

全长转录组测序在植物中的应用研究进展

赵陆滢, 曹绍玉, 龙云树, 张应华, 许俊强

(云南农业大学云南省滇台特色农业产业化工程研究中心, 昆明 650201)

摘要: 全长转录组测序 (Iso-Seq) 是指使用第三代测序 (TGS) 技术获得 mRNA 的全长序列。随着以单分子实时 (SMRT) 测序为代表的 TGS 技术逐渐发展成熟, 以及该技术在基因组和转录组 *de novo* 测序等方面的优势, 全长转录组测序技术已经广泛被应用于动物、植物、微生物测序研究中, 并挖掘出大量相关数据。本文综述了基于 Iso-Seq 技术的原理及其在模式植物和非模式植物研究中的应用, 期望为研究植物全长转录组相关分析提供参考。

关键词: 全长转录组测序; Iso-Seq; TGS; 单分子实时测序; 模式植物和非模式植物

Applications and Research Progresses of Full-length Transcriptome Sequencing in Plants

ZHAO Lu-yan, CAO Shao-yu, LONG Yun-shu, ZHANG Ying-hua, XU Jun-qiang

(Dian-Tai Engineering Research Center for Characteristic Agriculture Industrialization of Yunnan Province, Yunnan Agricultural University, Kunming 650201)

Abstract: Full-length transcriptome sequencing (Iso-Seq) is a technique method by deployment of the third generation sequencing (TGS) technology to obtain the full length sequence of mRNA. With the development of TGS technology which is represented by single molecule real-time (SMRT) sequencing, is becoming more and more mature, and its application advantages in genome and transcript *de novo* sequencing, full-length transcriptome sequencing technology has been widely used in animals, plants and microorganisms, and produced a large number of related data. In this paper, the principle of Iso-Seq technology and the applications of full-length transcriptome sequencing in model and non-model plants are reviewed, which aims to provide reference in analysis of full-length transcripts in plants.

Key word: full-length transcriptome sequencing; Iso-Seq; TGS; single-molecule real-time sequencing; model and non-model plants

基因组和转录组测序是生命科学领域的基础性工作^[1]。高通量测序技术正向着高通量、低成本、长读取长度的方向发展。一代^[2-3]、二代^[4-7]测序技术, 加快了高通量测序时代的到来, 也从单个基因位点的研究上升到全基因组层面的研究^[8]。然而, 人们对高通量测序技术的研究并没有停止, 因此又产生了第三代测序技术 (TGS, third-generation sequencing) 的单分子测序技术 (Single-

molecule sequencing), 其中 PacBio 公司的单分子实时 (SMRT, Single-molecule realtime) 测序技术已经商业化^[9-10]。全长转录组测序 (Iso-Seq, isoform sequencing) 技术即单分子实时测序技术^[11]。该技术具有超长读长, 无需模板扩增、运行时间较短、直接检测表观修饰位点、直接测转录本、较高但完全随机的测序错误等特点^[9-10, 12], 除了单独使用 Iso-Seq 技术外, 还可采用混合测序策略, 如二代测序 (RNA-

收稿日期: 2019-03-26 修回日期: 2019-04-30 网络出版日期: 2019-05-20

URL: <http://doi.org/10.13430/j.cnki.jpgr.20190326001>

第一作者研究方向为园艺蔬菜遗传与育种研究, E-mail: 1215590936@qq.com

通信作者: 许俊强, 研究方向为蔬菜育种与分子生物学研究, E-mail: xujunqiang101@163.com

基金项目: 国家自然科学基金 (31560560); 云南省应用基础研究基金 (2015FD019)

Foundation project: The National Natural Science Foundation of China (31560560), Yunnan Applied Basic Research Fund (2015FD019)

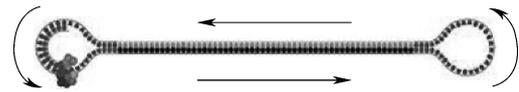
seq) 加三代测序 (Iso-Seq) 方法。TGS 技术提供了许多仅通过 SGS (二代测序) 技术无法获得的信息^[13-14]。

目前全长转录组测序的研究内容包括可变剪接分析、转录本结构研究、基因合成途径及差异表达分析、基因的多样性分析、参考基因组组装、完善基因注释、新基因预测、转录组组装、构建全长转录组数据库、转录组测序、基因组重头测序、DNA 甲基化等方面。

1 全长转录组测序的原理

全长转录组测序技术是基于边合成边测序的原理, SMRT 测序在目标 DNA 分子复制过程中捕获序列信息。被称为 SMRTBell 的芯片, 是一个封闭的单链环状 DNA, 它是通过将类似哑铃的结构连接到目标双链 DNA (dsDNA) 分子的两端而产生的 (图 1)。当一个 SMRTBell 样本被加载到一个称为 SMRT 单元的芯片上时, 一个 SMRTBell 扩散到一个名为零级波导 (ZMW, Zero-Mode Waveguide) 的排序单元中, 它为光检测提供了最小的可用体积^[13]。这项技术的核心便在于使用了 ZMW (图 2A)。ZMW 是一种小孔, 其直径只有几十个纳米,

由于 ZMW 底部的纳米小孔要比激光的单个波长还要短, 当 ZMW 的底部被激光照射时, 纳米小孔并不能让激光直接穿过, 而会在小孔处产生光的衍射, 只照亮零级波导底部的小片区域。在 ZMW 底部的这小片区域内, DNA 聚合酶被锚定, 当 DNA 聚合酶上加载到单个的脱氧核苷酸并形成新的化学键时, 脱氧核苷酸上的荧光物质就会被激活, 从而发光并被检测到。由于这个纳米孔很小, 因此, 在孔内 DNA 链周围, 脱氧核苷酸能被标记到的就很有有限, 其中 4 种被荧光标记的脱氧核苷酸很快地从孔外进入到孔内又出去。由于形成的背景荧光信号非常稳定且很弱, 才能在含有大量溶液的背景上, 得以检测到单个荧光标记的核苷酸^[13, 15] (图 2B)。



发夹适配器连接到双链 DNA 分子的末端, 形成一个封闭的圆圈。

聚合酶锚定在 ZMW 的底部, 并将碱基合并到读链中

Hairpin adaptors are ligated to the end of a doublestranded DNA molecule, forming a closed circle. The polymerase is anchored to the bottom of a ZMW and incorporates bases into the read strand

