扩增子和宏基因组数据分析实用指南

刘永鑫[1,2,3]✉, 秦媛[1,2,3,4], 陈同[5], 卢美萍[6], 钱旭波[6], 郭晓璇[1,2,3], 白洋[1,2,3,4]✉

1. 中国科学院遗传与发育生物学研究所，植物基因组学国家重点实验室
2. 中国科学院大学，生物互作卓越创新中心
3. 中国科学院遗传与发育生物学研究所，中国科学院-英国约翰英纳斯中心植物和微生物科学联合研究中心
4. 中国科学院大学，现代农学院
5. 中国中医科学院中药资源中心
6. 浙江大学医学院附属儿童医院风湿、免疫和变态反应科

✉通讯作者: yxliu@genetics.ac.cn（刘永鑫）、ybai@genetics.ac.cn（白洋）

刘永鑫、秦媛、陈同为共同第一作者

摘要

近年来高通量测序技术的发展促进了一系列适合微生物组研究的技术发展，同时也积累了海量数据。然而，微生物组数据分析过程复杂、分析工具种类较多，这限制了广大研究者进入该领域。本文系统概述了微生物组常用测序技术——扩增子和宏基因组等方法的优缺点，推荐了常用软件、分析流程和数据库，以便研究者选择恰当的分析工具和方法。此外，我们还介绍了微生物组下游分析通用的统计和可视化方法，包括多样性分析、物种组成分析、差异分析、相关分析、网络分析、机器学习、进化分析、来源追溯以及常用可视化方法。最后我们还介绍了可重复分析方法。我们希望研究者通过阅读此文能学会分析工具的选择方法和高效的数据分析方法，进而有效地挖掘数据背后的生物学意义。

关键字

微生物组、宏基因组、标记基因、高通量测序、分析方法、分析流程、可重复性、可视化

1. 概述

微生物组是指整个微生境，包括微生物、基因组和周围环境。随着高通量测序（high-throughput sequencing, HTS）技术和数据分析方法的发展，近年来微生物组在人类、动物、植物和环境中所起的作用变得越来越清晰。这些研究成果彻底改变了我们对微生物组的理解。许多国家已经成功启动了国际微生物组研究计划，例如 NIH 人类微生物组计划（HMP）、人类肠道宏基因组学计划（MetaHIT）、整合 HMP（iHMP）和中国科学院微生物组计划（CAS-CMI）。这些项目都取得了令人瞩目的成就，将微生物组研究推向了黄金时代。

扩增子和宏基因组学分析的框架是在近十年建立的。但是，微生物组分析方法及标准在近年来有着迅猛地发展。例如，有人提出了在扩增子数据分析中用扩增子序列变异体（amplicon sequence variants, ASV）代替操作分类单位（operational taxonomic units, OTU）的建议。又如，最近发布了下一代微生物组分析软件 QIIME 2，它是一种可重复、交互式、高效、有社区支持的分析平台。此外，最近也提出了许多新的方法用于物种分类、机器学习和多组学整合分析。

高通量测序和分析方法的发展为深入了解微生物组的结构和功能提供了新的手段。但是，这些新的发展让研究人员，特别是那些没有生物信息学背景的研究人员在选择合适的软件和分析流程时中变得颇为费劲。为了让研究者了解本领域最新进展，并让大家快速掌握相应的软件选择和使用技巧，我们特推出此综述。在这篇文章中，我们将讨论广泛用于微生物组分析的软件包，总结他们的优点和局限性，并提供了示例代码和选择使用这些工具的窍门。

2. 高通量测序方法介绍

微生物组研究的第一步是了解高通量测序方法具体的优点和局限性。这些测序方法主要用于三个层面的分析：微生物、DNA 和 mRNA 层面（**图 1A**），研究者应根据样本类型和研究目标选择适合的方法。

培养组学（Culturome）是一种在微生物层面培养和鉴定微生物的高通量方法（**图 1A**）。微生物的获取方法如下。首先，将样品破碎，根据经验在液体培养基中将其稀释，然后分散至 96 孔细胞培养板或培养皿中。第二步，将培养板在室温下培养 20 天。第三步，对每个孔中的微生物进行扩增子测序，并选择纯度高、非冗余菌落孔中的样本作为候选样本。第四步，纯化候选样本并进行 16S rDNA 全长 Sanger 测序。最后保留纯化的分离株。培养组学是获得细菌种群最有效的方法，但它昂贵且劳动强度大（**图 1B**）。这种方法在人类、小鼠、海洋沉积物、拟南芥和水稻的微生物组研究中得到开展。这些研究不仅进一步完善了宏基因组学分析的物种分类和功能基因的数据库，而且还为开展功能实验验证提供了细菌基础材料。如想了解培养组学更多的信息请参阅这两篇文献(Lagier et al., 2018; Liu et al., 2019)。

DNA 易于提取、保存和测序，这使研究人员能够开发各种高通量测序的方法（**图 1A**）。微生物组最常用的高通量测序方法是扩增子和宏基因组测序（**图 1B**）。扩增子测序是微生物组分析中使用最广泛的测序方法，几乎可以应用于所有类型的样品。扩增子测序中使用的主要标记基因包括用于原核生物的 16SrDNA、用于真核生物的 18S rDNA 和内转录间隔区（internal transcribed spacers, ITS）。16S rDNA 扩增子测序是最常用的方法，但是目前可用引物列表较为混乱。选择引物的一个好方法是先评估其特异性和总体覆盖度，这个过程要么利用真实样品进行，也可以基于 SILVA 数据库和宿主因素（包括叶绿体、线粒体、核糖体和其他非特异性扩增的潜在来源）进行电子 PCR。另一个替代方案是研究者参考与自己研究相似的已发表论文中使用的引物，这样可以节省优化方法的时间，也便于研究结果之间的比较。两步 PCR 法通常用于扩增子文库制备，扩增同时向每个样品添加标签（barcode）和接头序列。样品测序通常在 Illumina MiSeq、HiSeq2500 或 NovaSeq6000 平台上进行，产出双端 250bp（PE250）的序列，每个样品含有 5~10 万条序列。扩增子测序可应用于低生物

含量标本或被宿主 DNA 污染的样品。然而，该技术只能达到“属”级分辨率（主要是由于测序片段较短，通常仅 300~500 bp）。此外，它的可靠性受引物和 PCR 循环次数影响，这可能导致下游分析出现假阳性或假阴性结果（**图 1B**）。

