

高通量测序技术在转座子研究中的应用

刘振, 徐建红

浙江大学农业与生物技术学院作物科学研究所, 浙江省作物种质资源重点实验室, 杭州 310058

摘要: 高通量测序技术极大地提高了测序效率, 大幅度降低了测序成本, 同时该技术具有特异性好、灵敏度高、精确性高等优势, 目前已被广泛应用于遗传变异、转录组学和表观组学等研究。近年来, 高通量测序技术也逐渐应用于转座子的研究, 并取得了丰硕的成果。本文主要综述了高通量测序技术在转座子研究中的应用, 包括转座子含量估算、靶点偏好性及分布、多态性及群体频率、稀有转座子的鉴定、转座子的水平转移以及转座子标签技术中的应用等, 并简要介绍了目前研究中采用的主要测序策略和算法, 及其存在的利弊和相应的解决方案。最后对高通量测序技术, 尤其是第三代测序技术的发展趋势和它们在转座子未来的研究中的应用进行了展望, 以期相关的科研人员提供一个全面的了解和参考。

关键词: 高通量测序; 转座子; 偏好性; 多态性; 水平转移

The application of the high throughput sequencing technology in the transposable elements

Zhen Liu, Jianhong Xu

Key Laboratory of Crop Germplasm of Zhejiang Province, Institute of Crop Science, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, China

Abstract: High throughput sequencing technology has dramatically improved the efficiency of DNA sequencing, and decreased the costs to a great extent. Meanwhile, this technology usually has advantages of better specificity, higher sensitivity and accuracy. Therefore, it has been applied to the research on genetic variations, transcriptomics and epigenomics. Recently, this technology has been widely employed in the studies of transposable elements and has achieved fruitful results. In this review, we summarize the application of high throughput sequencing technology in the fields of transposable elements, including the estimation of transposon content, preference of target sites and distribution, insertion polymorphism and population frequency, identification of rare copies, transposon horizontal transfers as well as transposon tagging. We also briefly introduce the major common sequencing strategies and algorithms, their advantages and disadvantages, and the corresponding solutions. Finally, we envision the developing trends of high throughput sequencing technology, especially the third generation sequencing technology, and its application in transposon studies in the future, hopefully providing a comprehensive understanding and reference for related scientific researchers.

Keywords: high throughput sequencing; transposable element; bias; polymorphism; horizontal transfer

收稿日期: 2015-04-02; 修回日期: 2015-05-14

基金项目: 国家自然科学基金项目(编号: 31171165)和中央高校自主科研计划(编号: 2014QNA6019)资助

作者简介: 刘振, 博士研究生, 专业方向: 作物遗传育种。E-mail: 13960797234@126.com

通讯作者: 徐建红, 博士, 教授, 研究方向: 基因组学与分子生物学。E-mail: jhxu@zju.edu.cn

DOI: 10.16288/j.ycz.15-140

网络出版时间: 2015-7-15 16:55:11

URL: <http://www.cnki.net/kcms/detail/11.1913.R.20150715.1655.002.html>

转座子是基因组中一类可移动的 DNA 片段,能够从基因组中的一个位置移动到另一个位置。转座子的发现打破了遗传物质在染色体上呈线性固定排列的传统观念,对遗传学和分子生物学的发展具有深远意义^[1]。转座子几乎存在于所有的真核生物中,并在基因组中占有较大比例,例如转座子占哺乳动物基因组的 50% 以上,而在植物基因组中甚至可高达 85% 以上^[2~4]。根据转座子的结构和转座机制将其分为两类: 型转座子,又称反转录转座子,以 RNA 为中间媒介进行转座,属于“复制-粘贴”型,每转座一次可增加一个拷贝,在植物基因组中的含量较为丰富; 型转座子,也称作 DNA 转座子,以 DNA 为中间体进行转座,属“剪切-粘贴”型,该类转座子在动物(特别是哺乳动物)基因组中占有较高的比例(图 1)。另外,根据结构的完整性和转座机制的不同又将转座子分为自主转座子和非自主转座子,前者本身能够转座,而后者只能依赖于相应的自主转座子进行转座^[5]。

转座子是物种发生遗传变异的重要来源,在生物的进化中发挥着重要作用^[6]。它们在基因组中不仅能够影响基因的表达,改变基因的结构,重排染

色体,改变基因组的大小,而且还能够维持基因组的稳定性,保持异染色质的沉默以及介导基因的水平转移等^[7~13]。此外,转座子在生物的遗传进化和功能基因组学的研究中同样也发挥着举足轻重的作用。例如,通过转座子插入多态性的研究不仅可以了解物种间的进化关系而且还可用于分子标记的开发和基因的图位克隆;将转座子作为分子标签也能研究基因的功能和基因间的互作^[14~16]。

为了充分了解转座子的生物学功能以及它们在遗传学和功能基因组学研究中的应用,以往通常采用 PCR、分子杂交、芯片以及 Sanger 测序的方法对转座子进行分析和研究。尽管这些方法在转座子的研究中已取得了丰硕的成果,但也存在着诸多缺陷,如工作量大、周期长、精度低、费用高等。而高通量测序技术的出现在很大程度上缓解了这些问题,而且还为转座子的研究开辟了新的领域,注入了新的活力。

高通量测序技术(也称为二代测序技术)能够同时对数百万个序列进行大规模的平行测序,极大地提高了测序效率,大幅度降低了测序成本^[17]。目前商业化的高通量测序平台主要有 Roche454、Illumina/ Solexa、Life/APG、Helicos BioSciences 以及 Pacific Biosciences 等^[18~22]。不同的测序平台通常拥有不同的特点,因而可满足不同的科研需求。例如, Roche454 可产生高达 500 bp 的测序读长,较适合全基因组的拼接和组装; Illumina/Solexa 和 Life/APG 测序平台每个运行周期能够产生相当大的数据量,适用基因组的重测序;而 Helicos BioSciences 可直接对 RNA 进行测序,省去了常规测序方法中的 RNA 反转录过程^[22,23]。高通量测序技术目前被广泛地应用于生物学的多个领域,如基因组测序和重测序、转录组测序(RNA-seq)、表观组测序(MeDIP-seq)以及染色质相关的研究(ChIP-seq、DNase-seq)^[19,23~29]。近年来,该技术也被普遍地应用到转座子的研究中并取得了许多重大成果:构建转座子的多态性图谱^[30];检测与癌症相关的稀有转座子插入^[31];揭示转座子进行水平转移的载体^[32];证实转座子偏好性分布的机制^[33]等。本文主要综述了高通量测序技术在转座子研究中的应用现状、采用的测序策略和算法、存在的利弊和相应的解决方案,以及测序技术将来的发展趋势,以期对相关科研人

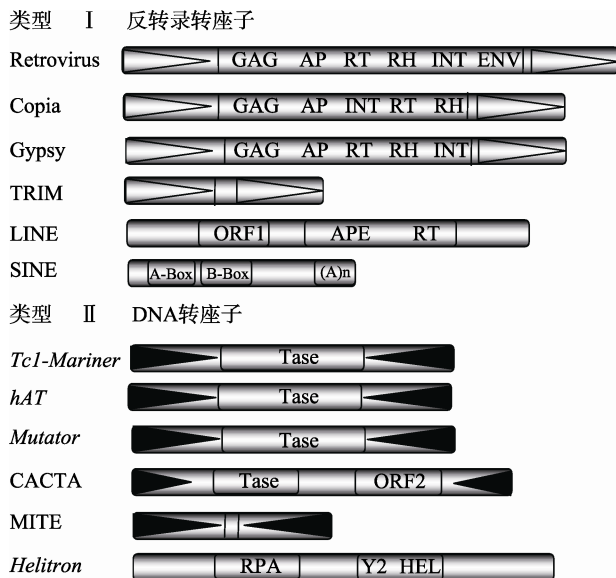


图 1 转座子的类型和结构(参考文献^[5]并修改)

AP: Aspartic proteinase; APE: Apurinic endonuclease; ENV: Envelope protein; GAG: Capsid protein; HEL: Helicase; INT: Integrase; RT: Reverse transcriptase; RH: RNase H; RPA: Replication protein A; Tase: Transposase; Y2: YR with YY motif; ▴: 表示 LTR (Long terminal repeat); ▴: 表示 TIR (Terminal inverted repeat)。

