Oxford NANOPORE Technologies

全基因组测序

弥合植物基因组的测序缺口

白皮书

2022年1月

nanoporetech.com/publications

nanoporetech.com

目录

引言

根据最新估计，目前科学界已知的植物物种约有40万种。然而，其中有40%的物种因气候变化、栖息地消失和病害等因素正濒临灭绝[1]。如何攻克这些难题，是植物研究的关注重点。这在依赖少数物种的作物生产领域尤为紧迫。据报道，世界上90%的耕地仅种植了20种植物，彰显出人类面对气候变化和植物病害的潜在影响时的脆弱程度[2]。此外，预计到2050年，全球人口将增长约30%，达到98亿[3]。不断增加的人口使得作物产量压力与日俱增。

在已知的40万植物物种中，仅有900个物种（0.2%）完成了基因组测序[1,4]。

植物对食物、燃料、药物和服装至关重要，因此不难理解，植物的遗传表征正在得到格外关注。遗传表征可以保护遗传多样性，并能用于识别有利于育种的特征。

随着研究植物基因组的重要性被广泛认可，问题也由此而来：为什么已被测序的植物如此之少？迄今为止，约900个植物物种完成了基因组测序[1,4]，这远远不及于迄今已完成30多万个测序的微生物菌株[5]。如下文所述，这个问题的答案在于植物基因组的大小和复杂性，使得传统技术很难对其进行测序。

本文概述了研究人员目前是如何通过纳米孔测序技术来应对上述难题，并生成高质量、高度连续的植物基因组组装，从而为植物保护和育种带来新的机遇。

纳米孔测序在植物基因组研究中的优势

简化大型基因组的测序工作

不同植物基因组的大小迥异，从61 Mb（螺旋狸藻*[Genlisea tuberosa]*）[5]到152 Gb（重楼百合*[Paris japonica]*）[6]不等，后者几乎比人类基因组大50倍。植物基因组还呈现出多样的倍性（染色体拷贝数），从二倍体（两个拷贝：例如拟南芥*[Arabidopsis thaliana]*）到十倍体（十个拷贝：例如择捉草莓*[Fragaria iturupensis]*）不等。传统的短读长测序技术通常生成的读长长度为150-300个核苷酸，因此从本质上难以对上述大型基因组进行测序。

相对而言，基于纳米孔的测序能对所呈现的DNA片段的整个长度进行处理。数以千计kb级大小的完整片段经常规化处理，已有长度超过4 Mb的超长读长[7]。纳米孔技术生成的长读长序列简化了基因组的组装过程（**图1**）。

大多数真核生物基因组测序项目仅能生成一个结合所有染色体拷贝数据的单倍型共有序列。这是由于受传统短读长测序的性质所限，同一染色体上的基因连接工作无法经济有效地完成。而如今，纳米孔技术提供的长读长序列使得单倍型分析更为可行，甚至也适用于多倍型物种。长片段单倍分型信息可以为植物进化和驯化研究带来新的见解，并有可能促进作物生产领域的未来进步[10,11]。

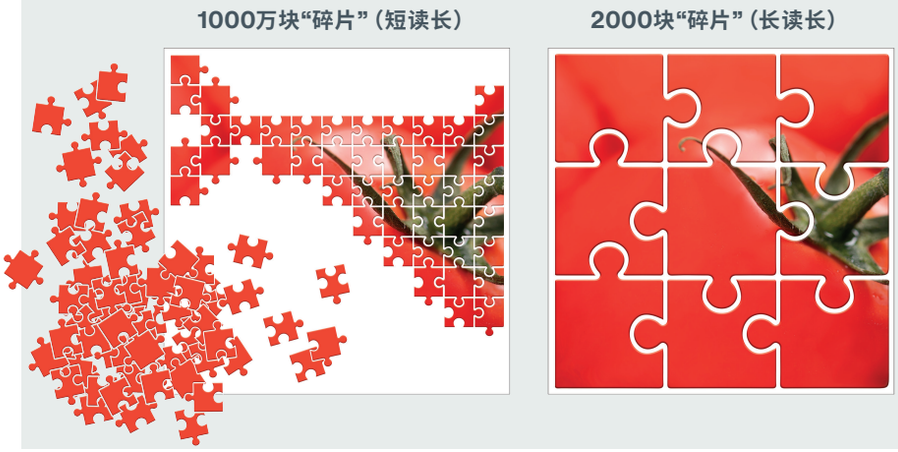
“我们证明，目前借助长读长测序技术，已实现无缺口的染色体序列端粒到端粒组装”[8]

“Oxford Nanopore能够在一周内为小型植物基因组测出具有参考组质量的基因组，效果优于耗费近20年的手工甄选结果。”[9]

纳米孔长读长测序也可与染色质构象技术结合，用于搭建scaffold和校正基因组组装。经证实，Oxford Nanopore Pore-C工作流程可显著改善组装连续性，还可为转录调控提供新见解（参见案例研究2）。

图1

像拼图组块大的拼图一样，长读长DNA比短读长DNA组装起来要容易。番茄基因组的长度大约为1 Gb，相当于1000万个100 bp的短读长或2000个500 kb的长读长。使用短读长测序时，高比例（60%）的重复DNA使得组装更复杂。



弥合缺口：解析重复区域和结构变异

重复DNA序列是大多数真核生物基因组的最大组成部分，这一点在植物中最为明显[12]，一些物种（如玉米）显示出超过80%的重复DNA[13]（**图2**）。重复序列可以分为三大类：转座子（TE）、串联重复和高拷贝数基因，其中转座子是最大的组成部分[14]。

这些重复DNA既可以复制又可以插入新的染色体位点，在基因组进化中发挥着重要作用[15]。因此，如果能进一步准确定位和表征这些转座子，将为植物进化和适应带来新的启示，对基因操作具有重要的实用价值。

> ***“我们证实，生成超长读长为复杂区域的组装带来重大突破。包含多个转座子的复杂区域常见于植物基因组中”[17]***

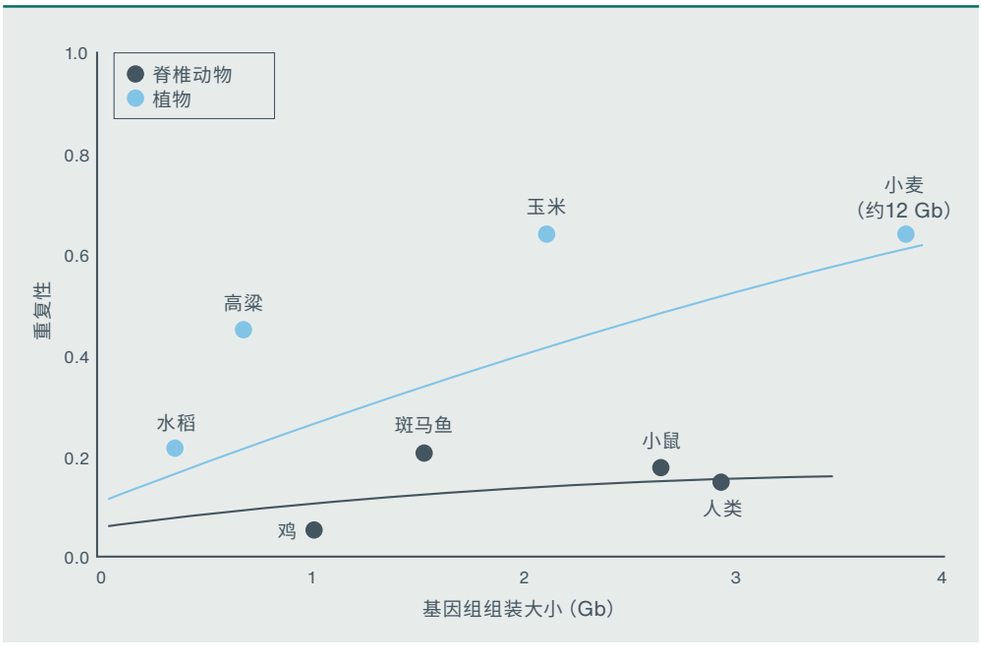
使用短读长测序方法时，重复DNA对基因组组装造成了巨大困难，在该方法下，测序读长不能完整识别所有重复区域。在大片段结构变异区域中也观察到类似情况：仅使用短读长测序技术无法顺利或经济有效地将其解析出来。以上因素导致，在现有的采用短读长技术创建的基因组组装中，大多数呈现出大量与重复区域和结构变异有关的缺口。

