PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2020,47(1):5~16
www.pibb.ac.cn

单分子测序技术及应用研究进展*

俞晓玲[1)] 姜文倩[1)] 郑 玲[1)] 石 杨[2,3)] 叶寒辉[1)] 林 峻[2,3,4)**]

([1)] 福建医科大学孟超肝胆医院感染科，福州 350025；[2)] 福州大学生物科学与工程学院，福州 350108；
[3)] 福州大学应用基因组学研究所，福州 350108；[4)] 福州大学福建省海洋酶工程重点实验室，福州 350108）

摘要 从DNA双螺旋结构的发现开始，生命科学研究进入分子水平，在20世纪70年代出现的测序技术为破译遗传密码作出了巨大贡献. 近几年出现的单分子测序技术，可以在单个分子水平读取核苷酸序列，也被称为第三代测序技术，主要代表有HeliScope、Nanopore和PacBio等. 与传统的第一代和第二代测序技术相比，第三代测序能够产生更长的碱基读长，能直接对RNA进行测序，无需逆转录，测序速度极快，同时其中某些技术所涉及的设备可以小型化，可便携至野外现场测序. 第三代测序技术在生命科学基础理论研究及生物医学临床实践中，具有广泛的应用. 本文重点介绍了各种单分子测序技术的原理、优缺点，及其应用研究进展.

关键词 单分子测序，HeliScope测序，纳米孔测序，PacBio测序
中图分类号 Q7，Q3 **DOI**：10.16476/j.pibb.2019.0167

1 概 述

测序技术用于确定核酸、蛋白质、多糖等生物大分子的一级结构，较为常见的是核酸测序，确定核苷酸在核酸序列中的顺序，包括DNA和RNA测序. 目前，DNA测序已经从第一代DNA测序技术发展到第三代DNA测序技术.

Maxam-Gilbert测序技术、Sanger双脱氧测序技术、荧光自动测序技术和杂交测序技术统称为第一代测序技术. 20世纪70年代中后期，测序技术逐渐开始发展. 1975年，Sanger和Coulson [1] 发表了“加减法”技术，该技术使用DNA聚合酶和放射性标记核苷酸进行测序. 1977年，Maxam和Gilbert [2] 建立了基于化学断裂法的Maxam-Gilbert测序. 同年，Sanger等 [3] 提出了双脱氧链终止法. 1986年，Smith等 [4] 开发了一种基于Sanger和荧光检测原理的DNA序列分析半自动化方法. 目前，一代测序中广泛应用的DNA测序仪就是基于毛细管电泳和荧光标记等原理开发的. 20世纪80年代末，具有部分二代测序特点的杂交测序技术出现 [5].

人类基因组计划是DNA测序的一个里程碑，随着人类基因组计划的完成，传统测序方法的通量已经无法满足基因测序的需求. 20世纪90年代中后期，第二代测序（next-generation sequencing，NGS）技术出现. 随着NGS的发展，出现了不同的二代测序平台. 2005年，出现了第一个商用二代测序平台——454 Roche GS FLX [6]，该平台目前已经被淘汰. 随后出现了Illumina（2006年）、SOLiD（2007年）、Ion Torrent（2010年）和华大基因的BGISEQ（2013年）. 除了SOLiD是基于连接法测序（sequence by ligation，SBL）外，其他平台都采用边合成边测序（sequence by synthesis，SBS）技术. 由于测序试剂及原理的不同，不同的测序平台在读长、通量、基因组覆盖率、错误率、成本和运行时间等方面有所差异 [7-8].

2008~2009年，采用与第二代平台不同方法的测序技术首次被描述为“第三代” [9]. 一代测序只

* 福州市感染性疾病医学中心（2018080306），福州市临床重点专科建设项目经费（201510301），福建省临床重点专科建设项目，福州市卫生健康科技创新平台建设项目（2019-S-wp6）和福州市卫生健康科研创新团队培育项目（2019-S-wt4）资助.
** 通讯联系人.
Tel：0591-22863805，E-mail：jun@fzu.edu.cn.
收稿日期：2019-07-20，接受日期：2019-12-02

能对纯培养物进行鉴定、培养及测序耗时长；二代宏基因组测序读长短（Short-Read），多轮PCR扩增容易发生交叉污染，且不能直接对RNA进行测序；不同于二代测序需要将长链DNA打断，第三代测序在单分子水平读取核苷酸序列，因此也被称为单分子测序，主要有HeliScope[10]、Nanopore[11]和PacBio[12]. 与目前的测序技术相比，第三代测序能够产生更长的读长[13]、能直接对RNA进行测序（无需逆转录），其便携性与测序速度也是重要优势[14]. 本文主要综述HeliScope、Nanopore和PacBio这三种单分子测序技术的原理、优势和缺陷，以及应用研究进展.

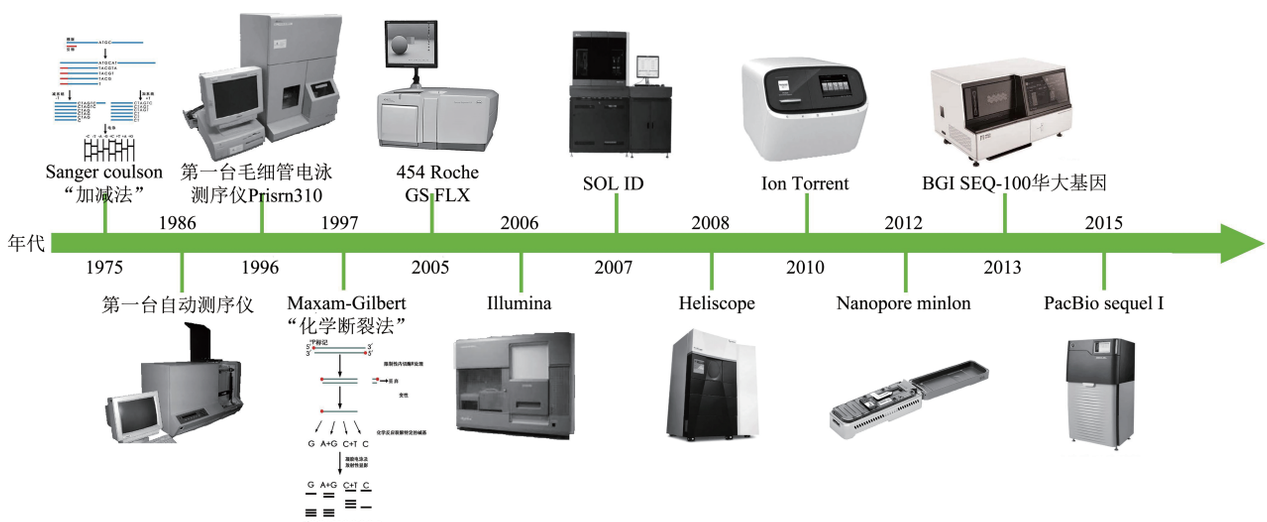

Fig. 1　History of nucleic acid sequencing technology

图1　核酸测序技术发展历史

从1975年Sanger发明加减法测序开始，到2015年PacBio sequel Ⅰ测序仪的出现，测序技术不断发展进步，从最开始的一代测序技术发展到如今的第三代单分子测序技术.

