

倪晓鹏,高志红. 园艺作物基因组测序研究进展[J]. 江苏农业科学,2016,44(2):9-13.

doi:10.15889/j.issn.1002-1302.2016.02.003

# 园艺作物基因组测序研究进展

倪晓鹏,高志红

(南京农业大学园艺学院,江苏南京 210095)

**摘要:**回顾了园艺作物全基因组测序的发展历程,介绍 3 代测序技术的特点和应用现状。总结了葡萄、番木瓜、草莓等果树,黄瓜、白菜、番茄等蔬菜以及莲、康乃馨等花卉在内的 25 种园艺作物基因组测序简况。重点介绍了全基因组测序在基因注释、比较基因组学研究、重测序、全基因组关联分析、转录组学、SNP 芯片开发等方面的应用。最后讨论了全基因组测序研究的难点和今后的研究方向。

**关键词:**园艺作物;基因组测序;基因组学

**中图分类号:** Q78 **文献标志码:** A **文章编号:** 1002-1302(2016)02-0009-04

园艺业是农业种植业的重要组成部分,对丰富人类营养和美化、改善人类生存环境有重要意义。我国是世界园艺大国,园艺是我国农业的重要组成部分,其中蔬菜播种面积由 1990 年的 600 万  $\text{hm}^2$  增加到 2013 年的 2 100 万  $\text{hm}^2$ ,产量由 1990 年的 1.95 亿 t 增加到 2013 年的 7.35 亿 t<sup>[1]</sup>。种植面积和产量分别占世界的 43% 和 49%,居世界第一位。随着动植物全基因组测序的不断发展,2001 年由美国能源部推动的人类基因组工作草图的发表被认为是人类基因组计划成功的里程碑<sup>[2]</sup>。破译人类遗传信息,将对医学、生物学乃至整个生物科学产生不可估量的影响。这一计划促成了大批动植物全基因组测序的完成,也有望从基因组水平上分析基因的结构、组成、调控和物种进化,从而避免传统分子生物学带来的种种弊端,大大促进全基因组测序技术在园艺作物上的应用发展。

## 1 全基因组测序的发展历程及其在园艺植物测序中的应用

距世界第一例模式植物拟南芥全基因组测序发表已有十余年<sup>[3]</sup>,在此期间测序技术飞速发展,每年发表的测序文章都稳定增加。拟南芥的测序就是基于第一代测序技术 Sanger 测序完成的,运用了链终止和断裂技术。第一代测序技术已经规模化,且具有测序读长较长、测序准确率高等特点,但是由于其时间长、成本高、通量低等缺点而无法满足现在实验需求。

2004 年,美国国家人类基因组研究所(NHGRI)发起了一项融资计划<sup>[4]</sup>,目标是在未来 10 年将人类基因组测序的费用减少到 1 000 美元,这刺激了商业化的新一代测序(next generation sequencing,NGS)技术的快速萌芽。第二代测序技术大幅提升基因组测序的输出与成本比,相较第一代测序技术

涵盖范围更加广泛,可以同时测定多个平行 DNA 片段,从而输出更多测序阅读量,但一般长度较短、质量低。第一个 NGS 技术是罗氏公司(Roche)公司的 454 焦磷酸测序法<sup>[5]</sup>,Solexa 和 Illumina 测序平台在 1 年后也进行商业化应用。第二代测序技术具有成本低、单次数据量大、消耗时间短等特点,故又被称为第二代高通量测序技术,并渐渐成为大规模全基因组测序技术的主导。

Roche 454 测序平台是基于焦磷酸测序,由于低数据率和相对较短的读取,最初适用于细菌基因组测序,随后技术改进,利用 Roche 454 与 Sanger 测序结合起来运用于更复杂的基因组测序,完成苹果基因组测序<sup>[6]</sup>,取代 Sanger 测序作为主要数据源完成可可基因组<sup>[7]</sup>和甜瓜<sup>[8]</sup>基因组测序。Illumina 公司提供的 Solexa 系统比 Roche 454 系统数据产出量多且花费较少的费用。自推出以来,读取长度获得明显改善,并成功运用到黄瓜的测序工作中<sup>[9]</sup>。

