微生物学报
Acta Microbiologica Sinica
2017, 57(6): 932-941
http://journals.im.ac.cn/actamicrocn
DOI: 10.13343/j.cnki.wsxb.20170070

马俊才，博士，中国科学院微生物研究所微生物资源与大数据中心主任，正高级工程师，世界微生物数据中心主任，世界微生物菌种保藏联合会(WFCC)理事会执委，科技部人类遗传资源管理专家委员会委员。国家“863 计划” “微生物数字化信息系统集成关键技术的研发”项目的首席科学家。主要研究领域包括：微生物资源和生物技术领域信息化、基于云环境的微生物大数据管理和分析平台。

微生物组大数据管理与分析

亓合媛，孙清岚，马俊才*

中国科学院微生物研究所，北京 100101

摘要：高通量技术的迅猛发展促使微生物生态学研究获得了重大突破，掀起了元基因组学(Metagenomics)研究的热潮。元基因组学通常被定义为对未培养的环境样本中微生物群体的 DNA 序列分析。随着微生物组学数据的日益剧增，微生物大数据的高效管理与分析越来越受到研究者的关注。如何从海量的微生物组数据中挖掘出具有科研价值的数据信息并应用于实际问题成为当前的研究热点。目前已有很多计算生物学程序工具及数据库用于元基因组数据的分析与管理。本文主要综述了随着高通量测序技术的进步，国际上主要的微生物组计划及微生物组数据平台，如人类微生物组项目(human microbiome project，HMP)、地球微生物组项目(earth microbiome project，EMP)、欧盟的肠道微生物组计划(metagenomics of human intestinal tract，MetaHIT)、MG-RAST、iMicrobe、整合微生物组(integration microbial genomes，IMG)以及 EBI Metagenomics 等；介绍了微生物数据分析的主要流程与工具；提出了建设多源异构的微生物生态数据管理与分析系统的必要性。

关键词：微生物组，元基因组学，数据分析，高通量测序

早在约 35 亿年前，微生物作为古老的生命形式出现在地球上，如今，微生物早已遍布地球的每一个角落，参与地球的碳与养分循环，与植物和动物的健康和疾病状态息息相关。当前，99%的数万亿存在的微生物尚未被发现与鉴定[1]。微生物群落一般具备如下几点特征：(1) 个体微小，形态差异特征少。即便在显微镜下也难以区分

基金项目：国家重点研发计划(2016YFC0901702，2016YFB1000605)；国家“863 计划”(2015AA020108)
*通信作者。Tel：+86-10-64807422；Fax：+86-10-64807426；E-mail：ma@im.ac.cn
收稿日期：2017-02-16；修回日期：2017-04-04；网络出版日期：2017-04-11

不同的物种，因此不易通过直接观察鉴定物种类型。(2) 密度大。Silvay 等人研究发现，15 mL 室内空气中可培养出(9–13)×10^7 个真菌菌落[2]。由此可见，1 mL 空气可能含有多达数十亿的微生物。(3) 丰度差异显著。不同种属的微生物在不同群落中的丰度具有显著差异。这在一定程度上促进了微生物群落的功能特异性，然而这种分布特征也增加了统计优势种、非优势种以及稀有种的难度。(4) 复杂度高。高密度、高丰度差异的微生物群体造成了微生物群落的复杂性和多样性，群落微生物的高度多样性进一步使得微生物群落鉴定更加困难[3]。

针对如此复杂多样的微生物生态系统，微生物组学的研究将精准解析特定环境/条件下样品中微生物的群落结构特征及功能，关联其与时空、理化特征、宿主健康状态，从而探索微生物与微生物、宿主、环境等因素之间的相互关系网络。这一系列的分析研究将依赖于更加精准与恰当的技术。

