应用昆虫学报 *Chinese Journal of Applied Entomology* 2013,50(2):293 - 297. DOI:10.7679/j.issn.2095 - 1353.2013.039

线粒体基因组测序策略和方法*

沙 淼 林立亮 李雪娟 黄 原**

(陕西师范大学生命科学学院 西安 710062)

摘 要 本文综述了线粒体基因组测序策略和方法,在传统测序方法中介绍了基于物理分离线粒体 DNA 的克隆文库测序方法和基于 PCR 扩增产物的直接测序方法,后者重点介绍了基于长 PCR 扩增产物的引物步移法和基于总 DNA 的引物步移法;应用新一代高通量测序方法有基于总 DNA 样品的方法,包括需要预扩增 mtDNA 的多物种平行高通量和无需预扩增 mtDNA 的高通量方法,基于总 RNA 样品的转录组测序方法等。在实际工作中,选择哪种方法取决于研究规模、样品大小和保存状态、经费情况等。总的来说,基于长 PCR 扩增产物的引物步移法尤其适合小规模昆虫线粒体基因组研究,而对于大规模线粒体基因组研究来说,NGS 技术无疑是省时省力的最佳选择。

关键词 线粒体基因组,长 PCR,引物步移法,高通量测序

Strategy and methods for sequencing mitochondrial genome

SHA Miao LIN Li-Liang LI Xue-Juan HUANG Yuan**

(College of Life Science, Shaanxi Normal University, Xi'an 710062, China)

Abstract This paper reviews strategy and methods for sequencing mitochondrial genomes. Two traditional sequencing methods are introduced; the clone library-based sequencing method from physically purified mtDNA, and the PCR amplicon-based sequencing method, which includes primer-walking sequencing from long PCR amplicons and primer-walking sequencing from total DNA. For the application of next-generation sequencing strategies, there are total DNA-based methods, which include the parallel tagged sequencing (PTS) method based on pre-amplifying mtDNA and total DNA, and the transcriptomic sequencing method based on total RNA. From a practical perspective, which method is chosen depends on the scale of the research, size and preservative status of the sample, and financial support. The primer walking sequencing method based on long-PCR amplicons is particularly suitable for small scale research on insect mitochondrial genomes. For large scale mitochondrial genome research, NGS technology is the best choice for saving time and effort.

Key words mitochondrial genome, long-PCR, primer-walking, next-generation sequencing

1 前言

线粒体基因组作为研究 DNA 复制和转录的良好模型,同时也是遗传和进化上广泛使用的分子标记,近年来被广泛应用于生物学的许多领域。后生动物线粒体基因组大小一般在 16 ~ 17 kb 之间,编码 13 个蛋白质基因,2 个 rRNA 基因和 22 个 tRNA 基因。自从第一个线粒体基因组(人类 mtDNA)序列 1981 年被测定以来,截止 2012 年 12 月底在 NCBI 基因组数据库中登记的线粒体全基因组序列的后生动物有 2 924 种,其中昆虫 346 种。线粒体基因组的研究方法在近几十年中不断发展,尤其最近高通量测序技术的出现,为线粒体全基因组测序提供了新的策略。本文对目前常用的传统和高通量线粒体基因组测序策略和方法进行了综述。

* 资助项目:国家自然科学基金(31172076,30970346)。

** 通讯作者,E-mail:yuanh@snnu.edu.cn

收稿日期:2013-01-01,接受日期:2013-01-10

2 传统线粒体基因组测序方法

传统线粒体基因组测序技术主要是基于克隆文库或 PCR 扩增产物的 Sanger 测序方法,其中基于 PCR 扩增产物的引物步移法仍然是目前小规模线粒体基因组研究的主流方法。

2.1 基于物理分离线粒体 DNA 的克隆文库测序方法

物理分离 mtDNA 是通过氯化铯密度梯度离心或差速离心法将细胞的不同组分分开获得纯度较高的 mtDNA,然后再选用切割位点较少且分布较均匀的一种或几种核酸限制性内切酶对 mtDNA 进行部分或完全酶切,获得长度为数百碱基适合用于测序的短片段,或者利用不同强度的超声波将获得的 mtDNA 随机打断,形成不同长度的短片段,然后将其克隆到质粒载体中进行测序(Tamura and Aotsuka,1988)。