宏基因组测序比扩增子测序提供了更多的信息，但是这种技术较昂贵。对于人类粪便等“纯”样本，每个样本可接受的测序数据量从 6~9 GB 不等。文库构建和测序的相应价格在 100~300 美元之间（700~2000 元，测序量和耗材纯净级别对价格影响很大）。对于包含复杂微生物群（如土壤）或受宿主 DNA 污染的样品，每个样本所需的测序量需要 30~300 GB。总之，16S rDNA 扩增子测序可以用于研究细菌和/或古菌的组成，如果需要更高的物种分类学分辨率和功能信息，则可随机抽取部分样本进行宏基因组测序。当然，假设有足够的可用资金，宏基因组测序可直接用于样本量较小的研究。

转录组测序可以分析微生物组中的 mRNA，量化基因表达水平，并可提供微生物群落的功能信息。值得注意的是，为了获得微生物组的转录信息，需要有效去除宿主 RNA 和所有种类的 rRNA（**图 1B**）。

由于病毒有 DNA 或 RNA 作为其遗传物质，从技术上讲，宏病毒组研究包含宏基因组和宏转录组（图 1A/B）。由于样品中病毒的生物含量较低，因此要获得足够数量的病毒 DNA 或 RNA 进行分析，必须进行病毒富集或去除宿主 DNA/RNA（**图 1B**）。

测序方法的选择取决于科学问题和样本类型。整合使用不同的方法是很多学者推荐的，因为多组学方法可同时获得微生物组分类和功能的信息。实际上，由于时间和成本的限制，大多数研究人员只能选择一种或两种测序方法。尽管扩增子测序只能提供微生物群的物种分类学组成信息，但它具有成本效益优势（每个样品仅需 150~350 元），可用于大规模研究。此外，扩增子测序产生的数据量相对较小，分析快速且容易进行。例如，使用普通的便携式计算机在一天之内就可以完成数百个扩增子样品的数据分析。因此，扩增子测序通常用于探索性研究。与扩增子测序相反，宏基因组测序不仅将分类学分辨率扩展到“种”或“株”的水平，而且还提供了潜在的功能信息。宏基因组测序也使得从短片段组装微生物基因组成为可能。但是，对于低生物含量或被宿主基因组严重污染的样品，宏基因组测序并非目前的最佳选择（**图 1B**）。

图 1 各种高通量测序方法的优势和局限性。A 介绍了用于不同分析层次的测序方法。在分子水平上，微生物组研究分为三种层面：微生物、DNA 和 mRNA。相应的研究技术包括培养组、扩增子、宏基因组、宏病毒组和宏转录组。B 用于微生物组研究测序方法的优缺点。

3. 分析流程

“分析流程”指的是特定程序或脚本，该程序或脚本以一定顺序整合了数个甚至数十个软件程序，用来完成复杂的分析任务。截至 2020 年 6 月 23 日，在 Google 学术中提及“amplicon”和“metagenome”分别超过 24 万和 4 万次。由于他们的广泛使用，我们将讨论用于扩增子和宏基因组分析的当前最佳流程。研究人员应该熟悉 Shell 环境和 R 语言，这在我们之前的综述中曾经讨论过（刘永鑫等，2019，遗传）。

3.1 扩增子分析

扩增子分析的第一阶段任务是将原始序列（一般是 fastq 格式）转换为特征表。原始序列通常以双端 250 bp（PE250）形式从 Illumina 测序平台生成。本文未讨论如 Ion Torrent，PacBio 和 Nanopore 等测序平台，他们产生的数据可能不适合下面讨论的分析流程。分析时，首先将原始序列根据标签（barcode）进行分组，这个过程又叫“拆分”（demultiplexing）。然后将序列合并以获得扩增子序列，并去除标签和引物。通常还需要质量控制步骤以去除低质量的扩增子序列。所有这些步骤都可以使用 USEARCH 或 QIIME 完成，或者也可选择测序服务公司提供的纯净扩增子数据用于下一步分析（**图 2A**）。

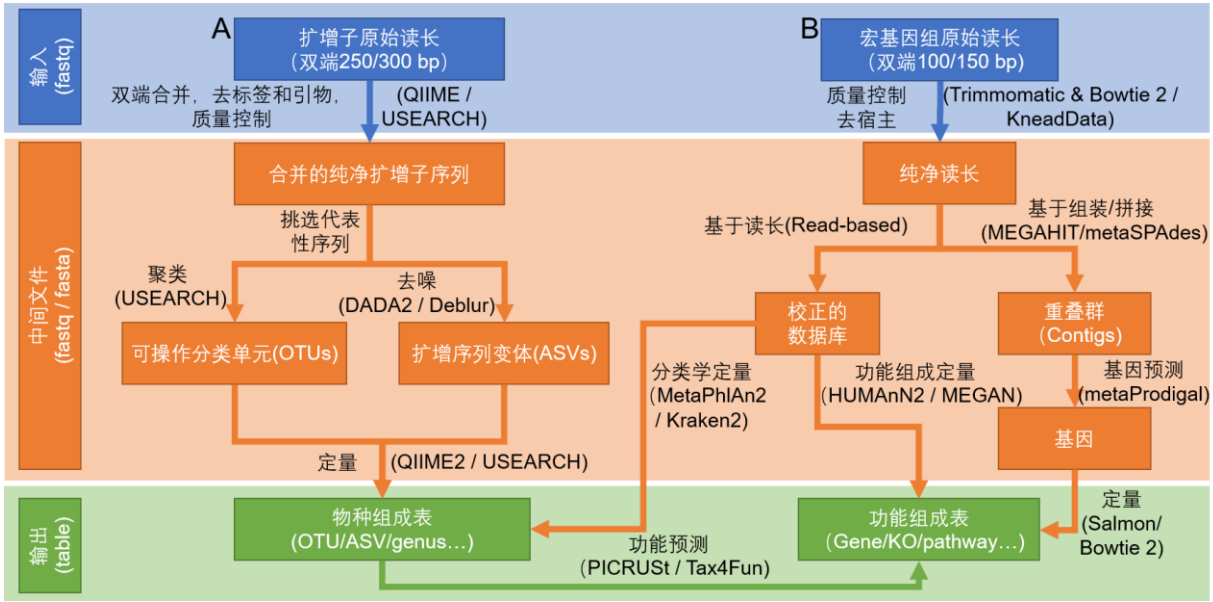

图 2 扩增子(A)和宏基因组(B)测序常用流程。蓝色、橙色和绿色分别代表输入、中间及输出文件。箭头边上的文字代表方法，括号中的是常用软件。物种和功能表格统称为特征表。关于这方面更详细的信息请阅读表 1。

