DOI: 10.13376/j.cbls/2018107
文章编号：1004-0374(2018)08-0896-10

· 技术与应用 ·

长读长测序技术在宏基因组学研究中的应用

孙 丹[1]，袁文亮[2]，彭司华[1]*
(1 上海海洋大学水产与生命学院发育生物学系，水产种质资源发掘与利用教育部重点实验室，农业部国家水生动物病原库，科学技术部海洋生物科学国际联合研究中心，上海 201306；2 上海理工大学光电信息与计算机工程学院，上海 200093)

摘 要：由于很多微生物无法单独分离培养，研究微生物群落整体的宏基因组学是目前揭示微生物多样性的重要方法。长读长测序技术可以覆盖重复序列和复杂结构，获得短读长无法检测的基因组信息。现着重介绍了两类长读长测序技术，即基于第三代测序技术的单分子长读长测序技术和基于片段相互联系的合成长读长测序技术，并进一步介绍了长读长测序技术在宏基因组学领域的应用。

关键词：宏基因组；长读长测序技术；合成长读长；第三代测序技术；微生物
中图分类号：Q78；Q933-3 **文献标志码**：A

Application of long-read sequencing technologies in metagenomics research

SUN Dan[1], YUAN Wen-Liang[2], PENG Si-Hua[1]*
(1 Department of Developmental Biology of College of Fisheries and Life Science, Key Laboratory of Exploration and Utilization of Aquatic Genetic Resources of Ministry of Education, National Pathogen Collection Center for Aquatic Animals of Ministry of Agriculture, International Research Center for Marine Biosciences of Ministry of Science and Technology, Shanghai Ocean University, Shanghai 201306, China; 2 School of Optical-Electric and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Because most microorganisms cannot be isolated and cultured alone, the study of the metagenomics of microbial communities is important to reveal microbial diversity. Long-read sequencing technology may cover repetitive sequences and complex structures, obtaining genomic information that cannot be detected by short-read sequencing. This review focuses on two types of long-read sequencing technology: single-molecule long-read sequencing technology based on third generation sequencing technology and synthetic long-read sequencing technology based on interconnection of DNA fragment. Furthermore, the application of long-read sequencing technology in metagenomics is also reviewed.

Key words: metagenomics; long-read sequencing; synthetic long-read; third generation sequencing; microbe

微生物是地球生物圈的重要组成部分，并且在生态系统的过程和功能中扮演了独特角色[1]。另外，微生物与人体的健康状况息息相关，不仅与感染性疾病的治疗和预后有关[2]，而且肠道菌群也影响着人体代谢[3]、肠道营养吸收[4]和人体免疫系统调节[5]。与动植物共生的微生物也可以带来农业水产和生物技术方面的经济价值，尤其是在生物能源和生物降解等领域[6-7]。尽管微生物的分离培养方面取得了很多进展[8-9]，但是自然界中仍然有很多微生物难以分离培养[10]。无法单独培养和测序的微生物被称为“暗物质”[11]，对不能分离培养的微生物的研究存在很大的困难，于是Schmidt等[12]和Handelsman等[13]率先提出宏基因组学(metagenomics)，其特点是应用现代基因组学的技术直接研究自然状态下的微生物群落，而无需分离培养。

10多年来，新一代测序技术(next generation

收稿日期：2018-05-07；**修回日期**：2018-06-01
基金项目：上海市自然科学基金项目(15ZR1420800)；上海海洋大学博士科研启动项目(A2-0203-00-100313)
***通信作者**：E-mail: shpeng@shou.edu.cn

sequencing, NGS) 在 DNA 测序领域得到了广泛的应用，并且随着计算机技术发展掀起了基因组领域的革命 [14]。这些技术进步不但使人们能快速获取并分析基因组数据，而且也使得进行基因组学和宏基因组学等大型项目成为了可能。然而，基于 NGS 的基因组数据存在重复序列、不均匀分布、测序错误和株间差异等问题，使宏基因组的 *de novo* 拼接十分困难 [15-17]。更长的测序读长就有可能跨过复杂区域，甚至覆盖完整微生物基因组，从而能够更好地解密微生物群落。长读长 (long-read) 是个不断发展的概念。2007 年 Bench 等 [18] 和 2008 年 Wommack 等 [19] 率先注意到宏基因组研究中的测序读长问题，将测序长度超过 600 bp 称为长读长。早期第三代测序平均读长也仅仅为 1 300 bp。随着测序技术的发展，很多高质量的长读长测序方法出现。2015 年，Koren 和 Phillippy[20] 将理论读长 7 kbp 以上 (黄金阈值) 的测序技术定义为长读长测序技术 (long-read sequencing technologies)。本文将综述长读长测序技术的基本原理和最新进展，以及其在宏基因组学领域的最新应用，并讨论其在宏基因组学研究中的优越性和未来发展趋势。

1　长读长测序技术

宏基因组学被广泛用于微生物研究，可以从微生物群落重构个体基因组、预测基因功能与网络和进行生物多样性分析等 [1,21-22]。在宏基因组学研究中，7 kb 的测序读长是一个重要临界值，被视作“黄金阈值”(golden threshold)[20]。这是因为微生物的 rDNA 操纵子的全长在 5~7 kb 之间，也是目前 77% 已知微生物基因组的最大重复序列 [23]。当长读长超过 7 kb 时，可以实现通过全长 rDNA 序列研究微生物多样性，不只是比对部分 rDNA 序列，这样更能够发现 rDNA 序列结构的差异。7 kb 的长读长使得在拼接基因组时能够克服构建个体完整基因组最主要的困难 —— 重复序列，并有机会将大多数微生物组装出完整的个体基因组序列。这样可以得到绝大多数操纵子的完整信息，用于微生物基因组注释，并可分析种内水平变异造成的调控差异。