野生草莓<sup>[10]</sup>作为第一个采用 Roche 454、Illumina、SOLiDTM 三大平台共同完成测序的植物,标志着测序技术向多平台合作而不是独立运行的模式发展。近几年来,Illumina 测序成为第二代测序平台的主导,截至 2014 年 11 月,已为多种园艺作物如中国白菜<sup>[11]</sup>、马铃薯<sup>[12]</sup>、香蕉<sup>[13]</sup>、橙<sup>[14]</sup>和西瓜<sup>[15]</sup>基因组测序(表 1)。

近年来,新一代测序平台已经出现并被称为“第三代测序技术”,与第二代测序方法相比有着进一步模式转变,有 2 个突出特点,一是测序前不再需要 PCR 扩增,二是荧光或电流信号都可以在互补链加核苷酸的酶反应中被检测到。2 种测序平台现在已经投入商业化运营,分别是 Pacific Biosciences 公司的 PacBio RS 平台(<http://www.pacificbiosciences.com>)和 Ion Torrent 公司的 Personal Genome 平台(Life Technologies,<http://www.iontorrent.com>)。PacBio 测序平台使用实时零模式波导检测单个 DNA 聚合酶的活性<sup>[31]</sup>,大规模并行单分子实时(SMRT)测序保证了高通量测序。其突出特点是序列读长,经报道其单程长读取对的准确率为 84.2%~97.8%<sup>[32]</sup>,相较于第二代测序技术,第三代测序技术可以读取更长序列、提高测序通量、节省试剂成本,但其错误率比第二代测序技术高,3 代测序技术优缺点比较见表 2。

收稿日期:2015-02-06

基金项目:江苏省“青蓝工程”项目;江苏省科技支撑计划(编号:BE2012361);农业部种质资源保护项目(编号:2015NWB026)。

作者简介:倪晓鹏(1991—),男,从事果树分子生物学研究。

E-mail:2014104018@njau.edu.cn。

通信作者:高志红,博士,教授,从事果树种质资源与分子生物学研究。E-mail:gaozhihong@njau.edu.cn。

表 1 25 种全基因组测序园艺作物信息

作物	物种	科、属	测序材料	测序方法	基因组大小 (Mb)	期刊	发表时间 (年)
果树	葡萄 <sup>[16]</sup>	葡萄科葡萄属	PN 40024	Sanger, WGS	475	Nature	2007
果树	番木瓜 <sup>[17]</sup>	番木瓜科番木瓜属	Sun Up	Sanger, WGS	372	Nature	2008
蔬菜	黄瓜 <sup>[9]</sup>	葫芦科甜瓜属	“Chinese long” 9930	Sanger, Illumina GA	367	Nature Genetics	2009
果树	苹果 <sup>[6]</sup>	蔷薇科苹果属	Golden Delicious	Sanger, Roche 454, WGS	742	Nature Genetics	2010
果树	草莓 <sup>[10]</sup>	蔷薇科草莓属	Hawaii 4	454, SOLiD, Illumina WGS	240	Nature Genetics	2010
果树	可可 <sup>[7]</sup>	梧桐科可可属	Belizean Criollo	Sanger, Roche 454, Illumina WGS	430	Nature Genetics	2011
果树	海枣 <sup>[18]</sup>	棕榈科刺葵属	Dongzao	Illumina WGS	658	Nature Biotechnology	2011
果树	甜瓜 <sup>[8]</sup>	葫芦科甜瓜属	PI192	Roche 454	454	BMC Genetics	2011
蔬菜	马铃薯 <sup>[12]</sup>	茄科茄属	DM1 – 3 516 R44	Illumina, Roche 454 WGS	844	Nature	2011
蔬菜	白菜 <sup>[11]</sup>	十字花科芸薹属	Chiifu – 401 – 42	Illumina GA II	485	Nature Genetics	2011
蔬菜	番茄 <sup>[19]</sup>	茄科茄属	Heinz 1706	Roche 454	900	Nature	2012
果树	香蕉 <sup>[13]</sup>	芭蕉科芭蕉属	DH – Pahang Sanger	Roche 454, Illumina WGS	523	Nature	2012
果树	梨 <sup>[20]</sup>	蔷薇科梨属	Dangshansuli Illumina	HiSeq2000 BAC – by – BAC	527	Genome Research	2012
蔬菜	西瓜 <sup>[15]</sup>	葫芦科西瓜属	Chinese elite watermelon inbred line 97103	Illumina	425	Nature Genetics	2011
果树	甜橙 <sup>[14]</sup>	芸香科柑橘属	Illumina	GAII WGS	367	Nature Genetics	2012
果树	梅 <sup>[21]</sup>	蔷薇科李属	No. BJFU1210120008	Illumina WGS	280	Nature Communications	2012
果树	桃 <sup>[22]</sup>	蔷薇科梅属	Lovell	Sanger, WGS	265	Nature Genetics	2013
花卉	莲 <sup>[23]</sup>	睡莲科莲属	sacred lotus	Illumina, Roche 454	929	Genome Biology	2013
花卉	醉蝶花 <sup>[24]</sup>	醉蝶花科白菜花属	Purple Queen	Illumina, WGP	290	Plant Cell	2013
果树	猕猴桃 <sup>[25]</sup>	猕猴桃科猕猴桃属	Hongyang	Illumina	616	Nature Communications	2013
花卉	康乃馨 <sup>[26]</sup>	石竹科石竹属	Francisco	HiSeq 1000 + GS FLX	622	DNA Research	2013
蔬菜	辣椒 <sup>[27]</sup>	茄科辣椒属	Zunla – 1, Chiltepín	Illumina, WGS	3 480	Phas	2014
蔬菜	萝卜 <sup>[28]</sup>	十字花科萝卜属	Sayatori 26704, Aokubi S – h	Illumina	402	DNA Research	2014
蔬菜	茄子 <sup>[29]</sup>	茄科茄属	Nakate – Shinkuro, AE – P03, EPL – 1, LS1934, WCGR112 – 8	Illumina WGS	1 093	DNA Research	2014
花卉	蝴蝶兰 <sup>[30]</sup>	兰科蝴蝶兰属	<i>Phalaenopsis equestris</i>	Illumina WGS	1 160	Nature Genetics	2014

