基因组学与应用生物学, 2023 年, 第 42 卷, 第 3 期, 第 233-241 页

述评
Review

# 纳米孔测序技术在基因组学中的应用研究进展

雷文龙[1] 雷思茹[1] 陈帅[2] 孔维龙[2] 张兴坦[2] 唐海宝[1*]
1 福建农林大学生命科学学院, 福州, 350000; 2 中国农业科学院深圳农业基因组研究所, 深圳, 518000
*通信作者, tanghaibao@gmail.com

**摘 要** 牛津纳米孔技术(Oxford nanopore technologies, ONT)是新兴的第三代单分子测序技术, 以其超长读长、直接测序、实时测序等技术优势在生命科学领域发挥了重要作用, 尤其在解决测序读长的技术瓶颈方面有着突破性的进展, 并广泛应用于基因组学的研究中。本综述简要介绍了 ONT 测序的基本原理及优势, 并重点阐述了 ONT 测序在基因组从头组装、基因组结构变异鉴定、三维基因组结构研究及碱基修饰检测领域中的应用及研究进展, 同时就 ONT 测序未来面临的机遇和挑战进行了讨论与展望。

**关键词** 牛津纳米孔技术(ONT); ONT 应用; 基因组学

# Advances in Application of Nanopore Sequencing Technology in Genomics

LEI Wenlong[1] LEI Siru[1] CHEN Shuai[2] KONG Weilong[2] ZHANG Xingtan[2] TANG Haibao[1*]
1 College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, 350000; 2 Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, 518000
*Corresponding author, tanghaibao@gmail.com

DOI: 10.13417/j.gab.042.000233

**Abstract** Oxford nanopore technologies (ONT) is an emerging third-generation single-molecule sequencing technology, which offers various technical advantages of ultra-long read length, direct sequencing and real-time sequencing. ONT has made significant contributions to the life science field, especially in addressing the technical limitations of short sequencing read length with breakthrough progress and is widely used in genomics research. This review briefly introduced the principles and advantages of ONT sequencing, with a primary focus on the application and research progress of ONT sequencing in the *de novo* genome assembly, genome structure variation detection, study of three-dimensional genome structures and base modification detection. The future opportunities and challenges of ONT sequencing are also discussed and prospected.

**Keywords** Oxford nanopore technologies(ONT); ONT application; Genomics

1986 年, 美国遗传学家 Thomas Roderick 首次提出了基因组学的概念:基因组学是对生物体内的所有基因进行集体表征和定量研究, 以及在不同的基因组之间进行比较的一门交叉生物学学科。基因组学研究的基础是通过使用高通量 DNA 测序技术和生物信息学分析来组装和分析整个基因组的功能和结构。牛津纳米孔技术(Oxford nanopore technologies, ONT)作为最新一代的高通量测序技术, 于 20 世纪 80 年代由生理与医学诺贝尔奖获得者 Neher 和 Sakamann 首次提出, 经过几十年的发展, 英国 Oxford Nanopore Technologies 公司于 2012 年发布 MinION 测序仪, 通过不断的改进, 于 2015 年将其商业化(Wang et al., 2021)。此后, 英国 Oxford Nanopore Technologies 公司还分别发布 GridION 和 PromethION 测序仪, 自此, ONT 测序技术在基因组学中的应用有了大幅增长(Leggett and Clark, 2017)。作为最新一代的测序技术, ONT 测序在测序读长上有了显著提升, 并且在基因组从头组装、碱基修饰检测、基因组结构变异鉴定及三维基因组研究领域中发挥重要作用。

基金项目: 本研究由国家自然科学基金项目(32222019)资助。
引用格式: 雷文龙, 雷思茹, 陈帅, 等, 2023.纳米孔测序技术在基因组学中的应用研究进展.基因组学与应用生物学, 42(3): 233-241. [LEI W L, LEI S R, CHEN S, et al., 2023. Advances in application of nanopore sequencing technology in genomics, Genomics and Applied Biology, 42(3): 233-241.]

## 1　测序技术的发展

测序技术自二十世纪 70 年代发展至今，一共经历了三次重大变革，在测序读长和测序通量上都得到了极大提升。1953 年，沃森(James Waston)和克里克(Francis Crick)破译了 DNA 双螺旋结构，自此，分子生物学技术飞速发展，生物学家一直致力于研究 DNA 双螺旋结构中的 ATGC 四种碱基的排列顺序，碱基属于纳米级分子，并且嘌呤与嘧啶之间的化学结构高度相似，其增加了碱基识别的难度，但这一难题恰恰引起了全世界从事分子生物学研究的科学家的关注，在一定程度上推动了测序技术的发展。桑格和卡尔森开创的链终止法(Sanger and Coulson，1975)以及沃特·吉尔伯特(Walter Gibert)发明的链降解法(Maxam and Gilbert，1977)标志着第一代测序技术的诞生，他们借助该技术破译了第一个噬菌体(phage)基因组序列，自此便开启了基因组学研究的大门。链终止法的核心原理是 ddNTP 的 2′和 3′端不含羟基，核酸链在合成的过程中无法形成磷酸二酯键，从而终止了 DNA 合成反应。在测定核酸序列时，向反应体系中加入一定比例的带有放射性标记的 4 种 ddNTP，利用 DNA 聚合酶来延伸结合在待测核酸模板上的引物，直到掺入一种链终止核苷酸为止，最终得到的终止产物为一组长度各相差一个碱基的链，然后利用高分辨率变性凝胶电泳进行分离并根据其长度排序，最后利用 X-光胶片放射自显影进行检测，从而确定目的核酸片段的碱基排列顺序。第一代测序技术的主要特点是测序读长较长(普遍读长在 700~1 000 bp)，测序准确率可达 99.99%，高于二、三代测序。但第一代测序技术测序通量很低，相应的获得大量序列的测序成本极高。