图 1 SMRTBell 模型^[13]

Fig.1 SMRTbell template^[13]

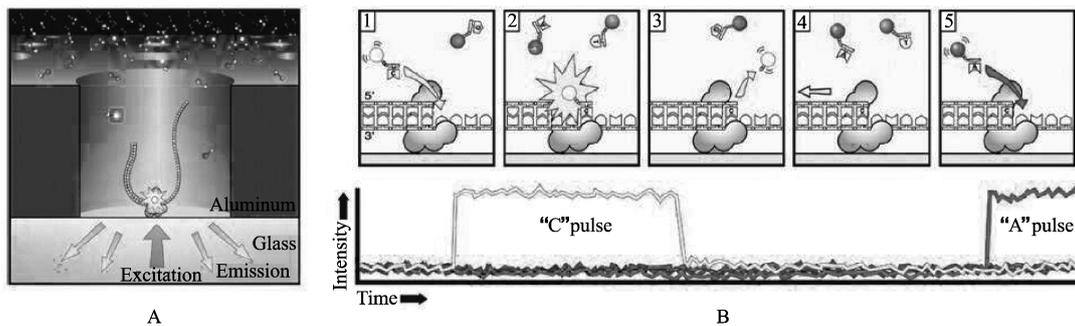


图 2 SMRT 测序原理^[13]

Fig.2 Principle of single-molecule real-time DNA sequencing^[13]

2 全长转录组测序在植物中的应用

目前, 在植物的研究中全长转录组测序技术已得到广泛的应用, 包括拟南芥、玉米、小麦、高粱、丹参、矮牵牛等。

2.1 全长转录组测序在模式植物中的应用

2.1.1 拟南芥 可变剪接 (AS, alternative splicing) 不仅影响植物发育过程中的生物钟和开花时间, 还参与应激反应^[16-18]。Zhu 等^[18] 结合 RNA-Seq、Iso-Seq 及蛋白质组学鉴定方法, 研究了模式植物拟

南芥 (*Arabidopsis thaliana* (L.) Heynh.) 经脱落酸 (ABA) 处理后的 AS 机制。研究发现, 大约 83.4% 的内含子基因有 AS 现象。ABA 处理会引起广泛的 AS 事件的差异, 表现为非常规剪接位点的增多, 这些结果增强了我们对正常条件下的可变剪接的理解, 以及对 ABA 处理的响应。

2.1.2 丹参 药用模式植物丹参 (*Salvia miltiorrhiza* Bunge) 的生物活性成分是丹参酮^[19-20], 先前的转录组研究是基于二代测序, 部分转录本不能代表全长 cDNA 序列。Xu 等^[21] 首次联合应用了 RNA-seq 和 Iso-Seq 技术对丹参不同的根组织 (周皮、韧皮

部、木质部)进行转录本测序,获得了16241个高质量的全长转录本,并进行了AS分析,发现具有AS现象的基因序列约40%,其中部分基因与萜类化合物和类异戊二烯代谢有关,这些结果为丹参的进一步研究提供了一个完整的转录组信息。徐志超^[22]结合二代和三代(Iso-Seq)测序技术进行了丹参基因合成途径及差异表达分析,得出根周皮丹参酮IIA含量分别是韧皮部和木质部的17倍和185倍,根组织的3个部位转录组差异表达显示参与萜类合成的主要上游基因在根周皮中显著高表达,丹参酮合成关键基因[*Copalyl diphosphate synthase*(*CPSI*)、*Kaurene synthase-like*(*KSL1*)和*CYP76AHI*]在根周皮中显著高表达。基于丹参酮的合成和积累规律,预测15个CYP450s、5个SDRs和1个20DDs参与丹参酮下游的生物合成。

2.1.3 野草莓 Li等^[23]针对相同的蔷薇科模式植物野草莓^[24](*Fragaria vesca* L.)样本进行二代高通量测序(Illumina)和三代全长转录组测序(PacBio RS),对结果中发生的AS事件进行比较。全长转录组测序中发现的可变剪接事件为57.67%,而利用二代高通量测序进行可变剪接分析的获得值为33.48%,表明全长转录组测序在可变剪接的分析上更具效率。

2.1.4 玉米 玉米(*Zea mays* L.)是研究遗传进化和高光效的C4模式植物^[25-26]。Wang等^[27]利用Iso-Seq技术从玉米自交系B73的6个组织(根、花粉、胚、胚乳、幼穗和幼雄穗)中得到了111151个转录本,进行了玉米的转录本结构研究。捕捉到了玉米RefGen_v3基因组中注释的约70%的基因,其中62547个(57%)转录本是全新的,有些是一些已知基因的组织特异性的转录异构体,有3%是来自全新基因位点的转录本。还发现了很多新的长链非编码RNAs和融合转录本,并发现DNA甲基化对于产生多种多样的转录异构体的重要作用。Zhou等^[28]利用Iso-Seq技术对W64A自交授粉后16d胚乳进行16个SMRT cell的全长转录组测序,共得到1057799条原始序列,筛选出166693条高质量全长转录本,进行了转录本结构的分析。通过与参考基因组(B73 RefGen_v4.0)对比,鉴定出3399个新基因位点,同时找到590条W64A特异转录本序列。结合W64A与其他自交系构建的F₂群体和回交群体,开发了基于全基因组基因座的1051个SSR标记和243个InDel标记,其中63.0%和

58.8%的分子标记得到验证。针对预测和报道玉米胚乳硬质/粉质相关基因,利用转录本信息预测其编码区,比较了32个基因编码区的变异,其中一些基因编码区的提前终止可能对籽粒性状影响较大。

2.1.5 高粱 高粱(*Sorghum bicolor* (L.) Moench)是C4植物研究的模式植物^[29-30]。Abdel-Ghany等^[31]采用Iso-Seq技术对高粱BTx623品系的种子进行了AS分析,发现有10053个异构体经历了AS事件,其中2950个异构体与现有的基因模型吻合,超过7000个未报道AS事件。在27860个转录本中,11342个(40.7%)AS事件是新发现的,7065个(25.4%)是全长异构体。69.9%的基因只有1个异构体,只有1%左右的基因有5个以上的异构体。此外,非生物胁迫下,异构体呈现出表达量或数目的差异性。全长转录组测序数据表明在高粱转录组中AS事件比之前的认知要普遍很多。Wang等^[32]利用三代Iso-Seq技术对玉米和高粱的11对匹配组织转录组进行了比较,揭示了两个物种中大量新的基因转录异构体。进化上比较年轻的基因往往会在生殖组织中产生,与进化上古老的基因相比具有更少的异构体。不同组织和不同物种间均观察到了可变剪切和活性的异同点。玉米的亚基因不存在异构体偏性,然而在B套基因组上的基因更偏向于在花粉组织中表达,而在A套基因组上的基因更偏向于在胚乳组织中表达。同时还鉴定了玉米和高粱高度保守的剪切事件。另外,构建了这两个物种高分辨率的poly(A)位点图谱,显示了两物种在mRNA剪切上的异同点。

