandt and Goebel 1994; Edgar 2018)。而且OTU这种方法不能检测到物种或菌株之间的细微差异，错过了真实的生物学序列变异(Qian*等，* 2020)。近几年已经开发出以扩增子序列变体 (amplicon sequence variants, ASVs)为载体的新方法。ASV方法对原始数据进行去噪 (denoise)，无需设定阈值，相当于100%聚类(Tikhonov*等，* 2015)。相对于OTU方法有更好的特异性和敏感性，并且能够更好的区分生态模式(Callahan*等，* 2017)。目前基于ASV方法使用最广泛的3个包为DADA2,UNOISE3和Deblur(Nearing*等，* 2018)。

(1) DADA2：Divisive Amplicon Denoising Algorithm (DADA)于2012年首次发表，2016年发表了功能更加强大的DADA2(Rosen*等，* 2012; Callahan*等，* 2016)。DADA2是一个R包 (<https://github.com/benjjneb/dada2>)，用于校正Illumina测序中扩增子错误，可以识别出更多真实的突变体，与主流三大软件 (mothur、QIIME、UPARSE)相比，可以以最少的假阳性结果检测到最全的代表性序列(Callahan*等，* 2016)。DADA2现已发展成为完整的扩增子分析流程，可使用R语言对原始序列进行过滤、去嵌合、拼接等，同时也可在QIIME2中使用。

(2) UNOISE3：DADA2问世后，Robert在2016年发表了UNOISE2算法，并指出与DADA2相比具有相当或更高的准确性(Edgar 2016)。该算法目前已更新到UNOISE3版本，并整合在USEARCH v10当中。UNOISE3针对Illumina测序产生的序列，采用一次聚类策略，可分为两个阶段：去除测序和PCR过程中的点错误以及去除嵌合体，从而生成zero-radius OTUs (zOTUs)。

(3) Deblur：2016年Rob Knight团队开发出了一种新颖的基于sub-operational-taxonomic-unit (sOTU)的算法，称为Deblur (Amir*等，* 2017)。ASVs, sub-OTUs, zero-radius OTUs具有相同含义，可统一称为ASV(Nearing*等，* 2018)。Deblur法与DADA2和UNOISE2的去噪方法一样，可以从Illumina数据中获得达到单核苷酸精度的分辨率，不同的是Deblur不仅可以对混合样本进行操作，还能对每个样本进行独立操作，但之后需单独进行去除嵌合体操作。在地球微生物组计划 (The Earth Microbiome Project, EMP)中就使用该方法处理了全球数据 (Thompson*等，* 2017)。

DADA2、UNOISE2、UNOISE3和Deblur使用了不同的算法处理相同的概念，都能更接近真实的生物序列，但之间仍存在差异。在稳定性方面，Deblur优于DADA2；在运行速度上，UNOISE2最快，Deblur次之，DADA2最慢，之间均相差一个数量级(Amir*等，* 2017)。由于低丰度序列更可能为错误序列，故而被舍弃：UNOISE2默认去掉丰度小于4的序列；UNOISE3默认去掉小于8的序列；DADA2默认去掉singletons；Deblur默认去掉所有样本中和小于10的序列和每个样本中的singletons。Jacob T. Nearing等人采用模拟群落、土壤和宿主相关的群落对这三种去噪算法进行了评估，结果表明这三种算法在每个样本的组成上分析一致；对于真实土壤数据和其他两个与宿主相关的数据集，DADA2与其他两个去噪方法相比可以发现更多的ASVs，表明它在发现稀有生物方面可能更好，但可能有假阳性；运行速度上同样UNOISE3比DADA2和Deblur分别快1200倍和15倍以上(Nearing*等，* 2018)。Andrei Prodan等人对模拟群落和荷兰阿姆斯特丹六个民族的成年个体粪便样本，同样使用DADA2, Qiime2-Deblur和USEARCH-UNOISE3进行了比较研究，结果表明DADA2敏感性最好，但特异性降低；UNOISE3显示了分辨率和特异性之间的最佳平衡(Prodan*等，* 2020)。