表 2 几种测序平台优缺点比较

测序平台	优点	缺点
Sanger	测序读取片段长, 准确性较高	技术复杂, 耗时、耗财, 组装存在缺口与错误
Roche 454	读取片段长(最大 1 kb), 更易映射到参考基因组, 易于 de novo 基因组组装, 运行时间短(23 h 以内)	相对输出量少(大约 100 万个阅读对, 700 Mb 测序数据), 试剂耗费高, 同聚物重复出错率高
Illumina/Solexa	输出量最高, 费用最低, 读取长度高达 300 bp, 兼容几乎所有应用	样品装载量低, 重叠集群超载和序列质量差, 序列复杂度要求高
SOLiD	输出量市面上仅次于 Illumina, 错误率低(0.06%)	所有平台最短的读取长度(最大 75 nt), 运行时间长, 样品准备试剂盒和服务发展不完善
Ion Torrent	半导体技术, 无需光学扫描和荧光核苷酸。运行速度快, 应用范围广	与 454 测序平台相同缺点(同聚物重复出错率高)
PacBio	读取长度极长(20 kb), 基因组组装的理想工具, 运行速度快	测序成本高, 总错误率高(~14%), 输出量最少(最小~50 Mb), 应用范围狭窄

2 全基因组测序的应用

2.1 基因注释

基因注释是利用生物信息学方法给测序完成的物种序列附上生物学信息的过程, 通过识别不编码蛋白质的基因片段, 识别基因上的元素(基因预测), 给元素附上生物学信息的手段进行重复序列的识别、非编码 RNA 的预测、基因结构预测和基因功能注释, 发现与农艺性状相关基因, 例如花期调控、植物生长习性、耐寒性、休眠、果实性状与品质等。Fukuoka 等通过鉴定苹果 (*Malus × domestica*) 的 146 个 *MADS – box* 基因, 与拟南芥和水稻 *MADS – box* 基因聚成 6 个亚组, 预测 *MADS – box* 基因在 17 条染色体上的密度, 在根、茎、叶、花组