相对而言，纳米孔测序提供的长读长能够识别这些区域，从而得到更为连续的基因组组装[18,19]。因此，研究人员现在正在使用纳米孔技术对参考基因组进行*de novo*（从头）测序，并在某些情况下，对已使用短读长方法创建的现有基因组组装进行校正[19]。

J·克雷格·文特尔研究所（J. Craig Venter Institute）的一个团队使用一个1000美元的MinION™测序芯片，在短短四天内为生物模型拟南芥创建了高度连续的基因组组装[19]。不仅如此，该组装的连续性比现有的“黄金标准”参考基因组更高，而碱基准确度却毫不逊色。首席研究员Todd Michael教授表示：“［长读长纳米孔组装］*揭示了我们用短读长技术看不到的微小变异*”[19]。

图2

植物中可能存在包含高比例重复DNA的超大基因组。因此，使用传统短读长测序技术时，很难确保基因组组装的准确性。图片来自Jiao和Schneeberger[16]并经过调整。



从实验室到现场：按需使用且可扩展的测序技术

跨国界共享种质（可以种植新植物的活体组织）对培育增强型植物株系至关重要。然而，运输此类材料的监管要求日趋复杂，正阻碍着该领域的发展脚步[20]。资源有限的环境可能受到的影响最大，他们已减少在资本密集型基因组技术方面的投入，而这些技术是加强植物育种或保护工作所需要的。

传统测序平台需要大量资本性投入（>5万–100万美元）[21]、重要的基础设施工作和专业工程师校准。与传统测序平台不同，Oxford Nanopore公司的MinION启动包仅售1000美元起（包括测序试剂），由笔记本电脑上的USB端口供电。这款特别的便携式、经济型MinION和MinION Mk1C（包括一体化触摸屏和强大的集成计算功能，可实时分析数据）可为任何研究人员在实验室和现场环境中提供能够进行长读长测序的实时测序技术（**图3**）。亚琛工业大学（RWTH Aachen University）的Björn Usadel教授表示："[MinION] *意味着小实验室现在也可以进行测序并组装基因组*"[18]。J·克雷格·文特尔研究所的Todd Michael教授也表达了同样的观点，他评论道："*研究人员不再需要将样本送到核心实验室或服务提供商。他们现在可以自行操作，并且在一周之内得到问题的答案。他们应该这样做*"[22]。

台式GridION™和PromethION™设备拥有更高的输出量和通量，使用户能以经济高效的方式扩大研究规模，以满足极大或多个大型基因组方面的测序需求，例如对种子库和转基因植物株系进行表征（**图3**）。以上设备都无需动用资本性支出，并且每碱基成本与传统测序平台相当。此外，测序芯片可独立使用。该功能使得多个项目可以同时运行，从而获得高效的按需测序结果。

Oxford Nanopore公司还提供Flongle™，它是MinION和GridION设备的测序芯片适配器，专门为规模较小、频次较高的检测和实验提供更具成本效益的分析（**图3**）。可能的应用方向包括在启动大型基因组项目之前，对测序性能进行快速、低成本的优化和质量检查。

图3

Oxford Nanopore测序平台（从左上到右下）：Flongle，一种在MinION和GridION上使用的测序芯片适配器；便携式MinION和MinION Mk1C；GridION，可容纳5张Flongle或MinION测序芯片；PromethION P2和P2 Solo，能够运行两张独立的高容量测序芯片*；高通量PromethION（P24或P48）平台，可容纳多达24（P24）或48（P48）张高产量测序芯片。

* PromethION P2和P2 Solo设备目前可供预订，提早预订的设备预计于2022年发货。

植物相关问题

众所周知，由于植物细胞中的次级代谢物（如多糖和多酚）含量很高，很难从中提纯高质量的高分子量（HMW）DNA。这种污染物会对后续分析工作造成很大影响，也是测序产出低的最常见原因之一。荷兰KeyGene的科学家Alexander Wittenberg博士指出："*没有单一的实验方案能涵盖所有植物物种、组织或条件*"[23]。

因此，经验丰富的植物研究人员通常建议花时间仔细优化样本制备方法。不仅针对每个物种，而且针对给定物种的每种组织类型。与该做法类似地，德国亚琛工业大学的Anthony Bolger博士建议研究人员"*每次运行时仅小幅调整一个参数……不要频繁更换参数，不要添加超过50%的物质，只应进行微调*"[24]。

用不同的制备参数对多个测试样本进行测序会迅速增加实验成本，因此我们推出了Flongle，旨在解决这一潜在问题（参见上一节内容）。

RNA测序、转录组学、基因注释

RNA测序是基因发现和转录本表达分析的有力工具。如前所述，大多数植物基因组尚未得到测序，已组装的基因组通常不完整，可能缺失许多重要基因。而RNA测序提供了一种有效的方法，用于识别给定生物体中的所有基因，以便进行基因表达分析并支持基因组注释。

传统的RNA测序方法通常使用50–100个核苷酸的短读长序列，并必须使用复杂的计算技术进行组装。但是，Steijger等人[25]的一项研究显示，在所分析的超过半数的转录本中，自动转录本组装方法无法识别所构成的所有外显子。而且在已测出所有外显子的转录本中，超过半数的组装是错误的[25]。

在纳米孔测序下，读长长度等同于RNA（或DNA）片段长度，可明确分析全长转录本，使得基因注释更简便，并能够对异构体层面的基因表达进行准确表征和定量（**图4**）。使用纳米孔测序已经生成了长度超过20 kb的全长转录本[26]。

Oxford Nanopore公司不仅提供聚合酶链式反应（PCR）和无PCR的全长cDNA测序试剂盒，还提供RNA直接测序试剂盒。该试剂盒可以对无扩增RNA进行测序。这种独特的技术消除了两种潜在的偏差因素，即PCR和反转录，更重要的是，它还可以保留RNA碱基修饰，并在之后直接识别。