员提供全面的了解。

1 高通量测序技术在转座子研究中的应用

1.1 估算转座子的含量

不同真核生物的基因组大小存在着巨大的差异,尤其在高等植物中,基因组从螺旋狸藻(*Genlisea margaretae* L.)的 63 Mb 到重楼百合(*Paris japonica* L.)的 150 Gb 不等^[34, 35]。物种间基因组的差异主要取决于两方面的因素:染色体的多倍化和转座子等重复序列的扩增^[36~40]。通常较大的基因组往往含有较高比例的转座子等重复序列,即使在亲缘关系较近的物种内,转座子的含量也存在着很大的差异。如澳洲野生稻(*Oryza australiensis* L.)中 3 个反转录转座子家族的急剧扩增使其基因组的大小是栽培种的两倍^[36]。因此,对不同物种中转座子含量的分析是基因组学研究中一个至关重要的方面,将有助于揭示基因组的结构组成以及系统进化。

高通量测序技术为基因组的分子组成和结构特征的研究提供了一个便利的工具,尤其是对于较为复杂的基因组(多倍性或高重复序列含量)更是如此。该技术不仅适用于具有全基因组参考序列的物种(如水稻、玉米)的转座子研究,而且对于仅存在有限基因组序列的物种(如香蕉、豌豆)依然适用^[41~44]。通过对向日葵(*Helianthus annuus* L.)基因组进行 454 测序,发现重复序列在基因组中的含量高达 81%,并且其中大部分为 *Gypsy* 和 *Copia* 类型的反转录转座子^[45];通过对玉米(*Zea mays* L.)自交系 B73 和繁茂玉米(*Zea luxurians*)进行低覆盖度的 Illumina 测序,分别估算了这两个物种中转座子的含量,并证实它们之间 70% 的基因组大小差异是由转座子所引起^[42];同样,对其他两个玉米品系的测序分析发现转座子的扩增贡献了高达 76% 的序列差异^[46]。另外,部分物种的基因组呈现出染色体倍性高且基因组中重复序列比重大的特性,这无疑对全基因组的测序、拼接组装以及转座子的研究提出了严峻的挑战。面包麦(*Triticum aestivum* L.)的基因组大小约为 17 Gb,重复序列所占比重高达 90%,其中 80% 来自于转座子序列^[47, 48]。然而,将高通量测序技术同染色体分离技术相结合可以对多倍体物种进行单条染色体的测序,从而大大降低序列拼接组装的难度。目前采用该策略已成功地对普通小麦 5B 染色体进行了测序,

其中的转座子约占 70%,并且 *Cereba* 反转录转座子家族在该染色体长短臂间的含量存在差异^[49]。

利用高通量测序数据获得的转座子含量的信息与传统方法分析所得结果有着较高的一致性^[41, 42],说明该方法可靠性较好,当然也存在着一定的弊端。在玉米基因组的高通量数据分析中,发现转座子相对应的读序列数同整个家族在基因组中的含量之间呈正相关,但是这些读序列数同转座子的拷贝数之间并不存在显著的相关性,可能是由于不同拷贝间的大小差异所引起^[42]。该方法目前仅适用于转座子序列注释较好的模式生物,并且转座子数据库的完整性和准确性对转座子含量的估算产生较大的影响。另外,很多转座子家族(如 Helitron、Pack-MULE)经常携带基因片段,转座子与基因间的界限很难明确区分,因此利用该方法计算出的转座子含量仅是估计值。虽然利用高通量数据估算重复序列的含量存在一些缺点,但随着算法的优化以及转座子注释的完善,这些问题将会得到妥善的解决。

1.2 转座子的靶点偏好性及分布

不同类型的转座子通常采用不同的机制整合到基因组中,因此对整合位点的选择往往具有偏好性^[50~54]。了解转座子整合位点的性质将有助于阐明转座机制。对转座子整合位点的研究一般是通过分析插入位点处侧翼序列来实现的,目前用于该方面研究的转座子拷贝无非来源于人工诱导或自然发生的转座事件。利用人工诱导的转座子拷贝进行靶点偏好性分析时,可提供大量的整合位点数据,但是并不能确定由此产生的结果能否真实的反映出自然发生的转座特征,并且该方法仅适用于少数有活性的转座子家族。而利用自然发生的转座子拷贝进行分析时,可供分析的样本量通常较小,并且不能够获取转座子插入之前的序列信息,然而此类信息对于确定目标位点重复(Target site duplication, TSD)的长度以及目标位点模体(Target site motif, TSM)的序列往往至关重要^[55]。由此可见,现存的方法均存在着一定局限性,因而有必要开发新的研究方法用于转座子靶点偏好性的分析。

近年来,利用高通量测序数据分析转座子靶点偏好性的方法不仅突破了传统方法的局限性,而且还拥有其他多个方面的优势。Linheiro 等^[55]利用果

蝇(*Drosophila melanogaster*)的群体基因组学数据从 166 个果蝇品系中共鉴定到了 8000 个非参考的转座子插入位点,并基于这些位点分析了 22 个转座子家族(主要为 TIR 转座子和 LTR 反转录转座子)的靶点偏好性,结果表明同一分支的转座子家族通常具有相似的 TSD 和 TSM,并且这些转座子家族的 TSM 均呈现出回纹结构并延伸到 TSD 序列之外。这些结果为分析已知的转座子家族提供重要的特征信息,或许还适用于未研究的转座子家族。与此同时,该研究证实高通量测序数据应用于转座子靶点偏好性的分析具有如下的优点:(1)此方法基于自然发生的转座事件,因此可真实地反映出转座的特性;(2)能够从众多的品系中鉴定出大量的转座子拷贝,因而可供分析的样本量较大,提高了结果的精确性;(3)可对转座子插入前后的序列进行比较,从而为精确地确定 TSD 和 TSM 提供了一个重要的参考;(4)利用同一测序数据和算法可同时对几乎所有的转座子家族进行分析,实现了分析的高通量。

此外,高通量测序数据也被用于研究转座子在基因组中的分布。大量研究表明,许多转座子家族偏好于插入到基因组的异染色质区域^[56-60]。当转座子插入到富含基因的常染色质区域时,它们更可能引起不利的突变,因而受到较大的自然选择压,最终使该区域的转座子拷贝发生丢失^[61]。而新插入的转座子拷贝由于插入时间较短,受到的自然选择压有限,因而它们的分布主要由插入的特性来主导^[62]。通过对大豆栽培种(*Glycine max* L.)和野生种(*Glycine soja* L.)的重测序数据分析^[33],揭示 LTR 反转录转座子在异染色质区域的富集主要是由于它们的偏好性插入所引起,而 DNA 转座子主要是由于常染色质区域受到更多的自然选择所导致。对人(*Homo sapiens* L.)LINE 转座子的研究发现,新插入的转座子更倾向于某个群体或者某个群体的某个种族中^[63]。随着越来越多的物种进行群体基因组的重测序,这些方法很可能揭示更多转座子靶点偏好性以及分布规律,从而为转座机理的研究提供依据和参考。

1.3 揭示转座子的多态性以及群体频率

转座子在物种间或物种内往往呈现出插入多态性,并且该特性可作为分子标记应用于遗传进化分析,基因的图位克隆以及 DNA 指纹图谱的构建等研

究中^[16,64,65]。目前基于转座子的多态性已开发出多种分子标记系统,如 SSAP(Sequence specific amplified polymorphism)^[66,67]、IRAP(Inter retrotransposon amplified polymorphism)^[68]、REMAP(Retrotransposon microsatellite amplified polymorphism)^[68]以及 RBIP(Retrotransposon based insertional polymorphism)^[69]等。然而,尽管生物体基因组中存在着大量的转座子,但并非所有的转座子家族都适用于分子标记的开发。因此,鉴定出具有适度插入多态性的转座子家族是一项必要而又棘手的研究。