2　HeliScope单分子测序

2.1　测序原理

Helicos遗传分析系统是首个基于荧光测序原理的商业化单分子测序平台. HeliScope单分子测序仪是第一台利用Helicos单分子测序（true single molecule sequencing，tSMS）技术直接测定DNA的基因分析仪. Helicos基因分析系统能够对几个到几千个碱基的核酸进行测序，但单位质量序列的产量取决于3'端羟基的数量，因此相对较短的模板测序效率较高. 对于超过1 000 nt的核酸，Helicos一般建议将核酸剪切至平均长度为100~200 nt[15].

Helicos基因分析系统由多个部分组成，它们作为一个集成系统一起工作. 由于标准的flow cell表面固定有多个末端带有荧光标记的oligo(dt)50，为了与flow cell表面的oligo(dt)50引物相容，必须在待测序分子的3'端合成至少50 nt的poly(dA)尾[16]. 用于测序的DNA大片段首先被断裂为成千上万的DNA短片段，并在其3'末端添加poly(dA)尾；带有poly(dA)尾的DNA链与一次性玻璃flow cell上的oligo(dt)50进行原位杂交，并通过oligo(dt)50带有的荧光信号精确定位；每一张标准的flow cell上有25条通道，一旦flow cell装载了适当的样品，它将与合成和成像测序所需的所有试剂一起插入Heliscope测序系统；flow cell插入后，拍摄第一张模板图像，通过oligo(dt)50末端的荧光标记精确定位杂交模板所处的位置；然后逐一加入荧光标记的单色末端终止子和聚合酶孵育，清洗未结合的单色末端终止子，在系统激光激发荧光部分后，用4个激光通过共焦显微镜拍摄图像，从图像中可以确定每个DNA序列中的一个核苷酸；随后，切割荧光染料和抑制基团，切割后，分离的荧光染料被洗掉，然后加入新的聚合酶和一个单荧光核苷酸，拍摄另一张图像，重复这个过程，直到片段完成测序[15]，测序原理如图2所示.

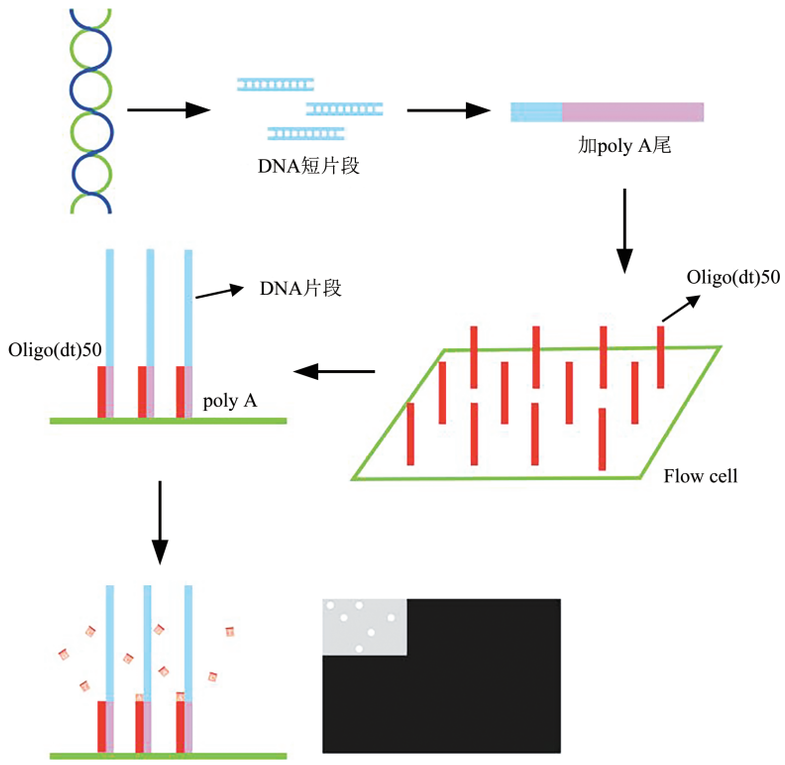

Fig. 2　Principle of Heliscope sequencing

图2　Heliscope测序原理

首先，DNA大片段被断裂为成千上万的DNA短片段，并加上与flow cell表面的oligo(dt)50引物相容的至少50 nt的poly A尾；随后，DNA分子与flow cell的oligo(dt)50进行原位杂交，拍摄第一张模板图像，精确定位杂交模板所处的位置；然后逐一加入荧光标记的单色末端终止子和聚合酶孵育，清洗，拍摄图像；随后，切割荧光染料和抑制基团，切割后，分离的荧光染料被洗掉，然后加入新的聚合酶和一个单荧光核苷酸，拍摄另一张图像，重复这个过程，直到片段完成测序.

2.2　测序优势及缺陷

Helicos单分子测序简化了DNA样品的制备过程，HeliScope灵敏度高，可以读取单个分子，由于不需要进行PCR扩增，避免了PCR扩增产生的偏倚和误差，适合拷贝数变异的检测，同时降解或修饰的分子可以直接用作测序模板[15]. HeliScope可用于RNA-seq或RNA直接测序[10]. 对具有极端GC含量的DNA或基因组，单分子测序效率更高，因为它对GC含量的敏感性更低[17]. 与其他技术相比，Helicos单分子测序只是简单的片段化DNA、加poly A尾再进行杂交、测序，所需的试剂和操作步骤较少.

但是HeliScope读长较短（55～70个碱基）且数据输出低（20 Gb）[18-19]，序列中碱基越多，样本中可使用链的百分比就越低，因为一些链在测序过程中不再伸长. 由于噪音的影响，测序的错误率较高是其主要缺陷. 虽然这一缺陷可以通过重复测序克服，但是在给定的精确度下增加了每个碱基的成本，抵消了试剂成本降低带来的一些收益.

2.3　应用

Helicos单分子测序已经成功地进行了各种各样的应用. a. 在病毒基因组测序方面，Harris等[18]用Helicos单分子测序对M13噬菌体进行测序，测序平均深度大于150×且覆盖率为100%. b. 在拷贝数变异研究方面，Pushkarev等[20]使用Helicos单分子方法对单个人类基因组进行测序，测定了约280万个单核苷酸多态性（single nucleotide polymorphisms，SNPs）（总有用深度为28×）并通过单独分析覆盖深度确定了752个拷贝数变异（copy number variation，CNV）区域. c. 在古老DNA测序方面，Orlando等[21]使用Helicos的HeliScope测序平台对保存在多年冻土中的更新世马骨进行测序，获得了115.9 Mb数据，并与Illumina测序结果进行比较，发现Helicos测序结果

中来自马的内源性DNA序列的百分比高于Illumina测序结果，Helicos只需要1×的测序深度就能生成一份古马基因组草图. 实验结果表明，通过结合现有的第二代和第三代测序方法，古基因组可以以前所未有的方式进行测序.