长读长测序技术分为两大类，分别是合成长读长 (synthetic long-read, SLR) 和单分子长读长 (single molecule long-read, SML)[24]。其中，单分子长读长测序技术是基于目前主流的第三代测序技术，包括 PacBio SMRT 技术 [25] 和 ONT nanopore 技术 [26]。它们可以实时长读长检测单分子核酸，其中 ONT nanopore 更是宣称能够实现“超长读长”(ultra-long reads，约 1 Mb)[27]。合成长读长测序是通过建库时在长片段 DNA 打断过程中加入特殊标记，通过标记将短读长测序结果合并成长读长测序结果。目前合成长读长测序技术主要基于 Illumina 第二代测序技术，但这些策略也有潜力应用于第三代测序。这些长读长测序技术的测序平台各有不同的通量和原始读长，如表 1 所示。不同测序平台由于各自不同的原理缺陷会产生不同程度的测序错误。Illumina 的测序错误主要源于聚合酶合成错误以及 GC 偏差 [28]。而第三代测序技术则普遍存在均聚物错误、插入删除错误 (indel) 和随机性错误 [29]。

1.1　长片段核酸的提取

无论是合成长读长还是单分子长读长测序，获得有效长读长的先决条件是提供给测序仪器的核酸片段足够长。目前在提取微生物宏基因组过程中，常常使用物理方法来破碎厚壁或者多层细胞壁，如氧化锆珠法和热酚法，但这些方法会产生相对原位裂解法更短的核酸片段 [30]。即便是原位裂解法，仍然会破坏核酸，限制测序读长，特别是对于难以裂解的革兰氏阳性菌 [24]。市面上普遍使用的离心柱法试剂盒，虽然简单快速易用，但在许多时候也会导致无法得到 10 kb 的长片段核酸。虽然这些方法可以针对某一种生物进行优化，但这并不适用于组成

表1 测序平台基本情况比较

测序平台	测序读长	测序反应时间	测序反应产量	上市时间
Illumina Miseq系列	2*300 bp	4~55 h	15 Gb	2011年
Illumina Hiseq系列	2*150 bp	7 h~6 d	1 500 Gb	2012年
Iilumina Novaseq系列	2*150 bp	19~40 h	6 000 Gb	2017年
PacBio RS II	250 bp~40 kb	0.5~6 h	400 Mb	2013年
PacBio Sequel	30 kb	4 d	10 Gb	2015年
ONT MinION	5~200 kb	1 min~48 h	10~20 Gb	2015年
ONT PromethION	<1 Mb	1 min~64 h	4.3~6 Tb	目前尚未上市

*，双端测序，一个从3'测序，一个从5'测序。

复杂的宏基因组。所以，需要使用更温和的方法，如基于磁珠分离、电泳分离、琼脂糖包埋或者鸡尾酒酶溶解法等手段[31-33]。同时，必须使用广口吸头并缓慢吸取，才能得到完整长片段核酸。

此外，随着 NGS 相关技术的发展，连续稀释和微流控技术改变了 NGS 测序前准备工作。例如，ONT 推出的 VolTRAX 微流控系统，可以在一个系统中全自动完成溶解、核酸提取等 nanopore 测序前准备工作。

1.2 长片段文库制备

NGS 测序文库制备有多种方法，包括连接法、转座酶法和标记法等[30]。制备文库的流程包括核酸的片段化、末端修复、连接接头 (adapter)、文库片段纯化、扩增 (如果需要) 等步骤。其中，核酸片段化的步骤对于长读长测序来说至关重要。使用 Covaris 开发出的 g-TUBE 能够实现 6~20 kb 的片段化。g-TUBE 配合兼容的台式离心机，依靠离心力使样品通过一个精确制造的孔口，产生对核酸片段的剪切力。通过调整离心转子转速和孔口的流速，从而实现剪切力的变化，最终获得合适的 DNA 片段尺寸。另外，现在也出现了一些酶法片段化和快速建库的方法，如利用转座酶建库 (Nextera，Illumina)，同时完成片段化和标记的过程[34]，这种方法大幅节省了文库制备的时间。“片段化酶”通过控制反应时间来获得所需片段长度[35]。酶法相对机械法的优点在于自动化程度高，从而避免了人为因素产生误差。该法更能保证测序长读长的稳定实现。然而，酶法片段化的缺陷在于酶切具有一定程度偏好性[36]，对于宏基因组这样物种组成复杂的情况，需要谨慎使用。

1.2.1 单分子长片段文库

在宏基因组研究中，运用单分子长读长技术的终极目标是完整地展现整个微生物基因组以及其携带的质粒。微生物基因组规模大约在 1~10 Mb[24,37]。目前 PacBio 的 SMRT 技术和 ONT 的 nanopore 技术尚不能完整测序这样的序列长度，仍然需要打断建库分别测序。其中 PacBio 的建库，修复末端后，将双链序列片段两端分别连接环状单链。环状单链两端分别与双链序列片段的正负链连接，得到一个类似哑铃 (“套马环”) 的结构，称为 SMRT Bell[38]。ONT 推出的全自动测序前制备系统 VolTRAX 前面已经介绍，不再赘述。在此基础上，Mostovoy 等[39]发明的 Bionano 高通量单分子基因组距离分析系统，可以实现单分子长片段的进一步组装。Bionano 的核心技术基于光学图谱和微流控，其原理是完整的单分子 DNA 在正电荷作用下缓慢拉直通过纳米流体孔，接着被拽入纳米管道并被附着到透明支架上。限制性内切酶将 DNA 特异性剪切并荧光标记，最后进行光学成像。如此能够在单分子层面定位长片段 DNA，简化了基因组的组装。