研究证实利用高通量的测序数据能够鉴定到大量具有多态性的转座子位点或非参考的转座子位点,并且基于这些位点可系统地研究转座子在基因组中的偏好性分布,突变频率以及群体分离等性质^[33,62,63,70-73]。通过对 185 例人类基因组的高通量数据进行分析,鉴定到 7380 个多态性的转座子插入位点,其中有 1/3 的位点在之前的研究中从未涉及^[30];对果蝇高通量的测序数据进行分析,在 146 个已测序的品系中共鉴定到多达 23 087 个多态性的转座子插入位点^[74];另外,通过对 17 个小鼠(*Mus musculus* L.)品系的基因组进行重测序同样也发现了数千个 SINE、LINE、ERV 转座子的多态性插入位点^[75]。同时,采用高通量测序技术研究转座子的插入多态性不仅具有通量高的优势,而且在特异性和精确度方面较传统方法也有很大的改善。例如,利用反转录转座子中高保守性的 PBS(Primer binding site)序列进行测序文库的构建,可大幅地提高多态性转座子鉴定的灵敏度,尤其对于传统方法难以捕捉到的位点来讲更是如此^[70]。本课题组利用玉米高通量测序数据分析了一个 TRIM 家族在不同品系中的插入情况,发现该 TRIM 家族的部分拷贝在 75 个玉米品系中呈现出丰富的插入多态性;并且基于这些多态性插入位点进一步分析了品系间的系统进化关系,不仅能将不同的群体,不同的品系正确地区分开来,而且还可将来自同一群体的品系在进化树中较好地聚类在一起。由此暗示通过高通量数据能够研究转座子的多态性和系统进化关系^[76]。此外,由于在高通量数据分析的过程中常会遇到多种问题和挑战,为了更好地将该技术应用于转座子多态性方面的研究,目前已开发出了多种算法和软件,如 T-lex^[77]、RelocaTE^[78]、ME-Scan^[79]、VariationHunter^[80]、Tangram^[81]等。尽管这

些软件或许有着不同的适用条件和范围,但它们的出现无疑促使了该技术的广泛应用。

高通量测序数据也被普遍地应用于研究转座子的群体频率^[77, 82]。例如, Kofler 等^[82]通过对 113 个果蝇品系进行全基因组重测序分析了 7843 个转座子位点的群体频率,结果表明不同转座子家族的群体频率往往存在较大的差异,并基于该频率数据进一步鉴定到了 13 个可能受到正向选择作用的转座子位点。相对于传统的方法,该方法主要有以下两方面的优势:(1)对群体频率的估算不存在偏好性。以往采用 PCR 方法估算转座子的群体频率时,往往倾向于扩增群体频率较高的转座子拷贝从而引入了一定的偏好性,而高通量测序的方法能够同等地对待参考和非参考的转座子拷贝从而避免了偏好性的产生;(2)拥有较高的精度。利用原位杂交的方法对转座子的群体频率进行分析时虽然不存在偏好性的问题,但是这种方法的精度有限并且仅能发现完整的转座子拷贝,然而高通量测序的方法能灵敏的捕捉到转座子的插入位点并且对转座子的完整性要求较低^[82]。鉴于上述优势,可预见高通量测序技术在转座子群体频率的研究方面将拥有更广阔的应用前景。

1.4 鉴定稀有的转座子拷贝

稀有转座子通常是指在整个群体中分布频率较低,甚至仅存在于某一品系或个体中的转座子拷贝,它们往往是由具有活性的转座子家族在最近的进化时期内发生转座的结果。研究已证实稀有转座子普遍地存在于真核生物的基因组中,并且它们在群体中的总量甚至远远高于普通的转座子^[72,74,82,83]。由于稀有转座子插入基因组中的时间通常较短,并且未完全受到自然选择的影响,因而在人与疾病相关的可能性更大,由此暗示了稀有转座子或许拥有更重要的研究价值^[84~86]。目前大量的研究证实稀有转座子的插入往往会伴随着多种复杂性状或疾病的发生(尤其是癌症,如肺癌、肠癌、皮肤癌、子宫癌等)^[31,83,87]。然而出于技术的限制鉴定稀有转座子的难度往往较大,因而相关方面的进展也较为缓慢。

高通量测序技术的出现为稀有转座子的研究提供了新的契机,利用它可方便地从群体基因组数据中鉴定到大量的稀有转座子拷贝。例如,通过对 5 种癌症类型的 43 例样本进行全基因组深度测序,鉴

定到了 194 个可信度高的稀有转座子拷贝,并且这些拷贝多插入于与癌症相关的基因中从而改变基因的表现修饰或表达水平。进一步的研究表明,稀有转座子在不同类型的癌症之间呈现显著偏好性分布,如它们多集中在皮肤癌样本中,而血癌和脑瘤中几乎不存在^[31]。此后, Helman 等^[87]对 11 种癌症的 200 例样本进行了相似的研究,同样也发现了多达 810 个稀有转座子的插入位点,并且其中的大多数位点来自于肺癌和头颈癌样本的基因组。另外,研究人员还采用了外显子测序技术对其他的 767 个肿瘤样本进行深度测序,结果鉴定到了 35 个插入到基因外显子中的稀有转座子拷贝。这一系列的研究证实了稀有转座子的确普遍存在于基因组中,并且积极参与相关基因的表达。相对于传统的方法,该技术的最大优势在于具有较高的灵敏性。例如,通过改进的捕获测序技术在 iPS(Induced pluripoten stem)细胞中鉴定到了 7 个极为稀有的 LINE-1 转座子拷贝,由于这些位点在群体中出现的频率极低很难用 PCR 等传统的方法进行捕捉^[88]。

1.5 探索转座子的水平转移

转座子能够在不同的物种之间发生转移并整合到新的基因组中,该现象称之为转座子的水平转移。转座子通过水平转移不仅可使其自身免于灭绝的灾难,而且对基因组的进化也有着深远的影响。在真核生物中转座子的水平转移现象较为普遍,已有大量的相关报道。在果蝇中发现至少 21 个不同的转座子家族介导近百例转座子的水平转移事件^[89]。通过对果蝇科中 3 个物种的基因组进行比较,发现近 1/3 的自主转座子来源于物种间的水平转移^[90]。同样,在其他无脊椎动物、脊椎动物以及植物中也发现了大量的转座子水平转移事件^[91~97]。目前用于证实转座子水平转移的依据主要有以下 3 个方面:(1)在分化程度较高的物种间转座子的序列反而呈现较高的相似性;(2)转座子间的进化关系同宿主基因组间的进化关系高度不协调;(3)转座子在系统进化中呈现非连续的分布^[98~100]。对这些方面的研究首先都需要获取物种的基因组序列,而高通量测序技术的出现恰好迎合了该方面的需求。

目前利用高通量测序技术对转座子水平转移已有大量的研究。对 40 个植物物种高保守性 LTR 反

转录转座子的研究发现,在 26 个物种(65%)中存在多达 32 例转座子的水平转移事件,并且这些水平转移事件甚至发生在亲缘关系相对较远的物种之间,如棕榈(*Elaeis guineensis* L.)和葡萄(*Vitis vinifera* L.)、番茄(*Solanum lycopersicum* L.)和扁豆(*Phaseolus vulgaris* L.)、杨树(*Populus alba* L.)和桃树(*Prunus persica* L.)等^[101]。另外,通过对果蝇基因组的比较不仅发现了一系列新的转座子水平转移事件,而且还揭示了不同类型的转座子通常拥有不同程度的水平转移倾向,如 LTR 反转录转座子和 DNA 转座子更易于发生水平转移,而非 LTR 反转录转座子发生转移的频率则相对较低^[90]。已有研究证实,即使低覆盖度的高通量测序数据也能够较好的用于转座子水平转移方面的研究。例如将 0.4 倍的中华按蚊(*Anopheles sinensis* L.)的高通量测序数据同蚊的参考基因组进行比较,发现 MJ1 转座子在分化程度较高的物种之间存在着水平转移现象^[102]。尽管在大量的物种之间存在转座子的水平转移现象,但是其发生的机制仍不是很清楚,例如转座子以何种物质作为载体进行转移、转移频率的高低以及转移过程中所需要的条件等^[89]。利用高通量测序技术对杆状病毒的群体进行深度测序,不仅明确地证实了杆状病毒可作为水平转移的载体介导转座子在物种内或物种间的基因组中进行穿梭,而且还精确地推算出了水平转移发生的频率^[32]。随着测序技术更广泛的应用,类似的研究可被进一步地推广到更多的物种或转座子家族,这无疑将有助于阐明转座子水平转移发生的条件和机制。