织和果实发育的 5 个阶段进行了分析, 表明 *MADS – box* 基因参与了苹果的生理和发育过程的各个方面<sup>[33]</sup>。Yu 等将超过 6 000 个茄子组织和处理组的 cDNA 克隆组成 1 个含有 16 245 个独特序列的单基因组, 对单基因组进行测序并采用 SAGE 策略对茄子单基因集在转录组中的功能分析, 相当数量的短序列标签被成功注释<sup>[34]</sup>。

2.2 比较基因组学研究

比较基因组学研究是基于基因组图谱和测序技术, 对已知的基因和基因组结构进行比较以了解相关基因的功能、表达机制和物种亲缘关系。

番茄测序小组完成了对栽培番茄及其近缘野生种醋栗番茄全基因组的精细序列分析。在解码的番茄基因组中共鉴定

出 34 727 个基因,其中 97.4% 的基因已经精确定位到染色体上。通过比较基因组分析发现了番茄果实进化和发育的基因组学基础,番茄基因组经历的 2 次三倍化使基因家族产生了特异控制果实发育及营养品质的新成员<sup>[19]</sup>。Zhang 等研究发现在梅花基因组中存在和抗病相关的 *PR* 基因家族,由 *PR* 编码的蛋白质可介导植物产生防御反应,以此来抵抗病害和不良环境。例如:梅花基因组中存在的 *PR10* 与梅花的叶和根抗盐度、抗干旱和抗病菌的机制相关<sup>[21]</sup>。

### 2.3 重测序

通过基因组重测序,可以对栽培种和野生种基因组之间的差异进行比较,从而揭示物种起源以及驯化过程,为鉴定有价值的遗传资源以及园艺作物育种提供重要参考。

Qi 等对 115 个黄瓜品系进行了深度重测序,鉴定出 112 个假定的驯化延伸,其中 1 个区域包含着参与黄瓜果实苦味消失过程的 1 个基因,以及 1 个重要的驯化特性,也调查了栽培品种中分离度的基因组基础,在  $\beta$ -胡萝卜素羟化酶发现了一个自然遗传变异,可以用来培育含有更高营养价值的黄瓜,揭露黄瓜进化的遗传历程,为未来遗传育种提供了保证<sup>[35]</sup>。

### 2.4 全基因组关联分析

全基因组关联分析 (genome-wide association study, GWAS) 是应用基因组中数以百万计的单核苷酸多态性 (single nucleotide polymorphism, SNP) 作为分子遗传标记,进行全基因组水平上的对照分析或相关性分析,通过比较发现影响复杂性状的基因变异的一种新策略。随着基因组学研究以及基因芯片技术的发展,人们已通过 GWAS 方法发现并鉴定了大量与复杂性状相关联的遗传变异。近年来,这种方法在农业动物重要经济性状主效基因的筛查和鉴定中得到了应用。

Kenta 等为了研究栽培番茄遗传和表型多样性的关系,重测序 6 个栽培番茄品系,基于全基因连锁不平衡分析表明,连锁不平衡的延长依赖于染色质的性质,全基因组关联分析表明所鉴定的 SNP 与农艺重要的性状有着密切的关系,3 个基因 (*FAS*、*SP*、*U*) 被发现与性状变异相关,证明通过重测序得到的大量 SNP 进行全基因组关联分析可以鉴别特异性位点,表明栽培番茄在基因组学和遗传学发展上又迈进了一大步<sup>[36]</sup>。

### 2.5 与转录组数据比较分析

转录组研究是基因功能及结构研究的基础和出发点,是全基因组测序完成后首先要面对的问题。最近科学家们将高通量测序技术应用于转录组分析开发出了 RNA 测序技术 (RNA-Seq),该技术能够在全基因组范围内检测基因表达情况,进行差异基因筛选分析。由于 RNA-Seq 技术具有通量高、可重复性好、检测范围宽、定量准等特点,已经广泛应用于细菌、拟南芥、水稻和人类等生物转录组的研究。

Zenoni 等利用 RNA-Seq 技术对葡萄栽培种 Corvina 浆果发育 3 个阶段的转录组变化进行了研究,对比 Pinot Noir 40024 参考基因组,分析测量基因表达水平,共检测到 17 324 个基因在果实发育过程中表达,其中 6 695 个在不同阶段会特异性表达,这表明了基因功能的多样性与表达特异性以及转录组的高度复杂性<sup>[37]</sup>。