* 1. 功能预测软件

通常情况下，扩增子分析只能获得菌群分类组成的信息，而功能信息由宏基因组研究较为准确。虽然宏基因组近年来的价格处于下降趋势，但与扩增子相比价格稍高，对于大规模测序分析来说使用宏基因组方法不占优势。因而现已开发出多种可用软件包来预测潜在的功能信息，这种预测的原理是将16S rDNA序列或分类信息与文献中的功能描述联系起来(Liu*等，* 2020)。主要有以下几种：

(1) PICRUSt：全称Phylogenetic Investigation of Communities by Reconstruction of Unobserved States，2013年由Curtis Huttenhower团队开发的最早的一款预测微生物群落功能的工具，可以使用在线版 (<http://huttenhower.sph.harvard.edu/galaxy>)，也可在Linux和Mac OS X系统上下载安装使用(Langille*等，* 2013)。该软件以Greengenes数据库 (McDonald*等，* 2012)为参考序列的OTU信息作为输入，用于预测京都基因和基因组百科全书 (Kyoto Encyclopedia of Genes and Genomes, KEGG) (Kanehisa and Goto 2000)通路或COGs (Clusters of Orthologs Groups) (Tatusov*等，* 1997)的宏基因组功能组成。在预测结果方面，PICRUSt1对人肠道微生物和土壤样本的预测结果最好，对哺乳动物肠道样本有较大波动，而对高盐微生物样本的预测准确度最低 (Langille*等，* 2013)。由于参考OTUs的这一限制，默认的PICRUSt1工作流程无法使用具有更好分辨率的ASVs作为输入序列，此外，PICRUSt1使用的Greengenes参考数据库自2013年以来就没有更新过，缺乏上千个最近增加的基因家族，所以PICRUSt2在2020年发表，弥补了之前版本的不足(Douglas*等，* 2020)。PICRUSt2可以使用OTU或ASV直接进行功能预测，并且一条命令即可完成全部分析，数据库也扩大了10倍，使得使用扩增子数据进行功能预测的精确性更高。

(2) Tax4Fun：是2015年发布的一个基于16S rRNA数据预测微生物群落功能的开源R包，适用于从SILVAngs网络服务器或QIIME应用程序获得的输出，没有在线的网页版，只可进行线下分析。Tax4Fun与PICRUSt的区别主要有两方面：一是数据库差异，Tax4Fun是基于SILVA数据库，而PICRUSt是基于Greengenes数据库；二是测序原理的差异，PICRUSt中一定比例OTU的基因组是经祖先状态重构算法 (ancestral-state reconstruction algorithm) 预测出来的，而Tax4Fun是基于KEGG库中已测序注释的原核基因组信息。因而Tax4Fun开发者指出与PICRUSt工具相比，使用Tax4Fun进行功能预测的结果与宏基因组图谱的相关性更高(Aßhauer*等，* 2015)。随着测序技术和数据库的不断发展，Tax4Fun在2020年发表了新版本Tax4Fun2 (Wemheuer*等，* 2020)。Tax4Fun2是用于从16S rRNA序列预测原核生物群落功能谱和功能基因冗余的R包，但也可以用于真核生物的合并。默认的参考数据集包括275个古菌基因组和12,002个细菌基因组。为了研究微生物群落是否包含了功能上的冗余成员以及程度如何，得以使之在面对多变环境时保持生态系统的相对稳定，Tax4Fun2针对单个功能通路引入了功能冗余指数 (functional redundancy index, FRI)的概念。除此之外Tax4Fun2的一个新特点是可以合并用户自定义的数据集，提高了功能预测的准确性和稳健性 (Wemheuer*等，* 2020)。

(3) Bugbase：于2017年发表的一个使用全基因组鸟枪或标记基因测序数据预测复杂微生物群表型的算法，包括网页版 (<http://bugbase.cs.umn.edu>) 和免费下载版 (<https://github.com/knights-lab/BugBase>)，可根据革兰氏阳性 (Gram Positive)、革兰氏阴性 (Gram Negative)、氧的利用 (Oxygen Utilizing)、氧化胁迫耐受 (Oxidative Stress Tolerant)、生物膜形成 (Biofilm Forming)、致病性 (Pathogenic)、移动元件含量 (Mobile Element Containing) 七个方面对微生物进行分类(Ward*等，* 2017)。