3　纳米孔（Nanopore）单分子测序

3.1　测序原理

使用纳米孔作为生物传感器的想法最初是在20世纪90年代由Deamer等[22]提出的. 纳米孔生物传感器可分为固态纳米孔和生物纳米孔两大类，已有文献证明这两种类型的纳米孔都能够在单分子水平上检测生物和化学分子[23-24]. 固态纳米孔可以从多种材料如硼、铝、硅、石墨烯以及混合材料中制备[24]，其化学物理性质优异，它们可以在各种实验条件下工作，进行DNA测序和蛋白质检测[25]. 生物纳米孔是一种跨膜蛋白质通道，可以通过改变特定位点氨基酸残基的分子生物学技术进行基因工程改造[26].

2014年，Oxford Nanopore Technologies（ONT）推出纳米孔测序仪的第一个原型——MinION. MinION有512~2 000个纳米孔，每个纳米孔的测序速度为120~1 000个碱基/min[27]，每一条flow cell可以生成10~30 Gb的DNA序列数据. Nanopore根据测序产出数据量的不同可分为3类：Flongle测序数据量为2 Gb；MinION和GridION测序数据量为50 Gb；而PromehION测序数据量可达220 Gb.

纳米孔测序是基于电信号的测序技术[28]，该技术的核心是蛋白质纳米孔. 在两个电解液室之间形成一个纳米孔，两个电解液室之间有一层不透水的膜[27]，蛋白质纳米孔（微型的小孔，其本质是在膜上形成通道）被嵌入在合成膜上（该膜具有非常高的电阻），并浸没在电生理溶液中；当电压施加到电解液室中时，会产生穿过孔的稳态离子电流，进入纳米孔的单分子会对离子的流动造成阻碍，这被称为Nanopore信号，不同的碱基造成的阻碍大小是不同的. 因此，监测通过孔的电流波动信号可以实现分子传感[27]，分析电流波动信号可以反推出进过纳米孔的碱基. 将带电生物分子（如DNA或RNA）添加到其中一个电解液室中，当生物分子通过纳米孔时，会对电流产生干扰，引起电流信号的特征性改变，可以传递出样品的许多特性，如生物分子的大小、浓度和结构；在此过程中，信号被实时分析，用来确定正在通过该孔的DNA或RNA链的碱基序列，从而分析通过它的整个DNA或RNA片段[29-30]. 以DNA测序为例，DNA链接触性穿过纳米孔，纳米孔是对穿过的

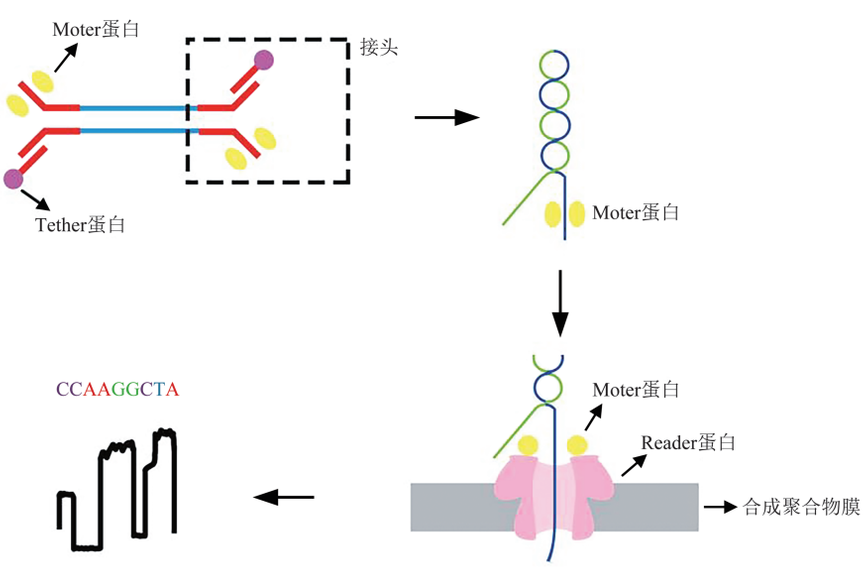

Fig. 3　Principle of Nanopore sequencing

图3　Nanopore测序原理

以DNA测序为例，Nanopore测序在文库构建时，在DNA片段上加了一个带有Moter蛋白（DNA解螺旋酶）和Tether蛋白（把DNA链吸附在测序芯片的膜上）的接头；Reader蛋白（构成纳米孔）插在一层电阻率很高的薄膜当中，并浸没在电生理溶液中；当电压施加到电解液室中时，会产生穿过孔的稳态离子电流；当DNA单链通过纳米孔时，会对离子流动造成阻碍，不同碱基造成的阻碍不同，通过分析电流信号可获得DNA序列.

DNA片段进行测序，而不是生成特定长度的序列. 文库构建时，将DNA序列与DNA解螺旋酶混合，当DNA酶复合物接近纳米孔时，单链DNA被拉过孔，DNA解螺旋酶一次让一个碱基通过纳米孔，可通过控制酶的运行速度，控制每秒产生的数据量；当DNA通过孔进行移动时，通过纳米孔那条链的核苷酸会产生特征性的电流干扰，纳米孔信号可用于确定该DNA链上的碱基序列，测序原理如图3所示.

3.2　测序优势及缺陷

Nanopore测序技术，有效地解决了二代宏基因组测序技术在病原学诊断领域的缺陷，与二代测序技术相比，纳米孔测序读长很长，能直接测定1 Mb以上的读长[31]，通过16 S rRNA测序可鉴定病原微生物精确到种. 纳米孔测序可以对单个DNA进行测序，而无需对样品进行PCR扩增或化学标记；对于RNA测序而言，无需把RNA逆转录成cDNA，节约了逆转录所需的操作和时间，大大降低测序的成本[32-33]；目前最新版Nanopore的RNA直接测序芯片运行一次可获得100万条全长RNA序列，测序通量很高. Nanopore的MinION是一款超小型的掌上测序仪，重量不足100 g，长度约10 cm，使用高速USB 3.0插入电脑，不需要额外的计算基础设施，便携性强；它不局限于实验室环境，可将测序仪带到如高山、丛林中、北极等地区直接进行测序[34-36]；单芯片的通量恰好适用于1~2个标本[37]，可随时开机测序.