### 2.6 SNP 芯片的开发

单核苷酸多态性 (SNP) 根据其在基因中的位置,可以分为基因编码区、基因非编码区、基因间隔区 (基因之间的区

域),在基因组中是多态性最丰富的 DNA 分子标记,具有数量众多、分布均匀、分型方便等特点,它可以识别遗传变异和关联的表型基因分型。

Chagné 等利用生物信息学工具,从包含苹果品系中 350 000 条序列的 EST 数据库中开发 SNP 标记,结果识别了 71 482 个假定 SNP 分子标记。设计了 464 个 PCR 引物对,对 PCR 产物进行测序,重新获得的 SNP 标记映射到苹果参考基因 (Royal Gala  $\times$  A689 - 24 杂交系和 Malling 9  $\times$  Robusta 5) SNP 基因分型采用高分辨率溶解 (HRM) 技术,共包含 210 个编码 SNP 的 93 个新标记被成功映射到参考图谱。此方法为使用数量性状定位 (QTL) 技术了解重要农艺性状和连锁不平衡分析的基因调控提供了借鉴,也成为物理和遗传图谱结合的有效标记<sup>[38]</sup>。

## 3 全基因组测序存在的挑战

测序任何基因组的挑战包括倍性和杂合度。与动物的基因组相比,植物基因组具有巨大的基因组大小、高度重复序列和全基因组或者片段基因组复制。

第二代测序平台 Roche 454、Illumina 和 SoLiD 不仅吞吐量显著增加,还减少了错误率,增加了阅读长度,使大范围物种测序都能采用测序平台经济地进行。然而,植物基因组的独特性给测序带来了许多挑战,如他们的重复特性,装配完整基因组是具有挑战性的,这是由于植物基因组转座因子的高拷贝数和扩张特性决定的。例如,玉米基因组测序通过 BAC-by-BAC 方法,基因组中 85% 含有转座元件测序<sup>[39]</sup>。植物基因组中全基因组、节段和串联重复的频率<sup>[40]</sup>也给旁系同源种类带来组装问题,如果是最近一次复制,其序列同源性会高。因此,尽管第二代测序平台能够测序一大部分基因组,但迄今为止的全基因组测序方案还是缺乏足够的代表性,质量评估在组装过程中还会丢失部分基因。

另一个问题是,不是所有植物品种都是纯合二倍体自交系,尽管使用 WGS 方法为杂合组织测序是可行的。然而高杂合度,包括交叉和无性繁殖的物种,如马铃薯,会阻碍 WGS 组装。为了解决葡萄基因组 (高度杂合二倍体) 问题,Jaillonet 等采用了高度近交的品种,减少了杂合程度,进而才完成了测序。使用长阅读可以提高一个基因组中组装分离单体型的能力<sup>[16]</sup>。

植物倍性也是在植物 denovo 测序和组装的一个巨大挑战,其结果取决于物种是一个同源多倍体是异源多倍体,迄今为止,所有多倍体测序依赖于倍性降低或者染色体的物理分离。例如,绝大多数马铃薯品系是四倍体,最初是取杂合二倍体马铃薯基因 (RH89 - 039 - 16) 进行测序从而避免同源染色体之间的高度杂合性,最后利用 1 个特殊的马铃薯基因型,1 个加倍的单倍体 (DM1 - 3 516 R44),含有纯合的 12 条染色体作为参考马铃薯基因组<sup>[12]</sup>。栽培草莓 (*Fragaria ananassa*) 是来自 4 个不同祖先的异源八倍体,野生草莓 (*Fragaria vesca*) 是二倍体,被用来测序是为了避免测序多个基因组带来的困难<sup>[10]</sup>。