挑选出代表性序列作为物种的代表是扩增子分析的关键步骤，主要包括聚类生成 OTU 和去噪生成 ASV 两类方法。UPARSE 算法将具有 97%相似性的序列聚类为 OTU，但此方法可能无法检测“种”或“株”之间的细微差异。DADA2 是最近开发的一种去噪算法，可挑选出更准确的代表性序列——ASV。QIIME 2 流程中有二种去噪方法可选，即 DADA2 插件的 *denoise-paired/single* 和 Deblur 插件的 *denoise-16S*，此外 USEARCH 中的*-unoise3* 也可用于高速去噪并挑选 ASV。最后，可以通过量化每个样本中特征序列的频率来获得特征表，即 OTU 或 ASV 表。同时，可以对特征序列进行分类，通常在界，门，纲，目，科，属和种的层级上进行分类，从而为微生物群提供了更多级别的降维视角。

通常，16S rDNA 扩增子测序只能用于获得有关物种分类组成的信息。但是近几年开发了许多可用的软件包来预测潜在的功能信息。该预测的原理是将 16S rDNA 序列或分类学信息与数据库中的

基因组或文献中的功能描述联系起来。PICRUSt 是一种基于 Greengenes 数据库 OTU 表的功能预测软件，可用于预测如 KEGG 通路的宏基因组功能组成信息。新开发的 PICRUSt2 软件包（https://github.com/picrust/picrust2）可以基于任意 OTU / ASV 表直接预测宏基因组功能。R 包 Tax4Fun 可以基于 SILVA 数据库预测微生物群的 KEGG 功能。原核生物分类功能注释（FAPROTAX）流程基于已发表微生物的代谢和生态功能执行功能注释，例如硝酸盐呼吸、铁呼吸、植物病原体、动物寄生虫或共生体，从而使其可用于对环境、农业和动物微生物组的功能分类和研究。BugBase 是 Greengenes 的扩展数据库，用于预测表型，例如需氧性、革兰氏染色和致病性，该数据库常用于医学研究。

3.2 宏基因组分析

宏基因组测序数据比扩增子测序提供了更高精度的物种组成信息，同时还能提供功能基因的信息，但数据量大、分析过程涉及软件众多，而且一般只能在高性能 Linux 服务器上开展分析。宏基因组相关软件安装推荐使用 Bioconda 频道，可有高效安装所需的软件和流程，并自动化解决依赖关系。宏基因组分析计算量大，多任务并行需要队列管理软件防止拥挤，如 GNU Parallel 软件。

Illumina HiSeqX/NovaSeq 系统产出原始数据一般是 PE150 数据，华大 BGI Seq500 产生的为 PE100 数据。宏基因组数据分析的第一个关键的步骤是质量控制和去除宿主污染，这些步骤需要 KneadData 流程（https://bitbucket.org/biobakery/kneaddata）或联合使用 Trimmomatic 和 Bowtie 2。Trimmomatic 是一种灵活的质量控制软件包，适用于 Illumina 测序数据，可进行修剪低质量序列、文库引物和接头序列。使用 Bowtie 2 软件，那些与宿主基因匹配的序列将被滤除。KneadData 是一个集成的分析流程，包括 Trimmomatic、Bowtie 2 和相关脚本，可用于质量控制，滤除宿主来源的序列，并输出纯净序列（**图 2B**）。

宏基因组学分析的主要步骤是使用基于序列和/或基于组装的方法将纯净数据转换为物种组成表和功能组成表。基于序列的方法是直接比对纯净序列至预定义的参考数据库，可直接获得特征表（图 2B）。MetaPhlAn2 是一种常用的生物物种分类学分析工具，可将宏基因序列与预定义的标记基因数据库比对，从而进行生物分类。Kraken 2 是基于精确 *k*-mer 匹配方法将序列与 NCBI 中非冗余序列数据库进行匹配，利用最低共同祖先（lowest common ancestor, LCA）算法进行物种分类。分类学的软件众较，其优缺点和适用范围详见 2019 年 Cell 发表的关于 20 种分类软件评测综述(Ye et al., 2019)。HUMAnN2 是一种广泛使用的功能定量分析软件，特色是可以用于探索样本内和样本间的贡献多样性（物种对特定功能的贡献）。MEGAN 是一种跨平台的图形用户界面（GUI）软件，可进行分类和功能分析（**表 1**）。此外，一些研究提供了特定研究对象的宏基因组参考集，他们可以实现更高的数据利用率和更高质量的物种和功能注释，例如人类肠道 2.0(Li et al., 2014)、小鼠肠道(Xiao et al., 2015)、鸡肠道(Huang et al., 2018)、海洋 OM-RGC.v2(Salazar et al., 2019)、柑橘根际(Xu et al., 2018)等。

基于组装的方法使用 MEGAHIT 或 metaSPAdes 等工具将纯净序列组装为长序列，即重叠群(contigs)（**图 2B**）。MEGAHIT 用于快速装配大量、复杂的宏基因组数据集，而且内存占用小；而 metaSPAdes 通常可以生成更长的重叠群，但需要更多的计算资源。metaGeneMark 或 Prokka 等软件可以识别出重叠群中的基因。已经装配好的重叠群中的冗余基因需要用 CD-HIT 等工具去除。最后，可以使用基于比对的工具 Bowtie 2 或非比对的方法 Salmon 来生成基因丰度表。宏基因组数据集中基因数量通常在百万级别，需要结合蛋白数据库的层级功能注释实现降维，如 KEGG 中的 KO、模块或通路表。

此外，宏基因组学数据可用于挖掘基因簇或组装微生物基因组草图。AntiSMASH 软件和数据库可以挖掘重叠群中潜在的生物合成基因簇，为挖掘新功能基因、酶、代谢通路、代谢物和抗生素等提供了非常重要的线索。分箱(binning)是一种恢复宏基因组数据中部分或完整细菌基因组的方法。可用的分箱工具包括 CONCOCT、MaxBin 2 和 MetaBAT2。分箱工具根据四核苷酸频率和丰度将重

