遗传 HEREDITAS (Beijing) 2009 年 8 月, 31(8): 799－808
ISSN 0253-9772 www.chinagene.cn

综 述

DOI: 10.3724/SP.J.1005.2009.00799

# 植物基因富集研究进展

曾少华[1, 2], 刘迪[1, 2], 王瑛[1]

1. 中国科学院武汉植物园, 武汉 430074;
2. 中国科学院研究生院, 北京 100049

**摘要**: 高等植物的基因组大小差异十分巨大, 在大基因组植物的基因组中, 各种重复序列占据了基因组中相当大一部分, 而低拷贝或单拷贝的基因序列仅占了很少一部分。对于大基因组物种而言, 大量的重复序列给基因组的研究工作带来很大困难, 使得大规模获得基因信息成为一项很有挑战性的工作。目前, 在基因组范围内富集基因的方法有 cDNA 文库、甲基化过滤文库、高 Cot 值文库、转座子标签富集法等。文章综述了这几种方法的技术原理和特性, 结合近年来国内外运用甲基化过滤技术的研究进展, 探讨了根据不同研究材料和研究目标, 如何高效选择适合的方法或者方法的组合。

**关键词**: 基因富集; 高 Cot 值文库; 甲基化过滤文库

# Advances of gene enrichment in plant genome

ZENG Shao-Hua[1, 2], LIU Di[1, 2], WANG Ying[1]

1. *Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan* 430074, *China*;
2. *The Graduate School, Chinese Academy of Sciences, Beijing* 100049, *China*

**Abstract:** The genome size varies greatly in higher plants. Repetitive sequences account for most of the large plant genomes while low-copy or single copy genic sequences, referred to as gene space, take up only a small portion of the genomes. Considering the large amount of repetitive sequences, it is a great challenge to obtain genic sequences using high-throughout methods in non-model plants bearing large genomes. Currently, several approaches have been developed for gene enrichment on a genome-wide scale, such as cDNA library, methylation filtration library, high Cot library and transposon tagging. Here, we reviewed the technical principles, advantages and disadvantages of these methods, as well as the recent development of methylation filtration technology. An in-depth discussion was performed for selection of one method or combination of methods according to the research objectives and plant materials, especially for plants with large genomes.

**Keywords:** gene enrichment; high Cot library; methylation filtration library

生命科学研究现已迈入后基因组时代, 对基因组功能和进化的研究已经开展得如火如荼。在过去的十几年中, 许多生命科学领域的科学家们都致力于高效测序技术的发展, 使人们获得的基因组序列信息量正以前所未有的速度海量增长。植物基因组学和比较基因组学领域的研究成果也随着越来越多

收稿日期: 2008–12–26; 修回日期: 2009–02–16
基金项目: 中国科学院农业基地知识创新工程重要方向项目(编号：KSCX2-YW-N-030), 中科院百人计划项目, 国家自然科学基金项目(编号：30800624), 中科院院长基金和武汉市晨光计划项目(编号：20055003059-45)资助
作者简介: 曾少华(1979–), 男, 博士, 研究方向：基因组学。Tel: 027-87510771; E-mail: shzeng@wbgcas.cn
通讯作者: 王瑛(1973–), 女, 博士, 研究员, 研究方向：比较功能基因组学。Tel: 027-87510675; E-mail: yingwang@wbgcas.cn

的植物全基因组序列的获得而愈加丰硕。目前，对全基因组进行测序主要有两种基本途径：一是“鸟枪法”(Whole-genome shotgun)，它是美国塞莱拉遗传公司(Celera Genomics Corporation)创始人克雷格·文特尔(Craig Venter)发明的，是目前常用的两种DNA测序法中较为快捷的一种，其基本原理是随机地将DNA片段打乱，构建随机片段文库，测序后通过强大的计算机运算方法对其进行排序。目前该方法已成功应用于水稻(*Oryza sativa* L.)[1]、番木瓜(*Carica papaya* L.)[2]、葡萄(*Vitis vinifera* L.)[3]、毛果杨(*Populus trichocarpa* (Torr. & Gray))[4]以及小立碗藓(*Physcomitrella patens*)[5]全基因组测序和芸苔属物种(*Brassica oleracea*)[6]基因组测序。另一种是BAC-by-BAC策略，其原理是先构建一个或多个该物种的基因组BAC文库，选择能覆盖全基因组范围且整体覆盖率能达到 10~20 倍的相互之间重叠最少的BAC克隆，再对每个BAC克隆进行亚克隆测序。

高等植物基因组大小差异悬殊，从十字花科某些物种的 40 Mb到百合科某些物种的 60 000 Mb。重要的经济作物如大麦(*Hordeum vulgare* L.)(4 900 Mb)、小麦(*Triticum aestivum* L.) (16 900 Mb)、高粱(*Sorghum bicolor* (L.) Moench) (735 Mb)和玉米(*Zea mays* L.)(2 500 Mb)等不仅基因组大而且还含有大量的重复序列，如玉米基因组的重复序列就高达整个基因组的80%。因此，要想从高等植物尤其是大量的经济作物中进行基因富集是一项很有挑战性的工作。对于小基因组的模式物种如水稻(430 Mb)和拟南芥(*Arabidospsis thaliana*)(130 Mb)等可以采用Whole genome shotgun 和 BAC-by-BAC的方法进行高效测序。但是对于大基因组物种而言，利用这两种传统的方法进行基因组测序，不仅昂贵而且由于重复序列的存在导致序列拼接困难，所以通过全基因组测序获得基因的方法不现实。因此，剔除基因组中的“垃圾”序列后进行基因组测序，成为大基因组物种获得基因信息的首选经济实用方法。目前已发展出几种基因富集测序策略如cDNA文库(cDNA library)、甲基化过滤文库(Methylation filtration library, MF)[7~9]和高Cot值文库(High Cot library, HC)[8, 10~12]以及包括利用转座子标签(Transposon tagging)富集基因[13~15]等方法。这些方法各有利弊，必须依据所要达到的目标和物种基因组特性，选择一种或者几种方法的组合进行基因组测序，才能更有利于经济、全面、客观地了解大基因组物种的基因信息。本文在综述各种基因富集基因组测序方法的基础上，讨论了如何高效利用这些方法或方法的组合。