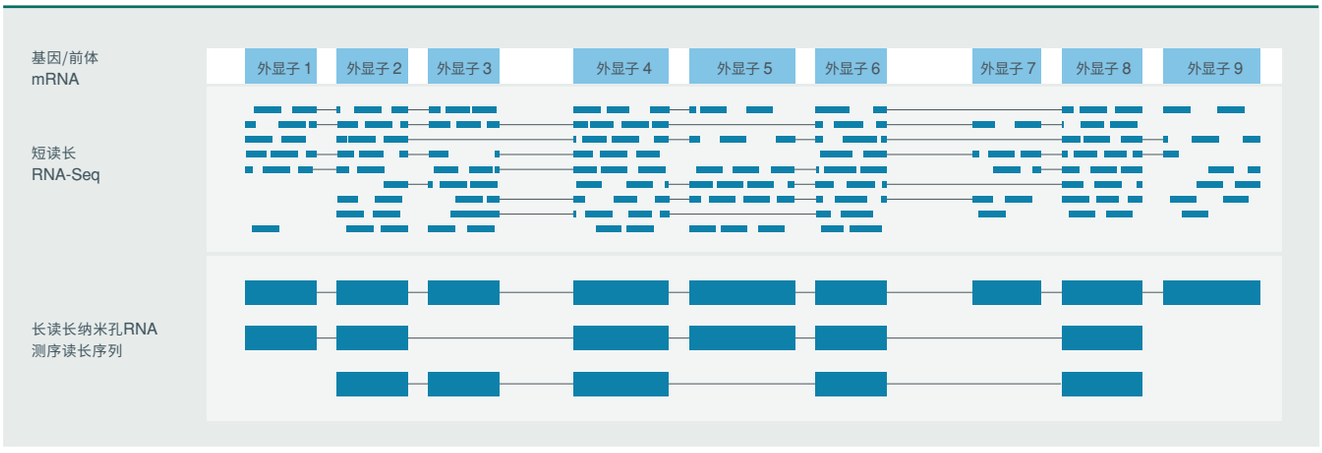


图4

选择性剪接会导致每个基因生成大量的mRNA异构体，进而改变蛋白质的组成和功能。传统RNA测序技术生成的短读长会丢失定位信息，增加了正确组装选择性mRNA异构体的难度。长纳米孔读长可以覆盖全长转录本，简化其鉴定工作。

直接分析的可能性

20世纪50年代，Barbara McClintock首次注意到植物中DNA甲基化的作用。她描述了一种可以在活跃和非活跃形式之间切换的新型可移动元素，称之为“相位变化”[27]。现在已知，植物中的DNA甲基化发生在CG、CHG和CHH（其中H可以是A、C或T）环境中[28]。在真核生物中已报道了200多种不同的RNA修饰。虽然大部分对DNA和RNA修饰的研究都是针对动物进行的，但现在植物也受到了越来越多的关注。近期研究正在探索N^6-甲基腺苷（m6A）、5-甲基胞嘧啶（m^5C）和尿苷化等修饰在植物发育中的作用[29,30]。

甲基化碱基在基因表达中发挥了关键作用，因此它在农业研究和植物育种研究中也越来越重要。例如，在对应激源（例如盐度、水短缺和农药）的反应中，已显示出差异甲基化。与传统测序技术不同，基于纳米孔的测序可以同时并直接检测DNA或RNA甲基化、其他碱基类似物和核苷酸序列，进一步提高了基因组表征水平及其详细程度。

“我们利用纳米孔技术的所有优势，通过直接RNA法检测修饰碱基和测序转录组学数据，以对基因组进行注释……”[16]

案例研究

案例研究1

解析拟南芥基因组：从着丝粒到T-DNA插入

近期对拟南芥（*Arabidopsis thaliana*）进行的多项长读长测序研究显示，其生成的丰富数据有助于开发高度连续的基因组组装和专业分析工具，从而能够解析着丝粒等高难度基因组区域。

在剑桥大学的Ian Henderson及其团队的努力下，纳米孔技术已成功应用于拟南芥基因组中的“黑洞”——着丝粒的相关难题[31]。

着丝粒在所有真核细胞中普遍存在，可能蕴含许多科学问题的答案。然而，由于它们具有高度重复性，自20多年前首次测序拟南芥基因组以来，仍有很多关于着丝粒的谜团尚未揭开。着丝粒存在于每条染色体中，它们的作用是在细胞分裂期间将一系列蛋白质（动粒）组装起来，这样染色体就可以被拖到分裂细胞的相反极上。着丝粒是一种高度保守功能的核心，但它出乎意料地成为基因组中进化最快且最具多样化的部分之一。虽然DNA进化迅速，但遗传信息却保持稳定。

将哥伦比亚型拟南芥（Col-CEN）作为模式物种时，使用R9和R10测序芯片生成的高深度数据集产生了131 Mbp的新基因组组装。这些数据集与之前被视为金标准的拟南芥参考基因组TAIR10高度一致。研究人员将所有着丝粒组装成一个重叠群（contig），并且对全部五个着丝粒进行了解析。

使用nanopolish和DeepSignal-plant分析Oxford Nanopore测序数据后，发现了关于着丝粒结构的新信息：每个着丝粒的私有卫星文库、致密的甲基化Mb级串联重复卫星阵列，以及Athila反转录转座子的侵入。丰富的数据将有助于研究重组途径。

剑桥的该团队还将染色质免疫沉淀技术（ChiP）与纳米孔测序相结合，用于研究表观基因组：他们对免疫沉淀DNA蛋白复合物中的DNA进行测序，以揭示表观遗传修饰信息如何在着丝粒的不同部分中保留。研究人员称，这“*凸显了那些有可能采用纳米孔技术的创新方法*”[32]。

植物遗传学方面的另一个长期研究挑战是，通过随机插入来自根癌农杆菌（*Agrobacterium tumefaciens*，一种植物病原体）的转移DNA（T-DNA）而创建出大量基因工程植物，并对其进行表征。目前已创建超过70万条T-DNA拟南芥插入株系，可用于研究单个基因的功能。然而，传统表征方法无法完成这项挑战。T-DNA株系的选择原理很简单，但同样来自剑桥大学的Boas Pucker称，“*现实要复杂得多*”[33]。