1.6 在转座子标签技术中的应用

转座子的插入突变为正向遗传学的研究提供了一种高效的分析方法,不仅能够揭示基因的功能、鉴定潜在必要的基因,而且还可为解析代谢网络、研究复杂的生物进程提供必要的信息。目前,已有多个转座子家族被应用于正向遗传学的研究中,并且这些转座子通常具有突变效率高及随机性插入等特性,如 piggyBac、Mariner、Mutator、Tnt1 和 Tos17^[103~109]。即使如此,鉴定与突变表型相关的转座子拷贝在一些情况下仍存在较大的难度,由此也限制了该技术更为广泛的应用。以往多基于 PCR 或 Southern 杂交的方法鉴定转座子的插入位点,但是这些方法通

常较为繁琐且费时费力,另外不同的插入拷贝之间存在着一定程度的序列分化,因此该技术在正向遗传学的研究中并未得到较为高效的应用。

目前高通量测序已被广泛地应用于转座子的标签技术中,并同多种捕获技术相结合建立了一系列可用于高效鉴定转座子插入位点的方法,如 DLA-454、Mu-Taq、Tn-seq、IN-seq、HITS、Mu-seq 以及生物素寡核苷酸标记测序等^[104, 110~115]。与传统方法相比,此类技术显著的优势在于它们不仅具有较好的特异性而且还拥有较高效率和成功率。例如,利用 Tn-seq 方法在绿脓假单胞菌(*Pseudomonas aeruginosa* L.)的转座子突变体库中鉴定到了十万多个 PAOI 转座子的插入位点,此数据几乎接近了该物种基因组发生饱和突变的情况^[116];利用 DLA-454 方法可识别玉米中 14 个 Mutator 转座子突变体中的 12 个位点^[110];利用生物素寡核苷酸标记测序的方法也能够鉴定到 80% 的插入突变位点^[104]。另外,通过对疟原虫(*Plasmodium berghei* L.)piggyBac 转座子突变体库 1 Gb 的高通量测序数据进行分析,鉴定到 47 个转座子的插入位点,通过反向 PCR 技术仅能鉴定到 10 个转座子的插入事件,由此证明高通量测序技术具有较高的灵敏性^[105]。除此之外,这些技术还被广泛地应用于基因适应性、基因互作以及微生物的复杂抗性等研究领域,并取得了较大进展^[117]。然而,尽管利用这些方法能够高通量地鉴定到转座子的插入位点并提高功能基因组学的研究效率,但是它们仍存在一些缺陷。例如,此类方法通常需要使用简并引物和进行多轮的 PCR 反应,因而一定程度上引入了扩增的偏好性和冗余性;另外部分方法在构建测序文库方面也较为繁琐,因此高通量测序技术在转座子标签方面的应用仍需要进一步的探索和改善。

2 测序策略和算法

近年来,高通量测序技术正以前所未有的速度取代传统的方法并广泛地应用于转座子相关领域的研究。目前,针对不同的研究目的、方案的可行性、以及实验成本等因素开发出了多种测序策略和分析算法。当前主要有两种类型的测序策略:靶位点测序和全基因组测序^[118]。

靶位点测序通常指仅针对基因组中目标片段而进行的测序。该测序策略在构建测序文库的过程中

通常需要使用一定的富集方法(如转座子展示技术、芯片杂交以及生物素寡核苷酸标记等)用于捕获基因组中的目标片段^[63,83,104,119]。例如,Helman等^[87]采用外显子捕获测序方法研究了大量肿瘤样本中转座子插入突变的情况,结果鉴定到了数十个插入到基因外显子中的转座子拷贝。相对于全基因组而言,靶点测序最大的优势在于它不但大幅度地提高了目标区段的测序深度而且还降低了测序的成本。但是,由于该策略只是针对某一家族或某一类型的转座子进行测序,因而此类测序数据通常不可用于其他方面的研究。

然而,全基因组测序数据不仅能够对几乎所有的转座子家族同时进行分析,而且还可对不完整的转座子拷贝进行分析。但是,由于高通量测序技术产生的序列具有读长短(35~500 bp)、数据量大等特点,因而往往需要开发专门的算法或软件进行转座子方面的研究。目前,主要存在以下3种常用的算法:Paired-ends、Split-reads、Depth of coverages(图2)^[4,120~122]。这些算法的应用都需要先将 Reads 定位到参考基因组中,然后分析转座子插入的情况。

2.1 Paired-ends 算法

paired-ends 算法中所分析的 paired-ends 序列是通过对已知长度的 DNA 片段从两端分别进行测序而产生的序列。由于两个序列之间的距离和相对方向是已知的,因此将这些序列比对到参考基因组中时,可通过两序列在参考基因组中的方向和跨度来判断此区段是否发生了转座子的插入、删除或倒置。若两序列在参考基因组中的跨度大于预期长度,则暗示被测样本基因组的该区段发生了片段的丢失或参考基因组中该区段内发生了转座子的插入(图2A);若两序列在参考基因组中的跨度小于预期长度,则暗示被测样本基因组的该区段发生了转座子的插入(图2B);若两序列在参考基因组中的相对方向发生了变化,则暗示基因组的该区段发生了转座子的倒置(图2C)^[123]。目前基于该算法开发出了大量的软件,如 VariationHunter^[80]、RetroSeq^[124]、HYDRA^[125]、BreakDancer^[126]等。

2.2 Split-reads 算法

Split-reads 算法的原理是将单条测序序列同参考基因组进行比对,若序列的一部分定位到基因组

中唯一性区段而另一部分定位到了转座子序列,则说明此位点处含有转座子的插入(图2D)。该算法的最大优势在于它能够明确地鉴定出转座子的具体插入位置。有研究采用该算法对 Roche/454 产生的高通量数据进行分析,从24个样本中共鉴定到了4000多个非参考插入的转座子位点^[30]。该算法对序列的读长提出了较高的要求,它通常适用分析 Roche/454 平台所产生的数据。然而,其他高通量测序平台由于产生的序列通常较短,因此在一定程度上限制了该算法的广泛应用。

2.3 Depth of coverages 算法

该算法假定基因组中所有片段的测序概率同等,同时它们对应 Reads 数目在理论上服从泊松分布的形式,并且同片段在基因组中出现的频率呈正相关。当基因组中的某些片段进行扩增时其相对应 Reads 的数目则会增多,反之将会减少。由此,可以根据 Reads 的覆盖深度来推算某些区段在基因组中的扩增或丢失情况(图2E)^[127]。由于引入了统计分析,该算法较适用于大片段插入或缺失突变的预测。另外,该方法并不适于对基因组中拷贝数目较高的片段(如转座子)进行估算,因而在实际应用中通常与其他算法相结合对潜在突变的位点进行预测。

尽管这些算法在应用中存在着各自的优势,但

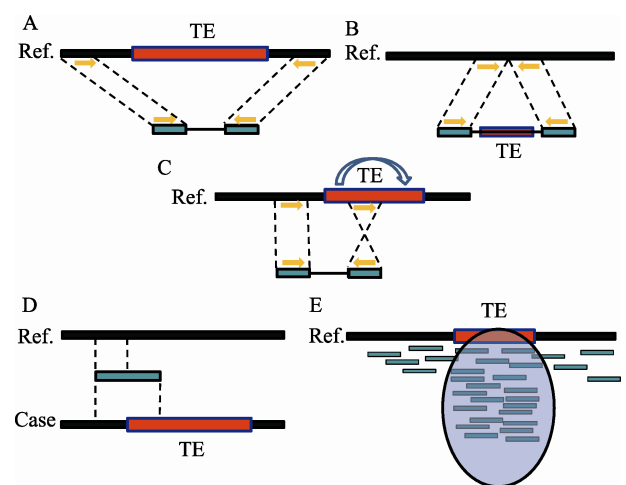


图2 利用高通量测序的读序列检测转座子插入的相关算法(参考文献^[120]并修改)