(4) FAPROTAX：全称Functional Annotation of Prokaryotic Taxa，是2016年发表的基于原核微生物分类的功能注释数据库，根据文献资料手动构建的，包含了80多个功能分组 (如硝酸盐呼吸、产甲烷、发酵、植物病原等)，超过7,600个功能注释，涵盖4,600个分类单元 ([www.zoology.ubc.ca/louca/FAPROTAX](http://www.zoology.ubc.ca/louca/FAPROTAX))。同时可以通过SILVA或Greengenes数据库生成的OTU表来对微生物群落功能进行预测(Louca*等，* 2016)。更适用于农业、环境等相关领域研究菌种功能 (Liu*等，* 2019)。

1. **16S扩增子分析常用数据库**

(1) SILVA：源于拉丁文，是一个最全面并定期更新的数据库，提供了细菌、古菌和真核生物三域的16S/18S小亚基 (small subunit rRNA Gene, SSU) 和23S/28S大亚基 (large subunit rRNA gene, LSU) 核糖体RNA (rRNA) 序列，自2007年第一次发布以来，SILVA项目已经发布了16个完整版本，2020年8月7日已更新到138.1版本，更正了SSU分类标准，并更新了LSU序列数据和分类标准，此外还包括在线分析工具SILVAngs (<https://www.arb-silva.de/ngs>) (Quast*等，* 2013)。

(2) Greengenes：是16S rRNA专用数据库，虽然自2013年之后没有再更新，但成为QIIME推荐数据库，PICRUSt1和Bugbase也基于该数据库 (McDonald*等，* 2012)。

(3) RDP：全称Ribosomal Database Project，提供了已注释的细菌和古菌小亚基rRNA基因和真菌大亚基rRNA基因，目前最新版本为2019年11月更新的11.5版，包括3,356,809条16S rRNA和125,525条真菌的28S rRNA，此外也包括在线分析流程 (<http://rdp.cme.msu.edu/>) (Cole*等，* 2014)。

(4) UNITE：是一个专注于真核生物ITS区的数据库和序列管理环境，包括了1,000,000个公共真菌ITS序列，在真菌ITS扩增子测序分析中用于嵌合体检测和物种分类(Nilsson*等，* 2019)。

1. **总结与展望**

下一代测序 (NGS) 技术的进步促进了微生物领域的快速扩展，包括人类微生物组计划 (Human Microbiome Project, HMP) 在内的多项国际合作基因组研究均已完成(Integrative 2019)。其中16S rRNA基因扩增子测序已成为研究细菌多样性和系统发育的基石，其以低成本、高效率的特点在人类(Cho and Blaser 2012)、土壤(Hartmann*等，* 2014)、海洋 (Moran 2015) 等各方面的研究中发挥了重要作用。但接踵而来的问题是无法从大量的测序数据中直接看出其中存在的现象规律，这就需要一系列计算工具和分析数据的方法对数据进行下游多样性、关联和相关性分析等。

目前对于16S rRNA扩增子分析来说，使用最多的三大分析软件为mothur、QIIME和USEARCH，引用均已过万。以Illumina平台下机数据为例，拿到的原始序列 (raw amplicon) 需要进行双端合并后去除barcode和引物，质控步骤去除低质量序列和嵌合体，得到干净的序列 (clean amplicon) 以进行后续分析，这些步骤均可以使用USEARCH和QIIME完成 (Liu*等，* 2020)。之后需要挑选代表性序列以减少测序错误带来的影响 (在Illumina测序中，每个核苷酸的错误率约为0.1%)，包括OTUs聚类和ASVs去噪两种方法。之前的方法是通过UPARSE等算法将相似序列 (通常阈值设为97%) 聚类成OTUs，但这种方法漏掉了细微而真实的生物序列变异，因而更推荐使用DADA2、QIIME2-Deblur、USEARCH、UNOISE3等去噪算法挑选代表性序列。其中DADA2和Deblur结果相似，但是Deblur支持并行处理，速度快且稳定，故而Rob Knight教授更推荐使用Deblur算法进行去噪分析(Knight*等，* 2018)。得到代表性序列后，需将这些序列比对到Greengenes、RDP和Silva等数据库当中获得序列的物种分类信息，该步骤可以通过例如QIIME和mothur等软件进行 (Knight*等，* 2018)。一般情况下，16S扩增子分析只能得到菌群分类组成上的信息，但由于PICRUSt、Tax4Fun、FAPROTAX、Bugbase等预测软件的出现，使得通过扩增子数据获得物种功能信息变成可能。