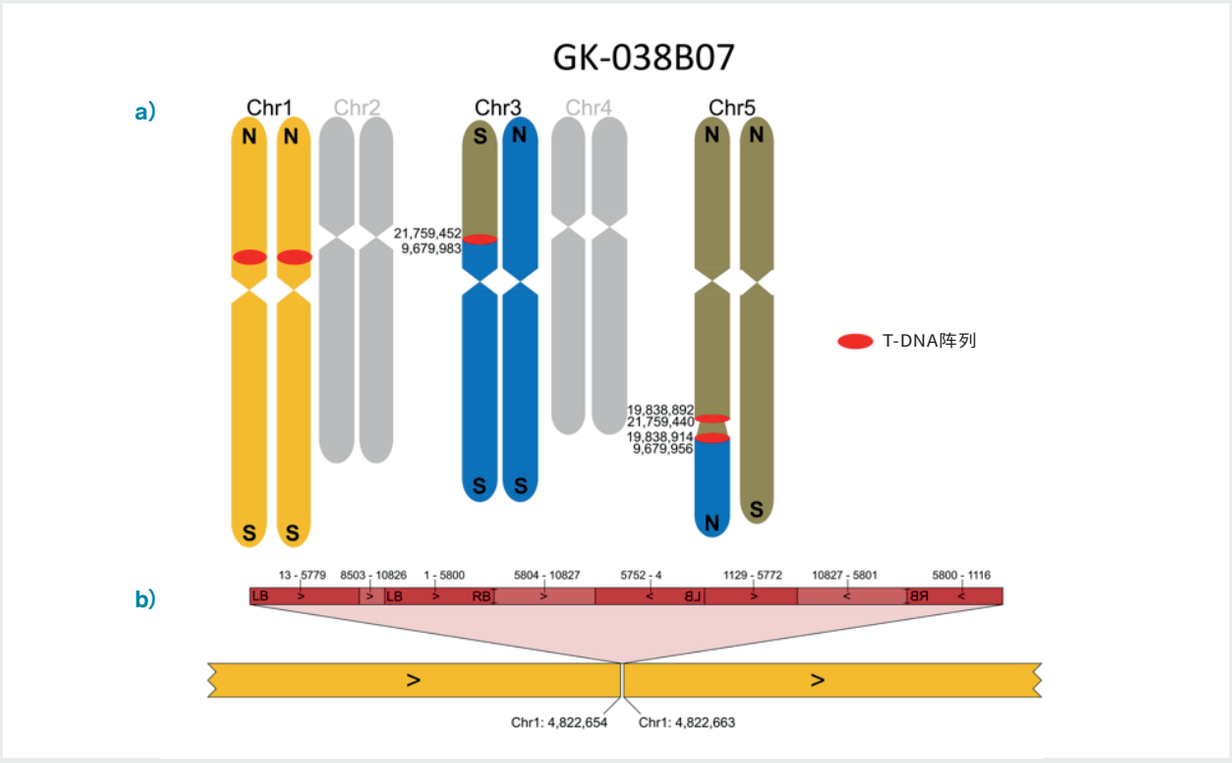


图5

使用GridION进行的纳米孔测序解析了转基因拟南芥株系GK-038B07中的易位、倒位和T-DNA插入。(a) 染色体示意图显示，3和5号染色体相互融合，并且在融合位点处的两个T-DNA阵列之间有2 Mbp倒位。数字代表根据TAIR9假染色体片段的终点。以灰色显示的染色体与Col-0参考序列匹配。(b) 通过局部组装解析的T-DNA插入的可视化示例。LB和RB代表T-DNA的左右边界；暗红色处代表T-DNA序列；浅红色处代表来自二元载体骨架的部分。图片来自Pucker, B., Kleinbölting, N., and Weisshaar, B. (2021)[34]并经过调整，可在知识共享许可协议下获取 (creativecommons.org/licenses/by/4.0) 。

Boas及其团队在表征复杂T-DNA插入株系时，展示了*“长读长测序的强大功能”*[34]。该技术不仅具有时效性（使用GridION仅需几小时即可提取数据并测序，产生的N50读长长度约为41 kb），也具有性价比（成本可控制在*“每个株系约200-300美元”*）。*“这种方法比PCR或短读长测序方法更快、更省力、更全面，而且……便宜得多”*，能够检测和纠正*“许多无着丝粒的组装错误”*。

该团队创建了loreta，这是一种基于长读长的T-DNA分析工具，可在GitHub上获取，专用于处理他们生成的海量数据。该工具能够解析复杂的T-DNA插入，并且鉴定先前未知的插入。此外，该工具易于研究真正的插入数，而传统方法仅用于表征一个插入：Boas鉴定出插入一个阵列的5.8 kb T-DNA的三个拷贝，从而将先前估计的每条株系约1.5个插入视为低估。该技术解释了染色体融合、易位、重复和倒位复制（**图5**）。值得注意的是，一些染色体重排不涉及T-DNA。该技术还可以表征与结构变异（SV）和整个基因组的重复相关的插入。丰富的数据激励了Boas及其团队，他们使用Canu创建了新的拟南芥基因组从头组装，然后使用Racon和Medaka进行矫正。该组装可在NCBI数据库[35]上找到。

案例研究2

使用Pore-C解锁植物3D基因组架构

一般认为，细胞核内的DNA空间组织会显著影响细胞行为，例如转录、复制、DNA修复、发育和疾病[36]。基因组具有特定结构，因此多个空间上不同的控制元件能够以物理形式存在于某个基因旁边，并对其造成影响。

过去十年来，科学家们在确定基因组的3D scaffold方面取得了巨大进展，但我们仍然不甚了解染色质结构的基础机制，以及它如何影响基因调节、细胞命运决定甚至是演化。染色质构象捕获（3C）技术已用于捕获基因组的物理拓扑结构，典型的下游处理包括扩增和短读长测序。然而，这些技术仅鉴定了基因组位点之间成对的相互作用。

Oxford Nanopore开发了Pore-C，这是一种端到端工作流程，可将染色质构象捕获方法与纳米孔测序技术的益处相结合（**图6**）：直接测序生成无扩增的长读长和超长读长，由此从多个基因组位点中获得接触点数据。使用Pore-C长读长，可以重建覆盖Mb级碱基的复杂重排，甚至涉及多条染色体[37]。此外，这项技术是无PCR的，可以保留甲基化等DNA修饰信息，并且能从纳米孔测序数据中直接获得该信息，无需追加测序运行。Pore-C测序还有一个优势，即能够进入重复区域和GC富集区域，有助于组装这些区域。

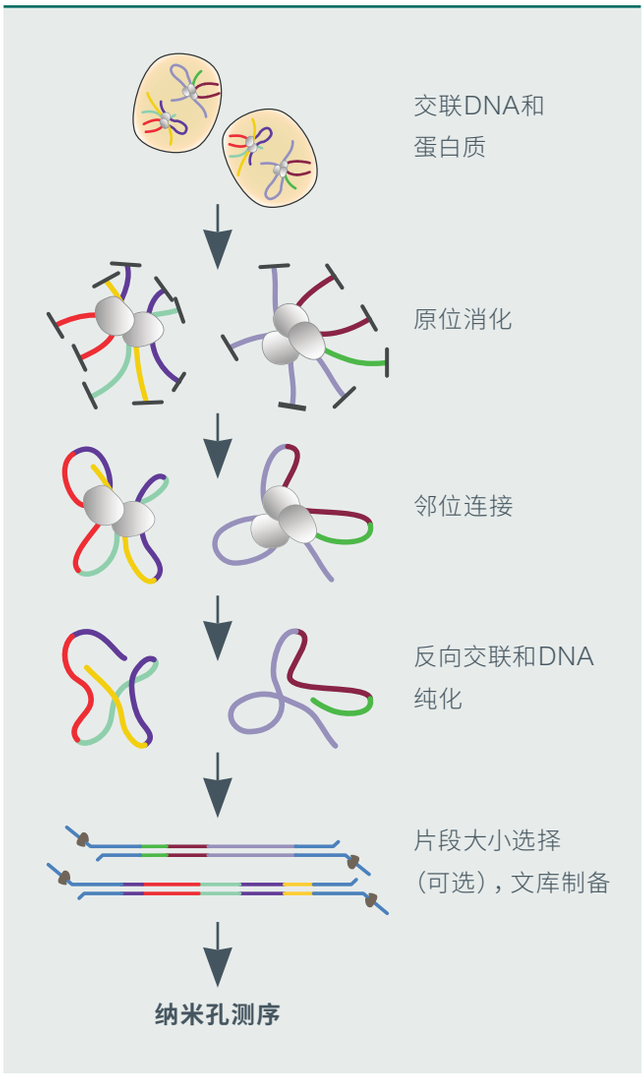


图6

Pore-C测序的实验室工作流程。使用甲醛处理，交联位于细胞核内的邻位DNA和蛋白质。然后，将限制性内切酶和邻位连接加入交联片段，以便分析其空间结构。图片摘自Oxford Nanopore Pore-C海报[38]。

	Pore-C	Hi-C （Dong等人，2018年）[39]
读长序列	104,539,406	2,227,694,973
测序的碱基对	121,866,117,653	445,538,994,600
过滤后的接触点	136,980,804	638,741,422

表1

在Pore-C测序研究和先前采用另一种测序方法的研究之间，比较读长序列数量和后续接触点数据。

不满足于传统短读长技术所能获得的解析度，美国纽约大学的Jae Young Choi及其团队使用Pore-C测序研究了水稻（*Oryza sativa*）中染色体的三维结构[36]。尽管水稻在经济文化方面具有重要意义，但目前仍然缺乏高质量的参考基因组，并且其进化史尚未完全揭晓。Jae认为，Pore-C的优势将为这一挑战带来最好的结果。利用PromethION的高产量，他们生成了约1亿个读长序列，并获得了约1亿个接触点。相较而言，一项先前研究使用的另一种测序技术需要超过20亿个读长序列才获得约6亿个接触点[39]，而使用Pore-C测序时，约二十分之一的读长序列数量就能获得约五分之一的接触数据（**表1**）。“*这是Pore-C的优势之一：用较少的读长序列获得更多的接触点*”[36]。由于使用纳米孔技术对整个连接产物进行测序，因此能检测到多个接触点。