2.1.6 矮牵牛 矮牵牛(*Petunia hybrida* Vilm.)是研究观赏植物生长和发育的模式植物^[33-34]。Ning等^[35]应用RNA-seq和Iso-Seq技术进行了矮牵牛的基因合成途径及差异表达分析,以RNA-seq构建的参考文库包含193749个拼接序列,在根部特异性高表达的转录本896个,在根部特异性低表达的转录本666个。以Iso-Seq构建的reads参考序列文库包含85571条序列,分别有2904、1618和3868个剪接体在根中的表达水平显著高于花、茎和叶组织,其中在根部特异性高表达的转录本有639个,分别有1967、1219和2780个亚型在根部的表达水平显著低于花、茎和叶组织,其中在根部一直处于特异性低表达的有869个。说明与二代RNA-seq相比,三代Iso-Seq可构建较高质量的参考序列文库,同时联合二代RNA-seq技术,并结合有效的序列算

法,从而获得具有更高可信度的差异表达分析和数据挖掘结果。

2.2 全长转录组测序在非模式植物中的应用

2.2.1 大豆 大豆 (*Glycine max* (L.) Merr.) 基因组测序已完成。张乐^[36]运用 Iso-Seq 技术进行了野生大豆和栽培大豆 *Glyma13g21630* 的基因多样性分析,在 133 份大豆品种中共检测到 29 个多态性位点,其中包括 7 个 InDel 和 22 个 SNP,多态性频率分别为 1InDel/434 bp 和 1SNP/138 bp。多态性位点在 *Glyma13g21630* 基因序列中表现为不均匀分布,其中变异富集区为内含子 3 和 5,其他区域变异则较小。在大豆由野生向栽培驯化的过程中,*Glyma13g21630* 基因由于正向选择作用而固定了有益变异,并表现出瓶颈效应。

2.2.2 小麦 小麦 (*Triticum aestivum* L.) 是异源六倍体,含有 A、B 和 D 3 个基因组,其基因组约 17 Gb,85% 以上基因组为重复序列。Ling 等^[37]通过构建 A 基因组 BAC 文库和 BAC 测序,结合 Iso-Seq 等技术,完成了乌拉尔图小麦 (G1812) 基因组测序和精细组装,并成功绘制出了小麦 A 基因组 7 条染色体的序列图谱,41507 个蛋白编码基因被注释。这些结果加速栽培小麦的遗传改良和分子设计育种,将利于小麦重要农艺性状基因的克隆、精细定位和功能解析。Dong 等^[38]利用 Iso-Seq 技术完善了小麦的基因注释。建立了 2 kb 以下和 2 kb 以上 2 个文库,借助于 Iso-Seq 的读长优势,不经过组装直接检测转录本的完整结构,成功地获取高可信度的剪切位点和转录本模型,并在此基础上进行基因剪切多态性分析,实现对已有注释结果的完善和提升。发现 3026 个新基因及 9591 个新的同源异构体,也找出了 72 个不同的小麦面筋蛋白基因的全长转录本。结合 4 个籽粒发育时期的二代转录组数据,发现了 6030 个基因在籽粒发育不同阶段中呈现同源异构体表达差异。

2.2.3 白菜 白菜 (*Brassica rapa* L.) 是重要的蔬菜作物^[39]。Zhang 等^[40]利用 Iso-Seq 等技术对白菜参考基因组进行了 *de novo* 组装和重新注释,并获得白菜参考基因组 v3.0 新版。v3.0 版基因组的序列总长度为 353.14 Mb, scaffold N50 值为 4.45 Mb, contig N50 值为 1.45 Mb, 仅含有 389 个 gaps 和 1301 条 scaffolds。v3.0 版的序列连续性提升近 30 倍,并且 gaps 数量和长度大大减少。在 v3.0 版中预测到 45985 个编码蛋白基因,低于 v2.5 版本中的 48826 个基因,原因是 v2.5 中存在

的组装错误。有 13318 个全长长末端重复反转座子 (LTR-RTs) 在 v3.0 版参考基因组中被预测到,远高于 v2.5 版 (4129 个) 和 v1.5 版 (801 个),并检测到 3 次转座子爆发事件,其中 1 次是新检测到的。该研究还准确地预测到了白菜基因组的着丝粒位置。

2.2.4 棉花 棉花 (*Gossypium* spp.) 是异源多倍体研究的理想模型,有 2 套高度相似的亚基因组 At 和 Dt。Wang 等^[41]利用 Iso-Seq 技术对海岛棉进行了 AS 分析,从 44968 个基因中鉴定了 176849 个全长转录本,鉴定了 15102 个棉花纤维中特异的 AS 事件,发现 51.4% 的亚基因组间的同源基因会通过 AS 产生结构差异的转录本。进一步研究发现,同一个基因由 AS 产生的转录本会受到 miRNA 的差异性调控,揭示了四倍体棉花中 AS 的复杂性和新的调控机制。

2.2.5 甘蔗 目前大多数甘蔗 (*Saccharum officinarum* L.) 转录组学研究受限于转录本的长度和转录本 isoform 的数量。Hoang 等^[42]利用 Iso-Seq 技术对采集的杂交甘蔗 (*Saccharum officinarum* × *S. spontaneum*) 22 个组织样本进行了研究,并完善了基因注释。共鉴定出 107598 (71%) 个单一转录异构体,其中与植物蛋白数据相匹配的系列有 92%,超过 2% 是新的转录本,超过 2% 是 lncRNA,全部序列的约 56% 注释到 GO 数据库,约 23% 注释到 KEGG 数据库。