1 高通量测序技术与微生物大数据

1.1 高通量测序技术

自 1997 年至 2005 年，Sanger 测序一直占据核酸测序技术的主导地位。然而，利用 Sanger 测序技术对一个细菌基因组进行完整测序需要花费数年的时间[4]。近年来，随着测序技术的不断发展，下一代测序技术(NextGen sequencing，NGS)不断涌现，测序通量大大提高，使得一个细菌基因组测序在几个小时之内即可完成[5]。NGS 技术包括多种不同的测序平台，主要有罗氏 GS20 454 测序仪[6]、Solexa (现在是 Illumina) Genome Analyzer (GA)[7]、Illumina Miseq、Hiseq、TruSeq[8]等。随着技术的进一步完善与发展，单分子测序在微生物组学研究中逐渐应用，该测序方法能以更短的时间测完一个细菌基因组并且提供更多的基因组信息。具有代表性的第三代单分子测序平台主要包括 PacBio[9]、Oxford Nanopore[10]等。

高通量测序技术的发展进步促进了对微生物及微生物群落结构和功能的发现与鉴定。例如，利用 16S rRNA 基因的高通量测序对细菌与古细菌进行系统进化树的分析；利用全基因组测序的方法对不同环境微生物测序，鉴定环境特异性基因。然而大量被鉴定的特异性基因的功能是未知的，这也进一步反映了微生物组的高度多样性，同时也提示研究人员环境微生物组的生物化学潜能亟待发掘。

1.2 微生物资源大数据

20 世纪 60 年代，世界菌种保藏联盟建立的世界微生物数据中心(world data centre of microorganisms，WDCM)，是全球微生物领域最重要的实物资源数据平台。该中心倡导的全球微生物菌种保藏目录(global catalogue of microorganisms，GCM)，整合了超过 37 万的微生物实物资源的详细信息，利用先进的数据挖掘手段，从全球超过 600 万已发表的微生物文献及专利中，进一步提取了微生物资源的后续研究和利用的信息[11]。该平台对于微生物资源从采集、保藏、跨国转移、学术和商业应用以及利益分享的各个环节都能提供有效的数据支持。2010 年，WDCM 落户中国科学院微生物研究所，这是我国生命科学领域的第一个世界数据中心。本团队以微生物领域相关的数字化资源整合为核心，突破微生物数字资源的知识挖掘和垂直检索、基于云技术的微生物信息系统构建等关键技术，建立了具有国际影响力的微

生物数据库，实现我国在微生物领域数字资源建设的突破。

针对微生物资源数据，关联整合高通量微生物测序数据，建立微生物采集、保藏-文献专利-基因/蛋白序列等多元化信息为一体的微生物大数据仓库，将为微生物的鉴定与溯源提供良好的支撑作用。

2 国际主要微生物组研究项目与平台

随着高通量测序技术的突飞猛进，测序成本的不断降低，针对不同环境微生物群落的高通量测序项目逐渐开展起来。研究人员试图利用海量序列数据探索微生物组的奥秘。当前，国际主要的微生物组计划包括美国的人类微生物组项目(human microbiome project，HMP)、地球微生物组项目(earth microbiome project，EMP)和欧盟的肠道微生物组计划(metagenomics of human intestinal tract，MetaHIT)。应这些项目数据管理与分析的需求，项目研究者也开发了诸多分析工具与数据管理平台，如 HMP 的数据分析与协调中心(data analysis and coordination center，DACC)、EMP 的 GA、EM-AG 及 EM-VIP 等。此外，其他研究人员也开发了较好的微生物组数据管理与分析平台，如 MG-RAST、IMG、iMicrobe 及 EMG 等。