在染色体规模的Pore-C可视化数据上，所显示的“*成束*”接触点代表拓扑相关结构域（TAD）：已发现3261个TAD，中位长度为80 kbp。该团队的下一步目标是，确定接触点数据是否与转录活性增强有关，还是仅为噪声。使用纳米孔读长序列的数据，可以很容易地进行功能性分析。通过纳米孔碱基识别技术，可以鉴定通常与基因活性相关的DNA修饰。他们通过分析，发现了与表观遗传修饰标记的缺陷和富集相关的TAD，说明“*Pore-C确实有检测生物学相关接触活性的功能*”。

由于Pore-C读长序列也可用于搭建scaffold和解析基因组组装，Jae及其团队使用该技术解析了铁心木（*Metrosideros polymorpha* var. *incana*）基因组的染色体结构，用于其生态式物种形成研究。他们使用GridION生成了一个完整的染色体级基因组组装，并获得接近300Mbp的序列（读长长度N50为25.9 Mbp）和一千出头的contig，最长长度达8.27 Mbp（N50为1.85 Mbp），对应于66x覆盖深度。放置上述所有序列仅需11个scaffold，对应铁心木（*Metrosideros*）的11条染色体[40]。

对包括3D scaffold在内的基因组进行深入了解后，将进一步了解到非编码区域如何通过改变顺式调节序列（如增强子、沉默子和绝缘子）来影响细胞行为。Pore-C测序使用少于其他技术的序列，即可检测出更多此类接触点，并且可以在千碱基对规模下鉴定接触点。

请访问nanoporetech.com/resource-centre/porec查看我们的Pore-C海报。

案例研究3

运用纳米孔测序的最新技术组装作物基因组

生成高质量、连续的植物参考基因组非常重要。然而，由于植物基因组的大片段、序列高度重复和多倍体性质，很难使用传统短读长测序技术进行测序和组装。纳米孔长和超长读长测序可增强植物基因组组装，解析结构变异、重复区域和转座子。荷兰KeyGene公司的Alexander Wittenberg及其团队计划开发一个易于使用、集成式的植物分析软件包，从而以“*前所未有的规模*”进行泛基因组分析[41]。

构建参考基因组时需要考虑四个重要方面：“*准确性、完整性、连续性和成本*”。Alexander发现，与其他长读长技术相比，纳米孔长读长具有“*大量*”优势，包括：每张测序芯片的产量更高，读长长度更长（超过50 kb，而其他技术约为15-20 kb），并行、灵活地生成数据，“*文库制备简单直接*”且输入量更低，以及“*更容易检测碱基修饰*”。

一直以来，研究人员会对长读长序列（尤其是在植物物种中）进行矫正，以提高共识准确度。为了规避矫正步骤，Alexander及其团队利用最新的Oxford Nanopore“Q20+”测序化学试剂，结合R10.3和R10.4测序芯片和他们内部开发的组装工具KeyGene STL，开发了一种工作流程。从不同作物物种中提取DNA，并评估原始和共识准确度。Alexander及其团队注意到“*不同植物物种在复杂性、组成和对人类和微生物物种的碱基修饰方面存在差异*”[41]，因此，他们与Oxford Nanopore合作，基于玉米基因组开发了一种用植物训练的碱基识别模型。经过训练的模型使玉米中的原始读长准确度提高了2.5%。在其他植物物种中使用此模型时，准确度也有所提高（**表2**）。

物种	原始读长准确度 (%)		提高率 (%)
	标准模型	经过训练的模型	
玉米	96.9	99.4	2.5
生菜	98.4	99.0	0.6
甜瓜	98.6	98.9	0.3

表2

使用标准Q20+碱基识别模型和经过植物训练的模型（用玉米开发而成），在三种不同植物物种中达到的碱基识别准确度百分比。

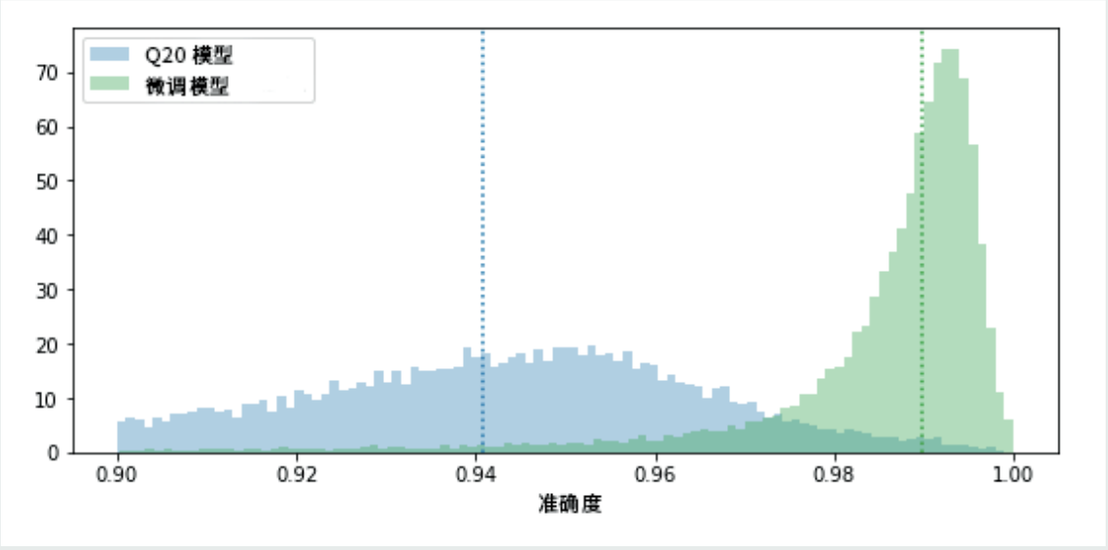


图7

使用标准Q20+碱基识别模型（蓝色）和经过植物训练的碱基识别模型（绿色）在玉米中获得的原始读长准确度。

研究人员还开发了其他重要作物的参考基因组：仅使用四张PromethION测序芯片对生菜进行测序，每张测序芯片产生约80 Gb的高覆盖深度数据集，平均读长长度为38 kb（N50为50.4 kb）。在30小时内总共组装了2.6 Gb的基因组组装，仅包含159个contig[41]。“*对于这个惊人的结果，极长的纳米孔读长功不可没*”[42]。与另一个长读长组装相比，他们的纳米孔测序工作流程产生了4倍高的连续性、非常高的共线性，并且大大提高了共识准确度。“*从目前公开可获得的数据来看，KeyGene STL ONT组装的连续性最高*”[41]。

KeyGene的团队还将其工作流程应用于甜瓜基因组，与其他测序技术相比，该流程同样显著降低了contig数量，说明其具有高度完整性。该团队还利用纳米孔测序的独特能力，生成了“世界首个”[41]仅有双链的组装。新型Oxford Nanopore Q20+化学试剂实现了双链测序，即对同一双链DNA的两个互补分子进行连续测序，以获得更高的共识准确度。在该团队的甜瓜数据集中，约10%包含双链读长序列。将35x基因组覆盖深度的双链数据与另一个长读长技术的35x深度数据比较，发现Oxford Nanopore数据的共识准确度“*显著更高*”。总结上述结果，Alexander评论道“*双链读长序列真令人惊喜*”[41]。