2.2.6 甜菜 甜菜 (*Beta vulgaris* L.) 是称之为石竹目 (Caryophyllales) 的一组开花植物中第 1 个完成基因组测序的代表性物种^[43]。Minoche 等^[44]利用 Iso-Seq 技术进行了甜菜及其同科菠菜 (*Spinacia oleracea* L.) 的新基因预测和验证,发现 98% 测序转录本包含完整的开放阅读框 (ORF),可协助二代转录组测序降低背景干扰,提高新基因预测灵敏度和精确度,改进了甜菜和第 1 个全基因组菠菜的基因集,为新测基因组非模式生物物种基因注释提供了思路。

2.2.7 毛竹 毛竹 (*Phyllostachys edulis* (Carrière) J.Houz.) 是生长最快的植物之一^[45],研究其快速生长机制具有重要的理论意义和实际应用价值。Wang 等^[46]结合 RNA-seq 和 Iso-Seq 技术,分析了 145522 个全长非嵌合 (FLNC) 系列,纠正了 2241 个错误基因,对先前的 8091 个同源位点进行鉴定。超过 42280 个差异剪接异构体是来源于 1286677 个内含子全长非嵌合系列,包括大量与地下茎系统相

关的 AS 事件。有 2154 个 AS 事件在没有装配的情况下被识别出来, AS 基因的丰度高于非 AS 基因。林泽忠^[47]进行了毛竹冬笋芽、鞭芽及鞭梢芽的基因差异表达分析, 平均得到 5G 的数据量, 经过差异表达分析后, 得出毛竹在冬笋芽、鞭芽及鞭梢芽不同生长时期的基因表达差异, 冬笋芽和鞭梢芽间差异表达基因 6863 个, 鞭芽和鞭梢芽间差异表达基因 5592 个, 笋芽和鞭芽间差异表达基因 7029 个。

2.2.8 黄芪 Li 等^[48]利用 Iso-Seq 技术构建了黄芪 (*Astragalus membranaceus* Bunge) 叶和根的全长转录组, 分别得到了 27975 条和 22343 条全长参考序列, 进行了 AS 分析。在叶和根里, 分别有 36.53% 和 34.46% 的基因含有 2 个以上的同源异构体, 分别检测出 6494 个和 4399 个可变剪切事件, 说明了黄芪叶和根转录组的复杂性。其中, 内含子保留类型的可变剪切事件最多。

2.2.9 茶树 朱兴正等^[49]采用 Iso-Seq 技术探讨了云茶 1 号茶树 (*Camellia assamica* (J.W.Mast.) H. T.Chang) 品种优异性状的遗传基础, 获得 Polished consensus 序列 213389 个, 预测到 CDSs 有 223120 个, 检测到 195062 个 SSR 位点, 完善了基因注释。有 170264 个同源序列在 NR 数据库中比对到 980 个物种, 在 KOG 数据库得到注释的有 103124 个, 根据其功能分为 26 类; 有 65524 个得到 GO 注释, 分为细胞组分、分子功能及生物学过程等 3 大类的 55 个功能组; 根据 KEGG 数据库, 105972 个得到了注释, 涉及到 216 个代谢途径分支, 包括茶叶品质、活性物质代谢以及抗逆等相关基因等, 还预测到隶属于 60 个转录因子家族的转录因子有 5785 个。

2.2.10 咖啡豆 Cheng^[50]等利用 Iso-Seq 探讨了四倍体阿拉比卡咖啡豆 (*Coffea arabica* L.) 的转录组多样性和复杂性, 完善了基因注释。在同型四倍体咖啡豆转录组中获得 95995 条差异转录本 (平均 3236 bp), NCBI 总共注释到 88715 条非冗余植物蛋白系列 (92.42%), 包括 34719 份高质量注释。中果咖啡豆 (*Coffea canephora* Pierre ex A.Froehner) 编码序列与 UTR、阿拉比卡咖啡豆 EST 和 RFAM 中有 1213 个没有命中的序列, 是咖啡豆中潜在的新基因。捕捉到更长的 UTRs, 特别是 5' 端 UTRs, 它便于识别上游的开放阅读框。Iso-Seq 测序还揭示了该基因组关键的咖啡因和蔗糖代谢基因中越来越长的转录本变异体。发现大于 10 k 碱基的长序列是不正确的注释, Iso-Seq 技术表明了以往研究的局限性。

2.2.11 松树 火炬松 (*Pinus taeda* L.) 的基因组有 22Gb。2014 年发表的数据完全基于短读^[51-52], 长度 100~250 bp, 包含 1100 万个 contig, 其 N50 值为 8206 bp。Zimin 等^[53]使用 Iso-Seq 技术, 在长读中产生了大约 12 倍的覆盖范围。将两结果进行基因组组装, 结果 N50 contig 值为 25361, N50 scaffold 值为 107821。兰伯氏松 (*Pinus lambertiana* Douglas) 基因组约 31Gb。Gonzalez-Ibeas 等^[54]采用 Iso-Seq 技术对兰伯氏松各种组织类型进行深度测序获得转录组, 产生超过 25 亿条的短读, 通过 Iso-Seq 产生的长读首次应对转录本从头组装。通过 HiSeq 获得转录本条数为 75175, MiSeq 为 45524, PacBio RSII 为 158113。

2.2.12 无油樟 Liu 等^[17]运用 Iso-Seq 技术进行了无油樟转录组测序及无参情况下的 AS 事件分析, 将 Iso-Seq 所获的 428 个 AS 事件与 RNA-Seq 和 PASA (Program to Assemble Spliced Alignments) 所得结果进行比较, 分别被 PASA 和 RNA-Seq 支持的可变剪接事件有 160 和 11 个, 同时获 PASA 和 RNA-Seq 支持的可变剪接事件 157 个。从 AS 事件中随机挑选 61 个进行验证, 有 41 个获得数据支持, 其余 20 个被认为是“false positive”可变剪接事件中, 有 3 个能被 RT-PCR 所验证, 4 个映射到无油樟基因组草图 gap 区域, 发生在非经典剪接位点的可变剪接事件 10 个, 这 10 个 AS 事件既未通过 RT-PCR 验证, 也不能被 RNA-Seq 或 PASA 所支持。表明全长转录组测序技术预测无参转录组可变剪接事件准确率达到 66%-76%。