A~C 分别表示 Paired-ends 算法中重测序基因组发生转座子的丢失、插入和倒置;D 表示 Split-reads 算法中重测序基因组发生转座子的插入;E 表示 Depth of coverages 算法,当重测序基因组发生转座子的插入时,该位点所对应的读序列数据相对较多。

也存在着一定的缺陷。例如 Paired-ends 算法不能较好地解决模糊匹配的问题, 并且它们对变异位点的精确性定位在很大程度上依赖于较为严谨的片段长度分布^[122]; Split-reads 算法仅局限于基因组中唯一性区段的分析; 而 Depth of coverages 算法通常易受测序偏好性的影响, 并且不能获取有关变异位点的具体信息^[128, 129]。由于不同的算法采用不同的原理并且拥有各自的偏好性, 因此即使对于同一样本数据通常也会产生各自迥异的结果。为此, 将这些算法进行整合将有利于提高检测的特异性和灵敏度^[120]。目前已开发出的多款软件整合了上述的算法, 如 TE analyzer(Tea)首先采用 Paired-ends 算法筛选出潜在的转座子插入位点, 然后利用 Split-reads 算法鉴定出这些转座子拷贝的确切插入位置^[31]; 而 SPANNER 软件则将 Paired-ends 和 Depth of coverages 算法进行整合从而用于遗传变异的研究^[130]。

3 结语与展望

高通量测序技术使转座子的研究发生了巨大的变化, 拓宽了转座子研究的广度和深度。该技术不但使转座子在含量的估算、整合位点的偏好性分析以及插入多态性等方面的研究实现了高通量, 而且相对于传统的方法它们在转座子的水平转移、稀有转座子的鉴定等方面通常还具有更高的特异性和灵敏性等优势。尽管如此, 该技术仍存在一些固有的缺陷阻碍了其更广阔地发展。由于高通量测序技术产生的序列具有读长短、数据量大等特点, 不仅为常规的基因组分析增加了难度, 而且也作为基因组中转座子等重复序列的研究提出了严峻的挑战, 为此往往需要开发复杂的算法以及高效的生物信息学工具来进行辅助的研究。另外, 相对于 Sanger 测序该技术在测序成本方面尽管有了较大的改观, 但仍不能够满足部分研究领域的需求, 例如在群体基因组学的研究中通常需要对大量的样本进行全基因组测序, 由此产生的高昂费用对于独立的研究团队仍是难以承受, 因此对于高通量测序技术将来的发展也提出了更高的要求。

第三代测序技术的开发将致力于序列更长、质量更高、成本低廉等方面的研究。目前以 Pacific Biosciences 为代表的第三代测序平台能够产生可高达 10 kb 的序列读长, 由于采用单分子测序技术不

需要进行 PCR 反应因而避免了扩增的偏好性, 但该平台产生的序列存在错误率较高等缺陷, 因此仍需进一步的完善^[22, 131, 132]。相信在不远的将来, 随着高通量测序技术的不断提高, 它们将更好地服务于转座子的研究从而为揭示转座子的转座机制和生物学功能发挥更重要的作用。

参考文献

- [1] McClintock B. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci USA*, 1950, 36(6): 344–355. [DOI]
- [2] Chénais B, Caruso A, Hiard S, Casse N. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. *Gene*, 2012, 509(1): 7–15. [DOI]
- [3] Huang CRL, Burns KH, Boeke JD. Active transposition in genomes. *Annu Rev Genet*, 2012, 46: 651–675. [DOI]
- [4] Vitte C, Fustier MA, Alix K, Tenaillon MI. The bright side of transposons in crop evolution. *Brief Funct Genomics*, 2014, 13(4): 276–295. [DOI]
- [5] Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*, 2007, 8(12): 973–982. [DOI]
- [6] Wessler SR. Transposable elements and the evolution of eukaryotic genomes. *Proc Natl Acad Sci USA*, 2006, 103(47): 17600–17601. [DOI]
- [7] Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet*, 2007, 8(4): 272–285. [DOI]
- [8] Ayarpadikannan S, Kim HS. The impact of transposable elements in genome evolution and genetic instability and their implications in various diseases. *Genomics Inform*, 2014, 12(3): 98–104. [DOI]
- [9] Bennetzen JL, Wang H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu Rev Plant Biol*, 2014, 65: 505–530. [DOI]
- [10] Gifford WD, Pfaff SL, Macfarlan TS. Transposable elements as genetic regulatory substrates in early development. *Trends Cell Biol*, 2013, 23(5): 218–226. [DOI]
- [11] Lee SI, Kim NS. Transposable elements and genome size variations in plants. *Genomics Inform*, 2014, 12(3): 87–97. [DOI]
- [12] Sentmanat M, Wang SH, Elgin SCR. Targeting heterochromatin formation to transposable elements in

- Drosophila*: potential roles of the piRNA system. *Biochemistry (Mosc)*, 2013, 78(6): 562–571. [DOI]
- [13] 赵美霞, 张彪, 刘胜毅, 马渐新. 白菜和甘蓝基因组转座子表达及其对基因调控的潜在影响. *遗传*, 2013, 35(8): 1014–1022. [DOI]
- [14] Dean C, Sjodin C, Bancroft I, Lawson E, Lister C, Scofield S, Jones J. Development of an efficient transposon tagging system in *Arabidopsis thaliana*. *Symp Soc Exp Biol*, 1991, 45: 63–75. [DOI]
- [15] Izawa T, Ohnishi T, Nakano T, Ishida N, Enoki H, Hashimoto H, Itoh K, Terada R, Wu C, Miyazaki C, Endo T, Iida S, Shimamoto K. Transposon tagging in rice. *Plant Mol Biol*, 1997, 35(1–2): 219–229. [DOI]
- [16] Kalendar R, Flavell AJ, Ellis THN, Sjakste T, Moisy C, Schulman AH. Analysis of plant diversity with retrotransposon-based molecular markers. *Heredity*, 2011, 106(4): 520–530. [DOI]
- [17] Ansorge WJ. Next-generation DNA sequencing techniques. *N Biotechnol*, 2009, 25(4): 195–203. [DOI]
- [18] Bennett S. Solexa Ltd. *Pharmacogenomics*, 2004, 5(4): 433–438. [DOI]
- [19] Bentley DR. Whole-genome re-sequencing. *Curr Opin Genet Dev*, 2006, 16(6): 545–552. [DOI]
- [20] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bembien LA, Berka J, Braverman MS, Chen YJ, Chen ZT, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu PG, Begley RF, Rothberg JM. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 2005, 437(7057): 376–380. [DOI]
- [21] Shendure J, Porreca GJ, Reppas NB, Lin XX, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 2005, 309(5741): 1728–1732. [DOI]
- [22] Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ, Holden D, Saxena R, Wegener J, Turner SW. Real-time DNA sequencing from single polymerase molecules. *Methods Enzymol*, 2010, 472: 431–455. [DOI]
- [23] Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. *J Appl Genet*, 2011, 52(4): 413–435. [DOI]
- [24] Kelly LJ, Leitch IJ. Exploring giant plant genomes with next-generation sequencing technology. *Chromosome Res*, 2011, 19(7): 939–953. [DOI]
- [25] Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. *Cell*, 2013, 155(1): 27–38. [DOI]
- [26] McGettigan PA. Transcriptomics in the RNA-seq era. *Curr Opin Chem Biol*, 2013, 17(1): 4–11. [DOI]
- [27] Chabbert CD, Adjalley SH, Klaus B, Fritsch ES, Gupta I, Pelechano V, Steinmetz LM. A high-throughput ChIP-Seq for large-scale chromatin studies. *Mol Syst Biol*, 2015, 11(1): 777. [DOI]
- [28] Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc*, 2010, 2010(2): pdb.prot5384. [DOI]
- [29] Zhao MT, Whyte JJ, Hopkins GM, Kirk MD, Prather RS. Methylated DNA immunoprecipitation and high-throughput sequencing (MeDIP-seq) using low amounts of genomic DNA. *Cell Reprogram*, 2014, 16(3): 175–184. [DOI]
- [30] Stewart C, Kural D, Stromberg MP, Walker JA, Konkel MK, Stutz AM, Urban AE, Grubert F, Lam HY, Lee WP, Busby M, Indap AR, Garrison E, Huff C, Xing J, Snyder MP, Jorde LB, Batzer MA, Korbel JO, Marth GT. A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet*, 2011, 7(8): e1002236. [DOI]
- [31] Lee E, Iskow R, Yang LX, Gokcumen O, Haseley P, Luquette LJ, III, Lohr JG, Harris CC, Ding L, Wilson RK, Wheeler DA, Gibbs RA, Kucherlapati R, Lee C, Kharchenko PV, Park PJ. Landscape of somatic retrotransposition in human cancers. *Science*, 2012, 337(6097): 967–971. [DOI]
- [32] Gilbert C, Chateigner A, Ernenwein L, Barbe V, Bézier A, Herniou EA, Cordaux R. Population genomics supports baculoviruses as vectors of horizontal transfer of insect transposons. *Nat Commun*, 2014, 5: 3348. [DOI]
- [33] Tian ZX, Zhao MX, She MY, Du JC, Cannon SB, Liu X, Xu X, Qi XP, Li MW, Lam HM, Ma JX. Genome-wide characterization of nonreference transposons reveals evolutionary propensities of transposons in soybean. *Plant Cell*, 2012, 24(11): 4422–4436. [DOI]
- [34] Greilhuber J, Borsch T, Müller K, Worberg A, Porembski S, Barthlott W. Smallest angiosperm genomes found in lentibulariaceae, with chromosomes of bacterial size. *Plant Biol*, 2006, 8(6): 770–777. [DOI]
- [35] Pellicer J, Fay MF, Leitch IJ. The largest eukaryotic genome of them all? *Bot J Linn Soc*, 2010, 164(1):