人类基因组计划的完成，标志着功能基因组时代的到来，第一代测序技术已经无法满足大规模基因组测序对于测序通量及效率的需求。为了弥补一代测序的技术缺陷，在 Sanger 测序技术的基础上通过技术变革，诞生了第二代测序技术，也称下一代测序技术(next generation sequencing，NGS)。最具代表性的二代测序平台有：基于焦磷酸测序法的罗氏(Roche)454 测序平台、美国 Illumina 公司的 Solexa 测序平台及美国 ABI 公司的 SOLiD 测序平台等。随着二代测序市场的激烈竞争，美国 Illumina 公司推出的 HiSeq 测序平台迅速成为主流，其核心思想是边合成边测序(sequencing by synthesis，SBS)(Fuller et al.，2009)，利用四种不同颜色的荧光标记 dNTP，当互补链在 DNA 聚合酶的作用下合成时，每次添加一种 dNTP 就会释放出不同的荧光，通过特定的计算软件处理荧光信号，从而获得待测 DNA 的序列信息。第二代测序的技术特点是通量高，准确度高达 99%，但同时存在一定的短板，即 Illumina 平台产生的 reads 读长较短，一般为 150 bp 左右，最长能达到 250~300 bp，并且二代测序具有明显的 GC 偏好性，基因组上(G+C)含量在 50%左右的区域更容易被检测到，覆盖到的 reads 更多，在高(G+C)含量或低(G+C)含量区域的覆盖度则较少(Tilak et al.，2018)。虽然第二代测序技术相较于第一代测序技术有了巨大的改进，但是其本质仍然是建立在 PCR 扩增的基础上，在 PCR 过程中有一定概率会引入错配碱基。

美国 Pacific Bioscience 公司发布了 SMRT 新型单分子荧光测序平台 PacBio RS(McCarthy，2010)，并于 2011 年进行商业发售。英国 Oxford Nanopore Technologies 公司推出 MinION 测序平台(Eisenstein，2012；Jain et al.，2016)，并于 2015 年将其商业化。至此，第三代测序时代来临。第三代测序技术弥补了短读长带来的基因组拼接难题以及避免了测序过程中由于 PCR 扩增带来的偏差，SMRT 测序技术也采用边合成边测序的策略，基本原理是利用四色荧光标记的 dNTP 和零模波导孔完成单个 DNA 的分子测序。而 ONT 平台则是利用电泳驱动单个分子逐一通过纳米孔来完成测序。与二代测序相比，第三代测序技术的优点主要体现在单分子实时测序，测序过程无须进行 PCR 扩增，并且在测序读长上具有明显的优势，ONT 测序平台产生的 reads 可达几百 kbp 甚至 Mb 级别。

## 2　ONT 测序的基本原理和特点

### 2.1　ONT 测序的原理

ONT 测序是基于纳米孔电信号的单分子实时测序技术，其核心是将一个纳米级的蛋白质孔(van Dijk et al.，2018)置于两个电解液室之间，而测序芯片中镶嵌有纳米孔通道的脂质双分子层将两个充满电解质溶液的隔室隔开，DNA 双链在马达蛋白的牵引下与锚定在生物膜上的八聚体纳米孔蛋白结合并解螺旋。由于化学性质不同，不同碱基通过纳米孔时，会引起不同电信号的变化。通过对这些电信号变化进行检测和解析，从而完成碱基序列的实时测定(图 1)。

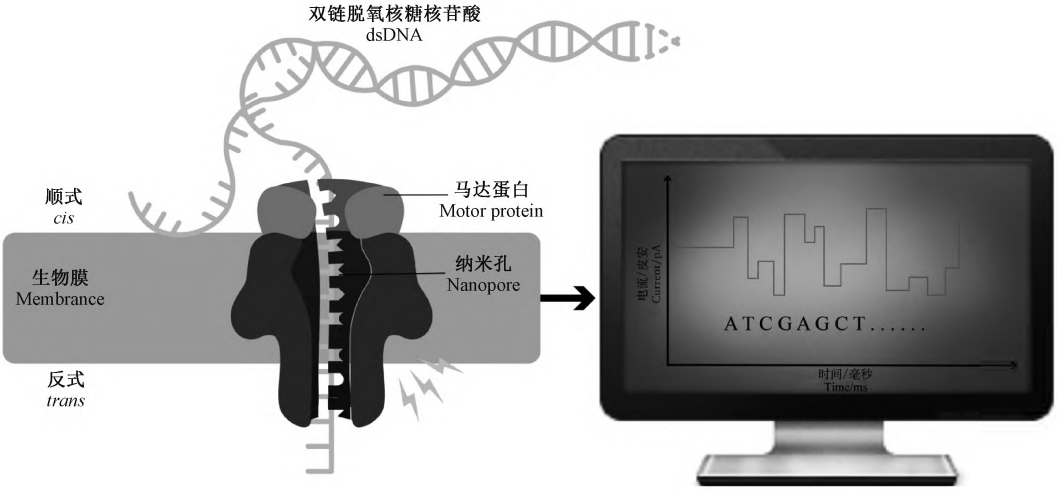


**图 1　ONT 测序原理**

**Figure 1　Sequencing principle of ONT**

### 2.2　ONT 测序的特点

超长读长。ONT 测序产生的 reads 长度不受测序设备的限制，reads 长度可以与输入片段长度保持一致，原则上可以检测通过纳米孔的全部核酸序列，因此，可以通过改变文库制备的方案来控制片段长度，其最长 read 可达 Mb 级别。这些超长 reads 具有诸多优势，可以跨越基因组高重复的区域，从而获得高质量、高连续的基因组组装。