2.2.13 苜蓿 苜蓿 (*Medicago sativa* L.) 以“牧草之王”著称。张寒^[55]应用 RNA-seq 和 Iso-Seq 技术对 119 个苜蓿品种 (系) 进行转录组测序和参考转录组组装, 建立了紫花苜蓿的参考转录组, 序列总长度为 597539277 bp, 转录本条数为 989620, 其中最和最短转录本长度分别为 16671 bp 和 139 bp, 平均长度为 603 bp, 中位数为 332 bp, N50 值为 868 bp, GC 值为 37.09%。通过二代测序与三代测序结果进行基因组拼接组装后, 转录本条数、系列总长、N50 及平均长度增加, 数据准确度得到了提高。

2.2.14 盐生草 汪军成^[56]以盐生草 (*Halogeton glomeratus* (M.Bieb.) C.A.Mey.) 为材料, 运用 Iso-Seq 技术构建了全长转录组数据库, 并获得 54835 条一致性异构体, 平均长度为 2663 bp, 与已知基因序列匹配的 isoform 序列共有 53230 (97.07%) 条,

其中成功获得了其编码序列 (CDS) 的异构体有 52141 (95.07%) 条。对幼苗进行 0、100、200、400 mmol/L NaCl 胁迫处理 3 d 后, 在盐生草的叶片中分别鉴定到盐胁迫响应的共表达上调、下调差异异构体 (DEIs) 11587 个, 并发现其编码蛋白质拥有跨膜结构域 (TMD) 的共表达上调 DEIs 有 16 个, 很可能包含 Na⁺ 跨膜转运相关蛋白。

2.2.15 猪笼草 Zulkapli 等^[57] 采用 Iso-Seq 技术分析了能以枯叶落物为食的瓶状猪笼草 (*Nepenthes ampullaria* Jack) 的全长 cDNA 序列, 并与食肉性莱佛士猪笼草 (*Nepenthes rafflesiana* Jack.) 及其食性杂交双色虎克猪笼草 (*Nepenthes* × *hookeriana* H. Low) 进行了比较, 构建了全长 CDNA 文库, 可以通过比较来自亲本和杂交种的猪笼草转录组, 来探究杂交影响猪笼草饮食习惯演变的机制。

3 问题与展望

全长转录组测序技术推动了植物领域的研究发展, 在模式植物和非模式植物的研究中也得到了广泛应用, 为部分无参基因植物的研究提供了可能。随着三代 PacBio 平台的飞速发展, 其测序长度和通量不断增加的潜力依然巨大。新一代测序技术仍存在一定的不足, 如测序技术中需要采用聚合酶及核酸外切酶等, 酶活性与稳定性的保持是一个重要的问题; 另一个问题是在 DNA 的固定上, 怎样保持 DNA 的延展性而避免二聚体结构的出现; 还有在需要荧光标记的情况下, 如何提高单分子信号灵敏度的同时又避免信号变成噪音增加荧光背景等问题。

Iso-Seq 是一种很有前景的测序技术, 未来的转录组测序是基于 PacBio 平台获得样品的全长转录组, 实践中也可根据遇到的问题来选用合理的混合测序 (二代加三代) 手段。由于 Iso-Seq 技术建立的时间还很短, 技术还不够成熟, 但潜力无限, 在建立标准流程、优化实验参数和质量控制方面的研究将是未来基因组学领域的一个重要研究方向。三代 Iso-Seq 技术也会被广泛应用到植物转录组学、基因组学、代谢组学和遗传学等研究中。

参考文献

- [1] 任毅鹏, 张佳庆, 孙瑜, 吴振峰, 阮吉寿, 贺秉军, 刘国卿, 高山, 卜文俊. 基于 PacBio 平台的全长转录组测序. 科学通报, 2016, 61 (11): 1250
Ren Y P, Zhang J Q, Sun Y, Wu Z F, Ruan J S, He B J, Liu

- G Q, Gao S, Bu W J. Full-length transcriptome sequencing on Pac Bio platform. Chinese Science Bulletin, 2016, 61 (11): 1250
- [2] Maxam A M, Gilbert W. A new method for sequencing DNA. Proceedings of the National Academy of Sciences of the United States of America, 1977, 74 (2): 560-564
- [3] Sanger F, Nicklen S, Coulson A R. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America, 1977, 74 (12): 5463-5467
- [4] Margulies M, Egholm M, Altman W E, Attiya S, Bader J S, Bemben L A, Berka J, Braverman M S, Chen Y J, Chen Z, Dewell S B, Du L, Fierro J M, Gomes X V, Godwin B C, He W, Helgesen S, Ho C H, Irzyk G P, Jando S C, Alenquer M L I, Jarvie T P, Jirage K B, Kim J B, Knight J R, Lanza J R, Leamon J H, Lefkowitz S M, Lei M, Li J, Lohman K L, Lu H, Makhijani V B, McDade K E, McKenna M P, Myers E W, Nickerson E, Nobile J R, Plant R, Puc B P, Ronan M T, Roth G T, Sarkis G J, Simons J F, Simpson J W, Srinivasan M, Tartaro K R, Tomasz A, Vogt K A, Volkmer G A, Wang S H, Wang Y, Weiner M P, Yu P, Begley R F, Rothberg J M. Genome sequencing in microfabricated high-density picolitre reactors. Nature, 2005, 437: 158-160
- [5] Bentley D R, Balasubramanian S, Swerdlow H P, Smith G P, Milton J, Brown C G, Hall K P, Evers D J, Barnes C L, Bignell H R, Boutell J M, Bryant J, Carter R J, Keira Cheetham R, Cox A J, Ellis D J, Flatbush M R, Gormley N A, Humphray S J, Irving L J, Karbelashvili M S, Kirk S M, Li H, Liu X, Maisinger K S, Murray L J, Obradovic B, Ost T, Parkinson M L, Pratt M R, Rasolonjatovo I M, Reed M T, Rigatti R, Rodighiero C, Ross M T, Sabot A, Sankar S V, Scally A, Schroth G P, Smith M E, Smith V P, Spiridou A, Torrance P E, Tzonev S S, Vermaas E H, Walter K, Wu X, Zhang L, Alam M D, Anastasi C, Aniebo I C, Bailey D M, Bancarz I R, Banerjee S, Barbour S G, Baybayan P A, Benoit V A, Benson K F, Bevis C, Black P J, Boodhun A, Brennan J S, Bridgham J A, Brown R C, Brown A A, Buermann D H, Bundu A A, Burrows J C, Carter N P, Castillo N, Chiara E Catenazzi M, Chang S, Neil Cooley R, Crake N R, Dada O O, Diakoumakos K D, Dominguez-Fernandez B, Earnshaw D J, Egbujor U C, Elmore D W, Etchin S S, Ewan M R, Fedurco M, Fraser L J, Fuentes Fajardo K V, Scott Furey W, George D, Gietzen K J, Goddard C P, Golda G S, Granieri P A, Green D E, Gustafson D L, Hansen N F, Harnish K, Haudenschild C D, Heyer N I, Hims M M, Ho J T, Horgan A M, Hoschler K, Hurwitz S, Ivanov D V, Johnson M Q, James T, Huw Jones T A, Kang G D, Kerelska T H, Kersey A D, Khrebtukova I, Kindwall A P, Kingsbury Z, Kokko-Gonzales P I, Kumar A, Laurent M A, Lawley C T, Lee S E, Lee X, Liao A K, Loch J A, Lok M, Luo S, Mammen R M, Martin J W, McCauley P G, McNitt P, Mehta P, Moon K W, Mullens J W, Newington T, Ning Z, Ling Ng B, Novo S M, O'Neill M J, Osborne M A, Osnowski A, Ostadan O, Paraschos L L, Pickering L, Pike A C, Pike A C, Chris Pinkard D, Pliskin D P, Podhasky J, Quijano V J, Raczky C, Rae V H, Rawlings S R, Chiva Rodriguez A, Roe P M, Rogers J, Rogert Bacigalupo M C, Romanov N, Romieu A, Roth R K, Rourke N J, Ruediger S T, Rusman E, Sanches-Kuiper R M, Schenker