2.1 主要国际微生物组研究计划

美国国立卫生院(NIH)于 2008 年发起了人类微生物组项目(human microbiome project，HMP，第一期 2008–2013 年；第二期 integrative human microbiome project，iHMP，2013 年至今)[12]，共资助 1.15 亿美元。目的是建立微生物基因组的参考数据库并描述人体微生物组的相关特征，探索人体微生物组的变化与疾病之间的关系，开发新的数据计算分析方法与工具。该项目从 554 个个体中共采取 12479 个 DNA 样本用于微生物组学分析。同年，欧盟也发布了肠道微生物组计划(metagenomics of human intestinal tract，MetaHIT，2008–2012 年)[13]，该项目共耗资 2120 万英镑，对 124 个欧洲个体进行了肠道菌群测序，获得了 576.7 Gb 的序列数据，鉴定了 330 万个非冗余的微生物基因。2013 年，由美国阿贡实验室发起并实施了地球微生物组项目(earth microbiome project，EMP，2010 年至今)[14]，该项目试图通过关联分析 20 万个来自不同环境的样本数据，创建一个完整的地球微生物数据库，最终利用微生物的群落结构与相互作用展现环境与生态系统。截至 2014 年，EMP 已完成了对 30000 个样本的元基因组测序。

这一系列重大的国际微生物组项目的发起与实施表明，微生物组的研究具有不容忽视的科研价值与应用前景。如何解析并利用海量的微生物组数据成为微生物组研究领域的热点问题。

2.2 国际主要微生物组数据管理与分析平台

随着众多微生物组计划的发起与实施，元基因组数据大量产出，对实验元数据及测序数据的存储、管理与分析成为微生物组研究的关键。前文所述的重要国际微生物组计划也都建立了相应的数据中心，以管理与分析项目产出的数据。例如，HMP 建立了数据分析与协调中心(data analysis and coordination center，DACC)用于存储项目产生的元数据以及高通量测序的序列数据(包括 16S rRNA 基因序列、全基因组序列等)，开发了可供大量数据高效组织、存储、获取、检索及注释的数据库系统。EMP 也建立了相关的数据平台，主

要包括：(1) GA (gene atlas)数据库用于存储 EMP 产出的所有数据；(2) EM-AG (earth microbiome assembly genome)用于存储经自动化分析工具组装、注释的基因组，并支持比较基因组学分析；(3) EM-VIP (earth microbiome visualization portal)用于提供数据的可视化交互，展示微生物组成、环境参数及基因组功能之间的关系网络；(4) EMMR (earth microbiome metabolic reconstruction)用于提供重建的代谢图谱。

目前较为广泛应用的微生物数据管理与分析平台主要包括美国阿贡实验室开发的 MG-RAST[15–16]、美国亚利桑那大学牵头的 iMicrobe、美国能源部联合基因组研究所建立的整合微生物基因组(integration microbial genomes，IMG)[17–18]以及 EBI 建立的 EMG (EBI metagenomics)[19]。

MG-RAST 建立了高通量数据分析工作流，为元基因组分析人员提供高性能的计算服务。该分析工作流将自动对提交的数据进行功能基因分析和物种分类鉴定，并对分析结果进行可视化展示。此外，MG-RAST 为用户提供了数据管理的功能，用户可自由选择私有数据是否共享。MG-RAST 具备的友好交互及便捷的一键式分析，赢得了诸多元基因组数据工作者的关注与使用。用户注册后上传数据，选择分析流程，即可提交任务。然而，由于元基因组数据量大，分析任务繁多，用户提交任务需要排队等待，这也就造成了使用 MG-RAST 分析数据的时间成本较高。

IMG 为用户提供结构和功能注释流程，同时也提供了多种分析和可视化工具以便于比较分析 IMG 数据集。用户可通过 MyIMG 菜单设定数据浏览偏好，在 MyJob 中跟踪提交的计算任务进度。IMG 重点分析组装的元基因组数据，不支持单一特定环境样本中新物种的鉴定与基因组重建、新基因簇的发现、可变遗传密码的鉴定、以及新病毒的鉴定等分析。

iMicrobe 支持不同生态环境来源的微生物数据检索、标记与共享，为用户提供用于元基因组分析的工具(APP)。此外，用户可通过 FTP 向 iMicrobe 提交数据(www.imicrobe.us)，然后选择相应的分析 APP 实现数据的解读分析。