对于动态单分子长片段技术来说，除能够实现长读长外，另一大优势在于没有 PCR 扩增。这样能够最小化杂信号的干扰，所以测序模板核酸是高保真、几乎无 GC 偏好的[40]。由于同时保留了核酸的表观遗传印记，并且 RNA 也不会因逆转录过程丢失信息或产生误差，这些将使得一般宏基因组研究进一步发展至表观宏基因组 (epimetagenomic)[30]、宏转录组 (metatranscriptomic)[41]，甚至表观宏转录组 (epi-metatranscriptomic)。

1.2.2 合成长片段文库

合成长读长测序则需要在文库制备过程中通过生物化学方法建立小片段 DNA 之间的联系。利用统计学和计算机的手段，追踪检测这些小片段，通过联系连接成合成长读长。这些合成长读长建库策略又可以分为稀释标记法和相邻标记法。

近年，Illumina 先后推出 Truseq 合成长读长测序文库 (Moleculo)[42] 和邻近保留转座测序文库 (CPT-seq)[43]。这些基于稀释标记的合成长读长测序策略，使得第二代测序技术在一定程度上克服了短读长的弱点。此外，基于类似原理的还有 Complete Genomics[44] 和 10X Genomics[39]。这些测序方法都是将溶液稀释成只存在个别大片段 DNA 分子的液滴；接着，将 DNA 分子打断并加入索引序列制备成子文库；最后，将子文库合并测序后进行亚组装。Moleculo 和 CPT-seq 分别使用 384 和 96×96 孔板稀释成数百，甚至数千份。而 10X Genomics 则利用微流体操控全自动生成超过一百万个油包水微滴，每个微滴都有一个独特的 16 nt 标签序列标记 1 个长片段 DNA。之后，这条长片段 DNA 被剪切，并连接上特定标签生成子文库。10X Genomics 提供了一种有效测序长片段 DNA 的方法，并且其具有高效全自动的特点。这使研究人员摆脱了多步 384 孔板加样稀释制备小文库的繁琐工作，避免了人为误差。因此，目前 10X Genomics 占有了合成长读长测序的绝大部分市场。

高通量 / 分辨率染色体构象捕获 (Hi-C)[45] 则是基于相邻标记法。与稀释标记法不同，这是一种保留大片段 DNA 分子连续结构信息的建库方法，其

原理是通过特异性切割后在核酸末端连接生物素标记，在其指导下序列被定位到基因组上。以此为依据对互作片段进行相应评分，构建一个包含基因组所有片段的连接频率矩阵。Hi-C 技术可以交联完整细胞内的相邻 DNA 分子，产生成对序列，描述超长片段基因组连续性。这项技术不但能够捕捉真核生物中多条染色体的相互作用，同样也被用于原核生物中建立质粒与宿主基因组的联系。成对序列只连接同时出现在同样细胞内的 DNA 分子，有助于宏基因组的去卷积和组装。Burton 等 [46] 报道了利用 Hi-C 技术，能够解决宏基因组样本中质粒与其宿主细菌基因组共出现问题。Stewart 等 [47] 利用 Hi-C 技术和其他宏基因组研究方法，从 43 头苏格兰牛瘤胃宏基因组中共组装出 913 个接近完整的微生物基因组草图，并解释 Hi-C 如何改善非监督聚类 (binning) 无法识别质粒与宿主关系的问题。

1.3　长读长测序的组装

组装 (assembly) 是将测序得到的序列片段 (reads) 拼接成一个连续重复覆盖序列 (contiguous subsequence, contig)。理想状况是一个 contig 描述一个完整的微生物基因组，采用长读长测序技术使之成为可能。目前组装算法有重叠排列一致法 (overlap-layout-consensus, OLC) 和 de Bruin 图法 (de Bruijn graph, DBG) 两种 [48]。OLC 法适用于 >1 kb 的长序列，计算所有序列的重叠，并简化结果覆盖图，直到获得一个长 contig[49]。OLC 组装的缺点是计算大量序列重叠对计算资源需求较高，而选择另一种 DBG 法处理这个问题是先把结果序列片段拆分 *k*-mers (*k*<127) 并将它们连接成图 [50]。de Bruijn 图法产生许多需要考虑的顶点，但其缺点是失去了长读长中蕴含的重要连续信息。

传统 OLC 法包括 Celera Assembler[49]、Falcon[51] 等，这些方法适合于错误率较低的长读长测序技术。而 TruSPAdes 软件 [52] 是在 SPAdes 软件的基础上开发出来的一种针对长读长测序的 de Bruijn 拼装方法。在此基础上，BIGMAC[53]、LoRMA[54] 等融合两者优点，利用图加快了重叠算法的运算效率。kd-tree-overlapper[55] 更是在此基础上，使用几何嵌入和 KD 树近似最近相邻 (ANN) 检索，可以获得 Mb 级别的单个 contig，而微生物个体基因组也仅仅是数 Mb。

对于合成长读长来说，还有一种被称为“读长云”(read cloud)[56] 的策略。当由于样本、技术、经费等限制，造成测序数据不足以得到高质量的长读长片段时，可以将获得的聚类短读长片段作为读长云。这里“云”指的是基因组可视化后，通常形成孤立聚类的序列片段。虽然无法生成长连续序列，但读长云技术仍然包含解决基因组重复序列的信号。Architect[56]、Minerva[57] 等算法软件可以提取这个信号，利用其改善宏基因组组装。Frank 等 [58] 早在 2016 年就结合 PacBio 单细胞测序和 Illumina 的 HiSeq 技术，显著改善了宏基因组测序数据拼装和分类 (binning) 效果。

为了更有效地改进宏基因组拼装效果，一种有前途的方案是结合单细胞测序、宏基因组测序、宏转录组测序技术的技术方案。2018 年，Xu 和 Zhao[59] 发表的综述文章给出了很详细的描述。我国学者赵方庆课题组 2017 年发表的成果表明，结合流式细胞仪和单细胞测序以及他们特有的算法，从宏基因组数据中得到了 75 个细菌的草图基因组，显示这是一种十分有竞争力的技术方案 [29]。