## 4 全基因组测序的展望

为了面对世界基因组测序技术发展如此迅猛的挑战,大量全基因组测序数据亟待深度挖掘,应制定长远的战略性基

基因组学研究计划,不仅局限于栽培物种,更应该深度开发我国重要野生近缘物种的测序,促进重要基因资源的挖掘、保护和利用。摆脱先前依赖外观表现型而转入到基因型依赖型的研究当中去,从单一基因研究深入到全基因组关联分析研究中。大力推动我国基因组学在基因克隆与分子育种领域的应用研究,提高园艺作物育种能力和水平。加强与转录组学、代谢组学、蛋白质组学和降解组学的相关研究,促进基因组学生物信息的共享与利用。加强生物信息学教育的投入,并将其应用于实践中。随着测序技术的不断发展,相信在不久的将来园艺作物全基因组测序将进入快速发展的阶段,为世界园艺产业带来巨大贡献。

#### 参考文献:

- [1] 中华人民共和国农业部. 中国农业统计资料:2013[M]. 北京: 中华人民共和国农业部,2014.
- [2] Collins F S, Morgan M, Patrinos A. The human genome project: lessons from large - scale biology [J]. Science, 2003, 300 ( 5617 ) : 286 - 290.
- [3] Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana* [J]. Nature, 2000, 408 ( 6814 ) : 796 - 815.
- [4] Schloss J A. How to get genomes at one ten - thousandth the cost [J]. Nature Biotechnology, 2008, 26 ( 10 ) : 1113 - 1115.
- [5] Margulies M, Egholm M, Altman W E, et al. Genome sequencing in microfabricated high - density picolitre reactors [J]. Nature, 2005, 437 ( 757 ) : 376 - 380.
- [6] Velasco R, Zharkikh A, Affourtit J, et al. The genome of the domesticated apple (*Malus × domestica* Borkh.) [J]. Nature Genetics, 2010, 42 ( 10 ) : 833 - 839.
- [7] Argout X, Salse J, Aury J M, et al. The genome of *Theobroma cacao* [J]. Nature Genetics, 2011, 43 ( 2 ) : 101 - 108.
- [8] Rodríguez - Moreno L, González V M, Benjak A, et al. Determination of the melon chloroplast and mitochondrial genome sequences reveals that the largest reported mitochondrial genome in plants contains a significant amount of DNA having a nuclear origin [J]. BMC Genomics, 2011, 12: 424.
- [9] Huang S, Li R, Zhang Z, et al. The genome of the cucumber, *Cucumis sativus* L. [J]. Nature Genetics, 2009, 41 ( 12 ) : 1275 - 1281.
- [10] Shulaev V, Sargent D J, Crowhurst R N, et al. The genome of woodland strawberry (*Fragaria vesca*) [J]. Nature Genetics, 2011, 43 ( 2 ) : 109 - 116.
- [11] Wang X, Wang H, Wang J, et al. The genome of the mesopolyploid crop species *Brassica rapa* [J]. Nature Genetics, 2011, 43 ( 10 ) : 1035 - 1039.
- [12] Potato Genome Sequencing Consortium, Xu X, Pan S, et al. Genome sequence and analysis of the tuber crop potato [J]. Nature, 2011, 475 ( 7355 ) : 189 - 195.
- [13] D'hont A, Denoeud F, Aury J M, et al. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants [J]. Nature, 2012, 488 ( 7410 ) : 213 - 217.
- [14] Xu Q, Chen L L, Ruan X, et al. The draft genome of sweet orange (*Citrus sinensis*) [J]. Nature Genetics, 2013, 45 ( 1 ) : 59 - 66.
- [15] Guo S, Zhang J, Sun H, et al. The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions [J]. Nature Genetics, 2013, 45 ( 1 ) : 51 - 58.
- [16] Jaillon O, Aury J M, Noel B, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla [J]. Nature, 2007, 449 ( 7161 ) : 463 - 467.
- [17] Ming R, Hou S, Feng Y, et al. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus) [J]. Nature, 2008, 452 ( 7190 ) : 991 - 996.