EMG 针对元基因组数据提供标准化分析与管理，可实现特定样本的物种多样性、功能及代谢网络的分析，同时支持同一项目下不同样本间的比较功能组学分析。另外，EMG 针对全球海洋微生物样本与海洋浮游生物多样性进行了详细的分析，为海洋微生物相关的元基因组学研究提供了参考数据集与成熟的分析流程。

综上所述，国际上针对微生物组数据的分析与管理平台已较为成熟，可对海量高通量测序的元基因组项目数据进行有效的分析与管理，不同的平台也各具特色，平台间的比较如表 1 所示，用户可根据自己的实际情况选择并使用。

3 微生物组大数据分析

3.1 基于 16S rRNA 基因序列的分析

基于 16S rRNA 序列的数据分析一般包括序列提取、质量控制、序列 OTU 聚类、种属分类鉴定、alpha 与 beta 多样性分析、以及其他特异性统计分析。OTU 聚类分析是 16S rRNA 基因序列分析的关键步骤之一，OTU 界定时所选取的算法工具对后期的分析结果的影响很大。常见的 OTU 聚类[20–21]方法主要包括 Uclust、cd-hit、Blast、mothur、usearch 和 prefix/suffix 等[22]。

表 1. 不同微生物组数据平台的特征分析

Table 1. Characteristics analysis of different microbiome platforms

Name	Organization	Taxonomy	Functional annotation	Comparison between samples	Comparison between projects	Advantanced Search	Connect to public database
MG-RAST	Argonne National Laboratory	YES	NO	YES	NO	NO	NO
IMG	DOE Joint Genome Institute	YES	YES	NO	NO	NO	NO
iMicrobe	University of Arizona	YES	NO	NO	NO	NO	NO
EMG	European Bioinformatics Institute	YES	YES	YES	YES	NO	YES

聚类后种属鉴定的速度与精度，一方面取决于序列比对算法的选择，另一方面依赖于高质量的参考数据库。高质量的参考数据库将提高鉴定结果的准确性。当前，较为广泛应用的参考数据库包括 Greengenes[23]、RDP[24]、Ez-Taxon[25–26]、SILVA[27]等。序列对对比(Pairwise Alignment)的工具[22]包括 BLAST[28–29]、BLAT[30]、Usearch/Uclust[31]、SINA aligner[32]及 RTAX[33]。其中，在可接受的计算成本下，BLAST 具有较高灵敏度和准确性，因此被广泛使用。BLAT 运行的速度至少比 BLASTn 快 2 个数量级，但是灵敏度下降了。在保证相同灵敏度的前提下，Usearch 的运行速度是 BLASTn 的 30–200 倍。该工具通常与参考数据库绑定在一起用于标记基因或功能基因的快速筛选。

针对 OTU 获得的典型序列进行多序列比对，比对方法主要有 ClustalW、MUSCLE[34]、Clustal Omega[35]、Kalign[36]、T-COFFEE[37]、COBALT[38]以及 FastTree[39]。此外，研究者也开发了多种软件或工具包用于系统发育树分析，如 MEGA、RAxML、MRBAYES、PhyML、TreeView、Clearcut、FigTree 及 ARB[22,40]。鉴于友好的用户交互及详细的说明文档，MEGA 的使用非常广泛。

对同一个样本的物种多样性(alpha 多样性)分析，主要采用 mothur[41]和 QIIME[21]，相比而言，QIIME 具备更强的整合能力，开放性强，用户更多。Mothur 经 Schloss 等人[41]改写后，相对封闭，使用者相对少了些。对不同样本间的物种多样性(beta 多样性)分析的工具主要包括 Unifrac、Bray-Curtis、Euclidean、Jaccard index、Yue & Clayton 及 Morisita-Hom[42]。Beta 多样性指标主要分为两类：数量和质量。数量指标主要采用 Bray-Curtis、加权 Unifrac 等，质量指标主要采用二进制 Jaccard index 进行计算。QIIME 工具包中整合了加权和未加权 Unifrac 距离计算方法，可用于不同物种或 OTU 序列的系统发育树分析，这也就使得 Unifrac 距离计算在群落比较中具备较好的优势。