直接、实时测序。ONT 测序文库制备简易快捷，无须对样品进行 PCR 扩增，可以直接对 DNA 或者 RNA 文库进行测序，并且能够完整保留碱基修饰的信息，既避免了 PCR 扩增过程中出现的碱基错配，也在一定程度上节约了时间和成本。而且可以边测序边输出结果，实现对测序数据的实时分析。

灵活便携。目前使用较为成熟的 MinION 纳米孔测序仪简单便携，重量不足 100 g，且操作系统简单，可以在各种复杂环境下完成实时测序。

## 3　ONT 测序的技术应用

### 3.1　ONT 测序在基因组组装中的应用

基因组包含了物种所有的生命遗传物质，为生命行使生物学功能提供指导（高胜寒 等，2018），完整解析物种基因组碱基序列是进行基因组学研究的基础。基因组组装一般可以分成三个层次：首先，是将全基因组测序生成的序列片段（reads）基于重叠区域进行拼接生成较长的连续片段（contig）；其次，通过不同长度的大片段（Mate-pair）文库对 contigs 进行排序从而获得更长的 scaffolds；最后，通过 gap 填充以及借助遗传图谱、光学图谱或染色质构象捕获技术（chromatin conformation capture，3C），将 scaffolds 锚定到染色体上，从而组装出染色体级别的基因组。

PacBio 和 ONT 测序技术尚未大量应用之前，二代测序平台承担了大部分物种的基因组组装和拼接工作，受限于二代测序的短读长，依赖于二代测序进行基因组组装仍然存在较大的困难，比如全基因组测序得到的短度远远小于基因组实际总长度，高通量测序得到的大批量数据会导致计算复杂度的增加以及计算资源的耗费，并且部分物种的基因组经历了全基因组复制、染色体重排、丢失以及重复序列大规模扩张之后，基因组高度重复且高度杂合（唐蝶和周倩，2021），二代测序产生的 reads 无法跨越基因组重复序列及高杂合区域，组装时会出现大量的分支（branch）和环化（loop），导致组装结果片段化。而 ONT 测序读长可达几百 kb，较二代测序提高了几十倍，并且相比于 PacBio 三代测序技术，ONT 平台独有的 Ultra-long 测序能够产生超长测序片段，可以跨越基因组高度重复区域，填补已有参考基因组存在的大量 gaps，尤其是在有丝分裂或减数分裂中起重要作用的着丝粒区域及与细胞衰老相关的端粒区域。并且在此基础上配合多轮纠错机制，在保证准确率的同时，能够较好地比对 DNA 序列，从而组装出高质量、高完整度的物种基因组。

借助 ONT 测序技术在基因组组装中的优势，端粒到端粒联盟（Telomere-to-Telomere）提供了目前最为完整的人类参考基因组 T2T-CHM13（Nurk et al.，2022），由此产生的新的参考基因组填补了 GRCh38

人类基因组 8%的空白。他们使用纳米孔 MinION 测序人类 CHM13hTERT 细胞系，获取了 50×的 ONT 数据，最终组装了 2.94 G 的基因组，contigN50 为 75 Mb，在完整性和连续性上超越了 contigN50 为 56 Mb 的 GRCh38 细胞系。此外，Miga 等(2020)借助 ONT Ultra-long 测序成功完成人类 X 染色体 T2T 组装，在该组装中，X 染色体中的 29 个缺口序列被全部填充，共计 1 147 861 bp。在重新组装的 3.1 Mb 的着丝粒中包括 CT-X 和 GAGE 等与癌症发生相关的区域，这些序列的补充促进了人类基因组相关疾病和癌症发生机理的解析。Logsdon 等(2021)同样使用 ONT Ultra-long 数据完成了人类 8 号染色体 T2T 组装，该组装成功填补了 5 个长期存在的 gaps，包括 2.08 Mb 的着丝粒序列，644 kb 的 β-防御素基因簇序列，863 kb 可变数目串联重复序列以及长臂和短臂末端的端粒区。此外，该研究利用相同的测序方式和组装方法完成了黑猩猩(*Pan troglodytes*)、猩猩(*Pongo pygmaeus*)和猕猴(*Macaca mulatta*)的 8 号染色体着丝粒的组装，并且对其进化历史进行解析，通过比较分析和系统发育分析发现黑猩猩的第 8 号染色体的着丝粒组织结构与人类最为相似。这些分析提供了多个灵长类动物同源染色体着丝粒区域的序列，为未来研究高度重复的基因组区域的遗传变异和进化提供了框架。

利用 ONT 测序数据同样使模式植物拟南芥(*Arabidopsis thaliana*)基因组取得重大突破，Naish 等(2021)首先基于 ONT Ultra-long reads 搭建了拟南芥染色体框架，并且结合 HiFi 数据进行拼接及矫正，成功组装出了拟南芥近 T2T 的基因组 Col-CEN v1.2，与拟南芥 TAIR10 基因组相比，Col-CEN v1.2 基因组中 5 条染色体的着丝粒区域被完整组装，其中 1 号、3 号和 5 号染色体完成了端粒到端粒的完整组装，2 号和 4 号染色体除短臂上富含 45 S 核糖体 DNA 和邻近的端粒区域外，也基本组装完成。他们发现，拟南芥的着丝粒 DNA 由 178 bp 的高度重复序列(CEN180)组成，同时基于新组装的着丝粒卫星序列，研究人员发现不同染色体上的 CEN180 序列具有明显的特异性，仅有 0.3%的 CEN180 序列是完全相同的，而同一染色体内 57%~69%的 CEN180 都具有多个拷贝，证明了拟南芥同一条染色体内的 CEN180 序列存在均质化的趋势。除此之外，他们还发现 ATHILA 逆转录转座子的插入会破坏了 CEN180 序列的遗传性和表观遗传学结构，促进 CEN180 序列的多样性，证明了拟南芥着丝粒的进化是由均质化和多样化共同决定的。