## 1 基因富集的方法

### 1.1 cDNA 文库法(cDNA library)

自 1976 年Hofstetter成功的构建了第一个cDNA文库以来，构建cDNA文库已成为研究功能基因组学的基本手段之一。经典cDNA文库的构建虽然高效、简便，但文库克隆的片段一般较小，单个克隆上的DNA片段太短，所能提供的基因信息很少，大多需要几个克隆才能覆盖一个完整的全基因的cDNA。为了克隆到完整的全长cDNA，建立富含全长cDNA的文库具有重要意义。从构建全长cDNA文库的历程来看，主要有CAPture[16]、Oligo-capping[17, 18]、SMART[19]、Cap-jumping[20]以及Cap-trapper[21~23]等方法。目前，广泛使用的方法是SMART技术。由于基因表达丰度差异很大，因此在构建的cDNA文库中，表达丰度高的基因将被重复测序，而表达丰度低的基因很可能丢失。为了提高低丰度基因的测序概率及降低成本，科学家们发明了均一化cDNA文库法(Normalized cDNA library)[24]。此外，为了满足不同需求如检测具有相同遗传背景的材料在不同发育时期或不同处理之间差异表达的基因，近年来发展的抑制性消减杂交技术[25]已成功应用于植物发育以及外界因子诱导组织细胞中相关应答基因的分析和克隆。目前对于大多数物种而言，全基因组测序是不现实的，为了快速、经济地获得基因序列、了解基因的功能以及基因组中基因数量等相关信息，构建cDNA文库是一种有效、简便且快速的可行方法。所以cDNA文库的构建已成为当前分子生物学研究和基因工程操作的基础。

### 1.2 高 Cot 值文库法(High Cot library, HC)

所谓Cot值指的是：在时间t时，由单链DNA复性成双链DNA时，待复性的单链DNA的起始浓度Co与复性时间t的乘积。在某一浓度条件下，重复序列DNA复性速度要快于单拷贝DNA，即重复序列DNA的Cot值比单拷贝DNA的Cot值小。并且在不同的盐离子浓度洗脱条件下，羟基磷灰石(HAP)具有将单

链DNA和双链DNA分离的功能。因此，结合HAP和DNA复性动力学特点即可将重复序列DNA和低拷贝DNA分离。在植物基因组中，通常存在大量的拷贝数成千上万的重复序列如转座子、反转录转座子、简单重复序列、卫星DNA等。与上述多拷贝重复序列不同，除了部分基因家族外，植物基因多以低拷贝存在。所以，根据DNA复性动力学，不同拷贝数的DNA片段在复性过程中速度不同，经HAP柱后可以将单链DNA(即基因序列)从双链DNA(即重复序列)中分离，从而达到富集基因的效果。目前，这一技术在基因组比较大的作物如玉米[8, 10]、高粱[11]和小麦[12]，以及鸡(*Gallus gallus*)[26]的基因组测序中应用很成功。

早在 20 世纪 60 年代末，Britten和Kohne[27]采用Cot值分析(或者称之为复性动力学)发现在真核基因组中存在大量的重复序列。后来科学家们将Cot/HAP技术应用于构建均一化的cDNA文库、从基因组中分离重复序列用于染色体原位抑制杂交和从基因组中克隆高拷贝的重复序列。直到 2002 年，Peterson等[11]将Cot/HAP技术与克隆测序技术相结合发展出一种新的基因组测序技术即CBCS(Cot-based cloning and sequencing)。Peterson[11]等以高粱为材料，采用CBCS技术依据不同Cot值建立了HR(Highly repetitive)、MR(Moderately repetitive)、SL(Single/low copy)3 种文库；其中HR文库捕获的DNA序列是高拷贝重复序列如反转录转座子、rDNA以及着丝粒重复序列，MR文库捕获的序列是中度重复序列，而SL文库中捕获的序列主要是低拷贝序列如基因。这一技术的问世，使Cot/HAP最终成功地应用于基因组测序并同时富集基因序列，也为其他大基因组、重复序列多、利用常规方法测序昂贵的物种进行基因组测序开辟了新的道路方向，更重要的是为其他大基因组物种进行基因组测序提供了成功的经验和一种极具可行性的新方法。此后，Yuan等[10]、Whitelaw等[8]和Lamoureux等[12]分别在玉米和小麦上进一步确证了利用Cot/HAP方法进行富集基因和基因组测序的可行性以及高基因富集效率。Yuan等[10]采用高Cot法富集玉米基因的效率由随机鸟枪法的 5%提高到 20%，效率提高 4 倍左右；并且随着Cot值的增高，重复序列反转录元件在HC文库中出现的概率降低。但是，是否Cot值越高，基因发现率就越高，关于这一点仍然存在诸多争议。在理论上提高Cot值可以提高基因富集效率，但是实际情况因实验条件和物种而异，例如，Lamoureux等[12]采用高Cot方法构建HC文库富集基因时发现，Cot值分别为 1189 和 1639 的HC文库在序列组成即重复序列和基因含量上没有明显的差异。此外，Lamoureux等[12]认为在理论上，选择大的DNA片段进行复性构建HC文库有利于提高单拷贝或低拷贝基因富集效率，但是并不是DNA片段越大基因富集效果越好，因为这样很可能将重复序列周围的低拷贝序列丢失。因此，在增加DNA片段大小构建HC文库时，需特别谨慎。值得注意的是：Lamoureux等[12]构建的HC文库与鸟枪法构建的文库相比，其优越性非常明显：基因富集效果提高了 13.7 倍，未知单拷贝序列富集效果提高了 5.8 倍，重复序列的剔除效果提高了 3 倍。因此，选择适宜的DNA片段大小范围以及合适的Cot值进行基因富集，是HC方法成功与否的关键。除了将Cot/HAP技术应用于挖掘基因和基因组测序外，该技术还应用于基因组重复序列的分布组成以及基因组进化的研究[26]。

### 1.3　甲基化过滤文库法(Methylation filtration library, MF)

迄今为此，研究学者发现 5mC(胞嘧啶 5 号位的甲基化)在植物基因组中普遍存在。20 世纪八、九十年代，科学家们发现在植物基因组中大部分重复元件具有甲基化的限制性位点，而对于基因而言，它们多处于亚甲基化状态。因此，科学家利用重复元件和基因间的甲基化水平不同，采用甲基化敏感的限制性内切酶构建富含基因的基因组文库。但是这种方法的使用受到限制性酶切位点的限制，导致某些基因在构建富集文库时丢失。直到后来，发现大肠杆菌的*Mcr*BC的限制修饰系统后，科学家们发展出了一种基于大肠杆菌*Mcr*BC的限制修饰系统的全新富集基因的方法，即甲基化过滤文库[9]。