当然，除 16S rRNA 基因外，也存在很多对特定功能基因的测序数据，其分析流程与 16S rRNA 基因序列分析流程类似，但用到的参考数据库会有所差别。

3.2 基于 WGS 的数据分析

高通量技术的发展与成本的降低，基于全基因组测序(whole genome sequencing，WGS)的元基因组数据越来越多，数据分析步骤主要包括序列的质量控制、高质量序列组装、组装(或不组装)后的序列与参考数据库比对并进行物种分类和丰度统计、样品间物种多样性的比较(如 PCA 分析、聚类分析、功能因子分析等)、基因组分分析(如前

噬菌体预测、功能基因预测)、功能注释(利用KEGG、CAZy、eggNOG 等数据库进行代谢通路、碳水化合物活性、同源性等分析)、耐药性分析等。鉴于微生物大数据的复杂性及多源异构性，研究人员通常需要根据自己的数据特性选择合适的分析工具并辅以自己开发的脚本来完成对元基因组的分析。表 2 中列出了元基因组分析中常用的软件工具及数据库等。

表 2. 元基因组分析常用的软件及数据库

Table 2. Analysis tools and reference databases for metagenomics

Function	Name
Quality control	FastQC, Trimmomatic, NGS QC Toolkit
Sequence assembly	Meta-Velvet, META-IDBA, IDBA-UD, Genovo, MetAMOS, SOAPdenovo
Sequence alignment	BLAST, MegaBLAST, BLAT, LAST, SSAHA2, Bowtie, BWA-SW, SOAP2, BWA, FR-HIT, MAUVE
Gene prediction	FragGeneScan, MataGeneAnnotator (MGA), Glimmer, Glimmer-MG, HMMer3, BLAST, RAPSearch2, RAST, XBASE, RDP classifier, NBC, CARMA3, MEGAN, Sort-ITEMS, GeneMark, MetaPhyler
Functional analysis	MEGAN, SmashCommunity, STAMP, shotgunFunctionalize R, ColVR-metagenomics, CloVR-microbe
Diversity analysis	QIIME, CloVR-16S, FANTOM
Statistic analysis	QIIME, cd-hit, mothur, EstimateS, SPADE, ShotgunFunctionalizeR, R-package, metagenomeSeq, GSEA, MetaPath, MetaPhyler, STAMP, HUMAnN
Database	RDP, SEED, COG, Greengenes, eggNOG, TIGRFAMs, NCBI NR/NT, Uniref100, FungiDB, KBase, Phantome, SILVA, InterPro, UniProt, PATRIC

4 总结和讨论

当前，大数据的产生速率远远高于数据处理效率。尽管如前文所述，国际上已有很多针对微生物组数据分析与管理的平台，但微生物组学数据具备的数据量大、高度不完整性和多源异构性的特点增加了数据分析的复杂性，同时微生物组各类大数据的综合分析，日益成为微生物组数据处理的瓶颈[43]。

针对微生物组大数据，如何存储、积累、提取关键信息并可视化展示数据，如何保证数据与分析的可重复性，这些都成为微生物大数据分析面临的重要挑战。同时建立微生物生态模型，并预测微生物群落的动态发展，预测相关的生物学效应，也是微生物组学研究的重要方向和关键应用。