微生物组分析方法和标准正在迅速发展，但还没有公认的统一的标准，故而无法确定微生物组学研究的最优法。尽管近期开发出了NIBSC (National Institute for Biological Standards and Control) DNA参考菌群Gut-Mix-RR和Gut-HiLo-RR，以及用于评估生物信息学工具流程偏差的四项措施框架，但还需要多方合作寻求特定的参考试剂确保正确的基准化 (Amos*等，* 2020)。

**基金项目**

“中央高校基本科研业务费专项资金资助(2018ZY21)”(supported by “the Fundamental Research Funds for the Central Universities(2018ZY21)”)

**参考文献**

1. Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., Kightley, E. P., Thompson, L. R., Hyde, E. R., Gonzalez, A. and Knight, R. (2017). Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems* 2(2). <https://doi.org/10.1128/mSystems.00191-16>

2. Amos, G. C. A., Logan, A., Anwar, S., Fritzsche, M., Mate, R., Bleazard, T. and Rijpkema, S. (2020). Developing standards for the microbiome field. *Microbiome* 8(1): 98. <https://doi.org/10.1186/s40168-020-00856-3>

3. Aßhauer, K. P., Wemheuer, B., Daniel, R. and Meinicke, P. (2015). Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data *Bioinformatics* 31(17): 2882-2884. <https://doi.org/10.1093/bioinformatics/btv287>

4. Behjati, S. and Tarpey, P. S. (2013). What is next generation sequencing? *Arch Dis Child Educ Pract Ed* 98(6): 236-238. <https://doi.org/10.1136/archdischild-2013-304340>

5. Berkman, P. J., Lai, K., Lorenc, M. T. and Edwards, D. (2012). Next-generation sequencing applications for wheat crop improvement. *Am J Bot* 99(2): 365-371. <https://doi.org/10.3732/ajb.1100309>

6. Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodriguez, A. M., Chase, J., Cope, E. K., Da Silva, R., Diener, C., Dorrestein, P. C., Douglas, G. M., Durall, D. M., Duvallet, C., Edwardson, C. F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J. M., Gibbons, S. M., Gibson, D. L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G. A., Janssen, S., Jarmusch, A. K., Jiang, L., Kaehler, B. D., Kang, K. B., Keefe, C. R., Keim, P., Kelley, S. T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M. G. I., Lee, J., Ley, R., Liu, Y. X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B. D., McDonald, D., McIver, L. J., Melnik, A. V., Metcalf, J. L., Morgan, S. C., Morton, J. T., Naimey, A. T., Navas-Molina, J. A., Nothias, L. F., Orchanian, S. B., Pearson, T., Peoples, S. L., Petras, D., Preuss, M. L., Pruesse, E., Rasmussen, L. B., Rivers, A., Robeson, M. S., 2nd, Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S. J., Spear, J. R., Swafford, A. D., Thompson, L. R., Torres, P. J., Trinh, P., Tripathi, A., Turnbaugh, P. J., Ul-Hasan, S., van der Hooft, J. J. J., Vargas, F., Vazquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K. C., Williamson, C. H. D., Willis, A. D., Xu, Z. Z., Zaneveld, J. R., Zhang, Y., Zhu, Q., Knight, R. and Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 37(8): 852-857. <https://doi.org/10.1038/s41587-019-0209-9>

7. Callahan, B. J., McMurdie, P. J. and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J* 11(12): 2639-2643. <https://doi.org/10.1038/ismej.2017.119>

8. Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. and Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* 13(7): 581-583. <https://doi.org/10.1038/nmeth.3869>

9. Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Pena, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J. and Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7(5): 335-336. <https://doi.org/10.1038/nmeth.f.303>

10. Cho, I. and Blaser, M. J. (2012). The human microbiome: at the interface of health and disease. *Nat Rev Genet* 13(4): 260-270. <https://doi.org/10.1038/nrg3182>

11. Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R. and Tiedje, J. M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res* 42(Database issue): D633-642. <https://doi.org/10.1093/nar/gkt1244>

12. Douglas, G. M., Maffei, V. J., Zaneveld, J. R., Yurgel, S. N., Brown, J. R., Taylor, C. M., Huttenhower, C. and Langille, M. G. I. (2020). PICRUSt2 for prediction of metagenome functions. *Nat Biotechnol* 38(6): 685-688. <https://doi.org/10.1038/s41587-020-0548-6>

13. Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19): 2460-2461. <https://doi.org/10.1093/bioinformatics/btq461>

14. Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* 10(10): 996-998. <https://doi.org/10.1038/nmeth.2604>

15. Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*. <https://doi.org/10.1101/081257>

16. Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* 34(14): 2371-2375. <https://doi.org/10.1093/bioinformatics/bty113>

17. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27(16): 2194-2200. <https://doi.org/10.1093/bioinformatics/btr381>

18. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J. and Turner, S. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* 323(5910): 133-138. <https://doi.org/10.1126/science.1162986>

19. Goodwin, S., McPherson, J. D. and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 17(6): 333-351. <https://doi.org/10.1038/nrg.2016.49>

20. Hartmann, M., Niklaus, P. A., Zimmermann, S., Schmutz, S., Kremer, J., Abarenkov, K., Luscher, P., Widmer, F. and Frey, B. (2014). Resistance and resilience of the forest soil microbiome to logging-associated compaction. *ISME J* 8(1): 226-244. <https://doi.org/10.1038/ismej.2013.141>

21. Integrative, H. M. P. R. N. C. (2019). The Integrative Human Microbiome Project. *Nature* 569(7758): 641-648. <https://doi.org/10.1038/s41586-019-1238-8>

22. Jain, M., Fiddes, I. T., Miga, K. H., Olsen, H. E., Paten, B. and Akeson, M. (2015). Improved data analysis for the MinION nanopore sequencer. *Nat Methods* 12(4): 351-356. <https://doi.org/10.1038/nmeth.3290>

23. Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28(1): 27-30. <https://doi.org/10.1093/nar/28.1.27>

24. Kechin, A., Boyarskikh, U., Kel, A. and Filipenko, M. (2017). cutPrimers: A New Tool for Accurate Cutting of Primers from Reads of Targeted Next Generation Sequencing. *J Comput Biol* 24(11): 1138-1143. <https://doi.org/10.1089/cmb.2017.0096>

25. Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L. I., McDonald, D., Melnik, A. V., Morton, J. T., Navas, J., Quinn, R. A., Sanders, J. G., Swafford, A. D., Thompson, L. R., Tripathi, A., Xu, Z. Z., Zaneveld, J. R., Zhu, Q., Caporaso, J. G. and Dorrestein, P. C. (2018). Best practices for analysing microbiomes. *Nat Rev Microbiol* 16(7): 410-422. <https://doi.org/10.1038/s41579-018-0029-9>

26. Kumar, K. R., Cowley, M. J. and Davis, R. L. (2019). Next-Generation Sequencing and Emerging Technologies. *Semin Thromb Hemost* 45(7): 661-673. <https://doi.org/10.1055/s-0039-1688446>

27. Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., Clemente, J. C., Burkepile, D. E., Vega Thurber, R. L., Knight, R., Beiko, R. G. and Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* 31(9): 814-821. <https://doi.org/10.1038/nbt.2676>

28. Leonelli, S. (2019). The challenges of big data biology. *Elife* 8. <https://doi.org/10.7554/eLife.47381>

29. Liu, Y., Qin, Y., Guo, X. X. and Bai, Y. (2019). [Methods and applications for microbiome data analysis]. *Yi Chuan* 41(9): 845-862. <https://doi.org/10.16288/j.yczz.19-222>