Raleigh等[28]和Dila等[29]发现*Mcr*BC识别DNA序列中[A/G]mC位点，随后Sutherland等[30]发现在大多数植物基因组中每隔 20~30 个碱基就有一个*Mcr*BC识别位点。因此，利用*Mcr*BC$^+$宿主菌构建基因组文库就可以剔除甲基化序列。通常植物基因组中重复序列是处于甲基化状态的，而基因是处于亚甲基化或未甲基化状态的，所以通过这种方法构建文库在

剔除重复序列的同时可以起到富集基因的效果。Rabinowicz等[9]在《Nature Genetics》上发表论文利用甲基化过滤的方法首次对富含重复序列(占整个基因组序列的 80%)的玉米基因组进行测序，发现与未经甲基化过滤的对照文库(鸟枪法文库)相比，甲基化过滤文库的基因富集率要高出 5~7 倍。Palmer等[7]采用相同的方法构建玉米甲基化过滤文库，基因富集效率提高了 6 倍，排除了 93%的重复序列对基因富集的影响。此外，Whitelaw等[8]综合甲基化过滤和高Cot值两种方法对玉米进行基因富集，与未经甲基化过滤文库相比，基因富集效率提高了 4 倍。除了在玉米上取得成功之外，该方法在高粱[31]上也获得成功，获得的550 000个MF 序列代表着96%的高粱基因，而且覆盖基因的平均长度的 65%。同时该方法也应用于其他植物[32]如小麦、大麦、大豆(*Glycine max* (Linn.) Merr.)、油菜(*Brassica napus* Linn)、马铃薯(*Solanum tuberosum* L.)、番茄(*Lycopersicon esculentum* Miller)、棉花(*Gossypium hirsutum* L.)、苔藓(*Ceratodon purpureus* Hedw.)、蕨(*Ceratopteris richardii* L.)和松树(*Pinus taeda* L.)。Rabinowicz等[32]认为与玉米和高粱相比，该方法在小麦和松树上的应用受到了限制，基因富集效率比较低。与其他植物相比，在六倍体小麦和松树中存在同源性序列，导致基因富集效率降低。

尽管Palmer等[7]和Whitelaw等[8]发现通过基因组过滤包括高Cot值和甲基化过滤的方法可以捕获大部分的玉米基因，而且Rabinowicz等[33, 34]发现用甲基化过滤的方法可以检测到 95%的玉米外显子，但是用基因组过滤获得的基因大部分是编码多肽的基因，而对于小分子蛋白和小RNA的基因以及植物基因组中普遍存在的串联重复序列而言，富集技术是否有效？此外，甲基化过滤方法是否可以区分大基因家族内进化为不同功能的不同拷贝基因？除此之外，甲基化过滤方法还有可能丢失一部分基因。值得注意的是，并不是所有物种都适合用甲基化过滤的方法进行测序。尽管大量文献显示，该方法在玉米[7~9, 33]和高粱[31]上获得了巨大成功，但是，也应该注意到在六倍体小麦和松树上，甲基化过滤的作用受到限制。Rabinowicz (私人交流)认为基因组大小与基因的富集效率具有相关性。Okagaki和Philips[35]认为要将基因组过滤方法应用于其他植物基因组测序的两个基本前提条件是：(1)该物种的大部分基因都是处于未甲基化的状态；(2)在该物种的基因内部没有或存在少量的重复序列。此外，还要考虑该物种的基因和基因组结构等因素。所以，尽管甲基化过滤法是一种很好的基因组测序方法，但是并不是所有物种都适用。对于基因组较大，基因序列处于亚甲基化状态的玉米[7~9]、小麦[36]和高粱[31]等物种而言，甲基化过滤法基因组测序是一种非常有效的方法。因此，在采用甲基化过滤法进行基因组测序时，有必要考虑基因组的大小、结构特异性、倍性、和物种差异等重要因素。

### 1.4 跨越甲基衔接物文库法 (Methylation-spanning linker library, MSLL)

用MSLL方法构建玉米文库富集基因是基于以下几点背景而发展出来的：(1)在禾本科植物基因组中，如大麦、小麦、玉米基因组中有 5~20 kb的基因区(Gene block)散状分布在有几个kb到 100 kb的DNA重复序列；在某些情况下，基因丰富的区域可以达到 50 kb以上；(2)在玉米基因组中，存在大量的重复序列，据估计：大约有 80%的DNA重复序列，其中LTR型反转录转座子(LTR-retrotransposons)占很大比例，而仅 20%的序列是基因序列[37]；(3)在高等植物核基因组中，广泛存在 5mC，而且通常在 5′-CG-3 和 5′-CNG-3 序列的胞嘧啶更加容易甲基化(5mC)[38]；(4)在成年组织中，大部分的LTR- retrotransposons在 5′-CG-3 和 5′-CNG-3 序列位点上是100%甲基化；而相同位置基因序列是处于未甲基化状态的[39]；(5)甲基化敏感性核酸内切酶*Hap* II 和*Sal* I 的发现为MSLL的实际操作提供了重要工具和可行性。*Hap* II 和*Sal* I 的识别位点序列分别为：5′-C^CGG-3 和 5′-G^TCGAC-3′。由于这两种酶是甲基化敏感的，所以当上述序列处于甲基化状态(如该序列位于重复序列中)时，这两种酶是不能酶切识别位点的；而当上述序列处于未甲基化状态(如该序列位于基因序列中)时，这两种酶是可以酶切识别位点的。正因为这种特性，导致MSLL克隆的两个末端序列是基因和重复序列之间的交界序列。因此，将其他方法获得的基因序列如MF、HC序列与MSLL序列拼接可以将基因定位在遗传图谱上。目前，该方法仅在玉米[40]上应用成功，在其他物种中尚未见报道。Yuan等[40]