2　长读长测序技术在宏基因组学研究中的应用

2.1　微生物多样性分析

2008 年，Wommack 等 [19] 报道了在两个微生物数据库和一个浮游病毒数据库中，利用 BLAST 和 COG 的分析结果比较短读长和长读长在宏基因组学的微生物多样性分析方面的优劣性，结果发现，短读长 (<400 bp) 表现出更少的同源性；这个现象在浮游病毒上表现更为显著，并且短读长会丢失在系统分类中远端序列的信息。研究结果证实了读长问题是摆在宏基因组研究面前的重要问题，只有长读长序列才能保证微生物多样性研究的精确性。由于宏基因组研究对象微生物的 rDNA 操纵子全长在 5~7 kb 之间 [60]，主流长读长测序技术 10 kb 的测序读长意味着可以完全覆盖整个 rDNA 操纵子而不仅是 16S rDNA 的部分区域。此外，由于长读长测序技术能够跨越 rDNA 序列，使得 rDNA 序列能够更好地与其基因组聚类到一个 bin 中。这是因为 rDNA 往往具有保守型，基于非监督学习方法的 binning 软件很难对其正确聚类 [61]。在本课题组未发表的研究结果中，长读长测序使得含有 16S rDNA 的 bin 数量增加近一倍。

Kuleshov 等 [62] 利用 Illumina 的 Truseq 合成长读长从人类肠道宏基因组中鉴别了 178 种细菌，其中 51 种是通过短读长方法无法发现的，利用长读长测序技术揭示出了前所未有的肠道宏基因组多样性，特别是种内的多样性。研究者发现菌株种内的

快速进化能够影响人类生理机能。而 Tedersoo 等 [63] 利用 PacBio 的 SMRT 测序技术评估了全长转录间隔序列 (ITS) 和更长的 rDNA 扩增子用作土壤真核微生物的宏条形码 (metabarcoding)，证明全长 ITS 和长 rDNA 序列显著提高了种间以及门水平的分类学识别。此外，国内孙志宏团队 [64] 利用 PacBio 的 SMRT 技术测序全长 16S rDNA 分析重庆萝卜泡菜生物多样性和种群结构，发现“胭脂红”比“春不老”有着更显著的生物多样性以及假单胞菌多样性，并发现了 3 种致病菌。目前这些研究较多基于 PacBio 测序技术，虽然较长的读长有助于同源识别，但同时存在低通量和 indel 错误倾向的问题 [65] 以及较低的测序准确率 [66]，所以，目前较多引入二代测序辅助测序碱基纠错。2018 年，Beaulaurier 等 [67] 也采用了 PacBio 的 SMRT 技术测序，对得到的长读长序列进行拼接后，创新地采用了基因组甲基化信息来实现宏基因组种和菌株级别的精确分类 (binning)。

另外，值得提出的是，PacBio 的 SMRT 技术与 16S RNA 和 18S RNA 测序数据结合，可以得到更多的全长 16S RNA 和 18S RNA 序列 [68]。

2.2 构建微生物个体基因组

构建完整基因组有助于检测基因组结构多样性，如大范围 indel 变异和微生物群落中水平基因转移 [69]。微生物对环境的适应是一个动态的过程，除了积累点突变之外，微生物可以利用转座酶和质粒完成快速的种内进化，在环境胁迫下获得新的基因或者增加基因拷贝数。然而，现有的通过短读长测序技术分析宏基因组数据的方法，无法描述亲缘关系密切的共发生菌株的基因组结构差异。例如，Moss 等 [70] 利用 10X Genomics 合成长读长技术研究了干细胞移植后白血病患者的粪便宏基因组。他们发现在经历强化疗法的巨大选择压力后，成为优势菌的粪便拟杆菌转座子集成位置的差异和基因组岛现象 (大区域转移)，并证实因此产生的抗生素抗性基因的过表达现象；而这些决定功能差异的微生物个体基因组结构差异，在以往碎片化的短读长组装结果中是无法得到的。

在宏基因研究中，构建微生物个体完整基因组有诸多困难。其中最主要的是重复序列，而 rDNA 是最广泛存在的大片段重复序列。在基因组拼接中，reads 通过片段重叠能够组装成一个更大的片段，称为 contig。多个 contigs 通过片段重叠，组成一个更长的 scaffold。而 gap 是得到的 scaffold 中含有的一定长度的未知序列。构建所谓完整基因组的理想情况就是没有 gap 和 scaffold，一个 contig 就能够描述完整的微生物基因组。理论上能够克服 7 kb“黄金阈值”障碍的长读长测序技术大大降低了构建个体基因组复杂性，可以将 80% 的微生物组装出完整的个体基因组序列 [20]。长短读长对 de Bruijn 图的影响对比如图 1 所示，超过黄金阈值的 *k* 值只用一个 contig 就构建了 *E. coli* K12 的个体基因组 [23]。

由于一般长读长测序技术存在测序通量较低和测序准确率不高的问题，目前较为常用的做法是通过增加测序一部分短读长高通量数据，用于矫正长读长数据的测序错误，这样的测序策略可以大大降低总的测序费用。White 等 [71] 利用 Illumina 的 Truseq