表 1 扩增子和宏基因组分析软件介绍

软件名称	链接	软件描述和软件优势	参考文献
QIIME	http://qiime.org	本领域最高引的分析流程（被引>2 万次），提供了丰富的分析流程、脚本和可视化方案，依赖关系复杂，安装有一定难度，2018 年起不再更新	(Caporaso et al., 2010)
QIIME 2	https://qiime2.org https://github.com/YongxinLiu/QIIME2ChineseManual	是 QIIME 的第二版，提供了命令行和可视化界面双运行模式，支持可重复分析，大数据和可交互的可视化，还提供中文版的帮助文档和视频教程	(Bolyen et al., 2019)
USEARCH	http://www.drive5.com/usearch https://github.com/YongxinLiu/UsearchChineseManual	是一种小巧、跨平台、高速计算的比对工具，含有超过 200 个的子命令，可实现扩增子分析，32 位版免费限<4GB 的数据，64 位版收费	(Edgar, 2010)
VSEARCH	https://github.com/torognes/vsearch	对标 64 位版 USEARCH，但它是免费的，可单独使用的跨平台扩增子分析流程，或作为 USEARCH 的补充，也可在 QIIME 2 中调用 vsearch 插件使用	(Rognes et al., 2016)
Trimmomatic	http://www.usadellab.org/cms/index.php?page=trimmomatic	用于宏基因组原始数据的质控，是基于 java 开发的 Illumina 数据低质量、引物和接头序列去除工具	(Bolger et al., 2014)
Bowtie 2	http://bowtie-bio.sourceforge.net/bowtie2	是一种快速的比对工具，常用于比对序列至参考数据库实现去除宿主污染或定量	(Langmead and Salzberg, 2012)
MetaPhlAn2	https://bitbucket.org/biobakery/metaphlan2	是一种物种分类工具，自带超过 10000 多个物种的标记基因数据库，输出结果为株水平相对丰度	(Truong et al., 2015)
Kraken 2	https://ccb.jhu.edu/software/kraken2	是一种物种分类工具，它使用精确 *k*-mer 方法匹配到 NCBI 数据库，按最低共同祖先原则分类，具有高精确和超高速分析的特点，输出结果为计数型数据	(Wood et al., 2019)
HUMAnN2	https://bitbucket.org/biobakery/humann2	基于 UniRef 蛋白数据库开发，可计算宏基因组和宏转录组数据的基因丰度、通路覆盖度和通路丰度，还能提供特定功能的物种贡献度	(Franzosa et al., 2018)
MEGAN	https://github.com/husonlab/megan-ce http://www-ab.informatik.uni-tuebingen.de/software/megan6	一个图形界面、跨平台软件，用于宏基因组物种和功能分析，提供多种可视化方案，如散点图、Voronoi 树图、聚类图、网络图等	(Huson et al., 2016)
MEGAHIT	https://github.com/voutcn/megahit	一个超快、省内存的宏基因组组装软件	(Li et al., 2015)
metaSPAdes	http://cab.spbu.ru/software/spades	高质量的宏基因组组装软件，可实现株水平组装，但对内存和计算资源消耗极大	(Nurk et al., 2017)
MetaQUAST	http://quast.sourceforge.net/metaquast	组装结果质量评估软件，可评估 N50 和错误组装事件，输出 PDF、可交互的 HTML 报告	(Mikheenko et al., 2016)
MetaGeneMark	http://exon.gatech.edu/GeneMark/	细菌、古菌、宏基因组和宏转录组的基因预测工具，支持 Linux/MacOSX 操作系统，还提供在线分析服务	(Zhu et al., 2010)
Prokka	http://www.vicbioinformatics.com/software.prokka.shtml	快速的原核生物基因组注释工具，调用 metaProdigal 进行宏基因预测，输出结果有核苷酸序列、蛋白序列和基因注释文件	(Seemann, 2014)
CD-HIT	http://weizhongli-lab.org/cd-hit	构建非冗余基因集	(Fu et al., 2012)
Salmon	https://combine-lab.github.io/salmon	基于 *k*-mer 的快速基因序列定量方法	(Patro et al., 2017)
metaWRAP	https://github.com/bxlab/metaWRAP	分箱流程，含有超过 140 个工具，支持 conda 安装，默认使用 MetaBAT、MaxBin 和 CONCOCT 进行分箱；提供分箱结果的提纯、定量、物种分类和可视化等功能	(Uritskiy et al., 2018)
DAS Tool	https://github.com/cmks/DAS_Tool	整合了 5 种分箱工具的分箱流程，提供结果提纯	(Sieber et al., 2018)

重叠群可分类为不同的“箱”（bins），类似于基因组草图。用多个软件优化分析结果和用重新组装的

方法可以获得更好的“箱”。我们建议使用 MetaWRAP 或 DAStool 分箱流程，他们集成了多个分箱工具，可以获得更优的分箱结果、更少的污染，更完整的基因组。这些流程还提供了有用的脚本，用于对箱进行评估和可视化。如想了解宏基因组实验和分析更全面的知识，我们推荐阅读这篇在 Nature Biotechnology 杂志发表的文献(Quince et al., 2017)。

4.统计和可视化

扩增子和宏基因组分析流程最重要的输出文件是物种和功能组成表（**图 2/3**）。研究人员可以使用这些分析技术回答的科学问题包括：微生物群中存在哪些微生物？不同组在 α 和 β 多样性上是否存在显著差异？哪些物种、基因或功能通路是各组的生物标记？为了回答这些问题，需要从整体和细节的角度熟悉统计分析和可视化方法。整体可视化可用于探索特征表中 α/β-多样性和物种分类学组成的差异。细节分析可能涉及生物标记识别、相关性分析、网络分析和机器学习等（**图 3**）。我们将在下文讨论这些方法，并提供示例和参考资料来帮助理解这些分析过程（**图 3 和表 2**）。