- M R, Seoane J M, Shaw R J, Shiver M K, Short S W, Sizto N L, Sluis J P, Smith M A, Ernest Sohna Sohna J, Spence E J, Stevens K, Sutton N, Szajkowski L, Tregidgo C L, Turcatti G, Vandevondele S, Verhovskiy Y, Virk S M, Wakelin S, Walcott G C, Wang J, Worsley G J, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin J C, Hurler M E, McCooke N J, West J S, Oaks F L, Lundberg P L, Klenerman D, Durbin R, Smith A J. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 2008, 456 (7218): 53-59
- [6] Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek J A, Costa G, McKernan K, Sidow A, Fire A, Johnson S M. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Research*, 2008, 18 (7): 1051-1063
- [7] 蒋会兵, 夏丽飞, 田易萍, 戴伟东, 孙云南, 陈林波. 基于转录组测序的紫芽茶树花青素合成相关基因分析. *植物遗传资源学报*, 2018, 19 (5): 967-978
- Jiang H B, Xin L F, Tian Y P, Dai W D, Sun Y N, Chen L B. Transcriptome analysis of anthocyanin synthesis related genes in purple bud tea plant. *Journal of Plant Genetic Resources*, 2018, 19 (5): 967-978
- [8] 曹晨霞, 韩琬, 张和平. 第三代测序技术在微生物研究中的应用. *微生物学通报*, 2016, 43 (10): 2269-2276
- Cao C X, Han W, Zhang H P. Application of third generation sequencing technology to microbial research. *Microbiology China*, 2016, 43 (10): 2269-2276
- [9] 柳延虎, 王璐, 于黎. 单分子实时测序技术的原理与应用. *遗传*, 2015, 37 (3): 259-268
- Liu Y H, Wang L, Yu L. The principle and application of the single-molecule real-time sequencing technology. *Hereditas*, 2015, 37 (3): 259-268
- [10] 唐勇, 刘旭. SMRT 测序技术及其在微生物研究中的应用. *生物技术通报*, 2018, 34 (6): 48-53
- Tang Y, Liu X. SMRT sequencing and its application in microorganism studies. *Biotechnology Bulletin*, 2018, 34 (6): 48-53
- [11] 王瑞娴, 李川. 全长转录组测序技术在非模式植物转录组学研究中的应用. *分子植物育种*, 2019, 17 (2): 502-508
- Wang R X, Li C. Application of full-length transcriptome sequencing in non-model plant transcriptome research. *Molecular Plant Breeding*, 2019, 17 (2): 502-508
- [12] Shin S C, Ahn D H, Kim S J, Lee H, Oh T J, Lee J E, Park H. Advantages of single-molecule real-time sequencing in high-GC content genomes. *PLoS One*, 2013, 8 (7): e68824
- [13] Rhoads A, Au K F. PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics*, 2015, 13 (5): 278-289
- [14] Ardui S, Ameer A, Vermeesch J R, Hestand M S. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Research*, 2018, 46 (5): 2159-2168
- [15] Treffer R, Deckert V. Recent advances in single-molecule sequencing. *Current Opinion Biotechnology*, 2010, 21 (1): 4-11
- [16] Pan Q, Shai O, Lee L J, Frey B J, Blencowe B J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 2008, 40 (12): 1413-1415
- [17] Liu X, Mei W, Soltis P S, Soltis D E, Barbazuk W B. Detecting alternatively spliced transcript isoforms from single-molecule long-read sequences without a reference genome. *Molecular Ecology Resources*, 2017, 17 (6): 1243-1256
- [18] Zhu F Y, Chen M X, Ye N H, Shi L, Ma K L, Yang J F, Cao Y Y, Zhang Y, Yoshida T, Fernie A R, Fan G Y, Wen B, Zhou R, Liu T Y, Fan T, Gao B, Zhang D, Hao G F, Xiao S, Liu Y G, Zhang J. Proteogenomic analysis reveals alternative splicing and translation as part of the abscisic acid response in *Arabidopsis* seedlings. *The Plant Journal*, 2017, 91: 518-533
- [19] 宋经元, 罗红梅, 李春芳, 孙超, 徐江, 陈士林. 丹参药用模式植物研究探讨. *药学报*, 2013, 48 (7): 1099-1106
- Song J Y, Luo H M, Li C F, Sun C, Xu J, Chen S L. *Salvia miltiorrhiza* as medicinal model plant. *Acta Pharmaceutica Sinica*, 2013, 48 (7): 1099-1106
- [20] 袁媛, 吴芹, 石京山, 陈修平. 丹参及其主要成分保肝作用的研究进展. *中国中药杂志*, 2015, 40 (4): 588-593
- Yuan Y, Wu Q, Shi J S, Chen X P. Advance in studies on hepatoprotective effect of *Salvia miltiorrhiza* and its main components. *China Journal of Chinese Materia Medica*, 2015, 40 (4): 588-593
- [21] Xu Z, Peters R J, Weirather J, Luo H, Liao B, Zhang X, Zhu Y, Ji A, Zhang B, Hu S, Au K F, Song J, Chen S. Full-length transcriptome sequences and splice variants obtained by a combination of sequencing platforms applied to different root tissues of *Salvia miltiorrhiza* and tanshinone biosynthesis. *The Plant Journal*, 2015, 82 (6): 951-961
- [22] 徐志超. 基于联合测序技术的丹参活性成分生物合成及调控机制研究. 北京: 北京协和医学院, 2016
- Xu Z C. Study on biosynthesis and regulation of active compounds in *Salvia miltiorrhiza* by hybrid sequencing. Beijing: Peking Union Medical College, 2016
- [23] Li Y, Dai C, Hu C, Liu Z, Kang C. Global identification of alternative splicing via comparative analysis of SMRT- and Illumina-based RNA-seq in strawberry. *The Plant Journal*, 2017, 90 (1): 164-176
- [24] Chen Solomenbb. 蔷薇科模式植物林丛野草莓的基因组测序完成. *农业生物技术学报*, 2010, 18 (1): 29
- Chen S. Genome of the woodland strawberry sequenced, a model system for rosaceae plants. *Journal of Agricultural Biotechnology*, 2010, 18 (1): 29
- [25] 闫蕾. 玉米隐花色素基因 CRY1a 的克隆及功能分析. 太原: 山西大学, 2016
- Yan L. Cloning and functional analysis of cryptochrome 1a in maize (*Zea mays* L.). Taiyuan: Shanxi University, 2016
- [26] 曾怡锦. 温、热带玉米特异转座子的验证及功能初探. 雅安: 四川农业大学, 2016
- Zeng Y J. Verification and function of specific transposable element in temperate and tropical Maize. Yaan: Sichuan Agricultural University, 2016
- [27] Wang B, Tseng E, Regulski M, Clark T A, Hon T, Jiao Y, Lu Z, Olson A, Stein J C, Ware D. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nature Communications*, 2016, 7: 11708
- [28] Zhou Y, Zhao Z X, Zhang Z Y, Fu M M, Wu Y R, Wang W Q. Isoform sequencing provides insight into natural genetic diversity in maize. *Plant Biotechnology Journal*, 2018, DOI:

- 10.1111/pbi.13063
- [29] 刘国庆, 杜瑞恒, 侯升林, 吕芃, 籍贵苏, 李素英. 高粱抗蚜研究进展. 植物学报, 2012, 47(2): 171-187
Liu G Q, Du R S, Hou S L, Lv F, Ji G S, Li S Y. Resistance to aphids in sorghum; a review. Bulletin of Botany, 2012, 47(2): 171-187
- [30] 张春来, 李艳锋, 赵威军, 赵靛, 王晨, 梁笃, 周福平. 高粱品质性状改良的分子遗传学基础. 植物生理学报, 2015, 51(5): 610-616
Zhang C L, Li Y F, Zhao W J, Zhao J, Wang C, Liang D, Zhuo F P. Molecular genetic basis for biotechnological improvement of grain quality characteristics in sorghum. Plant Physiology Journal, 2015, 51(5): 610-616
- [31] Abdel-Ghany S E, Hamilton M, Jacobi J L, Ngam P, Devitt N, Schilkey F, Ben-Hur A, Reddy A S. A survey of the sorghum transcriptome using single-molecule long reads. Nature Communications, 2016, 7: 11706
- [32] Wang B, Regulski M, Tseng E, Olson A, Goodwin S, McCombie W R, Ware D. A comparative transcriptional landscape of maize and sorghum obtained by single-molecule sequencing. Genome Research, 2018, 28: 921-928
- [33] 刘娜, 王琦, 刘同瑞, 熊枫, 张水明, 董丽丽. 矮牵牛 *PhTPS5* 基因的克隆与表达分析. 西北植物学报, 2018, 38(9): 1620-1624
Liu N, Wang Q, Liu T R, Xiong F, Zhang S M, Dong L L. Identification and expression analysis of *PhTPS5* in petunia. Acta Botanica Boreali-Occidentalia Sinica, 2018, 38(9): 1620-1624
- [34] 杨伟苑. 矮牵牛 *PhCESA3* 基因功能鉴定. 广州: 华南农业大学, 2016
Yang W Y. Functional characterization of *PhCESA3* in petunia. Guangzhou: South China Agricultural University, 2016
- [35] Ning G, Cheng X, Luo P, Liang F, Wang Z, Yu G, Li X, Wang D, Bao M. Hybrid sequencing and map finding (HySeMaFi): optional strategies for extensively deciphering gene splicing and expression in organisms without reference genome. Scientific Reports, 2017, 7: 43793
- [36] 张乐. 绥农 14 全长转录本的获得和注释及 *Glyma13g21630* 基因多样性分析. 北京: 中国农业科学院, 2011
Zhang L. Collection and annotation of suinong14 full-length transcripts and gene diversity analysis of *Glyma13g21630*. Beijing: Chinese Academy of Agricultural Sciences, 2011
- [37] Ling H Q, Ma B, Shi X, Liu H, Dong L, Sun H, Cao Y, Gao Q, Zheng S, Li Y, Yu Y, Du H, Qi M, Li Y, Lu H, Yu H, Cui Y, Wang N, Chen C, Wu H, Zhao Y, Zhang J, Li Y, Zhou W, Zhang B, Hu W, Van Eijk M J T, Tang J, Witsenboer H M A, Zhao S, Li Z, Zhang A, Wang D, Liang C. Genome sequence of the progenitor of wheat a subgenome triticum urartu. Nature, 2018, 557(7705): 424
- [38] Dong L L, Liu H F, Zhang J C, Yang S J, Kong G Y, Chu J S C, Chen N S, Wang A W. Single-molecule real-time transcript sequencing facilitates common wheat genome annotation and grain transcriptome research. BMC Genomics, 2015, 16(1): 1039
- [39] 吴昊天. 白菜类作物开花时间相关基因的表达分析. 南京: 南京农业大学, 2016
Wu H T. Measurement of flowering time and analysis of gene expression in *Brassica rapa*. Nanjing: Nanjing Agricultural University, 2016
- [40] Zhang L, Cai X, Wu J, Liu M, Grob S, Cheng F, Liang J L, Cai C C, Liu Z Y, Liu B, Wang F, Li S, Liu F Y, Li X M, Cheng L, Yang W C, Li M H, Grossniklaus U, Zheng H K, Wang X W. Improved *Brassica rapa* reference genome by single-molecule sequencing and chromosome conformation capture technologies. Horticulture Research, 2018, 5(1): 50
- [41] Wang M, Wang P, Liang F, Ye Z, Li J, Shen C, Pei L, Wang F, Hu J, Tu L, Lindsey K, He D, Zhang X. A global survey of alternative splicing in allopolyploid cotton: landscape, complexity and regulation. New Phytologist, 2018, 217(1): 163-178
- [42] Hoang N V, Furtado A, Mason P J, Marquardt A, Kasirajan L, Thirugnanasambandam P P, Botha F C, Henry R J. A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and *de novo* assembly from short read sequencing. BMC Genomics, 2017, 18(1): 395
- [43] Dohm J C, Minoche A E, Holtgräwe D, Capella-Gutiérrez S, Zakrzewski F, Tafer H, Rupp O, Sörensen T R, Stracke R, Reinhardt R, Goesmann A, Kraft T, Schulz B, Stadler P F, Schmidt T, Gabaldón T, Lehrach H, Weisshaar B, Himmelbauer H. The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). Nature, 2014, 505(7484): 546
- [44] Minoche A E, Dohm J C, Schneider J, Holtgräwe D, Viehöver P, Montfort M, Sörensen T R, Weisshaar B, Himmelbauer H. Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. Genome Biology, 2015, 16(1): 184
- [45] Sun H F, Li Q, Lei Z F, Zhang J B, Song X Z, Song X Z. Ecological stoichiometry of nitrogen and phosphorus in moso bamboo (*Phyllostachys edulis*) during the explosive growth period of new emergent shoots. Journal of Plant Research, 2019, 132(1): 107-115
- [46] Wang T, Wang H, Cai D, Gao Y, Zhang H, Wang Y, Lin C, Ma L, Gu L. Comprehensive profiling of rhizome-associated alternative splicing and alternative polyadenylation in moso bamboo (*Phyllostachys edulis*). Plant Journal for Cell & Molecular Biology, 2017, 91(4): 684-699
- [47] 林泽忠. 利用高通量测序技术探索毛竹冬笋成笋机制. 福州: 福建农林大学, 2016
Lin Z Z. Through high-throughput sequencing technologies to explore mechanism of bamboo winter shoot of *Phyllostachys pubescens*. Fuzhou: Fujian Agriculture and Forestry University, 2016
- [48] Li J, Harata-Lee Y, Denton M D, Feng Q J, Rathjen J R, Qu Z P, Adelson D L. Long read reference genome-free reconstruction of a full-length transcriptome from *Astragalus membranaceus* reveals transcript variants involved in bioactive compound biosynthesis. Cell Discovery, 2017, 3: 17031
- [49] 朱兴正, 夏丽飞, 陈林波, 孙云南, 田易萍, 宋维希, 蒋会兵. 保护品种云茶 1 号茶树全长转录组测序分析. 茶叶科学, 2018, 38(2): 193-201
Zhu X Z, Xia L F, Chen L B, Sun Y N, Tian Y P, Song W X, Jiang H B. Full-length transcriptome analysis of protected