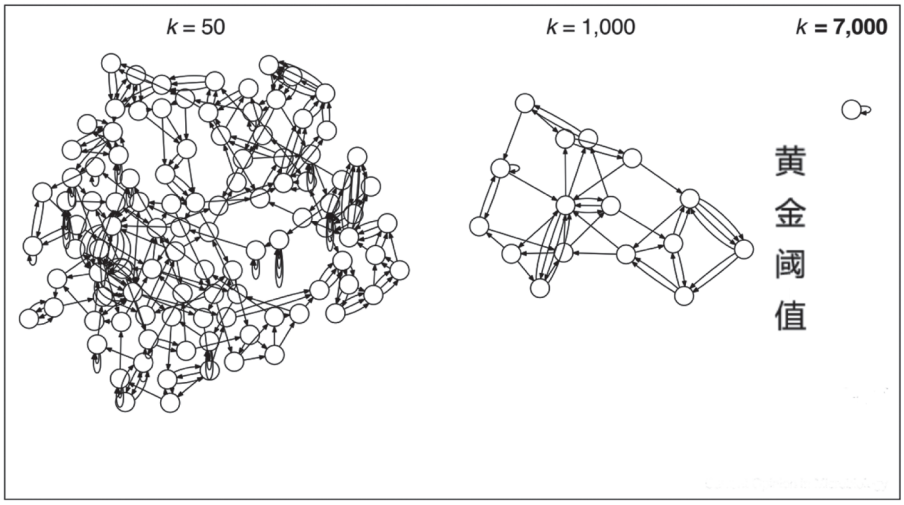

图1 *E. coli* K12基因组在不同*k*值情况下，contigs所组成的de Bruijn图[23]

合成长读长 + 短读长得到 >100 个微生物基因组，并构建了一个新的假单胞菌候选种的完整基因组。Tsai 等 [72] 利用 PacBio 的 SMRT 测序 + 短读长从宏基因组中构建了几株新发现的棒杆菌及其同伴噬菌体的完整基因组。而本课题组在研究青藏高原那根拉山土壤样本时，在使用 Illumina 高通量测序数据基础上加测了一部分 PacBio 的长读长数据。在未发表的结果中，个体基因组构建质量有了大幅的提升，90% 以上的完整度的微生物基因组从 12 个提升到 36 个，且所有基因组都包含完整 rDNA 操纵子。

现在也有一些宏基因组研究独立使用长读长测序技术，并且构建出完整的个体基因组。Bishara 等 [73] 基于读长云策略，使用 10X Genomics 合成长读长测序技术，从复杂的海洋沉积物样本中获得 23 个基因组，其中 9 个是完整基因组草图。Driscoll 等 [74] 利用 PacBio 的 SMRT 测序技术从淡水宏基因组中构建了 3 种固氮蓝细菌的共生菌的完整基因组，并发现其中一种拟杆菌有着全新的生物合成通路。Batovska 等 [75] 使用 MinION 非靶向检测了虫媒病毒。

此外，值得注意的是，所有基于第二代测序的测序方法都有着 GC 偏差的内在缺陷 [28]，不适合分析群落中包含个体基因组 GC 含量超过 60% 的微生物 [76]。极地和热液口等极端环境下的微生物往往有着较高的 GC 含量，因此，采自极端环境的宏基因组样本建议采用基于第三代测序技术的方法进行测序。Park 研究团队利用 SMRT 测序分别得到 GC 含量为 70.89% 和 71.8% 的南极链霉菌完整基因组 [40,77]。

2.3　注释基因功能与网络

基因注释是通过对两条 (或多条) 核酸序列进行排列比对，获得最大的相似程度，以此评估序列的同源性，并进一步推测序列的功能信息 [78]。对于一般宏基因组研究，首先将测序得到的原始数据组装成 contigs 或 scaffolds，并对其进行 binning 聚类。接着，软件预测注释基因及其蛋白，并且对种间同源和功能富集聚类。宏基因组学基因注释分析流程如图 2 所示。相对于传统的功能宏基因组，基于长读长测序技术的宏基因组数据分析，能够得到更完整的基因序列、上下游调控和系统发育信息。如 Oh 等 [79] 利用 PacBio 的 SMRT 测序技术，通过软件 CAT 比对糖类酶数据库的方法，从宏基因组中检测降解木质素的微生物酶，并发现了在温和条件下无需预处理的新木质素降解酶。得益于更长的读长，能够得到更完整的基因信息和更正确的群落组成。由此，宏基因组数据分析可以揭示出更完整的功能分子生态网络和群落互作 [1]。Slaby 等 [61] 利用 PacBio 的 SMRT 测序技术分析海绵微生物共生群落，发现了细菌防御 (如毒性 - 抗毒系统)、宿主寄生和细胞外基质利用相关基因的富集，揭示出微生物联合防御以及特殊的新陈代谢。