虽然纳米孔测序的优势很多，但也存在缺陷. 当一段DNA序列中有较少几个连续相同的碱基（如2个A）时，纳米孔测序识别A碱基的数量可能会产生误差（测序结果可能为1个A）；测序的误差率较大是其最大缺陷，纳米孔测序的误差率在15%~40%之间，测序错误主要由插入和删除碱基引起[38]，这对已经存在的适用于二代测序数据的生物信息学分析方法提出了新的挑战.

3.3　应用

纳米孔测序技术这两年已被广泛用于疫情爆发调查、环境监测、传染性病原体的检测和抗生素耐药性监测等领域.

3.3.1　在疫情爆发调查中的应用

在快速实时监测疫情爆发方面，研究人员通过小型便携式Oxford Nanopore MinION设备，对尼日利亚爆发的拉沙热疫情[39]和几内亚埃博拉疫情[40]进行了实时动态的基因组监测，一般20~25×的测序深度足以确定准确的基因型.

3.3.2　在环境监测中的应用

在环境监测方面，Samson等[41]利用MinION对恒河和亚穆纳河汇流处及其下游沉积物样本中的微生物群落进行了全基因组测序，功能遗传分析揭示了亚穆纳河汇流处和下游地区的原生微生物对外来化合物的降解、对有毒化合物的抗性以及抗生素抗性的影响.

3.3.3　在传染性病原体检测中的应用

在传染性病原体的检测方面，一般测序数据量为100~500 Mb. 匈牙利塞格德大学医学院的Prazsák等[42]使用纳米孔测序平台对水痘带状疱疹病毒（varicella zoster virus，VZV）的裂解转录组进行分析，揭示了VZV的复合转录组学结构. 美国疾病预防与控制中心的微生物学家John Barnes团队[43]使用纳米孔测序了完整的A型流感病毒RNA基因组. Moon等[44]使用来自痰液的16 S rRNA基因的深度测序来鉴定患有社区获得性肺炎患者中的流感嗜血杆菌，使用MinION Nanopore sequencer进行16 S rRNA扩增子测序，在韩国诊断出第一例胎儿弯曲杆菌性脑膜炎[45].

3.3.4　在抗生素耐药性监测中的应用

在耐药性方面，澳大利亚昆士兰大学的Bainomugisa等[46]使用纳米孔测序技术实现了结核分枝杆菌种属的快速全基因组测序，平均测序深度238×. Runtuwene等[47]使用纳米孔测序技术对疟疾寄生虫——恶性疟原虫进行基因分型并推断其耐药状态，测序深度≥50×. 英国Charalampous等[48]使用纳米孔测序技术快速鉴定出从下呼吸道感染患者中分离的细菌种属和细菌耐药基因，测序深度232.5×，纳米孔测序技术的临床开发与应用成为了精准病原检测的新里程碑.

4　PacBio单分子测序

4.1　测序原理

Pacific Biosciences是一家成立于2004年的生物技术公司，PacBio将其平台描述为基于零模波导（zero-mode waveguide，ZMW）特性的单分子实时测序（single-molecule real-time sequencing，SMRT）. ZMW是一种纳米光子封闭结构，由放置在透明二氧化硅基底上的铝包层薄膜中的一个圆孔组成[49]. ZMW孔的直径约为70 nm，深度约为100 nm，由于ZMW孔的直径小于光的波长，当光通过ZMW孔时，光场呈指数衰减[50]，在被照射

的ZMW孔内，产生一个足够小的发光观察体积，包含一个单核苷酸的DNA聚合酶的活性很容易检测出来. PacBio有两种测序模式：高精度长读（highly accurate long reads，HiFi reads）和连续长读（continuous long read，CLR）. HiFi reads采用circular consensus sequencing（CCS）模式，CCS从单个模板分子的多次传递中得到一致序列，从噪声的单个subreads中产生精确的reads，CCS主要应用于短于2 kb的DNA插入[51]，该模式精度＞99%[52]. CLR测序模式可以实现复杂基因组的高质量组装，预期1/2的数据读取量大于50 kb，最长的读取量可达175 kb，平均误差率为15%[52].

SMRT是一种并行的单分子DNA测序方法，以SMRT芯片为载体边合成边测序. SMRT测序从制备SMRTbell模板文库开始，SMRTbell模板是一种封闭的单链环状DNA，两端都有发夹接头；当SMRTbell加载到SMRT cell中时，SMRTbell扩散到ZMW的测序单元中，单个聚合酶锚定在每个ZMW的底部并与DNA模板、测序引物结合；随后4种不同荧光标记的dNTP底物随机进入ZMW底部，对于每种核苷酸碱基，都有一个相应的荧光染料分子，荧光染料分子附着在核苷酸的磷酸基团末端；当一种与正要合成的碱基一致的dNTP被DNA聚合酶结合后，激发光从ZMW底部照射进来，照射在被DNA聚合酶结合的dNTP上，使该dNTP发出荧光，在薄的区域产生对应于掺入碱基的光脉冲，每个脉冲具有其自己的颜色强度和持续时间用于识别碱基[53-54]，检测器检测核苷酸掺入的荧光信号，一个聚合反应循环完成后，dNTP上的磷酸基团被切割，连接在磷酸基团上的荧光基团也一起被切割下来，随后被切割的荧光染料分子扩散出检测体积，荧光信号不再被检测，通过不断循环这个过程完成测序，测序原理如图4所示.

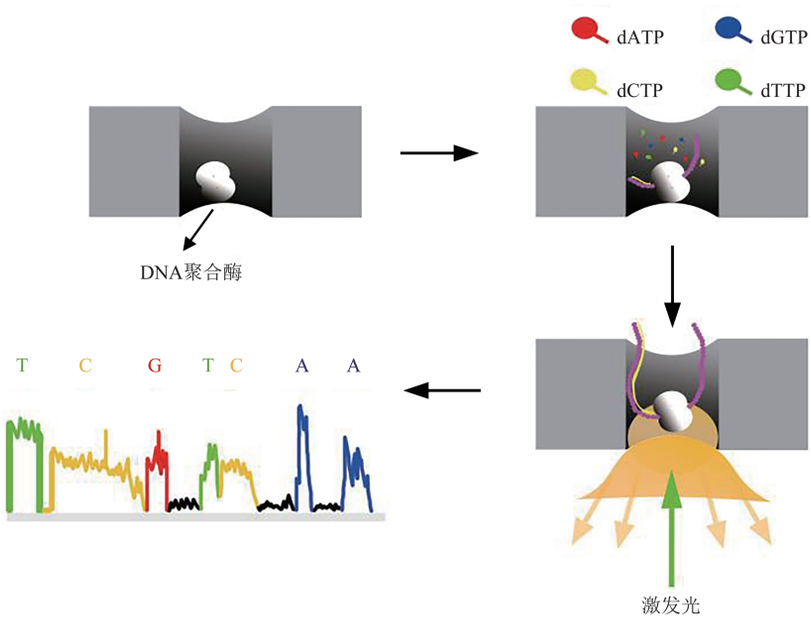

Fig. 4　Principle of PacBio sequencing

图4　PacBio测序原理

单个聚合酶锚定在每个ZMW的底部并与模板结合，种不同荧光标记的dNTP随机进入ZMW底部，当标记的核苷酸被DNA聚合酶结合后，在薄的区域产生对应于掺入的碱基的光脉冲，检测器检测核苷酸掺入的荧光信号，反应完成后荧光基团和磷酸基团被切割下来，随后被切割的荧光染料分子扩散出检测体积，荧光信号不再被检测.