马俊才博士领导的课题组在微生物数据集成与管理方面做了大量的工作：构建了 BOLD 国际镜像系统(BOLDmirror, www.boldmirror.net)，并积极地对 BOLDmirror 的数据同步功能进行升级，同步获取 BOLD 系统中发布的全部公开数据；自主开发了中国生命条形码信息管理系统(http://data.barcodeoflife.cn)，为中国的科研人员提供 DNA 条形码数据存储、数据管理和数据分析等服务。在国家“863 计划” “微生物数字化信息系统集成关键技术研发”的支持下，以微生物领域相关的数字化资源整合为核心，突破微生物数字资源的知识挖掘和垂直检索、基于云技术的微生物信息系统构建等关键技术，建立了具有国际影响力的微生物数据库，实现我国在微生物领域数字资源建设的突破。这为微生物大数据的深入挖掘与功能信息分析提供了强大的数据支持。

微生物组高通量数据的管理自动化与生物信息流程化是当今元基因组学研究的总体趋势。数据的流程化分析将有效提高数据分析的可靠性与

可比性，为研究者提供更多的具有科研价值的信息。然而，当前的数据分析流程(表 1)多数具有一定的局限性，所提供的分析仍不能满足当前对微生物组大数据的深入挖掘需求。

目前，针对微生物组数据的管理与分析需求，马俊才课题组正在建设国内微生物生态数据管理与云分析平台，在现有的微生物数据工作基础上，整合更多的国际与国内的微生物组数据，统一进行管理，并实现标准化的数据接口，实现对微生物组数据资源的高效管理与集成；对已整合的数据进行严格的质量控制，形成高质量的微生物组参考数据库，基于云平台建立微生物元基因组数据分析的标准化流程，该平台将在近期公布。

尽管如此，在微生物组大数据的收集、存储、功能挖掘、开发利用等关键技术方面，依然存在很多薄弱环节，制约着微生物组学研究的发展。因此，集成海量高质量、具有代表性的微生物生态大数据，建立高质量的微生物组参考数据库，实现高效的数据检索与分析；开发高效的微生物组数据分析流程，最终创建对微生物组数据的系统管理、高效分析及整合利用的大数据平台是微生物组数据研究的迫切需求。这一需求的实现，将极大地推动微生物学领域的研究。

参 考 文 献

[1] White III RA, Callister SJ, Moore RJ, Baker ES, Jansson JK. The past, present and future of microbiome analyses. *Nature Protocols*, 2016, 11(11): 2049–2053.

[2] Silva R, Havnar C. Microbial Analysis of indoor air quality at a community college. http://accounts.smccd.edu/case/air/rona_poster.pdf.

[3] Sheng HF, Zhou HW. Methods, challenges and opportunities for big data analyses of microbiome. *Journal of Southern Medical University*, 2015, 35(7): 931–934. (in Chinese)
盛华芳, 周宏伟. 微生微生物组学大数据分析方法、挑战与机遇. 南方医科大学学报, 2015, 35(7): 931–934.

[4] Heather JM, Chain B. The sequence of sequencers: the history of sequencing DNA. *Genomics*, 2016, 107(1): 1–8.

[5] Nickerson SL, Prosser DO, Lai SWS, Love DR. A comparison of benchtop high-throughput sequencing platforms in the diagnostic laboratory setting. *Pathology*, 2016, 48(S1): S96.

[6] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen ZT, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu PG, Begley RF, Rothberg JM. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 2005, 437(7057): 376–380.

[7] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Cheetham RK, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu XH, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IMJ, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu XL, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DMD, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Catenazzi MCE, Chang S, Cooley RN, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fajardo KVF, Furey WS, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Jones TAH, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo SJ, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning ZM, Ng BL, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O,

Paraschos LL, Pickering L, Pike AC, Pike AC, Pinkard DC, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Rodriguez AC, Roe PM, Rogers J, Bacigalupo MCR, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Sohna JES, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, vandeVondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang JW, Worsley GJ, Yan JY, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 2008, 456(7218): 53–59.