长读长测序技术的另一大优势在于能够有效地预测基因，而无需组装和 binning 聚类。而传统第

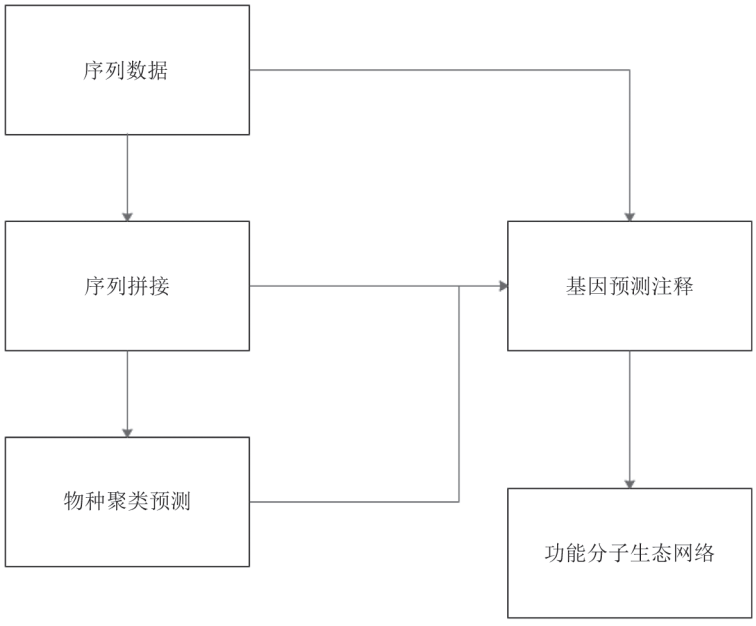

图2　宏基因组学基因注释分析流程

二代高通量测序技术的读长过短，无法单独进行基因的可靠识别[80]。随着NGS技术的发展，生命科学面临着原始数据泛滥的问题。从序列组装开始的传统宏基因组分析对于计算资源和人员要求都较高，如10 G以上的土壤宏基因组原始数据往往需要200~500 G内存、4线程运行2 d以上。因此，也出现了一些帮助宏基因组学数据分析的云计算平台，如Blacklight[6]。Sharon等[80]单独利用Illumina的Truseq合成长读长测序技术所产生的>5 kb的合成长读长原始宏基因组数据，注释基因及其蛋白，并进一步富集聚类了数千有着20成员的基因家族。此外，由于短读长序列组装依赖于足够的测序深度，任何基于短读长测序的分析都可能会遗漏群落中低丰度基因组中的重要基因信息[80]。长读长测序技术能够覆盖许多完整操纵子，有效注释基因功能与网络。为了从宏基因组原始数据中提取生物信息，Bras等[81]开发了Colib'read on galaxy软件包，可以得到系统树状图和热图的可视化结果。此外，AnnoTALE可以从长读长序列中筛选并注释转录激活因子样效应物核酸酶(transcription activator-like effector nucleases, TALENs)基因[82]。TALEN、ZFN和CRISPR/Cas是三大类基因组编辑核酸酶，可以靶向目的基因并改造基因组序列[83]。这些技术可以更加快速高效地从环境宏基因组中筛选出所需要的功能基因，特别是对于那些具有重复序列的基因。

3 结论与展望

由于宏基因组学中长读长相对短读长的优越性，合成长读长和单分子长读长测序技术被引入了宏基因组学研究之中。长读长可以跨越重复序列、复杂结构，能够显示短读长无法检测的序列结构和连接多样性，显著地改善宏基因组中短读长组装碎片化问题。

获得长读长序列可以利用多种算法组装成更长的contig，包含较长片段里完整的结构信息。长读长测序帮助研究者区分出单独使用短读长无法分辨的微生物多样性，特别是种内级别的系统发育。长读长测序技术能够构建出更完整的微生物基因组。非碎片化的数据不但能分析基因信息，更能体现基因组结构上的差异，从而解释微生物如何适应环境，如何快速进化。同时，更长的读长更能够节省计算资源，在注释基因过程中，无需预先组装，这样就大大加快了宏基因组研究中筛选功能基因的过程。随着测序费用的降低，长读长测序技术将会在宏基因组学研究中得到更广泛的应用。目前，长读长测序技术在通量和准确率上相对传统短读长高通量的第二代测序仍然处于劣势。此外，第三代测序技术在检测广泛存在甲基化修饰的极端环境微生物[84]时有着更显著的优势。基于隐马尔科夫模型的最新技术可以分析多种甲基化碱基[85]，使研究者获取到宏基因组中的表观遗传信息。进一步发展的长读长测序技术将在气候变化、抗生素耐药等重大科学问题的研究中具有更多的优势。

以PacBio为代表的长读长测序技术本身还有一些缺陷，如测序成本现在还大大高于Illumina为代表的二代测序技术。另外，长读长测序错误率较高，这不仅使基因组拼装质量受到影响，而且容易造成后续的基因预测片段化，给宏基因组注释带来一定困难。如果通过加大PacBio的测序深度，或用二代测序reads做测序数据矫正，无疑又增加了测序成本，相信随着三代测序技术的改进，这些问题会逐渐得到解决。

[参 考 文 献]

[1] Zhou J, Deng Y, Luo F, et al. Functional molecular ecological networks. MBio, 2010, 1: e00169-00110

[2] 袁易, 王铭杰, 张欣欣. 第三代测序技术的主要特点及其在病毒基因组研究中的应用. 微生物与感染, 2016, 11: 380-4

[3] 李旻. 人体肠道菌群结构与宿主代谢的相关性研究 [D]. 上海: 上海交通大学, 2009

[4] Goodman AL, McNulty NP, Zhao Y, et al. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. Cell Host Microbe, 2009, 6: 279-89

[5] Lee YK, Mazmanian SK. Has the microbiota played a critical role in the evolution of the adaptive immune system? Science, 2010, 330: 1768-73

[6] Couger MB, Pipes L, Squina F, et al. Enabling large-scale next-generation sequence assembly with Blacklight. Concurr Comput, 2014, 26: 2157-66

[7] Roumpeka DD, Wallace RJ, Escalettes F, et al . A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. Front Genet, 2017, 8: 23

[8] Seshadri R, Leahy SC, Attwood GT, et al. Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. Nat Biotechnol, 2018, 36: 359-71

[9] Bai Y, Mueller DB, Srinivas G, et al. Functional overlap of the Arabidopsis leaf and root microbiota. Nature, 2015, 528: 364-82

[10] Streit WR, Schmitz RA. Metagenomics--the key to the uncultured microbes. Curr Opin Microbiol, 2004, 7: 492-8

[11] Rinke C, Schwientek P, Sczyrba A, et al. Insights into the phylogeny and coding potential of microbial dark matter.