4.2　测序优势及缺陷

与前两种单分子测序相同，SMRT也能够在没有进行PCR扩增的情况下对单个DNA分子进行实时测序，而不同的是该技术能够通过DNA聚合酶直接观察DNA合成. SMRT的读长很长，测序数据中有1/2的数据读长大于20 kb，最长读长大于60 kb[53, 55]；当测序深度为30×时，测序准确率大于99.999%，并且无系统误差[54]；即使是高GC含

量的区域，测序的偏倚也较少[55]；SMRT还能在单碱基分辨率下直接检测DNA碱基修饰，包括某些类型的甲基化修饰[53，56]，展现序列的表观遗传学特征. 通过SMRT技术，可以对AT或GC富集区域以及大的结构变异，包括插入、缺失、倒位、易位、重复和串联重复等难以测序的区域进行测序[53，55，57-58].

但是PacBio机器巨大，硬件成本昂贵. PacBio的测序成本很高，以16 S rRNA测序为例，一般每个样品的测序数据量仅能达到5 000条序列（CCS）[59]，而在相同成本时，Nanopore直接RNA测序的通量可以高出很多. 同时，PacBio测序文库构建繁琐，不能直接测定RNA，需要逆转录成cDNA才能进行测序.

4.3　应用

目前PacBio在全基因组测序、微生物群落特征、病毒群体、体细胞变异、RNA测序、表观遗传学等方面都有所应用.

4.3.1　在全基因组测序中的应用

恶性疟原虫入侵人的红细胞是其存活及疟疾发病的关键，Campino等[60]利用全基因组测序数据进行连锁分析（测序深度100×），发现恶性疟原虫入侵途径中大多数表型变异是由一个包含*PfRh2a*和*PfRh2b*基因的基因座引起的，入侵途径的变异与恶性疟原虫系间*PfRh2a*和*PfRh2b*的表达差异显著有关，对疟疾疫苗的开发具有指导意义. Slager等[61]使用PacBio测序肺炎链球菌（*Streptococcus pneumoniae*）D39菌株的基因组序列，产生2个contig，分别为250~500×和5~25×，注释了89个新的蛋白质编码基因、34个小RNA（small RNAs）和165个假基因，揭示了一些先前未被短读测序检测到的倒位，加快了预防和治疗肺炎球菌新策略的制定. 除了病原微生物的全基因组测序，研究人员还利用PacBio对人类[62-63]、植物[64-65]、动物[66-67]的全基因组进行了测序，一般结构变异的测序深度为6~50×.

4.3.2　在微生物群落特征、病毒群体、体细胞变异研究中的应用

在微生物群落特征、病毒群体、体细胞变异研究方面，Wang等[68]基于SMRT技术对儿童龋齿口腔微生物群进行定性研究，获得702 304个修剪后16 S rRNA基因reads，应用PacBio测序对源于13个已知细菌门和110个属的876个种进行了检测，发现营养缺陷菌（*Abiotrophia* spp.）、奈瑟菌（*Neisseria* spp.）和韦氏杆菌（*Veillonella* spp.）可能与口腔微生物生态系统的健康有关，普雷沃氏菌（*Prevotella* spp.）、乳酸杆菌（*Lactobacillus* spp.）、小杆菌（*Dialister* spp.）和*Filifactor* spp.可能与龋病的发病机制和进展有关. Su等[69]通过SMRT对HIV-1的DNA和RNA进行测序，评估少数耐药变异体对病毒学结果的影响，有助于检测患者的预处理耐药性突变，对病毒学反应存在潜在影响，具有临床意义. Smith等[70]对来自白血病患者的单细胞和克隆进行靶向测序（平均测序深度3000×，最小深度500×），发现大多数急性髓细胞性白血病患者在获得对FLT3抑制剂quikartinib的耐药性的同时激活了FLT3的内部串联重复突变，表明quikartinib的临床耐药性是高度复杂的，反映了急性髓细胞性白血病潜在的克隆异质性.

4.3.3　在RNA测序中的应用

在RNA测序方面，Lian等[71]应用单分子长读长RNA测序和从头组装的短读长RNA测序对人类乳腺癌细胞的野生型和紫杉醇耐药型RNA进行测序，揭示了乳腺癌中紫杉醇抵抗的新靶点.

4.3.4　在表观遗传学中的应用

在表观遗传学方面，Hiraoka等[72]使用SMRT技术来揭示日本琵琶湖微生物群落的“宏表观基因组”，重建了来自不同细菌和古菌群的19个基因组草图，DNA化学修饰分析发现22个甲基化基序，其中9个是新发现，强调了宏表观基因组学是一种识别自然界中大量未经探索的原核生物DNA甲基化系统的有效方法.

5　测序数据分析

MinION是由MinKNOW软件控制的，该软件在MinION连接的电脑上运行，完成运行参数的选择、数据采集、实时信号分割和实验进程的反馈等核心任务. 对每一个reads，MinKNOW将信号分割结果和与测序过程相关的元数据存储在FAST5二进制文件中[26]. 目前已有一些软件能将FAST5文件解析和转换为更传统的FASTA或FASTQ序列格式，如HPG Pore、PoreTools、poRe、NanoOK、npReader等[26]. Nanopore数据分析的难点主要在于重测序数据分析. 重测序数据分析包括将测序实验产生的所有reads与参考基因组比对，以发现reads与参考基因组之间的差异. 由于Nanopore测序所得到的序列大小不一致（从kb到几十kb）、较高的测序误差以及和误差分布不均匀性，使得

Nanopore数据的比对特别具有挑战性[26]. 主要的计算问题是如何以与二代测序（second generation sequencing，SGS）比对方法相同的速度和灵敏度，将长读长与基因组的中度差异进行比对.