> ***“我们认为这确实是一项技术突破。值得强调的是，更长、更准确的Oxford Nanopore读长序列使之成为可能”[41]***

目前，虽然KeyGene公司的STL组装工具优于最新的公共工具，但他们打算微调计算工具以提高准确性，并使用多种植物物种来改进碱基识别模型（与Oxford Nanopore合作），同时增加双链读长序列的数量和长度。

从Nanopore社区中获取有关双链测序和植物训练碱基识别模型的最新信息：community.nanoporetech.com。

案例研究4

结构变异在作物育种中的重要性

甘蓝型油菜（*Brassica napus*）是世界范围内的主要油料作物，在烹饪、生物燃料和动物饲料中广泛使用。1.2 Gb甘蓝型油菜基因组是异源四倍体，其中一组染色体来自甘蓝（*B. oleracea*）（例如卷心菜，亚基因组C），另一组来自芜菁（*B. rapa*）（例如大头菜，亚基因组A）。这些基因组揭示了该生物体的进化史。

甘蓝型油菜基因组显示了广泛的基因层和染色体层的结构变异（SV），它们潜藏在重要的表型特征下，例如开花时间、抗病性和种子质量。准确解析这些结构变异有助于改良这些重要的经济作物。

短读长测序技术已被用于描述许多结构变异。然而，由于基因组的四倍体和短读长测序倾向于定位到一个以上的位点，使得亚基因组层的畸变解析极其困难。

为了更准确地解析甘蓝型油菜亚基因组层的结构变异，尤斯图斯·李比希大学（Justus Liebig University）的研究人员利用纳米孔读长序列来进行研究。由于其长度较长，大大降低了比对到多个位置（即“多重映射”）的可能性[43,44]。

该团队对取自世界各地的四种不同的甘蓝型油菜株系进行了测序，包括北美（N99，春季开花型）、中国（PAK85912，半冬季开花型）和欧洲（Express 617和R53，冬季开花型）。为了确保准确描述大片段的结构变异，研究人员执行了片段大小选择步骤，以减少长度小于40 kb的DNA片段。在最近的测序中，还利用核酸酶清洗来尽可能提高孔的可用性，从而提高测序产出。这些测序在单个MinION测序芯片上产出了超过30 Gb的数据，其中N50序列大约有40 kb。

> ***“[我们]识别到了插入，这在短读长技术下几乎不可能检测到。”[43]***

在使用Sniffles[46]算法识别结构变异前，使用NGMLR[45]将得到的测序数据与参考序列进行比对。

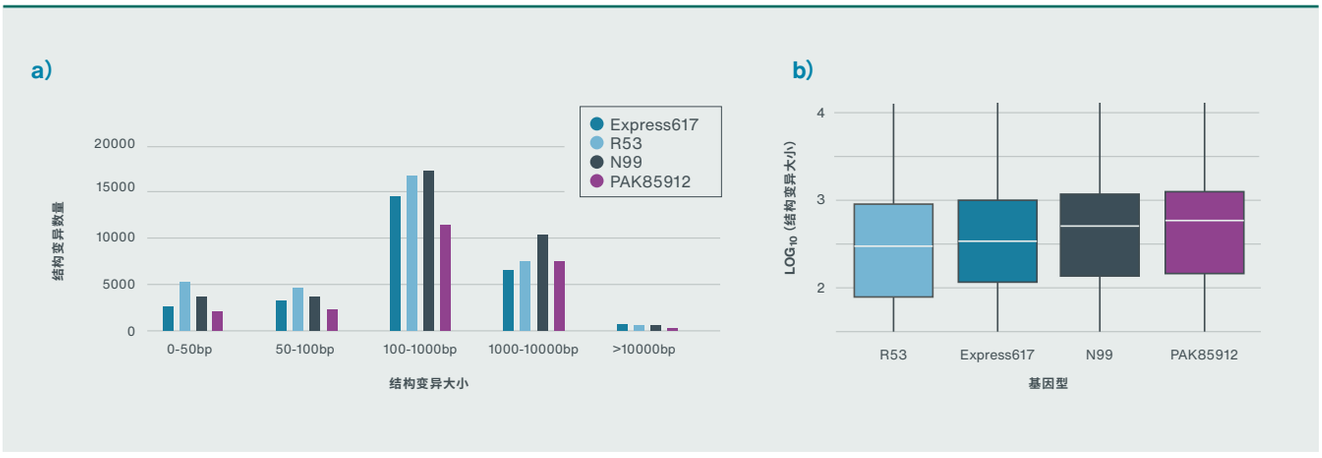


图8

在所有甘蓝型油菜株系中检测到的大多数结构变异的长度在100 bp到1000 bp之间 (a) 。总体而言，春季开花株系N99和PAK85912中包含较大的结构变异 (b) 。图片由德国尤斯图斯·李比希大学的Harmeet Singh Chawla提供[43]。

该团队观察到，所有植物株系中的大多数结构变异长度在100-1000 bp之间，首席研究员Harmeet Singh Chawla评论道，在短读长测序技术下"*几乎不可能*"检测到这种结构变异 (**图8a**) [39]。同样明显的是，与冬季开花基因型 (R53和Express 617) 相比，春季开花基因型 (N99和PAK85912) 中检测到的结构变异片段更大 (**图8b**) 。

值得注意的是，已发现5–8%的基因受到结构变异影响，并且与A亚基因组相比，C亚基因组中观察到的结构变异多样性较低。Harmeet认为，这可能反映了作物的育种历史，因为甘蓝 (C亚基因组) 的许多特征是经人工培育形成的，而相对来说，大头菜 (A-亚基因组) 几乎从未经过人工干预。

***"除了单核苷酸多态性 (SNP)，还有更多方式可以解释油菜籽中观察到的表型"*[43]**

通过检查已知与地理适应性有关的基因，该团队观察到了许多结构变异，包括在*BnVIN3* (一种与开花时间相关的基因) 中，识别出90 bp的插入。值得注意的是，该插入只在两个冬季开花品系之一中发现，即Express 617。