采用*Hap* Ⅱ和*Sal* Ⅰ两种甲基化敏感酶对玉米基因组DNA进行完全酶切后，将酶切片段连接转化到*Mcr*BC$^-$ *E.coli* DH10B感受态细胞中构建3个BAC文库(*Hap* Ⅱ BACs、*Sal* Ⅰ BACs(10~15 kb)、*Sal* Ⅰ BACs(15~25 kb))。挑选克隆测序发现: *Hap* Ⅱ BACs, *Sal* Ⅰ BACs(10~15 kb), *Sal* Ⅰ BACs(15~25 kb)3个文库对已知基因发现率分别为 5.5%、14%、18%, 加权平均值为 10.96%; 而对照*Eco*R Ⅰ BAC文库已知基因发现率与鸟枪法相近仅为 1.3%。由此可见，用MSLL方法富集基因是可行的，至少在玉米上是成功的。至于为何从 2002 年该方法发明以来，未见该方法在其他物种中报道，这可能与MSLL方法本身的局限性(如采用MSLL方法获得的克隆中仍然含有甲基化重复片段)以及对基因组背景知识(如甲基化敏感核酸内切酶的限制位点的数量和分布)的了解尚需深入有关。

### 1.5　亚甲基部分限制性文库法(Hypomethylated partial restriction library, HMPR)

在MSLL方法基础上，Emberton等[41]发明的HMPR方法也是采用甲基化敏感的限制性内切酶*Hpa* Ⅱ(5′-CCGG-3′)和*Hpy* CH4IV(5′-ACGT-3′)构建HMPR文库。其不同之处在于, MSLL是完全酶切，而HMPR为不完全酶切。Emberton等[41]同样也以玉米为材料，构建HMPR文库，测序发现: 大约 25%的HMPR序列与已知基因序列同源，而对照未过滤鸟枪法文库仅为 4%。因此，采用HMPR方法基因富集效率提高了 6~7 倍，与以前采用HC和MF方法相比富集效果一致。值得注意的是: 与HC和MF方法相比, HMPR剔除反转录转座子效果更好，仅残留 5%的反转录转座子序列。与MSLL类似，HMPR也有使用局限性: HMPR只适合于象玉米基因组那样基因序列区域处于非甲基化状态，而重复序列如LTR反转录转座子处于甲基化状态的物种。尽管HMPR和MSLL都可以获得大片段插入DNA，弥补MF/HC获得的拼接序列之间缺口，但是将该方法应用于对基因组的甲基化状态和程度有一定研究背景的物种更容易获得成功。

### 1.6　利用转座子富集基因法(Transposon tagging)

在玉米中，利用*Mutator*转座子倾向于插入单拷贝序列DNA的特性[42]，通过改造*Mu*转座子获得在整个基因组中转座活性更强的*RescueMu*转座子，可以从转座子的标签序列向两端测序挖掘基因[13, 15]。除了利用*RescueMu*转座子挖掘基因外，还有利用*Ac*转座子挖掘基因的报道[14]。由于转座子和反转录转座子在玉米中的研究比较深入，采用转座子方法富集基因是实际可行的。Settles等[43]在玉米中通过插入转座子进而得到一系列插入突变体和侧翼序列标签(Flanking sequence tags, FSTs)，并通过FSTs将这些插入突变定位到基因组的一些特异位点，为反向遗传学研究提供了广泛的资源。还有利用*Ac/Ds*转座子在胡萝卜 (*Daucus carota* L.)中挖掘基因的报道[44]。除了此前已报道过的*Jordan*转座子成功地在团藻(*Volvox carteri*)中标记到了控制团藻发育过程关键方面的一些基因之外，近期又在团藻中发现一种新的*Idaten*转座子，被认为是在团藻中除*Jordan*外又一个可用来标记重要发育基因的有力工具[45]。但是，由于可用于该方法的转座子数量有限，具有物种特异性，并且富集基因的方法效率较低，不能实现真正的高通量大规模富集基因，这也是限制是其广泛应用的主要因素。此外，对于基因组结构了解不是很清楚的物种而言，由于转座子的适用性有限，利用转座子大规模挖掘基因仍存在诸多困难。

## 2　各种基因富集方法的比较和综合利用

在上述众多基因富集方法中，目前应用比较广泛的是cDNA文库法、高Cot值法和甲基化过滤法。此外，构建不同组织、不同发育时期的cDNA文库获得表达基因序列也是当前富集基因的常用方法，对于大多数物种而言也是切实可行的方法。由于基因表达具有时空性即某些基因仅在某些特定的组织部位和/或者发育时期才表达，而且有些基因只在特定的胁迫环境条件下才表达，所以采用cDNA文库法富集基因时，需构建不同发育时期、组织部位和不同生长环境条件下的多个cDNA文库才可能获取完整的全基因组基因信息。此外，由于各个基因的表达丰度不同(如看家基因表达量高，而某些组织特异性的调控基因表达量相当低)，导致高表达的基因被多次重复的测序，而低表达的基因测序的可能性降低。因此，获得比较完整的表达基因谱势必提高测序成本。为了解决这一问题，大量文献报道采用各种方法构建均一化cDNA文库(Normalization cDNA li-

brary)[24]可有效降低高表达基因的代表性，提高低表达稀有基因的代表性，以期更经济地获得完整的基因组表达基因。在一定程度上，均一化cDNA文库降低了成本，为低丰度基因的克隆测序奠定基础。此外，SSH文库的发明为科学家挖掘相同遗传背景不同发育时期、生长环境之间的差异表达基因提供了高效率富集方法。随着基因组研究的发展及后基因组时代的到来，单个基因的CDS(Coding sequence)序列的获得已远远不能满足生命科学发展的需求；获得和了解与基因表达相关的调控序列成为研究热点。然而，cDNA文库是无法提供启动子和内含子等非编码序列信息的。相比之下，高Cot值法和甲基化过滤法、HMPR法和MSLL法可以弥补cDNA文库法的这一缺陷。