目前，PacBio测序的数据分析软件主要有3种. 第一种是Pacific Biosciences公司自己研发的SMRT Analysis，是SMRT Link软件的一部分，用于PacBio长度长测序数据分析. 第二种是PacBio DevNet，该软件是Pacific Biosciences公司与生物信息学专家合作，开发用于SMRT测序数据分析和注释的开源软件. 第三种是SMRT兼容分析软件，包括Biosoft Integrators、DNAnexus、PSSC Labs、Dovetail Genomics LLC等公司的产品（https：//www. pacb. com/products-and-services/analytical-software/）. PacBio数据具有较高的测序错误率，大部分错误是插入或缺失错误. 在基于比对的同源性搜索过程中，基因的插入或缺失错误会导致移码，只会产生边缘比对分数和短比对. 因此，很难区分准确比对和随机比对，这种模糊性会导致结构和功能注释上的错误[73]. 现有的移码校正工具被设计用于具有低得多的错误率的数据，对于PacBio数据没有优化.

三代测序的通用分析软件较为常用的有FALCON、Canu、MECAT和miniasm. FALCON是Pacific Biosciences公司开发的一款用于三代基因组从头组装（*De novo*）的软件. FALCON-Phase结合PacBio长读长和Hi-C数据进行二倍体基因组阶段性组装. 该方法建立在FALCON-Unzip的基础上，FALCON-Unzip是一个阶段性的二倍体基因组组装程序，在组装过程中生成单倍型解析的相位块[74]. FALCON-Unzip不需要父母信息，单倍型相位块的长度受杂合度的大小和分布、序列读取长度和读取覆盖范围的限制. Canu是Celera Assembler的一个分支，用于高噪声单分子测序的数据分析. 与Celera Assembler 8.2相比，Canu引入了对纳米孔测序的支持，将覆盖要求的深度减半，提高了组装的连续性，同时将大型基因组的运行时间减少了一个数量级[75]. MECAT是中山大学研究团队开发的一款三代测序数据分析软件. Falcon和Canu都是基于all-pair比对的方法来确定read pairs对之间的重叠进行校正，非常耗时，而MECAT是一种基于快速全局k-mer评分的比对过滤算法，可以减少非信息性匹配read pairs的数量，以及选择较小数量的信息性匹配reads，来进行read校正，从而显著降低计算成本和时间[76]. Miniasm不纠正测序错误，而是直接从原始读取重叠产生未修正的contig序列[77]，进一步节省了计算时间. 虽然Miniasm比其他长读Assembler快一个数量级，但产生的序列错误可能是其他方法的10倍以上[78].

6　展　　望

我们将第二代Illumina测序和第三代HeliScope、Nanopore和PacBio单分子测序在测序原理、读长、通量、优缺点、便携性以及准确率方面进行了比较，如表1所示：

Table 1　Comparison of Illumina sequencing and HeliScope，Nanopore and PacBio single molecular sequencing

表1　Illumina测序和HeliScope、Nanopore和PacBio单分子测序技术的比较

测序平台	原理	读长	通量	优点	缺点	是否便携	准确率	测序成本/GB
Illumina（Hiseq）	桥式PCR	2 × 150 bp	1 500 Gb	通量高，测序准确性高	读长短，PCR扩增易发生交叉污染，不能直接对RNA进行测序	否	＞99%	$ 20.5
Helicos	单色荧光边合成边测序	55~70 bp	20 Gb	无需PCR扩增，可直接对RNA进行测序	读长较短，通量较低，仪器昂贵	否	97%	$ 1 000
Nanopore（MinION）	纳米孔电流波动	1 Mb	50 Gb	超长读长，无需PCR扩增，可直接对RNA进行测序，便携	序列存在几个连续相同的碱基时准确率低	是	60%~85%	$ 42
PacBio	零模波导	CLR模式最长大于175 kb	160 Gb	无需PCR扩增，可以检测DNA碱基修饰	测序成本高，不能直接对RNA进行测序	否	85%（CLR），＞99%（CCS）	$ 1 500

近年来，单分子测序技术在不断发展并有了很大的改进. 与一代测序和二代测序相比，单分子测序技术显示了其快速、准确、高通量、长读长等优势，使宏基因组和转录组的研究取得了前所未有的进展，并且在感染性疾病的快速检测诊断中发挥了越来越重要的作用，长读长测序可能在不久的将来成为一种标准的医学诊断工具. MinION的便携性允许其在野外进行测序，这对疫情调查具有重要意义. 超长读长纳米孔测序可能在不久的将来允许人类基因组的完全、无间隙的组装，将进一步促进人类遗传学研究. 但是单分子测序仍然存在一些技术难题需要攻克，如针对纳米孔测序错误率较高的问题，需要找到减少错误率的方案；与纳米孔测序相比，PacBio必须提高读取长度和通量. 随着单分子测序技术的不断发展，相信不久的将来，这些技术会在各个生命科学研究领域大放异彩.

参 考 文 献

[1] Sanger F, Coulson A R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology, 1975, **94**(3): 441-448

[2] Maxam A M, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci USA, 1977, **74**(2): 560-564

[3] Sanger F, Nicklen S, Coulson A R. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA, 1977, **74**(12): 5463-5467

[4] Smith L M, Sanders J Z, Kaiser R J, *et al*. Fluorescence detection in automated DNA sequence analysis. Nature, 1986, **321**(6071): 674-679

[5] Drmanac R, Labat I, Brukner I, *et al*. Sequencing of megabase plus DNA by hybridization: theory of the method. Genomics, 1989, **4**(2): 114-128

[6] Margulies M, Egholm M, Altman W E, *et al*. Erratum: Corrigendum: Genome sequencing in microfabricated high-density picolitre reactors. Nature, 2006, **441**(7089): 120-120

[7] Metzker M L. Sequencing technologies - the next generation. Nat Rev Genet, 2010, **11**(1): 31-46

[8] Scholz M B, Lo C C, Chain P S. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. Curr Opin Biotechnol, 2012, **23**(1): 9-15

[9] Check Hayden E. Genome sequencing: the third generation. Nature, 2009, **457**(7231): 768-769.

[10] Taroncher-Oldenburg G. RNA direct. Science-Business eXchange, 2009, **2**(39): 1459-1459

[11] Clarke J, Wu H C, Jayasinghe L, *et al*. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol, 2009, **4**(4): 265-270

[12] Munroe D J, Harris T J. Third-generation sequencing fireworks at Marco Island. Nat Biotechnol, 2010, **28**(5): 426-428

[13] Bleidorn C. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. Systematics and Biodiversity, 2015, **14**(1): 1-8

[14] Schadt E E, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet, 2010, **19**(R2): R227-240

[15] Thompson J F, Steinmann K E. Single molecule sequencing with a HeliScope genetic analysis system. Curr Protoc Mol Biol, 2010, **92**(1): 7.10.1-7.10.14

[16] Bowers J, Mitchell J, Beer E, *et al*. Virtual terminator nucleotides for next-generation DNA sequencing. Nat Methods, 2009, **6**(8): 593-595

[17] Goren A, Ozsolak F, Shoresh N, *et al*. Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA. Nat Methods, 2010, **7**(1): 47-49