随着 ONT 测序技术的成熟发展及在基因组组装中凸显出的优势，白菜型油菜(*Brassica rapa*)，甘蓝型油菜(*Brassica oleracea*)和裂果焦(*Musa schizocarpa*)(Belser et al., 2018)、高粱(*Sorghum bicolor*)(Deschamps et al., 2018)、菊花(*Chrysanthemum nankingense*)、莱氏衣藻(*Chlamydomonas reinhardtii*)(Payne et al., 2023)、核桃楸(*Juglans mandshurica*)(Yan et al., 2021)等多个物种基因组依靠 ONT 测序平台得以更新，玉米(*Zea mays*)(Liu et al., 2020)、水稻(*Oryza sativa*)(Zhang et al., 2022)、西瓜(*Citrullus lanatus*)(Deng et al., 2022)、香蕉(*Musa acuminata*)(Belser et al., 2021)等物种依靠 Nanopore 的 Ultra-long 平台将基因组成功组装到近 T2T 水平。

### 3.2　ONT 测序在基因组结构变异鉴定中的应用

基因组变异形式多样，主要包括单点突变以及更大型的基因组结构变异(structural variations, SV)。SV 通常是指长度大于 50 bp 的序列变异，包含多种类型，主要有缺失、插入、重复、倒位及拷贝数变异(Roses et al., 2016)。相较于单核苷酸多态性(single nucleotide polymorphisms, SNP)，SV 在基因组变异中占比更大，广泛影响物种的表型，是物种进化的重要原因之一，因此，系统而准确地分析基因组的结构变异，对解析物种进化及遗传多样性具有重要的作用。

先前，绝大多数的变异研究集中在 SNP 上，这是由于在三代测序出现之前，SV 检测较为困难，一方面由于 SV 的类型较多，突变之间的差异较大，造成了检测的困难，另一方面，大片段的 SV 可以达到几百 kb，对于二代测序短片段来说，难以跨越整个 SV，因此，基于二代测序技术检测结构变异已无法满足现阶段的各项应用研究需求，而 ONT 测序技术在一定程度上解决了这一难题。ONT 测序产生的长测序片段可以较为完整而有效地保留重复片段的区域信息，长片段的 SV 也可以较为完整而准确地被鉴定。

利用 ONT 测序进行结构变异鉴定已被成功地应用于癌症检测中，Gong 等(2018)利用 ONT 技术成功鉴定了人乳腺导管癌细胞 HCC118T 基因组结构变异，得到了 34 100 个 SV 以及相近的 66 660 个断点，他们通过 PCR 以及多重 PCR 相结合的方法验证了 200 多个 SV 位点，证实了 ONT 鉴定 SV 具有较高的准确性，基于鉴定到的 SV，他们发现启动子区

域及调控区域的 SV 可选择性地激活原癌基因可能导致抑癌基因的失活，从而造成恶性肿瘤的发生。此外，利用 ONT 测序技术鉴定群体结构变异在基因组进化及功能塑造的过程中发挥重要作用，Quan 等(2021)利用 ONT 测序技术揭示了藏族人群的基因组结构变异使其更适应青藏高原的气候环境。他们首先对 25 个人类样本(10 个汉族，15 个藏族)进行 10×的 ONT 测序，绘制了中国藏汉人群的结构变异图谱，经过筛选，最终获得 38 216 个 SV，对比短读长测序的 SV 鉴定结果，结构变异图谱中有 27% 的 SV 是首次解析，他们进一步以纳米孔测序检测到的 SV 为标准，对 276 个二代短读长测序进行了 SV 分型并解析了藏汉人群的遗传结构，筛选出了 80 个与高原环境适应性相关的候选基因，该研究证实了 ONT 鉴定 SV 具有更高的灵敏度，并且揭示了 SV 在藏族人群高原适应过程中发挥重要作用。在冰岛人 SV 研究中，Beyter 等(2021)证实了结构变异在人类表型中的作用，他们利用 ONT 平台对 3 622 名冰岛人进行测序，通过筛选过滤，最终发现 133 886 个可靠的 SV 等位基因。在对冰岛人的表型直接测试及基于连锁不平衡(linkage disequilibrium)的方法探索 SV 对疾病和其他性状的影响，结果发现了多个影响表型形成的 SV，如胆固醇代谢调节因子 *PCSK9* 第一个外显子的罕见缺失与较低的低密度脂蛋白胆固醇(low-density lipoprotein cholesterol)水平有关；编码聚集蛋白聚糖(aggrecan)的 *ACAN* 基因可变数目的串联重复(variable number tandem repeat)与身高相关。