Nature, 2013, 499: 431-7
[12] Schmidt TM, DeLong EF, Pace NR. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J Bacteriol, 1991, 173: 4371-8
[13] Handelsman J, Rondon MR, Brady SF, et al. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol, 1998, 5: R245-9
[14] Koboldt DC, Steinberg KM, Larson DE, et al. The next-generation sequencing revolution and its impact on genomics. Cell, 2013, 155: 27-38
[15] Afiahayati, Sato K, Sakakibara Y. MetaVelvet-SL: an extension of the Velvet assembler to a *de novo* metagenomic assembler utilizing supervised learning. DNA Res, 2015, 22: 69-77
[16] Namiki T, Hachiya T, Tanaka H, et al. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res, 2012, 40: e155
[17] Peng Y, Leung HC, Yiu SM, et al. Meta-IDBA: a *de novo* assembler for metagenomic data. Bioinformatics, 2011, 27: i94-101
[18] Bench SR, Hanson TE, Williamson KE, et al. Metagenomic characterization of Chesapeake Bay virioplankton. Appl Environ Microbiol, 2007, 73: 7629-41
[19] Wommack KE, Bhavsar J, Ravel J. Metagenomics: read length matters. Appl Environ Microbiol, 2008, 74: 1453-63
[20] Koren S, Phillippy AM. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol, 2015, 23: 110-20
[21] Herlemann DP, Lundin D, Labrenz MF, et al. Metagenomic *de novo* assembly of an aquatic representative of the verrucomicrobial class *Spartobacteria*. MBio, 2013, 4: e00569-12
[22] Tyson GW, Chapman J, Hugenholtz P, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature, 2004, 428: 37-43
[23] Koren S, Harhay GP, Smith TP, et al. Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol, 2013, 14: R101
[24] Olson ND, Treangen TJ, Hill CM, et al. Metagenomic assembly through the lens of validation:recent advances in assessing and improving the quality of genomes assembled from metagenomes. Brief Bioinform, 2017 [Epub ahead of print]
[25] Eid J, Fehr A, Gray J, et al. Real-time DNA sequencing from single polymerase. Mol Sci, 2009, 323: 133-8
[26] Timp W, Mirsaidov UM, Wang D, et al. Nanopore sequencing: electrical measurements of the code of life. IEEE Trans Nanotechnol, 2010, 9: 281-94
[27] Jain M, Koren S, Miga KH, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol, 2018, 36: 338-45
[28] Rieber N, Zapatka M, Lasitschka B, et al. Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. PLoS One, 2013, 8: e66621
[29] Ji P, Zhang Y, Wang J, et al. MetaSort untangles metagenome assembly by reducing microbial community complexity. Nat Commun, 2017, 8: 14306
[30] Mason CE, Afshinnekoo E, Tighe S, et al. International standards for genomes, transcriptomes, and metagenomes. J Biomolecul Tech, 2017, 28: 8-18
[31] Maydan J, Thomas M, Tabanfar L. Electrophoretic high molecular weight DNA purification enables optical mapping. J Biomol Tech, 2013, 24: S57
[32] Nair S, Karim R, Cardosa MJ. Convenient and versatile DNA extraction using agarose plugs for ribotyping of problematic bacterial species. J Microbiol Methods, 1999, 38: 63-7
[33] Tighe S, Afshinnekoo E, Rock TM. Genomic methods and microbiological technologies for profiling novel and extreme environments for the Extreme Microbiome Project (XMP). J Biomol Tech, 2017, 28: 31-9
[34] Adey A, Shendure J. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. Genome Res, 2012, 22: 1139-43
[35] Dunham JP, Friesen ML. A cost-effective method for high-throughput construction of Illumina sequencing libraries. Cold Spring Harb Protoc, 2013, 2013: 820-34
[36] Tatsumi K, Nishimura O, Itomi K, et al. Optimization and cost-saving in tagmentation-based mate-pair library preparation and sequencing. Biotechniques, 2015, 58: 253-7
[37] 丁啸. 基于序列特征的宏基因组数据分析方法研究[D]. 南京: 东南大学, 2016
[38] 魏军, 赵志军. 下一代测序技术在分子诊断中的应用. 分子诊断与治疗杂志, 2013, 3: 145-51
[39] Mostovoy Y, Levy-Sakin M, Lam J. A hybrid approach for de novo human genome sequence assembly and phasing. Nat Methods, 2016, 13: 587-90
[40] Shin SC, Ahn DH, Kim SJ, et al. Advantages of single-molecule real-time sequencing in high-GC content genomes. PLoS One, 2013, 8: e68824
[41] Stolze Y, Bremges A, Maus I, et al. Targeted *in situ* metatranscriptomics for selected taxa from mesophilic and thermophilic biogas plants. Microb Biotechnol, 2018, 11: 667-79
[42] Kuleshov V, Xie D, Chen R. Whole-genome haplotyping using long reads and statistical methods. Nat Biotechnol, 2014, 32: 261-6
[43] Amini S, Pushkarev D, Christiansen L. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. Nat Genet, 2014, 46: 1343-9
[44] Peters BA, Kermani BG, Sparks AB, et al. Accurate whole genome sequencing and haplotyping from 10-20 human cells. Nature, 2013, 487: 190-5
[45] Lieberman-Aiden E, van Berkum NL, Williams L, et al. Comprehensive mapping of long range interactions reveals folding principles of the human genome. Science, 2009, 326: 289-93
[46] Burton JN, Liachko I, Dunham MJ, et al. Species-level