30. Liu, Y. X., Qin, Y., Chen, T., Lu, M., Qian, X., Guo, X. and Bai, Y. (2020). A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein Cell*. <https://doi.org/10.1007/s13238-020-00724-8>

31. Louca, S., Parfrey, L. W. and Doebeli, M. (2016). Decoupling function and taxonomy in the global ocean microbiome. *Science* 353(6350): 1272-1277. <https://doi.org/10.1126/science.aaf4507>

32. Lozupone, C. and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 71(12): 8228-8235. <https://doi.org/10.1128/AEM.71.12.8228-8235.2005>

33. Mahnert, A., Moissl-Eichinger, C., Zojer, M., Bogumil, D., Mizrahi, I., Rattei, T., Martinez, J. L. and Berg, G. (2019). Man-made microbial resistances in built environments. *Nature Communications* 10(1). <https://doi.org/10.1038/s41467-019-08864-0>

34. Mandal, S., Van Treuren, W., White, R. A., Eggesbo, M., Knight, R. and Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb Ecol Health Dis* 26: 27663. <https://doi.org/10.3402/mehd.v26.27663>

35. Marchesi, J. R. and Ravel, J. (2015). The vocabulary of microbiome research: a proposal. *Microbiome* 3: 31. <https://doi.org/10.1186/s40168-015-0094-5>

36. McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R. and Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* 6(3): 610-618. <https://doi.org/10.1038/ismej.2011.139>

37. Moran, M. A. (2015). The global ocean microbiome. *Science* 350(6266): aac8455. <https://doi.org/10.1126/science.aac8455>

38. Nearing, J. T., Douglas, G. M., Comeau, A. M. and Langille, M. G. I. (2018). Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ* 6: e5364. <https://doi.org/10.7717/peerj.5364>

39. Nilsson, R. H., Larsson, K. H., Taylor, A. F. S., Bengtsson-Palme, J., Jeppesen, T. S., Schigel, D., Kennedy, P., Picard, K., Glockner, F. O., Tedersoo, L., Saar, I., Koljalg, U. and Abarenkov, K. (2019). The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res* 47(D1): D259-D264. <https://doi.org/10.1093/nar/gky1022>

40. Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A. H., Nieuwdorp, M. and Levin, E. (2020). Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLoS One* 15(1): e0227434. <https://doi.org/10.1371/journal.pone.0227434>

41. Qian, X. B., Chen, T., Xu, Y. P., Chen, L., Sun, F. X., Lu, M. P. and Liu, Y. X. (2020). A guide to human microbiome research: study design, sample collection, and bioinformatics analysis. *Chin Med J (Engl)* 133(15): 1844-1855. <https://doi.org/10.1097/CM9.0000000000000871>

42. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glockner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 41(Database issue): D590-596. <https://doi.org/10.1093/nar/gks1219>

43. Rognes, T., Flouri, T., Nichols, B., Quince, C. and Mahe, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4: e2584. <https://doi.org/10.7717/peerj.2584>

44. Rosen, M. J., Callahan, B. J., Fisher, D. S. and Holmes, S. P. (2012). Denoising PCR-amplified metagenome data. *BMC Bioinformatics* 13: 283. <https://doi.org/10.1186/1471-2105-13-283>

45. Schadt, E. E., Turner, S. and Kasarskis, A. (2010). A window into third-generation sequencing. *Hum Mol Genet* 19(R2): R227-240. <https://doi.org/10.1093/hmg/ddq416>

46. Schloss, P. D. (2020). Reintroducing mothur: 10 Years Later. *Appl Environ Microbiol* 86(2): e02343-02319. <https://doi.org/10.1128/AEM.02343-19>

47. Schloss, P. D. and Handelsman, J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* 71(3): 1501-1506. <https://doi.org/10.1128/AEM.71.3.1501-1506.2005>

48. Schloss, P. D. and Handelsman, J. (2006). Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. *Appl Environ Microbiol* 72(10): 6773-6779. <https://doi.org/10.1128/AEM.00474-06>

49. Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J. and Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75(23): 7537-7541. <https://doi.org/10.1128/AEM.01541-09>

50. Stackebrandt, E. and Goebel, B. M. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16s rRNA Sequence Analysis in the Present Species Definition in Bacteriology *Int.J.syst.bacteriol* 44(4): 846-849. <https://doi.org/10.1099/00207713-44-4-846>

51. Tatusov, R. L., Koonin, E. V. and Lipman, D. J. (1997). A genomic perspective on protein families. *Science* 278: 631-637. <https://doi.org/10.1126/science.278.5338.631>

52. Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., Prill, R. J., Tripathi, A., Gibbons, S. M., Ackermann, G., Navas-Molina, J. A., Janssen, S., Kopylova, E., Vazquez-Baeza, Y., Gonzalez, A., Morton, J. T., Mirarab, S., Zech Xu, Z., Jiang, L., Haroon, M. F., Kanbar, J., Zhu, Q., Jin Song, S., Kosciolek, T., Bokulich, N. A., Lefler, J., Brislawn, C. J., Humphrey, G., Owens, S. M., Hampton-Marcell, J., Berg-Lyons, D., McKenzie, V., Fierer, N., Fuhrman, J. A., Clauset, A., Stevens, R. L., Shade, A., Pollard, K. S., Goodwin, K. D., Jansson, J. K., Gilbert, J. A., Knight, R. and Earth Microbiome Project, C. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 551(7681): 457-463. <https://doi.org/10.1038/nature24621>

53. Tikhonov, M., Leach, R. W. and Wingreen, N. S. (2015). Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *ISME J* 9(1): 68-80. <https://doi.org/10.1038/ismej.2014.117>

54. van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. and Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends Genet* 34(9): 666-681. <https://doi.org/10.1016/j.tig.2018.05.008>

55. Ward, T., Larson, J., Meulemans, J., Hillmann, B., Lynch, J., Sidiropoulos, D., Spear, J. R., Caporaso, G., Blekhman, R., Knight, R., Fink, R. and Knights, D. (2017). BugBase predicts organism-level microbiome phenotypes. *bioRxiv*: 133462. <https://doi.org/10.1101/133462>

56. Wemheuer, F., Taylor, J. A., Daniel, R., Johnston, E., Meinicke, P., Thomas, T. and Wemheuer, B. (2020). Prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. *Environmental Microbiome* 15(11): 1-12. <https://doi.org/10.1186/s40793-020-00358-7>

57. Westcott, S. L. and Schloss, P. D. (2015). De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* 3: e1487. <https://doi.org/10.7717/peerj.1487>

58. Yang, Y., Xie, B. and Yan, J. (2014). Application of next-generation sequencing technology in forensic science. *Genomics Proteomics Bioinformatics* 12(5): 190-197. <https://doi.org/10.1016/j.gpb.2014.09.001>

59. Yohe, S. and Thyagarajan, B. (2017). Review of Clinical Next-Generation Sequencing. *Arch Pathol Lab Med* 141(11): 1544-1557. <https://doi.org/10.5858/arpa.2016-0501-RA>

60. Zhang, J., Ding, X., Guan, R., Zhu, C., Xu, C., Zhu, B., Zhang, H., Xiong, Z., Xue, Y., Tu, J. and Lu, Z. (2018). Evaluation of different 16S rRNA gene V regions for exploring bacterial diversity in a eutrophic freshwater lake. *Sci Total Environ* 618: 1254-1267. <https://doi.org/10.1016/j.scitotenv.2017.09.228>

61. 吴悦妮, 冯凯, 厉舒祯, 王朱珺, 张照婧 and 邓晔 (2020). 16S-18S-ITS扩增子高通量测序引物的生物信息学评估和改进. *微生物学通报*. <https://doi.org/10.13344/j.microbiol.china.200054>

62. 张军毅, 朱冰川, 徐超, 丁啸, 李俊锋, 张学工 and 陆祖宏 (2015). 基于分子标记的宏基因组16S rRNA基因高变区选择策略. *应用生态学报* 26(11): 3545-3553. <https://doi.org/10.13287/j.1001-9332.20150812.005>