- 10–15. [DOI]
- [36] Piegu B, Guyot R, Picault N, Roulin A, Sanyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, Panaud O. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res*, 2006, 16(10): 1262–1269. [DOI]
- [37] SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL. Nested retrotransposons in the intergenic regions of the maize genome. *Science*, 1996, 274(5288): 765–768. [DOI]
- [38] Moghe GD, Shiu SH. The causes and molecular consequences of polyploidy in flowering plants. *Ann N Y Acad Sci*, 2014, 1320: 16–34. [DOI]
- [39] Weiss-Schneeweiss H, Emadzade K, Jang TS, Schneeweiss GM. Evolutionary consequences, constraints and potential of polyploidy in plants. *Cytogenet Genome Res*, 2013, 140(2–4): 137–150. [DOI]
- [40] 陈建军, 王瑛. 植物基因组大小进化的研究进展. 遗传, 2009, 31(5): 464–470. [DOI]
- [41] Macas J, Neumann P, Navrátilová A. Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genomics*, 2007, 8: 427. [DOI]
- [42] Tenailon MI, Hufford MB, Gaut BS, Ross-Ibarra J. Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. *Genome Biol Evol*, 2011, 3: 219–229. [DOI]
- [43] Hříbová E, Neumann P, Matsumoto T, Roux N, Macas J, Doležal J. Repetitive part of the banana (*Musa acuminata*) genome investigated by low-depth 454 sequencing. *BMC Plant Biol*, 2010, 10: 204. [DOI]
- [44] Sabot F, Picault N, El-Baidouri M, Llauro C, Chaparro C, Piegu B, Roulin A, Guiderdoni E, Delabastide M, McCombie R, Panaud O. Transpositional landscape of the rice genome revealed by paired-end mapping of high-throughput re-sequencing data. *Plant J*, 2011, 66(2): 241–246. [DOI]
- [45] Natali L, Cossu RM, Barghini E, Giordani T, Buti M, Mascagni F, Morgante M, Gill N, Kane NC, Rieseberg L, Cavallini A. The repetitive component of the sunflower genome as shown by different procedures for assembling next generation sequencing reads. *BMC Genomics*, 2013, 14: 686. [DOI]
- [46] Wang QH, Dooner HK. Remarkable variation in maize genome structure inferred from haplotype diversity at the *bz* locus. *Proc Natl Acad Sci USA*, 2006, 103(47): 17644–17649. [DOI]
- [47] Brenchley R, Spannagl M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, Kay S, Waite D, Trick M, Bancroft I, Gu Y, Huo NX, Luo MC, Sehgal S, Gill B, Kianian S, Anderson O, Kersey P, Dvorak J, McCombie WR, Hall A, Mayer KF, Edwards KJ, Bevan MW, Hall N. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, 2012, 491(7426): 705–710. [DOI]
- [48] Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, Schlub S, Le Paslier MC, Magdelenat G, Gonthier C, Couloux A, Budak H, Breen J, Pumphrey M, Liu SX, Kong XY, Jia JZ, Gut M, Brunel D, Anderson JA, Gill BS, Appels R, Keller B, Feuillet C. Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell*, 2010, 22(6): 1686–1701. [DOI]
- [49] Sergeeva EM, Afonnikov DA, Koltunova MK, Gusev VD, Miroshnichenko LA, Vrána J, Kubaláková M, Poncet C, Sourdille P, Feuillet C, Doležal J, Salina EA. Common wheat chromosome 5B composition analysis using low-coverage 454 sequencing. *Plant Genome*, 2014, 7(2): 16. [DOI]
- [50] O'Hare K, Rubin GM. Structures of P transposable elements and their sites of insertion and excision in the *Drosophila melanogaster* genome. *Cell*, 1983, 34(1): 25–35. [DOI]
- [51] Mori I, Benian GM, Moerman DG, Waterston RH. Transposable element Tc1 of *Caenorhabditis elegans* recognizes specific target sequences for integration. *Proc Natl Acad Sci USA*, 1988, 85(3): 861–864. [DOI]
- [52] Tudor M, Lobočka M, Goodell M, Pettitt J, O'Hare K. The pogo transposable element family of *Drosophila melanogaster*. *Mol Gen Genet*, 1992, 232(1): 126–134. [DOI]
- [53] Berry C, Hännenhalli S, Leipzig J, Bushman FD. Selection of target sites for mobile DNA integration in the human genome. *PLoS Comput Biol*, 2006, 2(11): e157. [DOI]
- [54] Levy A, Schwartz S, Ast G. Large-scale discovery of insertion hotspots and preferential integration sites of human transposed elements. *Nucleic Acids Res*, 2010, 38(5): 1515–1530. [DOI]
- [55] Linheiro RS, Bergman CM. Whole genome resequencing reveals natural target site preferences of transposable elements in *Drosophila melanogaster*. *PLoS One*, 2012, 7(2): e30008. [DOI]
- [56] Jiang N, Bao Z, Temnykh S, Cheng Z, Jiang J, Wing RA, McCouch SR, Wessler SR. Dasheng: A recently amplified nonautonomous long terminal repeat element that is a major component of pericentromeric regions in rice.