早期，科学家采用鸟枪法对基因组进行测序，为人们全面了解基因组信息提供了方法，尤其是基因组中的非编码和重复序列、基因数量(Gene content)等。但是由于采用鸟枪法基因组测序花费昂贵，对于一些基因组大的作物如玉米、高粱、小麦和大麦等而言，这种方法是不可行的。此外，HMPR和MSLL方法仅局限于基因组背景(甲基化位点的分布和数量)比较清楚的物种。而利用转座子标签方法富集基因的限制因素是：(1)可适用的转座子少，而且转座子具有物种特异性；(2)不能大规模，高通量的获得基因调控序列。由于这些不利因素的限制，导致这些方法在实践应用中受到限制，因此近期的文献鲜有报道。而高Cot值法和甲基化过滤法与鸟枪法相比在方法学上具有明显的优势。因为高Cot值法和甲基化过滤法在构建文库时，都将基因组中大量的高拷贝重复序列剔除，仅剩下占基因组比例比较小的低拷贝序列用于基因组测序(主要包括基因空间，Gene space)。与鸟枪法相比，综合HC和MF方法可以将基因发现率提高 4 倍[7, 8]，因此，将HC和MF方法应用于基因组较大，重复序列多的作物进行基因组测序可以大大节省成本，提高挖掘新基因效率，了解更多物种(除模式物种外)的基因组信息。这一优势在玉米[7, 8, 10, 32, 33, 46]、高粱[11, 31]、小麦[12, 36]、大麦、大豆、马铃薯、番茄、棉花[32]等重要作物中得到了充分体现，基因富集程度都达到或高于 2 倍，大麦甚至达到 18.7 倍，玉米达到了 13.2 倍。

但是对于基因富集而言，高Cot值法和甲基化过滤法也各有优劣。与甲基化过滤法相比，高Cot值法具有以下优点：(1)除获得单拷贝或低拷贝序列(即基因)外，同时还获得了不同重复度的重复序列，有利于了解基因组中重复序列的组成、分布、演化乃至整个基因组的进化[26]；(2)在构建HC文库挖掘基因时，不受材料由于组织部位和发育时期等因素导致甲基化水平和模式变更的影响；(3)具有转录活性的转座子和通过CpG镇压或者转录激活脱甲基化的重复元件也有可能包含在MF文库中。Whitelaw等[8]和Palmer等[7]发现在MF文库中仍然残留了比例较高的LTR反转录转座子序列，而这种序列在HC文库中极少存在；(4)由于细胞器DNA通常是处于未甲基化状态的，因此构建的MF文库中含有大量的胞质基因组DNA。尽管Rabinowicz[47]改进了MF的构建方法，但是在文库中有胞质基因组DNA出现仍然不可避免。与高Cot值法相比，甲基化过滤法具有以下优点：(1)基因家族和旁系基因有可能在构建HC文库时被均一化而丢失，而在MF文库中这些基因被保留下来；(2)在构建HC文库时，需要价格昂贵的仪器设备以及熟练的技艺，而构建MF则不受这些因素的限制；(3)构建HC文库时，采用的片段化DNA都比较小(一般在 300 bp左右)，所以使后续的序列拼接工作遭遇更多困难。而构建的MF文库时，选择的碎裂DNA片段大小一般在 0.5~4 kb之间，序列拼接问题可以得到有效改善。由上述可知，这两种方法各有优缺点，综合HC和MF两种方法，取其之长、避其之短才是今后基因富集的有效方法。

Fu等[48]从GenBank中提取 183 条GSSs(Genome survey sequences)(105 MF＋78HC)与 10 条对照基因序列比较发现：启动子、外显子和内含子在MF和HC文库中出现的比例(MF/HC)分别是 53%/19%、71%/56%和 37%/51%。由此可见，MF和HC两种方法对基因序列中启动子、外显子和内含子的捕获具有偏爱性。这种偏爱性表现在：与HC相比，MF更倾向于捕获基因 5′末端序列；而对于完整基因序列而言，MF和HC更倾向于捕获基因 3′末端序列[49]。这一点，与Whitelaw等[8]的报道相一致。Whitelaw等[8]发现用这两种方法捕获的基因序列中仅有大约 1/3 的序列是重叠的，而另外 2/3 的序列是不同的，这充分说明两种方法捕获玉米基因组不同部位的序列。用 78 个玉米全长cDNA序列与从NCBI(National Center for Biotechnol-

ogy Information)收集的 587 063 条MF序列、445 286 条HC序列、178 125 条RM (RescueMu)序列和 362 534 条RC(Random clone)序列进行比较，通过模拟发现95%的玉米基因可以用HC或者MF方法捕获；值得注意的是，综合两种方法后基因发现率为 100%[49]。因此，综合MF、HC两种方法既可以弥补两种方法捕获基因序列不同部位上的差异问题，为获得完整基因序列提供可靠依据；又可以提高基因组中基因发现率，为科学家们提供更多的基因组信息。由此可见，综合两种或多种方法进行基因组测序可能是未来对大基因组、重复序列多的物种基因组测序的主要选择方法之一。

## 3　甲基化过滤技术(Methylation filtration, MF)应用进展

近年来，由于测序技术的迅猛发展，越来越多重要的经济作物或农作物的基因组测序计划也开始了，而利用甲基化过滤方法对庞大复杂的基因组进行基因富集，是一个着手切入研究这些大基因组物种的有效途径。由于，具有降低测序成本，高通量获得基因序列和基因调控序列等优点，甲基化过滤技术已经在经济作物中得到广泛应用。Wang等[50]利用甲基化过滤的策略对番茄核基因组结构进行了研究，分析了基因组中的非甲基化序列构成和在基因序列区域的分布规律，其结果为番茄全基因组测序和其他茄科物种的基因组测序工作提供策略上的指导。在棉花基因组测序中，甲基化过滤技术和高Cot方法则被用于获取那些在EST序列中没有的基因组中的低拷贝DNA序列[51]。豇豆的基因组大小预计有 620 Mb，Chen等[52]通过甲基化过滤方法构建得到一个与未经甲基化过滤相比的基因富集率达 4.1 倍的基因组DNA文库，并估算豇豆基因组中亚甲基化的基因富集区的大小约是 151 Mb。烟草是一种重要的经济作物但是其基因组十分庞大，约有 4.5 Gb，甲基化过滤技术也成功地运用于烟草基因组的研究，获得了基因富集率达 10 倍，基因区域覆盖率约 1.0~1.4 倍的序列信息，这些序列标记到了烟草基因组中约90%的ORF[53]，并在此基础上深入挖掘分析，构建了一个烟草转录因子数据库TOBFAC[54]。