早期应用这种方法较多,如 Tzeng 等(1992)利用氯化铯密度梯度离心获得高纯度的樱口鳉(*Crossostoma lacustre*) mtDNA,用限制性内切酶 *Hind*Ⅲ将其切割为 9 个片段,然后单独克隆到 Bluescript KS + 质粒载体中,以大肠杆菌 JM101 或 JM109 为宿主进行扩增并用链终止法进行测序,经过序列拼接获得了该物种完整的线粒体基因组。鲁成等(2002)利用差速离心法获得家蚕 *Bombyx mori* 的 mtDNA,使用超声波将其随机片段化,然后回收 1.5 ~ 3.0 kb 的 DNA 片段克隆到 pUC18 载体上,转化入大肠杆菌 DH5α 中,挑取 300 个克隆,培养后提取重组质粒测序。

基于物理分离的线粒体基因组测序策略能够获得高纯度的 mtDNA,可以保证序列的准确性,但操作较繁琐,非常耗时,对实验材料消耗很大,需要大量新鲜样本,无法用于大多数物种的线粒体基因组测序。

2.2 基于 PCR 扩增产物直接测序的方法

该策略避免了从样本中直接提取纯化 mtDNA 的过程,而是基于包括核 DNA 和 mtDNA 的总 DNA 提取物,再通过 PCR 扩增出线粒体特定片段,纯化后直接测序。对于因为 mtDNA 异质性或复杂二级结构(主要是控制区序列)而导致的 PCR 产物无法直接测序的序列,需要将 PCR 产物克隆后,挑选 3 ~ 5 个克隆进行克隆测序。这种方法需要设计几十对相互重叠的引物来扩增覆盖线粒体全基因组的序列,故也称为引物步移法(primer-walking)。

基于 PCR 扩增产物直接测序的方法可以采用多种策略,大体可以归纳为基于长 PCR 的方法和无长 PCR 的引物步移法(即基于总 DNA 的引物步移法)。

(1)基于长 PCR 的方法　由于长 PCR 技术能够扩增 5 kb 以上的 DNA 片段,所以通过设计 2 ~ 5 对特异性长 PCR 引物进行长 PCR,可以从总 DNA 中特异性扩增出大片段的线粒体 DNA(叶维萍和黄原,2003)。然后将获得的大片段 DNA 通过以下方法片段化成适合测序的长度:

①克隆文库测序法:与离心法相同,利用一种或几种限制性内切酶对长 PCR 产物进行酶切,然后克隆到质粒载体中进行测序(Yamauchi *et al.*, 2004)。

②基于长 PCR 产物的引物步移法:以长 PCR 产物为模板进行两次 PCR,对得到的短片段进行直接测序(Machida *et al.*,2004)。

③以纯化的长 PCR 产物为测序模板,采用引物步移法直接测序。

基于长 PCR 方法的关键在于模板 DNA 质量和长 PCR 引物的设计。长 PCR 扩增的成功首先需要大分子 DNA 片段,所以要求样本来源于新鲜材料或保存条件较好的近几年的材料;此外,扩增片段越长,反应的特异性越低,所以长 PCR 对引物的要求更加严格。

基于长 PCR 策略的优点是需要的起始 DNA 样品很少,特别适合微小型昆虫;另外,大部分动物核基因组中都存在线粒体假基因,然而这些假基因通常都小于 1 kb,所以利用长 PCR 能够有效避免线粒体假基因的干扰。

(2)基于总 DNA 的引物步移法

由于目前在各类动物中已经测出的线粒体基因组很多,使得覆盖线粒体全基因组的保守引物的设计比较容易实现,这样如果从样本中提取的 DNA 量足够 50 次以上 PCR 反应使用的话,就可以直接从以 DNA 为模板进行引物步移法测定线粒体基因组。如 Song 和 Liang(2009)利用通用引物对半翅目碧蛾蜡蝉(*Geisha distinctissima*)线粒体基因组的 *cox1*、*cox2*、*nad5*、*cytb*、*lrRNA* 和 *srRNA* 基因进行 PCR 扩增并测序,然后在这些序列的基础