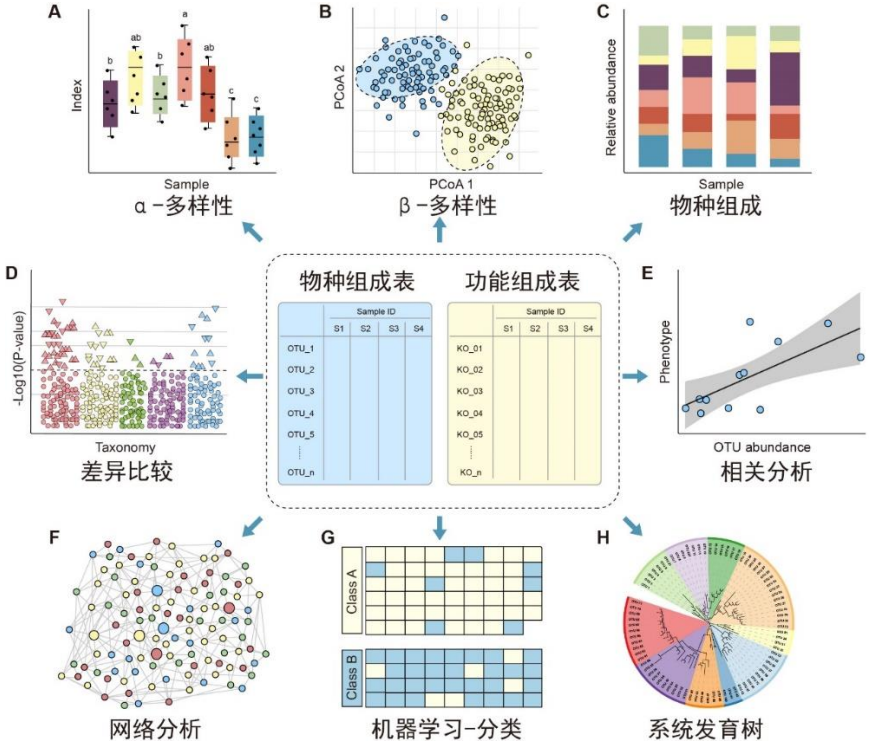

图 3 特征表统计和可视化方法概览。微生物组特征表的下游分析包括 α 多样性（A）、β 多样性（B）、分类学组成（C）、差异比较（D）、相关性分析（E）、网络分析（F）、机器学习分类（G）以及使用树图（H）展示的系统发育树。有关更多详细信息请参见表 2。

α 多样性评估样本内的多样性，包括物种数量（richness）和均匀度（evenness）的测量。可以计算 α 多样性的软件包括 QIIME、R 语言的 vegan 包和 USEARCH 等。通常使用箱线图直观地比较每组样品的 α 多样性值（**图 3A**）。可以使用方差分析（ANOVA）、Mann-Whitney U 检验（2 组比较的秩和检验）或 Kruskal-Wallis 检验（≥3 组的秩和检验）来统计评估组间的 α 多样性差异。注意，如果将每个组进行两次以上的比较，则应校正 P 值。**表 2** 中介绍了其他有关 α 多样性指数的可视化方法。

表 2 数据分析和可视化方法介绍

方法	科学问题	可视化方法	描述和例子的参考文献
α 多样性	计算样本内多样性	箱线图	展示组间 α 多样性分布(Edwards et al., 2015)或显著性差异(Zhang et al., 2019)
		稀释曲线	随着测序深度的增加伴随着样本多样性增加并趋于稳定(Beckers et al., 2017)
		韦恩图	共有或特有物种/分类单元的展示(Ren et al., 2019)
β 多样性	样本间或组间距离	非限制性 PCoA 散点图	展示不同组样本间距离(Zhang et al., 2019)或时间梯度变化(Zhang et al., 2018)
		限制性 PCoA 散点图	展示组间的主要差异(Zgadzaj et al., 2016)
		系统树图	通过层级聚类的办法展示样本间距离(Chen et al., 2019)
物种分类组成	特征的相对丰度	堆叠柱状图	每个样本(Beckers et al., 2017)或组(Jin et al., 2017)的物种构成
		冲击图	物种相对丰度随着季节(Smits et al., 2017)或时间序列改变(Zhang et al., 2018)
		桑基图	组间相对丰度变化及独有和共有物种展示(Smits et al., 2017)
差异比较	组间显著差异的物种	火山图	用类似于火山形状分布的散点展示 P 值、相对丰度和差异物种的数量(Shi et al., 2019)
		曼哈顿图	图形类似于曼哈顿地区高低的大楼，用于展示 P 值、物种分类情况等(Zgadzaj et al., 2016)
		扩展柱状图	为含有可信区间的柱状图(Parks et al., 2014)
相关	特征与元数据之间或特征间的相关性	有拟合曲线的散点图	展示特征随着时间(Metcalf et al., 2016)或其他数值型元数据(Salazar et al., 2019)之间的关系
		相关图	用颜色(Edwards et al., 2018)和/或形状(Zhang et al., 2018)可视化相关系数或距离的三角矩阵
		热图	展示特征相对丰度随着时间改变(Subramanian et al., 2014)
网络	特征相关性的总体展示	用颜色区分物种分类或模块	根据物种分类和/或模块着色特征(Jiao et al., 2016)
		用颜色凸显重要特征	凸显重要特征、展示他们的位置和连接度(Wang et al., 2018)
机器学习	用元数据预测分类型或连续型应变量	热图	用色块展示分类结果(Wilck et al., 2017)或在时间序列中的特征模式(Subramanian et al., 2014).
		柱状图	展示变量重要性或特征相对丰度(Zhang et al., 2019)，也可展示均方误差(Subramanian et al., 2014)
		克利夫兰点图	用点的高低代替条图，以便图片更简洁，用于展示变量重要性(Qian et al., 2020)
树图	系统进化树或物种分类层级	系统进化树或物种分枝图	用发育树(图 3H)展示物种进化关系(Levy et al., 2018)，物种分枝图可展示感兴趣的生物标记(Segata et al., 2011).
		环形树图	以不同颜色圆圈或气泡的方式展现特征层级(Carrión et al., 2019)

β 多样性评估样本之间微生物组的差异，通常将其与降维方法结合使用以便顺利实现可视化。降维方法有主坐标分析（PCoA）、非度量多维尺度（NMDS）或限制性主坐标分析（Constrained PCoA）等。这些分析可以在 R 包 vegan 中完成，并以散点图的方式可视化（**图 3B 和表 2**）。这些 β-多样性指数之间的差异可以用置换多元方差分析（PERMANOVA）进行计算，计算工具以 vegan 包中的 adonis()函数最常用。