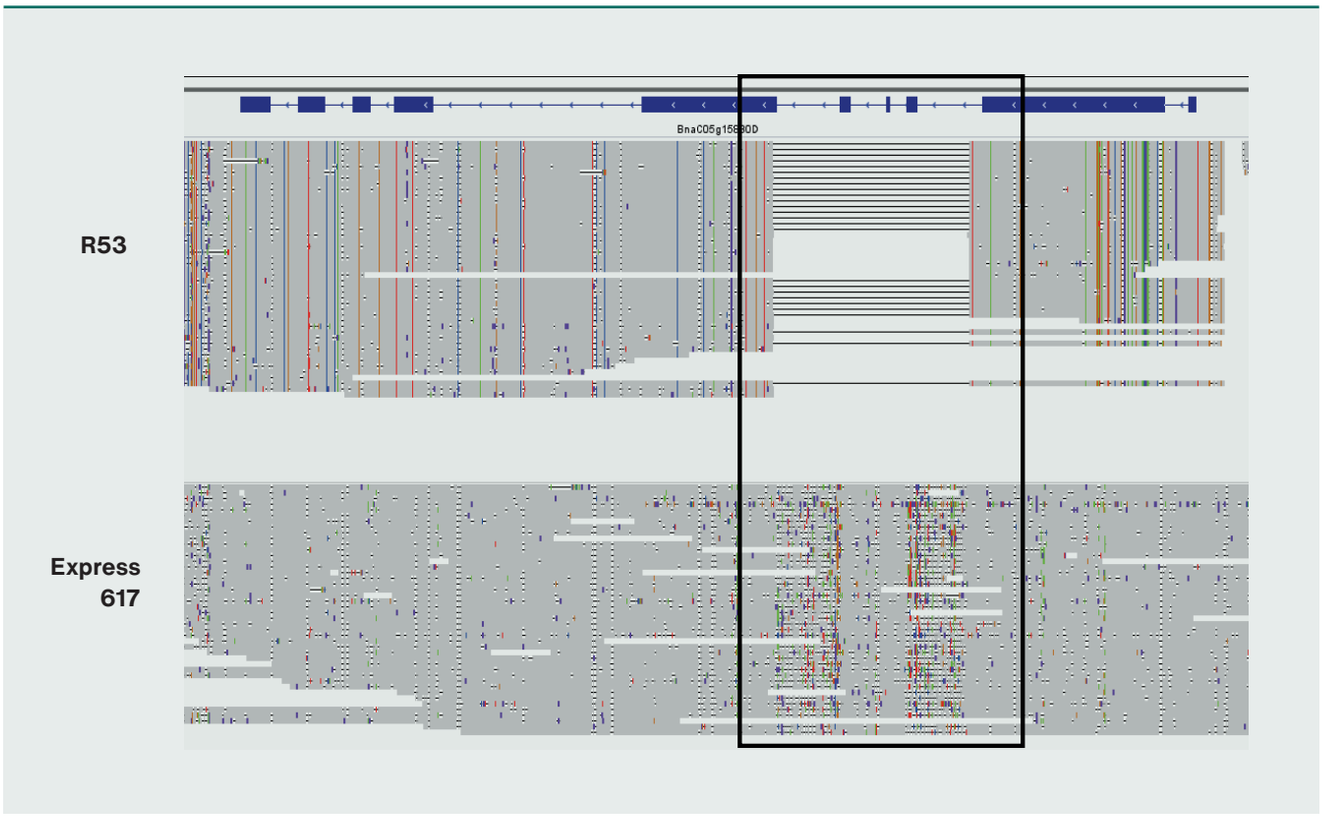


图9

在R53中观察到，纳米孔长读长测序能够识别出4-香豆酸:辅酶A连接酶基因中725 bp的缺失，但未在Express 617冬季开花型中观察到。图片由德国尤斯图斯·李比希大学的Harmeet Singh Chawla提供[43]。

研究还识别出与疾病抗性相关的结构变异，包括在R53株系的4-香豆酸:辅酶A连接酶基因中的725 bp的缺失（图9）。Harmeet认为，这一变异解释了20%的轮枝菌（*Verticillium*，甘蓝型油菜的主要真菌病原体）抗性表型[43]。

总之，该研究的研究人员认为，“*目前，长读长测序技术的成本更低、读长准确度大幅提高、平均读长长度显著增加。因此，该技术不仅可以对复杂植物基因组进行准确组装，对全基因组重测序而言，也越来越成为一个可行的选择*”[44]。

有关使用纳米孔长读长测序识别结构变异的更多信息，请访问：nanoporetech.com/structural-variation。

总结

2000年，拟南芥成为首个接受基因组测序和组装的植物。这一里程碑式的成就花了10年时间才完成，耗资约1亿美元。如今，使用单个MinION测序芯片就可在几天内实现这一壮举，成本仅需1000美元[17]。

纳米孔技术能够生成长读长序列和超长读长序列，与传统的短读长方法相比具有许多优势，它使研究人员能够组装和探索先前难以处理的植物基因组和转录组，从而为植物进化和育种策略带来新的重要见解。这些进展至关重要，它能经济有效地应对全世界一些最紧迫的难题，例如，如何在面临各类影响农业生产的情况下（气候变化、抗虫性、依赖于物种有限的基因同质作物），稳妥解决快速增长的人口的粮食问题。

植物研究的一个重要要求是能够表征和获取大型植物库（如种子库和转基因株系）中保存的遗传变异。现在，灵活的高通量PromethION平台已能满足这一需求。该平台能在72小时内产出高达14 Tb的数据*，使得研究人员能够完整、快速地对大量植物基因组进行表征。

* 系统以420 bp/秒的速度运行72小时的理论最大产出量。产出量可能因文库类型、化学试剂、运行条件等因素而异。

关于Oxford Nanopore Technologies

Oxford Nanopore Technologies推出了世界上首款纳米孔DNA测序仪——MinION，这是一种便携、实时、低成本的长读长设备。公司随后还推出了较大尺寸的GridION Mk1以及PromethION系列设备。

对于较小规模的分析，例如在大型基因组测序实验之前优化样本参数，Flongle（MinION和GridION平台的测序芯片适配器）是一种非常经济适用的选择。

通过使用长读长，可以分析重复DNA区域、大片段的结构变异，以及对全长RNA异构体进行表征。与传统短读长测序技术相比，纳米孔测序技术可提供更完整的基因组和转录组[19]。

公司现有多种测序平台可供选择，适用于不同大小的基因组以及更高的覆盖度和通量要求（**表3**）。

	Flongle	MinION和MinION Mk1C	GridION Mk1	PromethION P2和P2 Solo†	PromethION P24/P48
读长长度	片段长度 = 读长长度。当前最长读长>4 Mb[7]				
运行时间	1分钟 - 16小时	1分钟 - 72小时	1分钟 - 72小时	1分钟 - 72小时	1分钟 - 72小时
每台设备的测序芯片数量	1	1	5	2	24/48
每张测序芯片的DNA测序产出*	高达2.8 Gb	高达50 Gb	高达50 Gb	高达290 Gb	高达290 Gb
每台设备的DNA测序产出*	高达2.8 Gb	高达50 Gb	高达250 Gb	高达580 Gb	高达7/14 Tb
混样建库	1 - 96份样本	1 - >2000份样本	1 - >2000份样本	1 - >2000份样本	1 - >2000份样本

表3

我们提供多种纳米孔测序设备，适用于各类应用场景。
打印时数据正确。访问www.nanoporetech.com获取最新信息。

* 系统以420 bp/秒的速度运行72小时（或Flongle运行16小时）的理论最大产出量。产出量可能因文库类型、化学试剂、运行条件等因素而异。

†PromethION P2和P2 Solo设备目前可供预订，提早预订的设备预计于2022年发货。

有关使用纳米孔技术对植物基因组进行测序的最新信息，请访问nanoporetech.com/applications。

参考文献

1. Royal Botanic Gardens Kew. 2020. State of the world's plants and fungi. Available at: https://www.kew.org/science/state-of-the-worlds-plants-and-fungi [Accessed: 07 December 2021]

2. Finkers, R. Know your onion — The impact of long reads on large genomes. Presentation. Available at: https://nanoporetech.com/resource-centre/know-your-onion-impact-long-reads-large-genomes [Accessed: 07 December 2021]

3. United Nations. 2017. World population projected to reach 9.8 billion in 2050, and 11.2 billion in 2100. Available at: https://www.un.org/development/desa/en/news/population/world-population-prospects-2017.html [Accessed: 07 December 2021]

4. NCBI National Center for Biotechnology Information. Genomes information by organism. Available at: https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/ [Accessed: 07 December 2021]

5. Fleischmann, A. et al. Evolution of genome size and chromosome number in the carnivorous plant genus *Genlisea* (*Lentibulariaceae*), with a new estimate of the minimum genome size in angiosperms. Annals of Botany. 114 (8): 1651–1663 (2014).

6. Pellicer, J., Fay, M.F., and Leitch, I. J. The largest eukaryotic genome of them all? Botanical Journal of the Linnean Society. 164 (1) (2010).