总之，以上结果表明，ONT 测序在 SV 检测以及人类医学研究中占据越来越高的比重，必将持续引爆基因组遗传变异研究的热点。

### 3.3 ONT 测序在三维基因组研究中的应用

先前，对基因的表达调控研究主要以基因以及调控元件的线性关系为基础，但是“结构决定功能”已经是一个共识。而且近期的研究表明，基因表达的调控不单单是简单的线性调控关系，而是存在三维空间的复杂调控网络，所以，研究全基因组的三维空间结构和功能已成为基因组学的一个新发展趋势，Dekker 团队提出了用于研究三维基因组学的染色质构象捕获技术(3C)(Dekker et al., 2002)，用于检测染色质特定点到点之间交互作用，随后又发展出 4C(Simonis et al., 2006)、5C(Dostie et al., 2006)技术，分别用于测定一点到多点以及多点到多点的染色质交互作用。为了能在全基因组范围内研究染色质的相互作用，Lieberman-Aiden 团队以高通量测序(Schmitt et al., 2016)为手段，以 3C 技术为基础，衍生出了高通量染色质捕获技术(high-throughput chromosome conformation capture, Hi-C)(Lieberman-Aiden et al., 2009)，同年，Fullwood 团队发表了另一项重要技术，即配对末端标记测序技术分析染色质相互作用(Chromatin interaction analysis using paired end tag sequencing, ChIA-PET)(Fullwood et al., 2009)，Hi-C 和 ChIA-PET 技术成功应用在三维基因组学的研究中，解析了包括拟南芥(Tao et al., 2017)、水稻(Zhao et al., 2006)、玉米(Li et al., 2019)、棉花(*Gossypium*)(Wang et al., 2018)等在内的多个物种的染色质互作图谱，但是 Hi-C 和 ChIA-PET 技术也存在一定的局限性，即二代测序读长的限制以及双端测序的特点，使其一次只能检测两个基因组位点的空间互作，但是大量研究表明，基因组空间存在大量的高阶(两个以上的基因组位点)互作。

为了解决这一技术瓶颈，2019 年，英国 Oxford Nanopore Technologies 公司联合美国康奈尔医学院将 ONT 长读长测序技术与染色质构象捕获技术相结合，开发了 Pore-C 测序技术(Ulahannan et al., 2019)，实现了获取基因组内高阶互作信息的功能。他们利用 Pore-C 技术对人类 B 淋巴细胞(GM12878)DNA 进行测序，并与先前发布的黄金标准 Hi-C 数据进行比较，发现 Pore-C 在 A/B 区室、TAD 互作频率及边界鉴定上与 Hi-C 的结果相当甚至更好，并且，Pore-C测序检测基因组高阶互作的比例达到了 79%，远远高于现有的高阶三维基因组技术 SPRITE。此外，他们将 Pore-C 技术结合自主开发的 Chromunity 算法应用到人类 B 淋巴细胞系 GM12878 的三维基因组研究中，并通过高阶互作数据揭示了增强子和启动子之间相互作用的协同相关性。在对乳腺癌细胞系 HCC1954 的三维基因组研究中发现增强子的协同作用导致复杂乳腺癌 Tyfonas 的结构变异(Deshpande et al., 2022)。在植物领域研究中，Huang 等(2022)联合 Pore-C 技术及多种三维基因组分析技术，构建了目前精度最高的陆地棉(*Gossypium hirsutum* L.)TM-1 三维基因组图谱。通过分析发现，Pore-C 技术能够识别更多拓扑相关结构域之间的互作(TAD clique)，并且发现 TAD clique 的大小与包含的 B compartment 区域成正比例关系，与基因表达水平成反比例关系。此外，研究人员通过多

组学联合分析，鉴定了基因组上的非编码调控元件，剖析了亚基因组之间的同源基因表达不平衡与三维基因组结构差异的相关性（Huang et al., 2022）。

### 3.4　ONT 测序在甲基化鉴定中的应用

甲基化修饰是化学修饰的一种形式，是指在甲基转移酶的催化下将甲基基团添加到 DNA 或者 RNA 分子上，形成带有甲基化功能基团或羟甲基化功能基团的核苷酸分子。甲基化是机体内一种非常保守的表观遗传修饰机制（Zhang et al., 2018），与生物的生态适应密切相关。已有研究表明，甲基化在基因表达调控（Roundtree et al., 2017）、基因印记（Bewick et al., 2016）、转座子沉默（Wang et al., 2019）中起重要作用。最常见的甲基化修饰是 DNA 甲基化和 RNA 甲基化。

无论是在哺乳动物还是在植物中，比较常见并且研究较为透彻的 DNA 甲基化修饰是在胞嘧啶的五号位置上加入甲基化功能基团，进而形成的 5mC 甲基化修饰。目前，重亚硫酸氢盐测序（whole genome bisulfite sequencing，WGBS）（Kernaleguen et al., 2018）一直作为 5mC 检测的黄金标准，其基本原理是待测核酸首先经过重亚硫酸氢盐试剂进行处理，未甲基化的胞嘧啶在试剂的作用下进行脱氨基反应，并转化为尿嘧啶，但甲基化的胞嘧啶不会发生该反应，尿嘧啶经过 PCR 扩增之后会转化为胸腺嘧啶，由此甲基化胞嘧啶与未甲基化胞嘧啶得以区分开。但是这种测序方法也存在一定的缺点，即重亚硫酸氢盐的处理会将 DNA 分子降解，导致 DNA 分子的片段化，并且 WGBS 属于二代测序，产生的 reads 较短，因此难以覆盖高重复基因组的重复序列。而 ONT 测序通过修饰碱基和未修饰碱基的纳米孔 reads 产生的电流强度差异来检测甲基化，实现快速、长读长测序和单碱基单分子灵敏度。核酸分子通过纳米孔时不需要进行任何的预处理，不扩增，不富集，能够还原碱基修饰信息，并且依靠长读长的优势可以跨越基因组高度重复区域，避免了拼接比对错误，因此在甲基化修饰检测中具有独特的优势，基于 ONT 测序技术除了能鉴定 5mC 甲基化修饰之外，5hmC、6mA 等修饰也能够得到准确的鉴定。