物种的分类学组成（taxonomic composition）描述了微生物群落中存在的微生物群，通常使用堆叠柱状图对其进行可视化（**图 3C 和表 2**）。为简单起见，微生物通常以门或属的水平在图中展示。

差异比较用于识别组间特征差异，这些特征可以是物种、基因或通路等。可以使用的检验方法有 Welch's t 检验、Mann-Whitney U 检验、或 Kruskal-Wallis 检验等，也可用 ALDEx2、edgeR、STAMP 或 LEfSe 等包或软件完成分析。差异比较的结果可以使用火山图、曼哈顿图（**图 3D**）或扩展柱状图

图（**表 2**）可视化。注意，由于某些特征的相对丰度增加的同时伴有其他特征减少，这种数据的分析易于产生假阳性。研究者已经开发了几种方法来获得样品中的分类学绝对丰度，例如 HTS 结合流式细胞计数和“HTS+内参质粒+qPCR”的方法。

表 3 提供可重复分析的网站和工具

资源名称	链接	描述
组学原始数据归档库(GSA)	http://gsa.big.ac.cn	用于原始数据的存放和共享。优点是上传速度快，有中英文界面、QQ、邮件技术支持，而且被各种杂志认可
Qiita	https://qiita.ucsd.edu	提供扩增子数据存储、分析、跨研究/课题比较
MGnify	https://www.ebi.ac.uk/metagenomics	提供扩增子、宏基因组数据存储、共享、分析、跨研究课题比较
gcMeta	https://gcmeta.wdcm.org	提供扩增子、宏基因组数据存储、共享、分析
R Markdown	https://rmarkdown.rstudio.com	使用 RStudio 将文本、代码、图片排版在一起，生成格式精美的 pdf/html/docx 报告，此种方式在微生物组研究中越来越受欢迎
R Graph Gallery	https://www.r-graph-gallery.com	42 种图形的 R 代码
GitHub	https://github.com	用于存储代码的网站，具有检索功能

相关分析用于揭示特征和样本元数据（sample metadata）之间的关联（**图 3E**），常用于识别分类群与环境因素（例如 pH 值、经度和纬度）或临床指标之间的关联，或用于识别在时间序列中影响微生物群和动态分类群的关键环境因素。

网络分析从整体角度探讨了特征的共现性关系（**图 3F**）。相关网络的属性可能表示同时出现的分类单元或功能路径之间的潜在相互作用。相关系数和 *P* 值的计算可使用语言 R 中的 *cor.test()* 函数，或者用专门为稀疏型微生物组数据开发的 SparCC 包分析。网络的可视化和分析还可以使用 R 语言 igraph 包、CytoScape、Gephi 等软件实现。网络分析可参考几个很好的例子，例如探索门或模块分布的研究(Fan et al., 2019)或显示时间序列上趋势的研究(Wang et al., 2019)。

机器学习是人工智能的一个分支，可以从数据中学习、识别模式并做出决策（**图 3G**）。在微生物组研究中，机器学习用于特定特征的物种分类、β 多样性分析、分箱和组成分析（compositional analysis）。常用的机器学习方法包括随机森林，Adaboost 和深度学习等，他们可以用于“分类”和“回归”（**表 2**）。注：“分类”是指应变量为分类变量，比如患病和未患病；“回归”是指应变量为连续型变量，如时间。

树图被广泛用于微生物组系统发育树（图 3H）或物种层级分类注释的可视化。扩增子的代表性序列均为同源序列，非常适合于系统发育分析。我们建议使用适合大数据的 IQ-TREE 快速构建高可信度的系统树，并使用 iTOL 实现在线可视化。推荐使用 R 脚本 table2itol（https://github.com/mgoeker/table2itol）快速生成 iTOL 格式的树注释文件。此外，我们还推荐使用 GraPhlAn 构建高颜值的系统发育树或层级分类分支图(Cladogram)。

此外，研究人员可能对探明微生物来源感兴趣，以解决诸如肠道菌群起源、河流污染来源以及法医学检测等问题。FEAST 和 SourceTracker 用于揭示微生物群落的起源。如果研究人员希望关注宿主遗传信息与微生物之间的调控关系，则全基因组关联分析（GWAS）可能是一个不错的选择。

5. 可重复分析

可重复分析要求研究人员在出版论文的同时提交他们的数据和代码，而不仅仅是描述他们的方法。可重复性对于微生物组分析至关重要，因为如果没有原始数据、详细的样本元数据和分析代码，就不可能重现结果。如果读者可以运行代码，他们将更好地理解研究中所做的事情。我们建议研究人员使用以下步骤共享其测序数据、元数据、分析代码和详细的统计报告。

5.1 将原始数据和元数据上传至数据库（数据中心）。扩增子和宏基因组测序产生大量原始数据。

通常，原始数据必须在论文发表期间上载到数据中心，例如 NCBI、EBI 或 DDBJ。近年来在中国也建立了几个数据中心以提供数据存储和共享服务。例如，中国科学院北京基因组研究所建立的组学原始数据归档库（GSA）具有很多优势（**表 3**）。我们建议研究人员将原始数据上传到表 3 中的数据中心之一，这不仅可以提供数据备份，还可以满足论文发表要求。诸如《Microbiome》等期刊都要求在投稿之前将原始数据存储在数据中心。

5.2 与其他研究人员共享数据分析代码。分析代码可以帮助审稿人或读者评估实验结果的可重复性。我们在 https://github.com/YongxinLiu/Liu2020ProteinCell 提供了用于扩增子和宏基因组分析的示例分析流程。此外，还应提供分析中使用的运行环境和软件版本，以确保可重复性。如果使用 Conda 来部署软件，则命令"*conda env export -n environment_name > environment_name.yaml*"可以生成一个文件，其中包含所用软件及版本信息。对于不熟悉命令行的用户，可以使用 Qiita、MGnify 和 gcMeta 等在线服务器进行分析。但是，在线服务器分析方式的灵活性差强人意，因为他们提供的可调节步骤和参数较少。