7. Oxford Nanopore Technologies. Ultra-Long DNA Sequencing Kit. Available at: https://store.nanoporetech.com/ultra-long-dna-sequencing-kit.html [Accessed: 07 December 2021]

8. Belser, C. et al. Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing. Commun Biol. 4(1):1047 (2021).

9. Michael, T. The complex architecture of plant T-DNA transgene insertions. Presentation. Available at: https://nanoporetech.com/resource-centre/complex-architecture-plant-t-dna-transgene-insertions [Accessed: 07 December 2021]

10. Long, E.M. et al. Genome-wide imputation using the Practical Haplotype Graph in the heterozygous crop cassava. bioRxiv 443913 (2021).

11. Bhat, J.A. et al. Features and applications of haplotypes in crop breeding. Commun Biol. 4(1):1266 (2021).

12. Feschotte, C., Jiang, N. and Wessler, S.R. Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3(5):329-41 (2002).

13. Schnable, P.S. et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 326:1112–1115 (2009).

14. Lee, S.-I and Kim, N.-S. Transposable elements and genome size variations in plants. Genomics Inform. 12(3): 87–97 (2014).

15. Debladis, E., Llauro, C., Carpentier, M., Mirouze, M. and Panaud, O. Detection of active transposable elements in *Arabidopsis thaliana* using Oxford Nanopore Sequencing technology. BMC Genomics. 18:537 (2017).

16. Jiao, W.B and Schneeberger, K. The impact of third generation genomic technologies on plant genome assembly. Current Opinion in Plant Biology 36: 64-70 (2017).

17. Rousseau-Gueutin, M. et al. Long-read assembly of the *Brassica napus* reference genome Darmor-bzh. Gigascience. 9(12): giaa137 (2020).

18. Usadel, B. (2017) Complex tomato genomes: Easy with nanopores. Presentation. Available at: https://nanoporetech.com/index.php/talk/complex-tomato-genomes-easy-nanopores [Accessed: 07 December 2021]

19. Michael, T.P. et al. High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. Nat. Commun. 9(1):541 (2018).

20. International Food Policy Research Institute. International germplasm exchanges are getting more complicated, and here's why that matters to global food security. 2017. Available at: http://www.ifpri.org/blog/international-germplasm-exchanges-are-getting-more-complicated-and-heres-why-matters-global [Accessed: 07 December 2021]

21. Norris, A.L. et al. Nanopore sequencing detects structural variants in cancer. Cancer Biol. Ther. 17(3): 246-253 (2016).

22. Michael, T.P. (2017). Personal communication with Oxford Nanopore Technologies on 29 August 2017.

23. Wittenberg, A. PromethION sequencing of complex plant genomes. Presentation. Available at: https://nanoporetech.com/resource-centre/PromethION-sequencing-complex-plant-genomes [Accessed: 07 December 2021]

24. Bolger, A. LOGAN: Lossless graph-based analysis of NGS data sets. Presentation. Available at: https://nanoporetech.com/resource-centre/logan-lossless-graph-based-analysis-ngs-datasets [Accessed: 07 December 2021]

25. Steijger, T. et al. Assessment of transcript reconstruction methods for RNA-seq. Nat. Methods 10, 1177–1184 (2013).

26. Timp, W. & Jain, M. Direct RNA cDNA sequencing of the human transcriptome. Presentation. Available at: https://nanoporetech.com/resource-centre/direct-rna-cdna-sequencing-human-transcriptome [Accessed: 07 December 2021]

27. Ravindran, S. Barbara McClintock and the discovery of jumping genes. Proc Natl Acad Sci U S A. 109(50):20198-9 (2012).

28. He, H.-J., Chen, T. and Zhu, J.-K. Regulation and function of DNA methylation in plants and animals. Cell Res. 21(3): 442–465 (2011).

29. Gao, Y. et al. Quantitative profiling of N6-methyladenosine at single-base resolution in stem-differentiating xylem of *Populus trichocarpa* using nanopore direct RNA sequencing. Genome Biol. 22:22 (2021).

30. Ni, P. et al. Genome-wide detection of cytosine methylations in plant from nanopore data using deep learning. Nat Commun. 12(1):5976 (2021).

31. Naish, M. et al. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. Science. 374(6569) (2021).

32. Henderson, I. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. Presentation. Available at: https://nanoporetech.com/resource-centre/video/ncm21/the-genetic-and-epigenetic-landscape-of-the-arabidopsis-centromeres (2021) [Accessed: 8 December 2021]

33. Pucker, B. Effective characterization of T-DNA insertion lines through nanopore sequencing. Presentation. Available at: https://nanoporetech.com/resource-centre/video/lc21/effective-characterization-of-t-dna-insertion-lines-through-nanopore-sequencing (2021) [Accessed: 8 December 2021]

34. Pucker, B., Kleinbölting, N., and Weisshaar, B. Large scale genomic rearrangements in selected *Arabidopsis thaliana* T-DNA lines are caused by T-DNA insertion mutagenesis. BMC Genomics 22(599) (2021).

35. NCBI. Col-0_GKat-w. Genome assembly. Available at: https://www.ncbi.nlm.nih.gov/assembly/GCA_905067165.1 [Accessed: 8 December 2021]

36. Choi, J. Y. Unlocking the plant 3D genome architecture with Pore-C sequencing. Presentation. Available at: https://nanoporetech.com/resource-centre/video/ncm21/unlocking-the-plant-3d-genome-architecture-with-pore-c-sequencing [Accessed: 7 December 2021]

37. Ulahannan, N. et al. Nanopore sequencing of DNA concatemers reveals higher-order features of chromatin structure. bioRxiv 833590 (2019).

38. Oxford Nanopore Technologies. Pore-C: multi-contact, chromosome conformation capture for both genome-wide and targeted analyses. Poster. Available at: https://nanoporetech.com/resource-centre/porec. [Accessed: 17 December 2021]

39. Dong et al. Genome-wide Hi-C analysis reveals extensive hierarchical chromatin interactions in rice. Plant J. 94(6):1141-1156 (2018).

40. Choi, J. Y. et al. Ancestral polymorphisms shape the adaptive radiation of *Metrosideros* across the Hawaiian Islands. Proc. Natl. Acad. Sci. U.S.A. 118(37): e2023801118 (2021).

41. Wittenberg, A. Accuracy improvements in crop genome assembly using the Q20+ chemistry. Available at: https://nanoporetech.com/resource-centre/video/ncm21/accuracy-improvements-in-crop-genome-assembly-using-the-q20-plus-chemistry [Accessed: 14 December 2021]

42. KeyGene News. Available at: https://www.keygene.com/news-events/press-release-keygene-revolutionizes-crop-innovation-using-nanopore-sequencing-technology/ [Accessed: 14 December 2021]

43. Chawla, H.S. Long reads reveal small-scale genome structural variations in allotetraploid canola. Presentation at London Calling 2019.

44. Chawla, H.S. et al. Long-read sequencing reveals widespread intragenic structural variants in a recent allopolyploid crop plant. Plant Biotechnol J. 19(2):240-250 (2021).

45. GitHub. NGMLR. Available at: https://github.com/philres/ngmlr [Accessed: 07 December 2021]

46. GitHub. Sniffles. Available at: https://github.com/fritzsedlazeck/Sniffles [Accessed: 07 December 2021]

Oxford Nanopore Technologies
电话+44 (0)845 034 7900
电邮sales@nanoporetech.com
微信公众号nanoporetech
www.nanoporetech.com

WP_1052(CN)_V3_01Jan2022