[8] Voskoboynik A, Neff NF, Sahoo D, Newman AM, Pushkarev D, Koh W, Passarelli B, Fan HC, Mantalas GL, Palmeri KJ, Ishizuka KJ, Gissi C, Griggio F, Ben-Shlomo R, Corey DM, Penland L, White III RA, Weissman IL, Quake SR. The genome sequence of the colonial chordate, *Botryllus schlosseri*. *eLife*, 2013, 2: e00569.

[9] 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 XX, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma CC, 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, Turner S. Real-time DNA sequencing from single polymerase molecules. *Science*, 2009, 323(5910): 133–138.

[10] Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, Studholme DJ. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*, 2015, 3: 1–8.

[11] Wu LH, Sun QL, Desmeth P, Sugawara H, Xu ZH, McCluskey K, Smith D, Alexander V, Lima N, Ohkuma M, Robert V, Zhou YG, Li JH, Fan GM, Ingsriswang S, Ozerskaya S, Ma JC. World data centre for microorganisms: an information infrastructure to explore and utilize preserved microbial strains worldwide. *Nucleic Acids Research*, 2017, 45(D1): D611–D618.

[12] Aagaard K, Petrosino J, Keitel W, Watson M, Katancik J, Garcia N, Patel S, Cutting M, Madden T, Hamilton H, Harris E, Gevers D, Simone G, McInnes P, Versalovic J. The Human Microbiome Project strategy for comprehensive sampling of the human microbiome and why it matters. *The FASEB Journal*, 2013, 27(3): 1012–1022.

[13] Qin JJ, Li RQ, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li JH, Xu JM, Li SC, Li DF, Cao JJ, Wang B, Liang HQ, Zheng HS, Xie YL, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu HM, Yu C, Li ST, Jian M, Zhou Y, Li YR, Zhang XQ, Li SG, Qin N, Yang HM, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, MetaHIT Consortium, Bork P, Ehrlich SD, Wang J. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 2010, 464(7285): 59–65.

[14] Gilbert JA, Jansson JK, Knight R. The earth microbiome project: successes and aspirations. *BMC Biology*, 2014, 12(1): 69.

[15] Glass EM, Wilkening J, Wilke A, Antonopoulos D, Meyer F. Using the Metagenomics RAST Server (MG-RAST) for analyzing Shotgun metagenomes. *Cold Spring Harbor Protocols*, 2010, 2010(1): pdb.prot5368.

[16] Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA. The metagenomics RAST server-a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 2008, 9(1): 386.

[17] Kyrpides NC. Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects world-wide. *Bioinformatics*, 1999, 15(9): 773–774.

[18] Markowitz VM, Chen IMA, Chu K, Szeto E, Palaniappan K, Pillay M, Ratner A, Huang JH, Pagani I, Tringe S, Huntemann M, Billis K, Varghese N, Tennessen K, Mavromatis K, Pati A, Ivanova NN, Kyrpides NC. IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Research*, 2014, 42(D1): D568–D573.

[19] Mitchell A, Bucchini F, Cochrane G, Denise H, ten Hoopen P, Fraser M, Pesseat S, Potter S, Scheremetjew M, Sterk P, Finn RD. EBI metagenomics in 2016-an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Research*, 2016, 44(D1): D595–D603.

[20] Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*, 2010, 26(2): 266–267.

[21] Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K,

Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 2010, 7(5): 335–336.

[22] Ju F, Zhang T. 16S rRNA gene high-throughput sequencing data mining of microbial diversity and interactions. *Applied Microbiology and Biotechnology*, 2015, 99(10): 4119–4129.

[23] DeSantis TZ, Hugenholtz P, Larsen N, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, 2006, 72(7): 5069–5072.

[24] Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, McGarrell DM, Garrity GM, Tiedje JM. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Research*, 2005, 33(S1): D294–D296.

[25] Chun J, Lee JH, Jung Y, Kim M, Kim S, Kim BK, Lim YW. EzTaxon: a web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences. *International Journal of Systematic and Evolutionary Microbiology*, 2007, 57(10): 2259–2261.