5.3 提供详细的统计和可视化报告。用于特征表统计分析和可视化的工具包括 Excel、GraphPad 和 Sigma plot，但是他们是商业软件工具，而且较难快速重现结果。我们建议使用诸如 R Markdown 或 Python Notebooks 之类的工具来记录所有分析代码和参数，并将他们存储在诸如 GitHub 之类的版本控制管理系统中（**表 3**）。这些工具是免费、开源、跨平台的，并且易于使用。我们建议研究人员在 R markdown 文件中记录微生物组的所有统计分析和可视化方法。RMarkdown 文档可以包含代码、表格、图片，支持输出为 PDF/网页/Word 格式报告方便阅读。这种工作模式将大大提高微生物组分析的效率，并使分析过程透明且易于理解。R 可视化代码可以参考 R Graph Gallery（**表 3**）。可以将分析用的输入文件（特征表+元数据）、分析代码（*.Rmd）和输出结果（图、表和 HTML 报告）上载到 GitHub，这将方便同行重复您的分析或重用您的分析代码。ImageGP（http://www.ehbio.com/ImageGP）提供了 20 多种统计和可视化方法，对于没有 R 语言背景的研究人员而言是一个不错的选择。

6. 注意事项和展望

值得注意的是，实验操作对研究结果的影响远大于为选用的分析方法。最好将详细的实验过程记录为元数据，例如采样方法、时间、位置、操作员、DNA 提取试剂盒、批次、引物和标签序列等。元数据可以用于下游分析，而且可以帮助研究者确定是否这些操作差异导致了假阳性结果。实验中应收集的元数据等信息请参阅"标记基因序列的最少信息标准（MIMARKS）和宏基因组序列的最少信息标准"(Field et al., 2008; Yilmaz et al., 2011)，"细菌和古菌单个扩增基因组（MISAG）和宏基因组组装的基因组（MIMAG）的最少信息标准"(Bowers et al., 2017)，以及"未培养病毒基因组的最少信息标准"(Roux et al., 2019)。

一些特定的实验步骤可提供微生物组分析的独特视角。例如，开发和使用去除宿主 DNA 的方法可以有效增加植物内生菌和人类呼吸道感染样品中真实微生物组的比例。叠氮化丙锭可以物理去除土壤中大量残留的 DNA。此外，当使用微生物含量较低的样品时，研究人员必须格外小心，以免由于污染而导致假阳性结果。对于这些情况，可使用不含 DNA 的水作为阴性对照。在人类微生物组研究中，个体之间的主要差异是由于饮食、生活方式和药物使用等因素造成的，遗传对于微生物组的影响小于 2%。

鸟枪宏基因组测序可以深入了解菌株水平的微生物群落结构，但很难恢复高质量的全基因组。单细胞基因组测序显示了在微生物组研究中非常有前途的应用。基于流式细胞仪和单细胞测序，MetaSort 可以从分选的亚宏基因组样品中恢复高质量的微生物基因组(Ji et al., 2017)。最近开发的第三代测序技术已用于宏基因组分析，例如太平洋生物科学（PacBio）的单分子实时（SMRT）测序和牛津纳米孔（ONT）测序平台。随着测序数据质量的提高和成本的降低，这些技术将引发微生物组测序领域的技术革命，并将微生物组研究带入一个新时代。

7. 结论

在这篇综述中，我们讨论了各阶段分析扩增子和宏基因组数据的方法，包括测序方法的选择、分析软件和流程、统计分析和可视化以及可重复分析的实施等。其他方法，如宏转录组、宏蛋白组学和宏代谢组学等可提供微生物组的动态信息，但是由于这些方法成本高以及需要复杂的实验和分析方法且不够成熟，尚不及扩增子和宏基因组如此广泛应用，所以未在本文的讨论范围之内。将来随着这些技术的进一步发展，对于微生物组研究的视角将更全面。

参考文献（下面仅列出重点参考文献，详细文献请阅读英文版原文）：

Beckers, B., Op De Beeck, M., Weyens, N., Boerjan, W., and Vangronsveld, J. (2017). Structural variability and niche differentiation in the rhizosphere and endosphere bacterial microbiome of field-grown poplar trees. Microbiome 5, 25.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 37, 852-857.

Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nature Biotechnology 35, 725.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7, 335-336.

Carrión, V.J., Perez-Jaramillo, J., Cordovez, V., Tracanna, V., de Hollander, M., Ruiz-Buck, D., et al. (2019). Pathogen-induced activation of disease-suppressive functions in the endophytic root microbiome. Science 366, 606-612.

Chen, Q., Jiang, T., Liu, Y.-X., Liu, H., Zhao, T., Liu, Z., et al. (2019). Recently duplicated sesterterpene (C25) gene clusters in Arabidopsis thaliana modulate root microbiota. Science China Life Sciences 62, 947-958.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460-2461.

Edwards, J., Johnson, C., Santos-Medellín, C., Lurie, E., Podishetty, N.K., Bhatnagar, S., et al. (2015). Structure, variation, and assembly of the root-associated microbiomes of rice. Proceedings of the National Academy of Sciences 112, E911-E920.

Edwards, J.A., Santos-Medellín, C.M., Liechty, Z.S., Nguyen, B., Lurie, E., Eason, S., et al. (2018). Compositional shifts in root-associated bacterial and archaeal microbiota track the plant life cycle in field-grown rice. PLOS Biology 16, e2003862.

Fan, K., Delgado-Baquerizo, M., Guo, X., Wang, D., Wu, Y., Zhu, M., et al. (2019). Suppressed N fixation and diazotrophs after four decades of fertilization. Microbiome 7, 143.

Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., et al. (2008). The minimum information about a genome sequence (MIGS) specification. Nature Biotechnology 26, 541-547.

Franzosa, E.A., McIver, L.J., Rahnavard, G., Thompson, L.R., Schirmer, M., Weingart, G., et al. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. Nature Methods 15, 962-968.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150-3152.

Huang, P., Zhang, Y., Xiao, K., Jiang, F., Wang, H., Tang, D., et al. (2018). The chicken gut metagenome and the modulatory effects of plant-derived benzylisoquinoline alkaloids. Microbiome 6, 211.

Huson, D.H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., et al. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLOS Computational Biology 12, e1004957.

Ji, P., Zhang, Y., Wang, J., and Zhao, F. (2017). MetaSort untangles metagenome assembly by reducing microbial community complexity. Nature Communications 8, 14306.