deconvolution of metagenome assemblies with Hi-C-based contact probability maps. G3 (Bethesda), 2014, 4: 1339-46
[47] Stewart RD, Auffret MD, Warr A. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. Nat Commun, 2018, 9: 870
[48] 叶丹丹, 樊萌萌, 琼关, 等. 宏基因组研究的生物信息学平台现状. 动物学研究, 2012, 33: 574-85
[49] Myers EW, Sutton GG, Delcher AL, et al. A whole-genome assembly of *Drosophila*. Science, 2000, 287: 2196-204
[50] Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci USA, 2001, 98: 9748-53
[51] Korlach J, Gedman G, Kingan SB, et al. *De novo* PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. GigaScience, 2017, 6: 1-16
[52] Bankevich A, Pevzner PA. TruSPAdes: barcode assembly of TruSeq synthetic long reads. Nat Methods, 2016, 13: 248-50
[53] Lam KK, Hall R, Clum A, et al. BIGMAC: breaking inaccurate genomes and merging assembled contigs for long read metagenomic assembly. BMC Bioinform, 2016, 17: 435
[54] Salmela L, Walve R, Rivals E, et al. Accurate self-correction of errors in long reads using de Bruijn graphs. Bioinformatics, 2017, 33: 799-806
[55] Parkhomchuk D, Bremges A, McHardy AC. Fast and memory-efficient noisy read overlapping with KD-trees. bioRxiv, 2017, doi:10.1101/166835
[56] Kuleshov V, Snyder MP, Batzoglou S. Genome assembly from synthetic long read clouds. Bioinformatics, 2016, 32: i216-24
[57] Danko DC, Meleshko D, Bezdan D, et al. Minerva: an alignment and reference free approach to deconvolve linked-reads for metagenomics. bioRxiv, 2017, doi: 10.1101/217869
[58] Frank JA, Pan Y, Tooming-Klunderud A, et al. Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. Sci Rep, 2016, 6: 25373
[59] Xu Y, Zhao F. Single-cell metagenomics: challenges and applications. Protein Cell, 2018, 9: 501-10
[60] Treangen TJ, Abraham AL, Touchon M, et al. Genesis, effects and fates of repeats in prokaryotic genomes. FEMS Microbiol Rev, 2009, 33: 539-71
[61] Slaby BM, Hackl T, Horn H, et al. Metagenomic binning of a marine sponge microbiome reveals unity in defense but metabolic specialization. ISME J, 2017, 11: 2465-78
[62] Kuleshov V, Jiang C, Zhou W, et al. Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. Nat Biotechnol, 2016, 349: 64-9
[63] Tedersoo L, Tooming-Klunderud A, Anslan S. PacBio metabarcoding of fungi and other eukaryotes: errors, biases and perspectives. New Phytol, 2018, 2179: 1370-85
[64] Yang J, Cao J, Xu H, et al. Bacterial diversity and community structure in Chongqing radish paocai brines revealed using PacBio single-molecule real-time sequencing technology. J Sci Food Agric, 2018, 98: 3234-45
[65] Carneiro MO, Russ C, Ross MG, et al. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics, 2012, 13: 375
[66] Nguyen NP, Mirarab S, Liu B, et al. TIPP: taxonomic identification and phylogenetic profiling. Bioinformatics, 2014, 30: 3548-55
[67] Beaulaurier J, Zhu S, Deikus G, et al. Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. Nat Biotechnol, 2018, 36: 61-9
[68] Karst SM, Dueholm MS, McIlroy SJ, et al. Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. Nat Biotechnol, 2018, 36: 190-5
[69] Ikuta T, Takaki Y, Nagai Y, et al. Heterogeneous composition of key metabolic gene clusters in a vent mussel symbiont population. ISME J, 2016, 10: 990-1001
[70] Moss EL, Bishara A, Tkachenko E, et al. *De novo* assembly of microbial genomes from human gut metagenomes using barcoded short read sequences bioRxiv, 2017, doi: 10.1101/125211
[71] White RA 3rd, Bottos EM, Roy Chowdhury T, et al. Moleculo long-read sequencing facilitates assembly and genomic binning from complex soil metagenomes. mSystems, 2016, 1: e00045-16
[72] Tsai YC, Conlan S, Deming C, et al. Resolving the complexity of human skin metagenomes using single-molecule sequencing. MBio, 2016, 7: e01948-15
[73] Bishara A, Moss EL, Kolmogorov M, et al. Culture-free generation of microbial genomes from human and marine microbiomes bioRxiv, 2018, doi: 10.1101/ 263939
[74] Driscoll CB, Otten TG, Brown NM, et al. Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. Stand Genomic Sci, 2017, 12: 9
[75] Batovska J, Lynch SE, Rodoni BC, et al. Metagenomic arbovirus detection using MinION nanopore sequencing. J Virol Methods, 2017, 249: 79-84
[76] Chaisson MJ, Wilson RK, Eichler EE. Genetic variation and the *de novo* assembly of human genomes. Nat Rev Genet, 2015, 16: 627-40
[77] Kwon HT, Jang EH, Na SK, et al. Complete genome sequence of *Stenotrophomonas* sp. KCTC 12332, a biotechnological potential bacterium. J Biotechnol, 2017, 256: 27-30
[78] 吴清发. 基因组学研究中一些常用软件的概述. 遗传, 2003, 25: 708-12
[79] Oh HN, Lee TK, Park JW, et al. Metagenomic SMRT sequencing-based exploration of novel lignocellulose-degrading capability in wood detritus from torreya nucifera in bija forest on Jeju island. J Microbiol Biotechnol, 2017, 27: 1670-80

[80] Sharon I, Kertesz M, Hug LA, et al. Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. Genome Res, 2015, 25: 534-43
[81] Le Bras Y, Collin O, Monjeaud C, et al. Colib'read on galaxy: a tools suite dedicated to biological information extraction from raw NGS reads. GigaScience, 2016, 5: 9
[82] Grau J, Reschke M, Erkes A, et al. AnnoTALE: bioinformatics tools for identification, annotation, and nomenclature of TALEs from Xanthomonas genomic sequences. Sci Rep, 2016, 6: 21077
[83] 胡小丹, 游敏, 罗文新. 基因编辑技术. 中国生物化学与分子生物学报, 2018, 34: 267-77
[84] Ehriich M, Gama-Sosa MA, Carreira LH, et al. DNA methytation in thermophilic bacteria: N4-methylcytosine, 5-methylcytosine, and N6-methyladenine. Nucleic Acids Res, 1985, 13: 1399-412
[85] Schatz MC. Nanopore sequencing meets epigenetics. Nat Methods, 2017, 14: 347-8