[26] Kim OS, Cho YJ, Lee K, Yoon SH, Kim M, Na H, Park SC, Jeon YS, Lee JH, Yi H, Won S, Chun J. Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. *International Journal of Systematic and Evolutionary Microbiology*, 2012, 62(3): 716–721.

[27] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 2013, 41(D1): D590–D596.

[28] Matsuda F, Tsugawa H, Fukusaki E. Method for assessing the statistical significance of mass spectral similarities using basic local alignment search tool statistics. *Analytical Chemistry*, 2013, 85(17): 8291–8297.

[29] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of Molecular Biology*, 1990, 215(3): 403–410.

[30] Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Research*, 2002, 12(4): 656–664.

[31] Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 2010, 26(19): 2460–2461.

[32] Pruesse E, Peplies J, Glöckner FO. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*, 2012, 28(14): 1823–1829.

[33] Soergel DAW, Dey N, Knight R, Brenner SE. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *The ISME Journal*, 2012, 6(7): 1440–1444.

[34] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 2004, 32(5): 1792–1797.

[35] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li WZ, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 2011, 7(1): 539.

[36] Lassmann T, Sonnhammer EL. Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, 2005, 6(1): 298.

[37] Notredame C, Higgins DG, Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 2000, 302(1): 205–217.

[38] Papadopoulos JS, Agarwala R. COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics*, 2007, 23(9): 1073–1079.

[39] Price MN, Dehal PS, Arkin AP. FastTree 2-approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 2010, 5(3): e9490.

[40] Notredame C. UNIT 3.8 computing multiple sequence/structure alignments with the T-coffee package. *Current Protocols in Bioinformatics*, 2010, doi: 10.1002/0471250953.bi0308s29.

[41] Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 2009, 75(23): 7537–7541.

[42] Gobet A, Quince C, Ramette A. Multivariate Cutoff Level Analysis (MultiCoLA) of large community data sets. *Nucleic Acids Research*, 2010, 38(15): e155.

[43] Kyrpides NC, Eloe-Fadrosh EA, Ivanova NN. Microbiome data science: understanding our microbial planet. *Trends in Microbiology*, 2016, 24(6): 425–427.

Big data management and analysis of microbiome

Heyuan Qi, Qinglan Sun, Juncai Ma*

Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China

Abstract: Advances in high-throughput sequencing have allowed significant breakthroughs in microbial ecology studies. This has led to the rapid expansion of research in the field metagenomics, which is often defined as the analysis of DNA sequences from microbial communities in environmental samples without prior need for culturing. Microbiome data management and analysis have caused concern for microbial researchers because of the dramatic increase of metagenomics data. It has been a research hotspot that how to mine valuable information for further application form such big microbial data. Till now, many metagenomics computational tools and databases have been provided in order to allow the exploitation of the huge influx of data. In this review article, we provide an overview of the sequencing technologies and international microbiome projects as well as the platforms for microbial data archiving and analysis, such as Human Microbiome Project (HMP), Earth Microbiome Project (EMP), Metagenomics of Human Intestinal Tract (MetaHIT), MG-RAST, iMicrobe, Integration Microbial Genomes (IMG) and EBI Metagenomics and so on. We also discussed about the basic pipelines and main tools for metagenomics data. Finally, we proposed the necessity of establishing a platform for multi-source microbial data management and bioinformatic analysis.

Keywords: microbiome, metagenomics, data analysis, high throughput sequencing

(本文责编：张晓丽)

Supported by National Key Research and Development Plan (2016YFC0901702, 2016YFB1000605) and by the National High Technology Research and Development Program of China (863 Program) (2015AA020108)

*Corresponding author. Tel: +86-10-64807422; Fax: +86-10-64807426; E-mail: ma@im.ac.cn

Received: 16 February 2017; Revised: 4 April 2017; Published online: 11 April 2017