- Genetics*, 2002, 161(3): 1293–1305. [DOI]
- [57] Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, Carrington JC, Doerge RW, Colot V, Martienssen R. Role of transposable elements in heterochromatin and epigenetic control. *Nature*, 2004, 430(6998): 471–476. [DOI]
- [58] Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang HB, Wang XY, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang LF, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob ur R, Ware D, Westhoff P, Mayer KFX, Messing J, Rokhsar DS. The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, 2009, 457(7229): 551–556. [DOI]
- [59] Rizzon C, Marais G, Gouy M, Biémont C. Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome. *Genome Res*, 2002, 12(3): 400–407. [DOI]
- [60] Schmutz J, Cannon SB, Schlueter J, Ma JX, Mitros T, Nelson W, Hyten DL, Song QJ, Thelen JJ, Cheng JL, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu SQ, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du JC, Tian ZX, Zhu LC, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA. Genome sequence of the palaeopolyploid soybean. *Nature*, 2010, 463(7278): 178–183. [DOI]
- [61] Gaut BS, Wright SI, Rizzon C, Dvorak J, Anderson LK. Recombination: an underappreciated factor in the evolution of plant genomes. *Nat Rev Genet*, 2007, 8(1): 77–84. [DOI]
- [62] Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, Okumoto Y, Tanisaka T, Wessler SR. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature*, 2009, 461(7267): 1130–1134. [DOI]
- [63] Ewing AD, Kazazian HH. Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. *Genome Res*, 2011, 21(6): 985–990. [DOI]
- [64] Kumar A, Hirochika H. Applications of retrotransposons as genetic tools in plant biology. *Trends Plant Sci*, 2001, 6(3): 127–134. [DOI]
- [65] Pocza P, Varga I, Laos M, Cseh A, Bell N, Valkonen JP, Hyvönen J. Advances in plant gene-targeted and functional markers: a review. *Plant Methods*, 2013, 9(1): 6. [DOI]
- [66] Waugh R, McLean K, Flavell AJ, Pearce SR, Kumar A, Thomas BBT, Powell W. Genetic distribution of Bare-1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). *Mol Gen Genet*, 1997, 253(6): 687–694. [DOI]
- [67] Syed NH, Sureshsundar S, Wilkinson MJ, Bhau BS, Cavalcanti JJV, Flavell AJ. Ty1-copia retrotransposon-based SSAP marker development in cashew (*Anacardium occidentale* L.). *Theor Appl Genet*, 2005, 110(7): 1195–1202. [DOI]
- [68] Kalendar R, Schulman AH. IRAP and REMAP for retrotransposon-based genotyping and fingerprinting. *Nat Protoc*, 2006, 1(5): 2478–2484. [DOI]
- [69] Flavell AJ, Knox MR, Pearce SR, Ellis THN. Retrotransposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. *Plant J*, 1998, 16(5): 643–650. [DOI]
- [70] Monden Y, Fujii N, Yamaguchi K, Ikeo K, Nakazawa Y, Waki T, Hirashima K, Uchimura Y, Tahara M. Efficient screening of long terminal repeat retrotransposons that show high insertion polymorphism via high-throughput sequencing of the primer binding site. *Genome*, 2014, 57(5): 245–252. [DOI]
- [71] Sveinsson S, Gill N, Kane NC, Cronk Q. Transposon fingerprinting using low coverage whole genome shotgun sequencing in cacao (*Theobroma cacao* L.) and related species. *BMC Genomics*, 2013, 14: 502. [DOI]
- [72] Hormozdiari F, Alkan C, Ventura M, Hajirasouliha I, Malig M, Hach F, Yorukoglu D, Dao P, Bakhshi M, Sahinalp SC, Eichler EE. *Alu* repeat discovery and characterization within human genomes. *Genome Res*, 2011, 21(6): 840–849. [DOI]
- [73] Marchi E, Kanapin A, Magiorkinis G, Belshaw R. Unfixed endogenous retroviral insertions in the human population. *J Virol*, 2014, 88(17): 9529–9537. [DOI]
- [74] Cridland JM, Macdonald SJ, Long AD, Thornton KR. Abundance and distribution of transposable elements in two *Drosophila* QTL mapping resources. *Mol Biol Evol*, 2013, 30(10): 2311–2327. [DOI]
- [75] Nellåker C, Keane TM, Yalcin B, Wong K, Agam A, Belgard TG, Flint J, Adams DJ, Frankel WN, Ponting CP. The genomic landscape shaped by selection on transposable elements across 18 mouse strains. *Genome Biol*, 2012, 13(6): R45. [DOI]
- [76] Liu Z, Li XX, Wang TZ, Messing J, Xu JH. The *Wukong* terminal-repeat retrotransposon in miniature (TRIM)

- elements in diverse maize germplasm. *G3*, 2015, doi: 10.1534/g3.115.018317. [DOI]
- [77] Fiston-Lavier AS, Carrigan M, Petrov DA, González J. T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. *Nucleic Acids Res*, 2011, 39(6): e36. [DOI]
- [78] Robb SMC, Lu L, Valencia E, Burnette JM, Okumoto Y, Wessler SR, Stajich JE. The use of RelocaTE and unassembled short reads to produce high-resolution snapshots of transposable element generated diversity in rice. *G3*, 2013, 3(6): 949–957. [DOI]
- [79] Witherspoon DJ, Xing JC, Zhang YH, Watkins WS, Batzer MA, Jorde LB. Mobile element scanning (ME-Scan) by targeted high-throughput sequencing. *BMC Genomics*, 2010, 11: 410. [DOI]
- [80] Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, Eichler EE, Sahinalp SC. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, 2010, 26(12): i350–i357. [DOI]
- [81] Wu JT, Lee WP, Ward A, Walker JA, Konkel MK, Batzer MA, Marth GT. Tangram: a comprehensive toolbox for mobile element insertion detection. *BMC Genomics*, 2014, 15: 795. [DOI]
- [82] Kofler R, Betancourt AJ, Schlötterer C. Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. *PLoS Genet*, 2012, 8(1): e1002487. [DOI]
- [83] Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, Van Meir EG, Vertino PM, Devine SE. Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell*, 2010, 141(7): 1253–1261. [DOI]
- [84] Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev*, 2009, 19(3): 212–219. [DOI]
- [85] Cohen JC, Pertsemliadis A, Fahmi S, Esmail S, Vega GL, Grundy SM, Hobbs HH. Multiple rare variants in *NPC1L1* associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc Natl Acad Sci USA*, 2006, 103(6): 1810–1815. [DOI]
- [86] Zhu XF, Feng T, Li YL, Lu Q, Elston RC. Detecting rare variants for complex traits using family and unrelated data. *Genet Epidemiol*, 2010, 34(2): 171–187. [DOI]
- [87] Helman E, Lawrence MS, Stewart C, Sougnez C, Getz G, Meyerson M. Somatic retrotransposition in man cancer revealed by whole-genome and exome sequencing. *Genome Res*, 2014, 24(7): 1053–1063. [DOI]
- [88] Arokium H, Kamata M, Kim S, Kim N, Liang M, Presson AP, Chen IS. Deep sequencing reveals low incidence of endogenous LINE-1 retrotransposition in human induced pluripotent stem cells. *PLoS One*, 2014, 9(10): e108682. [DOI]
- [89] Loreto ELS, Carareto CMA, Capy P. Revisiting horizontal transfer of transposable elements in *Drosophila*. *Heredity*, 2008, 100(6): 545–554. [DOI]
- [90] Bartolomé C, Bello X, Maside X. Widespread evidence for horizontal transfer of transposable elements across *Drosophila* genomes. *Genome Biol*, 2009, 10(2): R22. [DOI]
- [91] Diao XM, Freeling M, Lisch D. Horizontal transfer of a plant transposon. *PLoS Biol*, 2006, 4(1): e5. [DOI]
- [92] Roulin A, Piegu B, Wing RA, Panaud O. Evidence of multiple horizontal transfers of the long terminal repeat retrotransposon RIRE1 within the genus *Oryza*. *Plant J*, 2008, 53(6): 950–959. [DOI]
- [93] Casse N, Bui QT, Nicolas V, Renault S, Bigot Y, Laulier M. Species sympatry and horizontal transfers of *Mariener* transposons in marine crustacean genomes. *Mol Phylogenet Evol*, 2006, 40(2): 609–619. [DOI]
- [94] de Boer JG, Yazawa R, Davidson WS, Koop BF. Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. *BMC Genomics*, 2007, 8: 422. [DOI]
- [95] Novikova O, Śliwińska E, Fet V, Settele J, Blinov A, Woyciechowski M. CR1 clade of non-LTR retrotransposons from *Maculinea* butterflies (Lepidoptera: Lycaenidae): evidence for recent horizontal transmission. *BMC Evol Biol*, 2007, 7: 93. [DOI]
- [96] Ray DA, Feschotte C, Pagan HJT, Smith JD, Pritham EJ, Arensburger P, Atkinson PW, Craig NL. Multiple waves of recent DNA transposon activity in the bat, *Myotis lucifugus*. *Genome Res*, 2008, 18(5): 717–728. [DOI]
- [97] Wright DA, Voytas DF. *Athila4* of *Arabidopsis* and *Calypso* of soybean define a lineage of endogenous plant retroviruses. *Genome Res*, 2002, 12(1): 122–131. [DOI]
- [98] Gilbert C, Schaack S, Pace JK II, Brindley PJ, Feschotte C. A role for host-parasite interactions in the horizontal transfer of transposons across phyla. *Nature*, 2010, 464(7293): 1347–1350. [DOI]
- [99] Kuraku S, Qiu H, Meyer A. Horizontal transfers of Tc1 elements between teleost fishes and their vertebrate parasites, lampreys. *Genome Biol Evol*, 2012, 4(9): 929–936. [DOI]
- [100] Wallau GL, Ortiz MF, Loreto ELS. Horizontal transposon transfer in eukarya: detection, bias, and perspectives. *Genome Biol Evol*, 2012, 4(8): 801–811. [DOI]
- [101] El Baidouri M, Carpentier MC, Cooke R, Gao DY, Lasserre E, Llauro C, Mirouze M, Picault N, Jackson SA, Panaud O. Widespread and frequent horizontal transfers