- [18] Al - Dous E K, George B, Al - Mahmoud M E, et al. De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*) [J]. Nature Biotechnology, 2011, 29 ( 6 ) : 521 - 527.
- [19] Sato S, Tabata S, Hirakawa H, et al. The tomato genome sequence provides insights into fleshy fruit evolution [J]. Nature, 2012, 485 ( 740 ) : 635 - 641.
- [20] Wu J, Wang Z, Shi Z, et al. The genome of the pear (*Pyrus bretschneideri* Rehd.) [J]. Genome Research, 2013, 23 ( 2 ) : 396 - 408.
- [21] Zhang Q, Chen W, Sun L, et al. The genome of *Prunus mume* [J]. Nature Communications, 2012, 3: 1318.
- [22] International Peach Genome Initiative, Verde I, Abbott A G, et al. The high - quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution [J]. Nature Genetics, 2013, 45 ( 5 ) : 487 - 494.
- [23] Ming R, Vanburen R, Liu Y, et al. Genome of the long - living sacred lotus (*Nelumbo nucifera* Gaertn.) [J]. Genome Biology, 2013, 14 ( 5 ) : R41.
- [24] Cheng S, van den Bergh E, Zeng P, et al. The *tarenaya hassleriana* genome provides insight into reproductive trait and genome evolution of crucifers [J]. The Plant Cell, 2013, 25 ( 8 ) : 2813 - 2830.
- [25] Huang S, Ding J, Deng D, et al. Draft genome of the kiwifruit *Actinidia chinensis* [J]. Nature Communications, 2013, 4: 2640.
- [26] Yagi M, Kosugi S, Hirakawa H, et al. Sequence analysis of the genome of carnation (*Dianthus caryophyllus* L.) [J]. DNA Research: an International Journal for Rapid Publication of Reports on Genes and Genomes, 2014, 21 ( 3 ) : 231 - 241.
- [27] Qin C, Yu C, Shen Y, et al. Whole - genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization [J]. Proceedings of the National Academy of Sciences of the United States of America, 2014, 111 ( 14 ) : 5135 - 5140.
- [28] Kitashiba H, Li F, Hirakawa H, et al. Draft sequences of the radish (*Raphanus sativus* L.) genome [J]. DNA Research: an International Journal for Rapid Publication of Reports on Genes and Genomes, 2014, 21 ( 5 ) : 481 - 490.
- [29] Hirakawa H, Shirasawa K, Miyatake K, et al. Draft genome sequence of eggplant (*Solanum melongena* L.): the representative solanum species indigenous to the Old World [J]. DNA Research: an International Journal for Rapid Publication of Reports on Genes and Genomes, 2014, 21 ( 6 ) : 649 - 660.
- [30] Cai J, Liu X, Vanneste K, et al. The genome sequence of the orchid *Phalaenopsis equestris* [J]. Nature Genetics, 2015, 47 ( 1 ) : 65 - 72.
- [31] Eid J, Fehr A, Gray J, et al. Real - time DNA sequencing from single polymerase molecules [J]. Science, 2009, 323 ( 5910 ) : 133 - 138.
- [32] Rasko D A, Webster D R, Sahl J W, et al. Origins of the *E. coli* strain causing an outbreak of hemolytic - uremic syndrome in Germany [J]. The New England Journal of Medicine, 2011, 365 ( 8 ) : 709 - 717.