中国的传统中药植物多为研究基础较弱并且基因组复杂的物种。目前武汉植物园已经启动了少数药用植物的基因组学研究，其中重点研究的物种淫羊藿的基因组约为 4 500 Mb(未发表结果)，且只有 6 条染色体，是研究大基因组结构和药用植物次生代谢的典型物种。我们实验室也正在运用甲基化过滤技术开展淫羊藿基因组结构的相关研究工作。初步的研究结果发现仅甲基化敏感的感受态细胞可以一定程度上过滤重复序列，促进基因组中的有用基因序列的发掘。深入的研究正在开展中，期望以此获得的基因组信息有利于淫羊藿的相关研究，为小檗科众多药用植物的相关研究提供技术和序列信息参考，并为其他大基因组药用植物和重要经济植物的研究提供科学的研究策略。

## 4　结 语

综上所述，目前对于大多数基因组结构背景不是很清楚的物种而言，采用 cDNA 文库法依然是一种高效，快捷的富集基因方法。但是由于 cDNA 文库法只能富集基因的编码序列，因此，为了获得基因的非编码序列如启动子、具有功能的内含子和基因间序列等以便于更好的了解基因组信息及分离克隆到目的基因，采用高 Cot 值法、甲基化过滤法、MSLL 和 HMPR 构建基因组文库成为比较合理的选择方案。又由于 MSLL 和 HMPR 方法必须对物种基因组背景信息比较了解，使得将这两种方法应用于大基因组物种基因富集受到限制。高 Cot 值法和甲基化过滤法各有优缺点且具有互补性(两种方法对基因的 5 和 3 末端具有不同的富集偏好性)，因此将这两种方法综合使用是今后高通量获取大基因组物种基因信息的首选方法。

## 参考文献(References):

[1] Yu J, Hu SN, Wang J, Wong GKS, Li SG, Liu B, Deng YJ, Dai L, Zhou Y, Zhang XQ, Cao ML, Liu J, Sun JD, Tang JB, Chen YJ, Huang XB, Lin W, Ye C, Tong W, Cong LJ, Geng JN, Han YJ, Li L, Li W, Hu GQ, Huang XG, Li WJ, Li J, Liu ZW, Liu JP, Qi QH, Liu JS, Li T, Wang XG, Lu H, Wu TT, Zhu M, Ni PX, Han H, Dong W, Ren XY, Feng XL, Cui P, Li XR, Wang H, Xu X, Zhai WX, Xu Z, Zhang JS, He SJ, Zhang JG, Xu JC, Zhang KL, Zheng XW, Dong JH, Zeng WY, Tao L, Ye J, Tan J, Ren XD, Chen XW, He J, Liu DF, Tian W, Tian CG, Xia HG, Bao QY, Li G, Gao H, Cao T, Zhao WM, Li P, Chen W, Wang XD, Zhang Y, Hu

JF, Liu S, Yang J, Zhang GY, Xiong YQ, Li ZJ, Mao L, Zhou CS, Zhu Z, Chen RS, Hao BL, Zheng WM, Chen SY, Guo W, Li GJ, Liu SQ, Tao M, Zhu LH, Yuan LP, Yang HM. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, 2002, 296(5565): 79–92.[DOI]

[2] Ming R, Hou SB, Feng Y, Yu QY, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KLT, Salzberg SL, Feng L, Jones MR, Skelton RL, Murray JE, Chen CX, Qian WB, Shen JG, Du P, Eustice M, Tong E, Tang HB, Lyons E, Paull RE, Michael TP, Wall K, Rice DW, Albert H, Wang ML, Zhu YJ, Schatz M, Nagarajan N, Acob RA, Guan PZ, Blas A, Wai CM, Ackerman CM, Ren Y, Liu C, Wang JM, Wang JP, Na JK, Shakirov EV, Haas B, Thimmapuram J, Nelson D, Wang XY, Bowers JE, Gschwend AR, Delcher AL, Singh R, Suzuki JY, Tripathi S, Neupane K, Wei HR, Irikura B, Paidi M, Jiang N, Zhang WL, Presting G, Windsor A, Navajas-Perez R, Torres MJ, Feltus FA, Porter B, Li YJ, Burroughs AM, Luo MC, Liu L, Christopher DA, Mount SM, Moore PH, Sugimura T, Jiang JM, Schuler MA, Friedman V, Mitchell-Olds T, Shippen DE, dePamphilis CW, Palmer JD, Freeling M, Paterson AH, Gonsalves D, Wang L, Alam M. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature*, 2008, 452(7190): 991–997.[DOI]

[3] Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyere C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pe ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quetier F, Wincker P, French-Italian Public. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 2007, 449(7161): 463–467.[DOI]

[4] Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Dejardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjarvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leple JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouze P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y, Rokhsar D. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, 2006, 313(5793): 1596–1604.[DOI]

[5] Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud PF, Lindquist EA, Kamisugi Y, Tanahashi T, Sakakibara K, Fujita T, Oishi K, Shin-I T, Kuroki Y, Toyoda A, Suzuki Y, Hashimoto S, Yamaguchi K, Sugano S, Kohara Y, Fujiyama A, Anterola A, Aoki S, Ashton N, Barbazuk WB, Barker E, Bennetzen JL, Blankenship R, Cho SH, Dutcher SK, Estelle M, Fawcett JA, Gundlach H, Hanada K, Heyl A, Hicks KA, Hughes J, Lohr M, Mayer K, Melkozernov A, Murata T, Nelson DR, Pils B, Prigge M, Reiss B, Renner T, Rombauts S, Rushton PJ, Sanderfoot A, Schween G, Shiu SH, Stueber K, Theodoulou FL, Tu H, Van de Peer Y, Verrier PJ, Waters E, Wood A, Yang LX, Cove D, Cuming AC, Hasebe M, Lucas S, Mishler BD, Reski R, Grigoriev IV, Quatrano RS, Boore JL. The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science*, 2008, 319(5859): 64–69.[DOI]

[6] Ayele M, Haas BJ, Kumar N, Wu H, Xiao YL, Van Aken S, Utterback TR, Wortman JR, White OR, Town CD. Whole genome shotgun sequencing of Brassica oleracea and its application to gene discovery and annotation in *Arabidopsis*. *Genome Res*, 2005, 15(4): 487–495.[DOI]

[7] Palmer LE, Rabinowicz PD, O'Shaughnessy AL, Balija VS, Nascimento LU, Dike S, de la Bastide M, Martienssen RA, McCombie WR. Maize genome sequencing by methylation filtrations. *Science*, 2003, 302(5653): 2115–2117.[DOI]