[18] Harris T D, Buzby P R, Babcock H, *et al*. Single-molecule DNA sequencing of a viral genome. Science, 2008, **320**(5872): 106-109

[19] Hart C, Lipson D, Ozsolak F, *et al*. Single-molecule sequencing: sequence methods to enable accurate quantitation. Methods Enzymol, 2010, **472**: 407-430

[20] Pushkarev D, Neff N F, Quake S R. Single-molecule sequencing of an individual human genome. Nat Biotechnol, 2009, **27**(9): 847-850

[21] Orlando L, Ginolhac A, Raghavan M, *et al*. True single-molecule DNA sequencing of a pleistocene horse bone. Genome Res, 2011, **21**(10): 1705-1719

[22] Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. Nat Biotechnol, 2016, **34**(5): 518-524

[23] Haque F, Li J, Wu H C, *et al*. Solid-state and biological Nanopore for real-time sensing of single chemical and sequencing of DNA. Nano Today, 2013, **8**(1): 56-74

[24] Feng Y, Zhang Y, Ying C, *et al*. Nanopore-based fourth-generation DNA sequencing technology. Genomics Proteomics Bioinformatics, 2015, **13**(1): 4-16

[25] Traversi F, Raillon C, Benameur S M, *et al*. Detecting the translocation of DNA through a nanopore using graphene nanoribbons. Nat Nanotechnol, 2013, **8**(12): 939-945

[26] Magi A, Semeraro R, Mingrino A, *et al*. Nanopore sequencing data analysis: state of the art, applications and challenges. Brief Bioinform, 2018, **19**(6): 1256-1272

[27] Stoloff D H, Wanunu M. Recent trends in nanopores for biotechnology. Curr Opin Biotechnol, 2013, **24**(4): 699-704

[28] Varongchayakul N, Song J, Meller A, *et al*. Single-molecule protein sensing in a nanopore: a tutorial. Chem Soc Rev, 2018, **47**(23): 8512-8524

[29] Squires A H, Gilboa T, Torfstein C, *et al*. Single-molecule characterization of DNA-protein interactions using Nanopore biosensors. Methods Enzymol, 2017, **582**: 353-385

[30] Tedersoo L, Tooming-Klunderud A, Anslan S. PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. New Phytol, 2018, **217**(3): 1370-1385

[31] Van Dijk E L, Jaszczyszyn Y, Naquin D, *et al*. The third revolution in sequencing technology. Trends Genet, 2018, **34**(9): 666-681

[32] Marinov G K. On the design and prospects of direct RNA sequencing. Brief Funct Genomics, 2017, **16**(6): 326-335

[33] Hayden E C. Nanopore genome sequencer makes its debut. Nature, 2012, doi: 10.1038/nature.2012.10051

[34] Shi Y, Tyson G W, Eppley J M, *et al*. Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean. ISME J, 2011, **5**(6): 999-1013

[35] Jain M, Fiddes I T, Miga K H, *et al*. Improved data analysis for the MinION nanopore sequencer. Nat Methods, 2015, **12**(4): 351-356

[36] Miles B N, Ivanov A P, Wilson K A, *et al*. Single molecule sensing with solid-state nanopores: novel materials, methods, and applications. Chem Soc Rev, 2013, **42**(1): 15-28

[37] Karamitros T, Magiorkinis G. Multiplexed targeted sequencing for Oxford Nanopore MinION: a detailed library preparation procedure. Methods Mol Biol, 2018, **1712**: 43-51

[38] Magi A, Giusti B, Tattini L. Characterization of MinION nanopore data for resequencing analyses. Brief Bioinform, 2017, **18**(6): 940-953

[39] Kafetzopoulou L E, Pullan S T, Lemey P, *et al*. Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak. Science, 2019, **363**(6422): 74-77

[40] Quick J, Loman N J, Duraffour S, *et al*. Real-time, portable genome sequencing for Ebola surveillance. Nature, 2016, **530**(7589): 228-232

[41] Samson R, Shah M, Yadav R, *et al*. Metagenomic insights to understand transient influence of Yamuna River on taxonomic and functional aspects of bacterial and archaeal communities of River Ganges. Sci Total Environ, 2019, **674**: 288-299

[42] Prazsak I, Moldovan N, Balazs Z, *et al*. Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus. BMC Genomics, 2018, **19**(1): 1-20

[43] Keller M W, Rambo-Martin B L, Wilson M M, *et al*. Author correction: Direct RNA sequencing of the coding complete influenza a virus genome. Sci Rep, 2018, **8**(1): 15746

[44] Moon J, Jang Y, Kim N, *et al*. Diagnosis of haemophilus influenzae pneumonia by Nanopore 16 S amplicon sequencing of sputum. Emerg Infect Dis, 2018, **24**(10): 1944-1946

[45] Moon J, Kim N, Lee H S, *et al*. Campylobacter fetus meningitis confirmed by a 16 S rRNA gene analysis using the MinION nanopore sequencer, South Korea, 2016. Emerg Microbes Infect, 2017, **6**(1): 1-3

[46] Bainomugisa A, Duarte T, Lavu E, *et al*. A complete high-quality MinION nanopore assembly of an extensively drug-resistant *Mycobacterium tuberculosis* Beijing lineage strain identifies novel variation in repetitive PE/PPE gene regions. Microb Genom, 2018, **4**(7). Doi: 10.1099/mgen.0.000188

[47] Runtuwene L R, Tuda J S B, Mongan A E, *et al*. Nanopore sequencing of drug-resistance-associated genes in malaria parasites, Plasmodium falciparum. Sci Rep, 2018, **8**(1): 8286

[48] Charalampous T, Richardson H, Kay G L, *et al*. Rapid diagnosis of lower respiratory infection using Nanopore-based clinical metagenomics. bioRxiv, 2018, **387548**

[49] Korlach J, Marks P J, Cicero R L, *et al*. Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. Proc Natl Acad Sci USA, 2008, **105**(4): 1176-1181

[50] Foquet M, Samiee K T, Kong X, *et al*. Improved fabrication of zero-mode waveguides for single-molecule detection. Journal of Applied Physics, 2008, **103**(3): 034301

[51] Wenger A M, Peluso P, Rowell W J, *et al*. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol, 2019, **37**(10), 1155 – 1162

[52] Wei Z G, Zhang S W. NPBSS: a new PacBio sequencing simulator for generating the continuous long reads with an empirical model. BMC Bioinformatics, 2018, **19**(1): 1-9

[53] Rhoads A, Au K F. PacBio sequencing and its applications. Genomics Proteomics Bioinformatics, 2015, **13**(5): 278-289

[54] Nakano K, Shiroma A, Shimoji M, *et al*. Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. Hum Cell, 2017, **30**(3): 149-161

[55] Shin S C, Ahn D H, Kim S J, *et al*. Advantages of single-molecule real-time sequencing in high-GC content genomes. Plos One, 2013, **8**(7): e68824