韩俊杰,王昊龙,李卫华. 关联分析及其在不同分子标记中的应用综述[J]. 江苏农业科学,2016,44(2):13-17.  
doi:10.15889/j.issn.1002-1302.2016.02.004

# 关联分析及其在不同分子标记中的应用综述

韩俊杰,王昊龙,李卫华

(石河子大学农学院/新疆生产建设兵团绿洲生态农业重点实验室,新疆石河子 832003)

**摘要:**关联分析技术在植物中的应用较晚,近些年才在植物数量性状和植物遗传育种研究中得以应用。关联分析以连锁不平衡作图(LD mapping)为基础,用来鉴定植物性状与目标基因之间的关系。关联分析有很明显的应用优势,它不仅不需要特殊的构图群体,而且还能够同时对多个基因位点进行检测,精度非常高。由于其良好的实用性,关联分析技术已经成为国际基因组学的研究热点。介绍了 LD 的定义和度量方法,综述了基于分子标记的关联分析在植物遗传中的应用,并对其在遗传学中的应用和发展作了展望。

**关键词:**植物;分子标记;关联分析;连锁不平衡;遗传学;应用

**中图分类号:** Q78      **文献标志码:** A      **文章编号:** 1002-1302(2016)02-0013-05

关联分析又称连锁不平衡作图(linkage disequilibrium mapping, LD mapping),是建立在连锁不平衡的基础上用于鉴定群体内目标性状与遗传标记相关关系的方法<sup>[1]</sup>。关联分析最早在人类疾病的研究中得以应用,并取得了丰硕的成果,目前在人类不同致病基因方面已有许多成功的报道<sup>[2]</sup>。在动物中的应用也相对较多,如在奶牛、鱼、羊、猪、鸡中的应用等。然而,由于目前对植物基因组 LD 结构的认识还不够全面,导致关联分析在植物中的应用相对较少<sup>[3]</sup>。但是,随着生物信息学、统计学和基因组学技术的快速发展,关联分析也已成功应用到许多植物上,如玉米、小麦、拟南芥、大麦、大豆等。近年来,许多植物的全基因组测序已经完成,这极大地促

进了基因组学的发展,特别是近年来大量单核苷酸多态性(single nucleotide polymorphisms, SNP)标记的开发,关联分析在植物关键基因的发掘上更是得到了广泛的应用,可为植物分子育种提供标记,在国际上已经成为热点研究方向。

关联分析通过群体上的进化历史和重组,可以在序列水平上来研究复杂的数量性状,因此与其他分析手段相比,关联分析有其独特的优势:(1)所需时间短,用自然群体作为材料,不需要专门构建,且有较大的选择范围,相对于数量性状基因座(QTL)至少需要 2 年以上的群体能节约大量的时间;(2)检测范围大,同一座位的不同基因可以用来同时检测,而 QTL 作图只能检测来自双亲的 2 个等位基因;(3)准确度高, QTL 作图精度一般在 10~30cM 左右,很少能在基因水平上被标记,而关联分析可达到单基因的水平<sup>[4-5]</sup>,如 Gebhardt 等利用关联分析方法将马铃薯抗晚疫病基因 *RI* 定位在 0.2cM 的基因组区段内<sup>[6]</sup>。在植物相关研究中,由于其杂交方式和重复类型都可以得到人为有效控制,因此关联分析在植物中的应用比在动物中的吸引力更高。

作物很多重要的农艺性状(如株高、产量、营养品质和抗

收稿日期:2015-01-05

基金项目:国家自然科学基金(编号:31260357)。

作者简介:韩俊杰(1989—),男,河南开封人,硕士,主要从事小麦品质遗传改良研究。E-mail:hanjunjie1208@sina.com。

通信作者:李卫华,博士,教授,主要从事小麦品质改良和分子机理研究。E-mail:lw\_h\_agr@shzu.edu.cn。

[33] Fukuoka H, Yamaguchi H, Nunome T, et al. Accumulation, functional annotation, and comparative analysis of expressed sequence tags in eggplant (*Solanum melongena* L.), the third Pole of the genus *Solanum* species after tomato and potato[J]. *Gene*, 2010, 450(1/2): 76-84.

[34] Yu Y, Liang Y, Lv M, et al. Genome-wide identification and characterization of polygalacturonase genes in *Cucumis sativus* and *Citrullus lanatus*[J]. *Plant Physiology and Biochemistry*, 2014, 74: 263-275.

[35] Qi J, Liu X, Shen D, et al. A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity[J]. *Nature Genetics*, 2013, 45(12): 1510-1515.

[36] Shirasawa K, Fukuoka H, Matsunaga H, et al. Genome-wide association studies using single nucleotide polymorphism markers developed by re-sequencing of the genomes of cultivated tomato[J]. *DNA Research; an International Journal for Rapid Publication of*

*Reports on Genes and Genomes*, 2013, 20(6): 593-603.

[37] Zenoni S, Ferrarini A, Giacomelli E, et al. Characterization of transcriptional complexity during berry development in *Vitis vinifera* using RNA-Seq[J]. *Plant Physiology*, 2010, 152(4): 1787-1795.

[38] Chagné D, Gasic K, Crowhurst R N, et al. Development of a set of SNP markers present in expressed genes of the Apple[J]. *Genomics*, 2008, 92(5): 353-358.

[39] Schnable P S, Ware D, Fulton R S, et al. The B73 maize genome: complexity, diversity, and dynamics[J]. *Science*, 2009, 326(5956): 1112-1115.

[40] Paterson A H, Bowers J E, Chapman B A. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2004, 101(26): 9903-9908.