[8] Whitelaw CA, Barbazuk WB, Pertea G, Chan AP, Cheung F, Lee Y, Zheng L, van Heeringen S, Karamycheva S, Bennetzen JL, SanMiguel P, Lakey N, Bedell J, Yuan Y, Budiman MA, Resnick A, Van Aken S, Utterback T, Riedmuller S, Williams M, Feldblyum T, Schubert K, Beachy R, Fraser CM, Quackenbush J. Enrichment of

gene-coding sequences in maize by genome filtration. *Science*, 2003, 302(5653): 2118–2120.[DOI]

[9] Rabinowicz PD, Schutz K, Dedhia N, Yordan C, Parnell LD, Stein L, McCombie WR, Martienssen RA. Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat Genet*, 1999, 23(3): 305–308.[DOI]

[10] Yuan YN, SanMiguel PJ, Bennetzen JL. High-Cot sequence analysis of the maize genome. *Plant J*, 2003, 34(2): 249–255.[DOI]

[11] Peterson DG, Schulze SR, Sciara EB, Lee SA, Bowers JE, Nagel A, Jiang N, Tibbitts DC, Wessler SR, Paterson AH. Integration of Cot analysis, DNA cloning, and high- throughput sequencing facilitates genome characterization and gene discovery. *Genome Res*, 2002, 12(5): 795–807.[DOI]

[12] Lamoureux D, Peterson DG, Li WL, Fellers JP, Gill BS. The efficacy of Cot-based gene enrichment in wheat (*Triticum aestivum* L.). *Genome*, 2005, 48(6): 1120–1126.[DOI]

[13] Fernandes J, Dong QF, Schneider B, Morrow DJ, Nan GL, Brendel V, Walbot V. Genome-wide mutagenesis of *Zea mays* L. using RescueMu transposons. *Genome Biol*, 2004, 5: R82.[DOI]

[14] Cowperthwaite M, Park W, Xu Z, Yan X, Maurais SC, Dooner HK. Use of the transposon Ac as a gene-searching engine in the maize genome. *Plant Cell*, 2002, 14(3): 713–726.[DO]

[15] Hanley S, Edwards D, Stevenson D, Haines S, Hegarty M, Schuch W, Edwards KJ. Identification of transposon-tagged genes by the random sequencing of Mutator-tagged DNA fragments from *Zea mays*. *Plant J*, 2000, 23(4): 557–566.[DOI]

[16] Edery I, Chu LL, Sonenberg N, Pelletier J. An efficient strategy to isolate full-length cDNAs based on an messenger-RNA cap petention procedure (capture). *Mol Cell Biol*, 1995, 15(6): 3363–3371.

[17] Suzuki Y, Sugano S. Construction of a full-length enriched and a 5'-end enriched cDNA library using the oligo- capping method. *Methods Mol Biol*, 2003, 221: 73–91.

[18] Suzuki Y, Yoshitomo-Nakagawa K, Maruyama K, Suyama A, Sugano S. Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library. *Gene*, 1997, 200(1-2): 149–156.[DOI]

[19] Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques*, 2001, 30(4): 892–897.

[20] Efimov VA, Chakhmakhcheva OG, Archdeacon J, Fernandez JM, Fedorkin ON, Dorokhov YL, Atabekov JG. Detection of the 5'-cap structure of messenger RNAs with the use of the cap-jumping approach. *Nucl Acids Res*, 2001, 29(22): 4751–4759. [DOI]

[21] Carninci P, Kvam C, Kitamura A, Ohsumi T, Okazaki Y, Itoh M, Kamiya M, Shibata K, Sasaki N, Izawa M, Muramatsu M, Hayashizaki Y, Schneider C. High- efficiency full-length cDNA cloning by biotinylated cap trapper. *Genomics*, 1996, 37(3): 327–336.[DOI]

[22] Carninci P, Shibata Y, Hayatsu N, Sugahara Y, Shibata K, Itoh M, Konno H, Okazaki Y, Muramatsu M, Hayashizaki Y. Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res*, 2000, 10(10): 1617–1630.[DOI]

[23] Carninci P, Westover A, Nishiyama Y, Ohsumi T, Itoh M, Nagaoka S, Sasaki N, Okazaki Y, Muramatsu M, Schneider C, Hayashizaki, Y. High efficiency selection of full-length cDNA by improved biotinylated cap trapper. *DNA Res*, 1997, 4(1): 61–66.[DOI]

[24] Soares MB, Bonaldo MD, Jelene P, Su L, Lawton L, Efstratiadis A. Construction and characterization of a normalized cDNA library. *Proc Natl Acad Sci USA*, 1994, 91(20): 9228–9232.[DOI]

[25] Diatchenko L, Lau YFC, Campbell AP, Chenchik A, Moqadam F, Huang B, Lukyanov S, Lukyanov K, Gurskaya N, Sverdlov ED, Siebert PD. Suppression subtractive hybridization: A method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc Natl Acad Sci USA*, 1996, 93(12): 6025–6030.[DOI]

[26] Wicker T, Robertson JS, Schulze SR, Feltus FA, Magrini V, Morrison JA, Mardis ER, Wilson RK, Peterson DG, Paterson AH, Ivarie R. The repetitive landscape of the chicken genome. *Genome Res*, 2005, 15(1): 126–136.[DOI]

[27] Britten RJ, Kohne DE. Repeated sequences in DNA. *Science*, 1968, 161(3841): 529–540.[DOI]

[28] Raleigh EA, Wilson G. Escherichia-coli K-12 restricts DNA containing 5 methylcytosine. *Proc Natl Acad Sci USA*, 1986, 83(23): 9070–9074.[DOI]

[29] Dila D, Sutherland E, Moran L, Slatko B, Raleigh EA. Genetic and sequence organization of the McrBC locus of Escherichia coli K–12. *J Bacteriol*, 1990, 172(9): 4888–4900.