[56] Kelleher P, Murphy J, Mahony J, *et al*. Next-generation sequencing as an approach to dairy starter selection. Dairy Sci Technol, 2015, **95**: 545-568

[57] Gordon D, Huddleston J, Chaisson M J, *et al*. Long-read sequence assembly of the gorilla genome. Science, 2016, **352**(6281): aae0344

[58] Pendleton M, Sebra R, Pang A W, *et al*. Assembly and diploid architecture of an individual human genome *via* single-molecule technologies. Nat Methods, 2015, **12**(8): 780-786

[59] Wagner J, Coupland P, Browne H P, *et al*. Evaluation of PacBio sequencing for full-length bacterial 16 S rRNA gene classification. BMC Microbiol, 2016, **16**(1): 1-17

[60] Campino S, Marin-Menendez A, Kemp A, *et al*. A forward genetic screen reveals a primary role for Plasmodium falciparum Reticulocyte Binding Protein Homologue 2a and 2b in determining alternative erythrocyte invasion pathways. Plos Pathog, 2018, **14**(11): e1007436

[61] Slager J, Aprianto R, Veening J W. Deep genome annotation of the opportunistic human pathogen *Streptococcus pneumoniae* D39. Nucleic Acids Res, 2018, **46**(19): 9971-9989

[62] Mizuguchi T, Suzuki T, Abe C, *et al*. A 12-kb structural variation in progressive myoclonic epilepsy was newly identified by long-read whole-genome sequencing. J Hum Genet, 2019, **64**(5): 359-368

[63] Audano P A, Sulovari A, Graves-Lindsay T A, *et al*. Characterizing the major structural variant alleles of the human genome. Cell, 2019, **176**(3): 663-675.e19

[64] Zhang J, Zhang X, Tang H, *et al*. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum L*. Nat Genet, 2018, **50**(11): 1565-1573

[65] Wittmeyer K, Cui J, Chatterjee D, *et al*. The dominant and poorly

penetrant phenotypes of maize unstable factor for orange1 are caused by dna methylation changes at a linked transposon. Plant Cell, 2018, **30**(12): 3006-3023

[66] Low W Y, Tearle R, Bickhart D M, *et al*. Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. Nat Commun, 2019, **10**(1): 260

[67] Kim B M, Kang S, Ahn D H, *et al*. The genome of common long-arm octopus Octopus minor. Gigascience, 2018, **7**(11). Doi: 10.1093/gigascience/giy119

[68] Wang Y, Zhang J, Chen X, *et al*. Profiling of oral microbiota in early childhood caries using single-molecule real-time sequencing. Front Microbiol, 2017, **8**(2244): 1-15

[69] Su B, Zheng X, Liu Y, *et al*. Detection of pretreatment minority HIV-1 reverse transcriptase inhibitor-resistant variants by ultra-deep sequencing has a limited impact on virological outcomes. J Antimicrob Chemother, 2019, **74**(5): 1408-1416

[70] Smith C C, Paguirigan A, Jeschke G R, *et al*. Heterogeneous resistance to quizartinib in acute myeloid leukemia revealed by single-cell analysis. Blood, 2017, **130**(1): 48-58

[71] Lian B, Hu X, Shao Z M. Unveiling novel targets of paclitaxel resistance by single molecule long-read RNA sequencing in breast cancer. Sci Rep, 2019, **9**(1): 6032

[72] Hiraoka S, Okazaki Y, Anda M, *et al*. Metaepigenomic analysis reveals the unexplored diversity of DNA methylation in an environmental prokaryotic community. Nat Commun, 2019, **10**(1): 159

[73] Du N, Sun Y. Improve homology search sensitivity of PacBio data by correcting frameshifts. Bioinformatics, 2016, **32**(17): i529-i537

[74] Kronenberg Z N, Rhie A, Koren S, *et al*. Extended haplotype phasing of *de novo* genome assemblies with FALCON-Phase. bioRxiv, 2019, **327064**

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

[76] Xiao C-L, Chen Y, Xie S-Q, *et al*. MECAT: an ultra-fast mapping, error correction and *de novo* assembly tool for single-molecule sequencing reads. bioRxiv, 2016, **089250**

[77] Li H. Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. Bioinformatics, 2016, **32**(14): 2103-2110

[78] Vaser R, Sović I, Nagarajan N, *et al*. Fast and accurate *de novo* genome assembly from long uncorrected reads. Genome Research, 2017, **27**(5): 737-746

The Application and Research Progress of Single Molecule Sequencing Technology*

YU Xiao-Ling[1)], JIANG Wen-Qian[1)], ZHENG Ling[1)], SHI Yang[2,3)], YE Han-Hui[1)], LIN Jun[2,3,4)**]

([1)]*Infectious Department, Mengchao Hepatobiliary Hospital of Fujian Medical university, Fuzhou* 350025, China;

[2)]*College of Biological Science and Engineering, Fuzhou University, Fuzhou* 350108, China;

[3)]*Institute of Applied Genomics, Fuzhou University, Fuzhou* 350108, China;

[4)]*Fujian Key Laboratory of Marine Enzyme Engineering, Fuzhou University, Fuzhou* 350108, China)

Abstract The discovery of the DNA double helix structure has turned life science research into the molecular level, and the sequencing technology that emerged in the 1970s has made a great contribution to the deciphering of the genetic code. Single-molecule sequencing technology, also known as third-generation sequencing technology, which has appeared in recent years, can read nucleotide sequences at the single molecular level. The single-molecule sequencing systems mainly include HeliScope, Nanopore and PacBio. Compared to traditional first- and next-generation sequencing technologies, third-generation sequencing can produce longer reads, sequence RNA directly without reverse transcription, and the speed is extremely fast. Meanwhile the equipment of some systems can be miniaturized and portable for in-field sequencing. The third-generation sequencing technology has numerous applications in the basic theroretical research of life science and the clinical practice in biomedicine. This paper focuses on the principles, the pros and cons, and the research progress and applications, of various single-molecule sequencing methods.

Key words single molecular sequencing, HeliScope sequencing, Nanopore sequencing, PacBio sequencing

DOI：10.16476/j.pibb.2019.0167

* This work was supported by grants from Clinical Medicine Center Construction Program of Fuzhou, Fujian, P.R.C（2018080306）, Key Clinical Specialty Discipline Construction Program of Fuzhou, P.R.C（201510301）, Key Clinical Specialty Discipline Construction Program of Fujian, P.R.C, Health and Technology Innovation Platform Construction Project of Fuzhou（2019-S-wp6）and Health Research Innovation Team Cultivation Project of Fuzhou（2019-S-wt4）.

** Corresponding author.

Tel：86-591-22863805, E-mail：jun@fzu.edu.cn.

Received：July 20, 2019 Accepted：December 2, 2019