- of transposable elements in plants. *Genome Res*, 2014, 24(5): 831–838. [DOI]
- [102] Diao YP, Qi YM, Ma YJ, Xia A, Sharakhov I, Chen XG, Biedler J, Ling EJ, Tu ZJ. Next-generation sequencing reveals recent horizontal transfer of a DNA transposon between divergent mosquitoes. *PLoS One*, 2011, 6(2): e16743. [DOI]
- [103] Klein BA, Tenorio EL, Lazinski DW, Camilli A, Duncan MJ, Hu LDT. Identification of essential genes of the periodontal pathogen *Porphyromonas gingivalis*. *BMC Genomics*, 2012, 13: 578. [DOI]
- [104] Williams-Carrier R, Stiffler N, Belcher S, Kroeger T, Stern DB, Monde RA, Coalter R, Barkan A. Use of Illumina sequencing to identify transposon insertions underlying mutant phenotypes in high-copy Mutator lines of maize. *Plant J*, 2010, 63(1): 167–177. [DOI]
- [105] Cao Y, Rui B, Wellems DL, Li MX, Chen BB, Zhang DM, Pan WQ. Identification of *piggyBac*-mediated insertions in *Plasmodium berghei* by next generation sequencing. *Malar J*, 2013, 12(1): 287. [DOI]
- [106] Brutnell TP. Transposon tagging in maize. *Funct Integr Genomics*, 2002, 2(1–2): 4–12. [DOI]
- [107] Grandbastien MA, Spielmann A, Caboche M. Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. *Nature*, 1989, 337(6205): 376–380. [DOI]
- [108] Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M. Retrotransposons of rice involved in mutations induced by tissue culture. *Proc Natl Acad Sci USA*, 1996, 93(15): 7783–7788. [DOI]
- [109] Yamazaki M, Tsugawa H, Miyao A, Yano M, Wu J, Yamamoto S, Matsumoto T, Sasaki T, Hirochika H. The rice retrotransposon *Tos17* prefers low-copy-number sequences as integration targets. *Mol Genet Genomics*, 2001, 265(2): 336–344. [DOI]
- [110] Liu SZ, Dietrich CR, Schnable PS. DLA-based strategies for cloning insertion mutants: cloning the *gl4* locus of maize using *Mu* transposon tagged alleles. *Genetics*, 2009, 183(4): 1215–1225. [DOI]
- [111] Howard TP III, Hayward AP, Tordillos A, Fragoso C, Moreno MA, Tohme J, Kausch AP, Mottinger JP, Delaporta SL. Identification of the maize gravitropism gene *lazy plant1* by a transposon-tagging genome resequencing strategy. *PLoS One*, 2014, 9(1): e87053. [DOI]
- [112] McCarty DR, Latshaw S, Wu S, Suzuki M, Hunter CT, Avigne WT, Koch KE. Mu-seq: sequence-based mapping and identification of transposon induced mutations. *PLoS One*, 2013, 8(10): e77172. [DOI]
- [113] van Opijnen T, Bodi KL, Camilli A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Methods*, 2009, 6(10): 767–72. [DOI]
- [114] Gawronski JD, Wong SMS, Giannoukos G, Ward DV, Akerley BJ. Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung. *Proc Natl Acad Sci USA*, 2009, 106(38): 16422–16427. [DOI]
- [115] Goodman AL, McNulty NP, Zhao Y, Leip D, Mitra RD, Lozupone CA, Knight R, Gordon JI. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe*, 2009, 6(3): 279–289. [DOI]
- [116] Gallagher LA, Shendure J, Manoil C. Genome-scale identification of resistance functions in *Pseudomonas aeruginosa* using Tn-seq. *MBio*, 2011, 2(1): e00315–10. [DOI]
- [117] van Opijnen T, Camilli A. Genome-wide fitness and genetic interactions determined by Tn-seq, a high-throughput massively parallel sequencing method for microorganisms. *Curr Protoc Microbiol*, 2015, 36: 1e.3.1–1e.3.24. [DOI]
- [118] Xing JC, Witherspoon DJ, Jorde LB. Mobile element biology: new possibilities with high-throughput sequencing. *Trends Genet*, 2013, 29(5): 280–289. [DOI]
- [119] Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M, Talbot RT, Gustincich S, Freeman TC, Mattick JS, Hume DA, Heutink P, Carninci P, Jeddleloh JA, Faulkner GJ. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature*, 2011, 479(7374): 534–537. [DOI]
- [120] Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet*, 2011, 12(5): 363–376. [DOI]
- [121] Abel HJ, Duncavage EJ. Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. *Cancer Genet*, 2013, 206(12): 432–440. [DOI]
- [122] Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods*, 2009, 6(Suppl. 11): S13–S20. [DOI]
- [123] Korb J, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen ZT, Tanzer A, Saunders AC, Chi JX, Yang FT, Carter NP, Hurler ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 2007, 318(5849): 420–426. [DOI]
- [124] Keane TM, Wong K, Adams DJ. RetroSeq: transposable element discovery from next-generation sequencing da-

- ta. *Bioinformatics*, 2013, 29(3): 389–390. [DOI]
- [125] Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang YJ, Hurles ME, Mell JC, Hall IM. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res*, 2010, 20(5): 623–635. [DOI]
- [126] Fan X, Abbott TE, Larson D, Chen K. BreakDancer-Identification of genomic structural variation from paired-end read mapping. *Current Protocols in Bioinformatics*, 2014, 15(6): 1–11. [DOI]
- [127] [127] Bailey JA, Gu ZP, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. Recent segmental duplications in the human genome. *Science*, 2002, 297(5583): 1003–1007. [DOI]
- [128] Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*, 2008, 36(16): e105. [DOI]
- [129] Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol*, 2009, 27(1): 66–75. [DOI]
- [130] Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. *Nature*, 2010, 467(7319): 1061–1073. [DOI]
- [131] Glenn TC. Field guide to next-generation DNA sequencers. *Mol Ecol Resour*, 2011, 11(5): 759–769. [DOI]
- [132] Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. *Hum Mol Genet*, 2010, 19(R2): R227–R240. [DOI]

(责任编辑: 胡松年)

• 综合信息 •

2016 年《遗传》征订启事

《遗传》创刊于 1979 年, 是中国遗传学会和中国科学院遗传与发育生物学研究所主办、科学出版社出版的学术期刊, 中文核心期刊, 中国精品科技期刊。已被 MEDLINE、生物学数据库 (BIOSIS)、生物学文摘 (BA)、医学索引 (Medical Index) 和美国化学文摘 (CA)、以及俄罗斯文摘杂志 (AJ) 等 20 多种国内外重要检索系统与数据库收录。主要刊登有创新性的研究论文、新技术与新方法、学科热点问题的综述、学术讨论、遗传学教学、遗传学家介绍、学术会议信息及科学新闻等, 内容涉及遗传学、基因组学、细胞生物学、发育生物学、分子进化、遗传工程及生物技术等领域。

本刊开辟绿色通道, 重要成果的研究论文可申请优先刊出。

欢迎投稿 欢迎订阅 欢迎刊登广告

月刊, 大 16 开本, 112 页, 定价 80 元/期, 全年 960 元。各地邮局发行。

邮发代号: 2-810。国内刊号 CN 11-1913/R, 国际统一刊号 ISSN 0253-9772。

地址: 北京市朝阳区北辰西路 1 号院中国科学院遗传与发育生物学研究所 2 号楼

邮编: 100101

网址: <http://www.chinagene.cn>; E-mail: yczz@genetics.ac.cn

电话: 010-64807669; 传真: 010-64807786

《遗传》编辑部
2015 年 9 月 10 日