- cultivation 'Yuncha 1' (*Camellia sinensis* var. *assamica*). *Journal of Tea Science*, 2018, 38(2): 193-201
- [50] Cheng B, Furtado A, Henry R J. Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. *Gigascience*, 2017, 6(11): 1-13
- [51] Neale D B, Wegrzyn J L, Stevens K A, Zimin A V, Puiu D, Crepeau M W, Cardeno C, Koriabine M, Holtz-Morris A E, Liechty J D, Martínez-García P J, Vasquez-Gross H A, Lin B Y, Zieve J J, Dougherty W M, Fuentes-Soriano S, Wu L S, Gilbert D, Marçais G, Roberts M, Holt C, Yandell M, Davis J M, Smith K E, Dean J F, Lorenz W W, Whetten R W, Sederoff R, Wheeler N, McGuire P E, Main D, Loopstra C A, Mockaitis K, deJong P J, Yorke J A, Salzberg S L, Langley C H. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology*, 2014, 15(3): 315-330
- [52] Zimin A, Stevens K A, Crepeau M W, Holtz-Morris A, Koriabine M, Marçais G, Puiu D, Roberts M, Wegrzyn J L, De Jong P J, Neale D B, Salzberg S L, Yorke J A, Langley C H. Sequencing and assembly of the 22-gb loblolly pine genome. *Genetics*, 2014, 196(3): 875-890
- [53] Zimin A V, Stevens K A, Crepeau M W, Puiu D, Wegrzyn J L, Yorke J A, Langley C H, Neale D B, Salzberg S L. An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. *Gigascience*, 2017, 6(1): 1-4
- [54] Gonzalez-Ibeas D, Martinez-Garcia P J, Famula R A, Delfino-Mix A, Stevens K A, Loopstra C A, Langley C H, Neale D B, Wegrzyn J L. Assessing the gene content of the megagenome: sugar pine (*Pinus lambertiana*). *G3 Genesgenetics*, 2016, 6(12): 3787-3802
- [55] 张寒. 基于 RNA-Seq 的苜蓿萌发期耐盐性状全基因组关联分析. 泰安: 山东农业大学, 2018
Zhang H. Genome-wide association study for salt tolerant traits of alfalfa at germination stage based on RNA-Seq. Taian: Shandong Agricultural University, 2018
- [56] 汪军成. 盐生草盐分区隔化耐盐机制研究. 兰州: 甘肃农业大学, 2017
Wang J C. Study on the salt tolerance mechanisms of ion compartmentation in halophyte *Halogeton glomeratus*. Lanzhou: Gansu Agricultural University, 2017
- [57] Zulkapli M M, Rosli M A F, Salleh F I M, Mohd Noor N, Aizat W M, Goh H H. Iso-Seq analysis of *Nepenthes ampullaria*, *Nepenthes rafflesiana* and *Nepenthes × hookeriana* for hybridisation study in pitcher plants. *Genomics Data*, 2017, 12: 130-131