目前，已有十几种分析工具使用 ONT 测序直接检测 DNA 甲基化（表 1），例如基于应用统计原理的 Tombo（Stoiber et al., 2016），基于隐马尔可夫模型算法的 Nanopolish（Simpson et al., 2017）、基于神经网络模型的 Megalodon、Guppy、Deepsignal（Ni et al., 2019）、DeepMod（Yuen et al., 2021），以及多工具相结合的 METEORE（Liu et al., 2019）。这些工具大多适用于检测人类等哺乳动物 DNA CpG 位点的甲基化，但是不同于哺乳动物，植物中的 DNA 甲基化不仅发生在 CpG 位点，而且也广泛存在于非对称性的 CHG、CHH 位点中，并调控植物的生理过程。为解决这一问题，王建峰团队在原有开发的工具 Deepsignal 的基础上，进一步提出了适合植物 DNA 甲基化鉴定的 Deepsignal-plant（Ni et al., 2021）工具，其基本原理是利用双向循环神经网络 BiLSTM 构建深度学习模型来处理纳米孔测序数据中目标位点 k-mer 的信号特征和序列特征，并且设计样本平衡和去噪策略处理训练样本数据，去除了训练数据中的假阳性样本，得到了高度可信的训练数据。通过拟南芥、水稻、黑芥（*Brassica nigra*）等数据集对训练模型进行性能测试，结果发现，Deepsignal-plant 在 3 个 motif（CpG、CHG、CHH）的预测结果与亚硫酸氢盐测序结果能够达到极高的相关性。并且，与亚硫酸氢盐测序的数据相比，ONT 测序凭借长读长优势在拟南芥中多鉴定到了 1% 的 5mC 位点，在水稻中多鉴定到 5% 的 5mC 位点，此外，通过对拟南芥和水稻重复区域片段进行甲基化鉴定时，呈现出差异甲基化水平，表明 ONT 数据在高重复基因组的 DNA 甲基化鉴定中具有可行性。

**表 1　基于 ONT 测序数据鉴定甲基化的软件比较分析**

**Table 1　Comparison of software for identifying methylation on ONT sequencing data**

| 工具<br>Tools | 基序<br>Motif | 原理<br>Principle | 优点<br>Advantages | 缺点<br>Disadvantages | 物种<br>Species |
|---|---|---|---|---|---|
| Tombo | 5mC/6mA/dam/dcm | 统计学<br>Statistics | 速度快、种类多<br>Fast and versatile | 假阳性高<br>High false-positive | 人、细菌<br>*Homo sapiens*, Bacteria |
| Nanopolish | 5mC/CpG | 隐马尔可夫模型<br>Hidden Markov Model | 速度快<br>Fast | 假阳性高<br>High false-positive | 人<br>*Homo sapiens* |
| mCaller | 6mA/5mCpG | 神经网络算法<br>Neural network algorithm | 准确性高<br>High accuracy | 物种限制<br>Species restrictions | 大肠杆菌<br>*Escherichia coli* |

续表
Continuing table

| 工具 Tools | 基序 Motif | 原理 Principle | 优点 Advantages | 缺点 Disadvantages | 物种 Species |
|---|---|---|---|---|---|
| Deepsignal | 6mA/5mCpG | 神经网络算法 Neural network algorithm | 速度快、准确率高 Fast and high accuracy | 准确性不稳定 Inconsistent accuracy | 人 *Homo sapiens* |
| Deepsignal-plant | 5mCpG/5mCHG/5mCHH | 神经网络算法 Neural network algorithm | 速度快、准确率高 Fast and high accuracy | 依赖其他软件 Reliance on other software | 拟南芥、水稻 *Arabidopsis thaliana*, *Oryza sativa* |
| NanoDisco | 6mA/4mC/5mC | 机器学习 Machine Learning | 灵敏度高 High sensitivity | 检测范围窄 Narrow test range | 细菌 Bacteria |
| METEORE | 5mCpG | 盘蛇 Snakemake | 准确率高 High accuracy | 依赖其他软件 Reliance on other software | |
| Megalodon | 6mA/5mCpG | 神经网络算法 Neural network algorithm | 准确率高 High accuracy | 速度慢、耗费资源 Slow and resource-intensive | 人 *Homo sapiens* |
| Guppy | 5mCpG/6mA | 碱基识别 Base calling | 检测种类多 Detection of a wide variety of bases | 耗费资源 Resource-intensive | 大肠杆菌、人 *Escherichia coli*, *Homo sapiens* |

ONT 可以对长链 RNA 直接测序，以实现 RNA 甲基化修饰的鉴定。FIO1 包含一个甲基转移酶结构域，与人类的甲基化转移酶 METTL16 同源，Xu 等(2022)发现 *FIO1* 突变导致 mRNA 的 $m^6A$ 修饰水平的下降，为了验证 *FIO1* 的功能，他们分别对 3 个野生型和 *FIO1* 突变体的拟南芥幼苗进行 Nanopore Direct RNA 测序，并且比较了野生型与 *FIO1* 突变体之间差异化的 $m^6A$ 修饰，证明了 FIO1 具有独特的 $m^6A$ 甲基转移酶功能，并优先在 CDS 区域进行 $m^6A$ 甲基化修饰，揭示了 *FIO1* 介导的 $m^6A$ 甲基化调控开花时间。该研究为理解 $m^6A$ 修饰在植物中的生物学功能提供了新的见解。