Jiao, S., Liu, Z., Lin, Y., Yang, J., Chen, W., and Wei, G. (2016). Bacterial communities in oil contaminated soils: Biogeography and co-occurrence patterns. Soil Biology and Biochemistry 98, 64-73.

Jin, T., Wang, Y., Huang, Y., Xu, J., Zhang, P., Wang, N., et al. (2017). Taxonomic structure and functional association of foxtail millet root microbiome. GigaScience 6, 1-12.

Lagier, J.-C., Dubourg, G., Million, M., Cadoret, F., Bilen, M., Fenollar, F., et al. (2018). Culturing the human microbiota and culturomics. Nature Reviews Microbiology 16, 540-550.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 357-359.

Levy, A., Salas Gonzalez, I., Mittelviefhaus, M., Clingenpeel, S., Herrera Paredes, S., Miao, J., et al. (2018). Genomic features of bacterial adaptation to plants. Nature Genetics 50, 138-150.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1674-1676.

Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., et al. (2014). An integrated catalog of reference genes in the human gut microbiome. Nature Biotechnology 32, 834-841.

Liu, Y.-X., Qin, Y., and Bai, Y. (2019). Reductionist synthetic community approaches in root microbiome research. Current Opinion in Microbiology 49, 97-102.

Metcalf, J.L., Xu, Z.Z., Weiss, S., Lax, S., Van Treuren, W., Hyde, E.R., et al. (2016). Microbial community assembly and metabolic function during mammalian corpse decomposition. Science 351, 158-162.

Mikheenko, A., Saveliev, V., and Gurevich, A. (2016). MetaQUAST: evaluation of metagenome assemblies. Bioinformatics 32, 1088-1090.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Research 27, 824-834.

Parks, D.H., Tyson, G.W., Hugenholtz, P., and Beiko, R.G. (2014). STAMP: statistical analysis of taxonomic and functional profiles. Bioinformatics 30, 3123-3124.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods 14, 417-149.

Qian, X., Liu, Y.X., Ye, X., Zheng, W., Lv, S., Mo, M., et al. (2020). Gut microbiota in children with juvenile idiopathic arthritis: characteristics, biomarker identification, and usefulness in clinical prediction. BMC Genomics 21, 286.

Ren, Z., Li, A., Jiang, J., Zhou, L., Yu, Z., Lu, H., et al. (2019). Gut microbiome analysis as a tool towards targeted non-invasive biomarkers for early hepatocellular carcinoma. Gut 68, 1014-1023.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ 4, e2584.

Roux, S., Adriaenssens, E.M., Dutilh, B.E., Koonin, E.V., Kropinski, A.M., Krupovic, M., et al. (2019). Minimum

Information about an Uncultivated Virus Genome (MIUViG). Nature Biotechnology 37, 29-37.

Salazar, G., Paoli, L., Alberti, A., Huerta-Cepas, J., Ruscheweyh, H.-J., Cuenca, M., et al. (2019). Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. Cell 179, 1068-1083.e1021.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068-2069.

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., et al. (2011). Metagenomic biomarker discovery and explanation. Genome Biology 12, R60.

Shi, W., Li, M., Wei, G., Tian, R., Li, C., Wang, B., et al. (2019). The occurrence of potato common scab correlates with the community composition and function of the geocaulosphere soil microbiome. Microbiome 7, 14.

Sieber, C.M.K., Probst, A.J., Sharrar, A., Thomas, B.C., Hess, M., Tringe, S.G., et al. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nature Microbiology 3, 836-843.

Smits, S.A., Leach, J., Sonnenburg, E.D., Gonzalez, C.G., Lichtman, J.S., Reid, G., et al. (2017). Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. Science 357, 802-806.

Subramanian, S., Huq, S., Yatsunenko, T., Haque, R., Mahfuz, M., Alam, M.A., et al. (2014). Persistent gut microbiota immaturity in malnourished Bangladeshi children. Nature 510, 417.

Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., et al. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods 12, 902-903.

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

Wang, J., Jia, Z., Zhang, B., Peng, L., and Zhao, F. (2019). Tracing the accumulation of in vivo human oral microbiota elucidates microbial community dynamics at the gateway to the GI tract. Gut, gutjnl-2019-318977.

Wang, J., Zheng, J., Shi, W., Du, N., Xu, X., Zhang, Y., et al. (2018). Dysbiosis of maternal and neonatal microbiota associated with gestational diabetes mellitus. Gut 67, 1614-1625.

Wilck, N., Matus, M.G., Kearney, S.M., Olesen, S.W., Forslund, K., Bartolomaeus, H., et al. (2017). Salt-responsive gut commensal modulates TH17 axis and disease. Nature 551, 585-589.

Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. bioRxiv, 762302.

Xiao, L., Feng, Q., Liang, S., Sonne, S.B., Xia, Z., Qiu, X., et al. (2015). A catalog of the mouse gut metagenome. Nature Biotechnology 33, 1103.

Xu, J., Zhang, Y., Zhang, P., Trivedi, P., Riera, N., Wang, Y., et al. (2018). The structure and function of the global citrus rhizosphere microbiome. Nature Communications 9, 4894.

Ye, S.H., Siddle, K.J., Park, D.J., and Sabeti, P.C. (2019). Benchmarking Metagenomics Tools for Taxonomic Classification. Cell 178, 779-794.

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J.R., Amaral-Zettler, L., et al. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature Biotechnology 29, 415.

Zgadzaj, R., Garrido-Oter, R., Jensen, D.B., Koprivova, A., Schulze-Lefert, P., and Radutoiu, S. (2016). Root nodule symbiosis in Lotus japonicus drives the establishment of distinctive rhizosphere, root, and nodule bacterial communities. Proceedings of the National Academy of Sciences 113, E7996-E8005.

Zhang, J., Liu, Y.-X., Zhang, N., Hu, B., Jin, T., Xu, H., et al. (2019). NRT1.1B is associated with root microbiota composition and nitrogen use in field-grown rice. Nature Biotechnology 37, 676-684.

Zhang, J., Zhang, N., Liu, Y.-X., Zhang, X., Hu, B., Qin, Y., et al. (2018). Root microbiota shift in rice correlates

with resident time in the field and developmental stage. Science China Life Sciences 61, 613-621.