[30] Sutherland E, Coe L, Raleigh EA. McrBC- a multisubunit GTP–dependent restriction endonuclease. *J Mol Bio*, 1992, 225(2): 327–348.[DOI]

[31] Bedell JA, Budiman MA, Nunberg A, Citek RW, Robbins D, Jones J, Flick E, Rohlfing T, Fries J, Bradford K, McMenamy J, Smith M, Holeman H, Roe BA, Wiley G, Korf IF, Rabinowicz PD, Lakey N, McCombie WR, Jeddeloh JA,

Martienssen RA. Sorghum genome sequencing by methylation filtration. *PloS Biol*, 2005, 3(1): 103–115.[DOI]

[32] Rabinowicz PD, Citek R, Budiman MA, Nunberg A, Bedell JA, Lakey N, O'Shaughnessy AL, Nascimento LU, McCombie WR, Martienssen RA. Differential methylation of genes and repeats in land plants. *Genome Res*, 2005, 15(10): 1431–1440.[DOI]

[33] Rabinowicz PD, McCombie WR, Martienssen RA: Gene enrichment in plant genomic shotgun libraries. *Curr Opin Plant Biol*, 2003, 6(2): 150–156.[DOI]

[34] Rabinowicz PD, Palmer LE, May BP, Hemann MT, Lowe SW, McCombie WR, Martienssen RA. Genes and transposons are differentially methylated in plants, but not in mammals. *Genome Res*, 2003, 13(12): 2658–2664.[DOI]

[35] Okagaki RJ, Phillips RL. Maize DNA-sequencing strategies and genome organization. *Genome Biol*, 2004, 5: 223.[DOI]

[36] Li WL, Zhang P, Fellers JP, Friebe B, Gill BS. Sequence composition, organization, and evolution of the core Triticeae genome. *Plant J*, 2004, 40(4): 500–511.[DOI]

[37] SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, MelakeBerhan 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] Gruenbaum Y, Navehmany T, Cedar H, Razin A. DNA seqenece specificity of methylation in higher-plant DNA. *Nature*, 1981, 292(5826): 860–862.[DOI]

[39] Gruenbaum Y, Stein R, Cedar H, Razin A. Methylation of CpG sequences in eukaryotic DNA. *FEBS Lett*, 1981, 124(1): 67–71.[DOI]

[40] Yuan YN, SanMiguel PJ, Bennetzen JL. Methylation-spanning linker libraries link gene-rich regions and identify epigenetic boundaries in *Zea mays*. *Genome Res*, 2002, 12(9): 1345–1349.

[41] Emberton J, Ma JX, Yuan YN, SanMiguel P, Bennetzen JL. Gene enrichment in maize with hypomethylated partial restriction (HMPR) libraries. *Genome Res*, 2005, 15(10): 1441–1446.[DOI]

[42] Cresse AD, Hulbert SH, Brown WE, Lucas JR, Bennetzen JL. Mu1-related transposable elements of maize preferentially insert into low copy number DNA. *Genetics*, 1995, 140(1): 315–324.

[43] Settles AM, Holding DR, Tan BC, Latshaw SP, Liu J, Suzuki M, Li L, O'Brien BA, Fajardo DS, Wroclawska E, Tseung CW, Lai JS, Hunter CT, Avigne WT, Baier J, Messing J, Hannah LC, Koch KE, Becraft PW, Larkins BA, McCarty DR. Sequence-indexed mutations in maize using the UniformMu transposon-tagging population. *BMC Genomics*, 2007, 8: 116.[DOI]

[44] Ipek A, Masson P, Simon PW. Genetic transformation of an Ac/Ds-based transposon tagging system in carrot (*Daucus carota* L.). *Eur J Hortic Sci*, 2006, 71(6): 245.

[45] Ueki N, Nishii I. Idaten is a New Cold-inducible Transposon of Volvox carteri that can be used for tagging developmentally important genes. *Genetics*, 2008, 180(3): 1343–1353.[DOI]

[46] Chan AP, Pertea G, Cheung F, Lee D,Zheng L, Whitelaw C, Pontaroli AC, SanMiguel P, Yuan YN, Bennetzen J, Barbazuk WB, Quackenbush J, Rabinowicz PD. The TIGR maize database. *Nucl Acids Res*, 2006, 34(Suppl. 1): D771–776.[DOI]

[47] Rabinowicz PD. Constructing gene-enriched plant genomic libraries using methylation filtration technology. *Methods Mol Biol*, 2003, 236: 21–36.

[48] Fu Y, Hsia A-P, Guo L, Schnable PS. Types and frequencies of sequencing errors in methyl-filtered and high Cot maize genome survey sequences. *Plant Physiol*, 2004, 135(4): 2040–2045.[DOI]

[49] Springer NM, Xu XQ, Barbazuk WB. Utility of different gene enrichment approaches toward identifying and sequencing the maize gene space. *Plant Physiol*, 2004, 136(2): 3023–3033.[DOI]

[50] Wang Y, van der Hoeven RS, Nielsen R, Mueller LA, Tanksley SD. Characteristics of the tomato nuclear genome as determined by sequencing undermethylated *Eco*R Ⅰ digested fragments. *Theor Appl Genet*, 2005, 112(1): 72–84.[DOI]

[51] Chen ZJ, Scheffler BE, Dennis E, Triplett BA, Zhang T, Guo W, Chen X, Stelly D, Rabinowicz PD, Town CD, Arioli T, Brubaker C, Cantrell RG, Lacape J, Ulloa M, Chee P, Gingle AR, Haigler CH, Percy R, Saha S, Wilkins T, Wright RJ, Van Deynze A, Zhu Y, Yu S, Abdurakhmonov I, Katageri I, Kumar PA, Mehboob R, Zafar Y, Yu JZ, Kohel RJ, Wendel JF, Paterson AH. Toward sequencing cotton (*Gossypium*) genomes. *Plant Physiol*, 2007, 145(4): 1303–1310.[DOI]

[52] Chen X, Laudeman TW, Rushton PJ, Spraggins TA, Timko MP. CGKB: an annotation knowledge base for cowpea (*Vigna unguiculata* L.) methylation filtered genomic genespace sequences. *BMC Bioinformatics*, 2007, 8: 129.[DOI]

[53] Rushton PJ, Bokowiec MT, Han S, Zhang H, Brannock JF, Chen X, Laudeman TW, Timko MP. Tobacco transcription factors: novel insights into transcriptional regulation in the Solanaceae. *Plant Physiol*, 2008, 147(1): 280–295.[DOI]

[54] Rushton PJ, Bokowiec MT, Laudeman TW, Brannock JF, Chen X, Timko MP. TOBFAC: the database of tobacco transcription factors. *BMC Bioinformatics* 2008, 9: 53.[DOI]