以上研究证明了甲基化修饰在基因组中的重要作用，这也将进一步促进 ONT 测序技术在鉴定甲基化修饰中的应用及进程。

## 4 讨论与展望

ONT 测序技术与第二代测序技术和 PacBio 的三代测序技术相比，具有便携、廉价、测序读长相对较长等方面的优势，自诞生以来，正在以极快的速度应用于基因组学研究，但是其本身仍然存在一些目前暂时无法改变的技术缺陷和局限性。不同于 Illumina 及 PacBio 的光信号原理，ONT 通过将电信号翻译成对应的碱基序列完成单分子测序，但是由于电信号的稳定性较差，并且噪声信号及随机误差也对碱基识别的准确率产生一定的影响，所以导致 ONT 在测序准确度上做出了很大的让步，远远低于二代测序和 PacBio 测序，并且其测序产生的错误并非随机，而是主要集中在是同聚物(homopolymer)和串联重复区域，即使经过自身纠错及二代数据校正，也难以达到较高的准确性，造成序列错误和真实变异难以区分，英国 Oxford Nanopore Technologies 公司为了提高准确性，采取了相应策略，比如改进纳米孔材料，改造马达蛋白酶及升级化学试剂，开发更为精确的碱基识别相关算法等。他们通过使用 R9 版纳米孔和 Kit10 试剂盒，在约两年的时间便将 92% 的原始读长准确度迭代到 98%，2021 年，该公司推出的 R10.4 纳米孔和 Kit12 试剂盒，将原始读长准确度提高到 99% 以上。

ONT 测序在读长上显著优于其他的测序平台。从理论上来讲，只要 DNA 链不发生断裂，就可以一直通过纳米孔。在未来的测序技术的发展趋势中，测序读长仍然是至关重要的一个指标，ONT 测序采取纯化高分子量 DNA、在原有技术的基础上通过更新纳米孔通道，以及升级相关化学试剂等方法增加读长将使其优势更加明显，进一步促进基因组组装的连续性，解决短读长难以攻克的复杂重复序列，对基因组结构变异的鉴定也具有更大的优势。

在解决了准确率的问题以及发挥更大读长优势之后，纳米孔测序技术具有非常诱人的应用前景，以 ONT 测序领衔的三代测序市场会越来越大，在基因组组装及后续的分析研究中应用也会越来越广泛。

## 作者贡献

张兴坦提供论文构思以及写作思路，并对论文提出修改意见；唐海宝对论文提出修改意见以及决定最终的定稿，雷文龙和雷思茹负责收集资料、整理资料；雷文龙负责论文的全部撰写；陈帅和孔维龙参与了论文修改和完善。全体作者都已阅读并同意最终的文本。

## 参考文献


高胜寒, 禹海英, 吴双阳, 等, 2018.复杂基因组测序技术研究进展.遗传, 40(11): 944-963.[GAO S H, YU H Y, WU S Y, et al., 2018. Advances of sequencing and assembling technologies for complex genomes. Hereditas, 40(11): 944-963.]

唐蝶, 周倩, 2021.植物基因组组装技术研究进展.生物技术通报, 37(6): 1-12.[TANG D, ZHOU Q, 2021. Research advances in plant genome assembly. Biotechnology Bulletin, 37(6): 1-12.]

BELSER C, BAURENS F C, NOEL B, et al., 2021. Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing. Commun. Biol., 4(1): 1047.

BELSER C, ISTACE B, DENIS E, et al., 2018. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. Nat. Plants, 4(11): 879-887.

BEWICK A J, JI L X, NIEDERHUTH C E, et al., 2016. On the origin and evolutionary consequences of gene body DNA methylation. Proc. Natl. Acad. Sci. USA, 113(32): 9111-9116.

BEYTER D, INGIMUNDARDOTTIR H, ODDSSON A, et al., 2021. Long-read sequencing of 3 622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nat. Genet., 53(6): 779-786.

DEKKER J, RIPPE K, DEKKER M, et al., 2002. Capturing chromosome conformation. Science, 295(5558): 1306-1311.

DENG Y, LIU S C, ZHANG Y L, et al., 2022. A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene disco-very and breeding. Mol. Plant, 15(8): 1268-1284.

DESCHAMPS S, ZHANG Y, LLACA V, et al., 2018. A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. Nat. Commun., 9(1): 4844.

DESHPANDE A S, ULAHANNAN N, PENDLETON M, et al., 2022. Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. Nat. Biotechnol., 40(10): 1488-1499.

DOSTIE J, RICHMOND T A, ARNAOUT R A, et al., 2006. Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res., 16(10): 1299-1309.

EISENSTEIN M, 2012. Oxford Nanopore announcement sets sequencing sector abuzz. Nat. Biotechnol., 30(4): 295-296.

FULLER C W, MIDDENDORF L R, BENNER S A, et al., 2009. The challenges of sequencing by synthesis. Nat. Biotechnol., 27(11): 1013-1023.

FULLWOOD M J, LIU M H, PAN Y F, et al., 2009. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature, 462(7269): 58-64.

GONG L A, WONG C H, CHENG W C, et al., 2018. Picky comprehensively detects high-resolution structural variants in nanopore long reads. Nat. Meth., 15(6): 455-460.

HUANG X H, TIAN X H, PEI L L, et al., 2022. Multi-omics mapping of chromatin interaction resolves the fine hierarchy of 3D genome in allotetraploid cotton. Plant Biotechnol. J., 20(9): 1639-1641.

JAIN M, OLSEN H E, PATEN B, et al., 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol., 17(1): 239.

KERNALEGUEN M, DAVIAUD C, SHEN Y M, et al., 2018. Whole-genome bisulfite sequencing for the analysis of genome-wide DNA methylation and hydroxymethylation patterns at single-nucleotide resolution. Methods Mol. Biol., 1767: 311-349.

LEGGETT R M, CLARK M D, 2017. A world of opportunities with nanopore sequencing. J. Exp. Bot., 68(20): 5419-5429.

LI E, LIU H, HUANG L L, et al., 2019. Long-range interactions between proximal and distal regulatory regions in maize. Nat. Commun., 10(1): 2633.

LIEBERMAN-AIDEN E, VAN BERKUM N L, WILLIAMS L, et al., 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 326(5950): 289-293.

LIU J N, SEETHARAM A S, CHOUGULE K, et al., 2020. Gapless assembly of maize chromosomes using long-read technologies. Genome Biol., 21(1): 121.

LIU Q, FANG L, YU G L, et al., 2019. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. Nat. Commun., 10(1): 2449.

LOGSDON G A, VOLLGER M R, HSIEH P, et al., 2021. The structure, function and evolution of a complete human chromosome 8. Nature, 593(7857): 101-107.

MAXAM A M, GILBERT W, 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA, 74(2): 560-564.

MCCARTHY A, 2010. Third generation DNA sequencing: pacific biosciences' single molecule real time technology. Chem. Biol., 17(7): 675-676.

MIGA K H, KOREN S, RHIE A, et al., 2020. Telomere-to-telomere assembly of a complete human X chromosome. Nature, 585(7823): 79-84.

NAISH M, ALONGE M, WLODZIMIERZ P, et al., 2021. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. Science, 374(6569): 7489.

NI P, HUANG N, NIE F, et al., 2021. Genome-wide detection of

cytosine methylations in plant from Nanopore data using deep learning. Nat. Commun., 12(1): 5976.

NI P, HUANG N, ZHANG Z, et al., 2019. DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. Bioinformatics, 35(22): 4586-4595.

NURK S, KOREN S, RHIE A, et al., 2022. The complete sequence of a human genome. Science, 376(6588): 44-53.

PAYNE Z L, PENNY G M, TURNER T N, et al., 2023. A gap-free genome assembly of *Chlamydomonas reinhardtii* and detection of translocations induced by CRISPR-mediated mutagenesis. Plant Commun., 4(2): 100493.

QUAN C, LI Y F, LIU X Y, et al., 2021. Characterization of structural variation in Tibetans reveals new evidence of high-altitude adaptation and introgression. Genome Biol., 22(1): 159.

ROSES A D, AKKARI P A, CHIBA-FALEK O, et al., 2016. Structural variants can be more informative for disease diagnostics, prognostics and translation than current SNP mapping and exon sequencing. Expert Opin. Drug Metab. Toxicol., 12(2): 135-147.

ROUNDTREE I A, EVANS M E, PAN T, et al., 2017. Dynamic RNA modifications in gene expression regulation. Cell, 169(7): 1187-1200.

SANGER F, COULSON A R, 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol., 94(3): 441-448.

SCHMITT A D, HU M, REN B, 2016. Genome-wide mapping and analysis of chromosome architecture. Nat. Rev. Mol. Cell Biol., 17(12): 743-755.

SIMONIS M, KLOUS P, SPLINTER E, et al., 2006. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat. Genet., 38(11): 1348-1354.

SIMPSON J T, WORKMAN R E, ZUZARTE P C, et al., 2017. Detecting DNA cytosine methylation using nanopore sequencing. Nat. Methods, 14(4): 407-410.

STOIBER M, QUICK J, EGAN R, et al., 2016. De novo identification of DNA modifications enabled by genome-guided nanopore signal processing. Bioinformatics, (2016-12-15) [2023-03-01]. DOI: 10.1101/094672.

TAO J F, ZHOU J Z, XIE T, et al., 2017. Influence of chromatin 3D organization on structural variations of the *Arabidopsis thaliana* genome. Mol. Plant, 10(2): 340-344.

TILAK M K, BOTERO-CASTRO F, GALTIER N, et al., 2018. Illumina library preparation for sequencing the GC-rich fraction of heterogeneous genomic DNA. Genome Biol. Evol., 10(2): 616-622.

ULAHANNAN N, PENDLETON M, DESHPANDE A, et al., 2019. Nanopore sequencing of DNA concatemers reveals higher-order features of chromatin structure. Genomics, (2019-11-07) [2023-03-01]. DOI: 10.1101/833590.

VAN DIJK E L, JASZCZYSZYN Y, NAQUIN D, et al., 2018. The third revolution in sequencing technology. Trends Genet., 34(9): 666-681.

WANG M J, WANG P C, LIN M, et al., 2018. Evolutionary dynamics of 3D genome architecture following polyploidization in cotton. Nat. Plants, 4(2): 90-97.

WANG P J, CHEN X J, GUO Y C, et al., 2019. Identification of CBF transcription factors in tea plants and a survey of potential CBF target genes under low temperature. Int. J. Mol. Sci., 20(20): 5137.

WANG Y H, ZHAO Y, BOLLAS A, et al., 2021. Nanopore sequencing technology, bioinformatics and applications. Nat. Biotechnol., 39(11): 1348-1365.

XU T, WU X W, WONG C E, et al., 2022. FIONA1-mediated $m^6A$ modification regulates the floral transition in *Arabidopsis*. Adv. Sci., 9(6): e2103628.

YAN F, XI R M, SHE R X, et al., 2021. Improved de novo chromosome-level genome assembly of the vulnerable walnut tree *Juglans mandshurica* reveals gene family evolution and possible genome basis of resistance to lesion nematode. Mol. Ecol. Resour., 21(6): 2063-2076.

YUEN Z W S, SRIVASTAVA A, DANIEL R, et al., 2021. Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. Nat. Commun., 12(1): 3438.

ZHANG H M, LANG Z B, ZHU J K, 2018. Dynamics and function of DNA methylation in plants. Nat. Rev. Mol. Cell Biol., 19(8): 489-506.

ZHANG Y L, FU J, WANG K, et al., 2022. The telomere-to-telomere gap-free genome of four rice parents reveals SV and PAV patterns in hybrid rice breeding. Plant Biotechnol. J., 20(9): 1642-1644.

ZHAO Z H, TAVOOSIDANA G, SJÖLINDER M, et al., 2006. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat. Genet., 38(11): 1341-1347.

（责任编辑　罗厚枚）