上设计出 17 对用于长 PCR 的特异性引物,将线粒体全基因组扩增出来并测序。

无论基于总 DNA 还是长 PCR 产物的引物步移法,对引物的数量和保守性都有一定的要求。在引物数量方面,一般至少需要 30 对以上的测序引物,这样双向测序后可以获得 42 kb 的有效序列(以单个测序反应的有效读长 700 bp 计算),如果以后生动物普遍的线粒体基因组大小为 17 kb 计算的话,为保证测序准确性平均 3 倍覆盖度要求的话,也需要至少 51 kb 的读序,也就是需要 37 对引物。

在设计引物步移法的 PCR 和测序引物时,需要相邻引物对之间的重叠区域较长(至少 50 bp 以上),或者设计可以在相邻引物对之间随意组合的引物。另外,为了满足引物的通用性,提高引物的使用效益,一套设计好的 PCR 通用引物最好能够实现对一类动物的全线粒体基因组测序,这样的引物大多数是位于线粒体保守区的简并引物,Simon 等(2006)提供了分布于昆虫线粒体全基因组范围的保守引物的位置和序列,是昆虫线粒体基因组引物设计的基本资料。刘念等(2006)利用 NCBI 核酸数据库中已有的 36 种昆虫的线粒体全基因组设计出 2 对长 PCR 通用引物从 10 种蝗虫总 DNA 中扩增出了线粒体 DNA。研究文献提供的这类引物,只要经过适当的修改就可以应用在其他类群上,避免了优化引物的麻烦。

3　高通量线粒体基因组测序技术

第二代测序(next-generation sequencing,NGS)技术的迅速发展为基因组测序带来了一场巨大的变革,NGS 在测序准确性、费用、耗时等方面表现出极大的优势。目前,广泛使用的 NGS 技术主要有 3 种,分别为 Roche/454 焦磷酸测序、Illumina/Solexa 聚合酶合成测序和 ABI/SOLiD 连接酶测序。与传统的 Sanger 测序相比,这 3 种测序技术单次运行产生的数据量大,因而被称为高通量测序技术。3 种测序平台在应用方面各有其优缺点,Roche/454 焦磷酸测序序列读长最长(200 ~ 300 bp),单次运行产生约 40 万条阅读序列,测序成本高;Illumina/Solexa 聚合酶合成测序和 ABI/SOLiD 连接酶测序序列读长较短(25 ~ 50 bp),单次运行可以产生几百万甚至几千万条读序列,测序成本低;而 ABI/SOLiD 连接酶测序则具有超高通量(Mardis,2008)。在线粒体基因组测序上应用较多的是 Roche/454 焦磷酸测序和 Illumina/Solexa 两种平台。

NGS 能够快速、高效的获得大量序列,使获得完整的线粒体基因组序列变得简便可行,能够在短时间内获得大量的线粒体基因组序列。另外,高通量测序技术对实验材料的消耗很小,同时能有效的使用长度很小的模版(特别适合于古 DNA 和降解 DNA),这对于许多无法获得大量纯化线粒体 DNA 的样品的测序工作,提供了行之有效的解决途径。

已经开发了多种应用 NGS 技术测定线粒体基因组的方法,基于总 DNA 样品的方法有需要预扩增 mtDNA 的多物种平行高通量和无需预扩增 mtDNA 的高通量方法,基于总 RNA 样品的转录组测序方法等。

3.1　基于预扩增 mtDNA 的高通量方法

该法也是首先设计几对长 PCR 引物从总 DNA 中扩增出 mtDNA,每个物种的扩增产物纯化后按相同摩尔比例混合在一起,进行并行标记高通量测序(parallel tagged sequencing,PTS),就是对每个物种的 mtDNA 片段连接一个唯一的序列标签,这样一次运行可以测定数个到数百个线粒体基因组。对测序得到的大量读序,首先根据标签进行分选,把同一物种的线粒体序列读序集中在一起,然后除去低质量序列的读序和标签序列,最后拼接出完整的线粒体基因组或几个叠连群,最后再设计引物通过 PCR 扩增填补缺口。PTS 策略是目前高效的大规模平行测序技术,既可以对鸟枪法产生的 DNA 文库进行高通量测序,还可以对 PCR 产物构建的文库进行快速测定。目前在 454 测序平台上使用 PTS 方法,可以在每次运行中测定 300 个平均长度为 17 kb 的线粒体全基因组序列。

这种方法已经成功应用于许多物种的线粒体基因组测序,如 Lloyd 等(2012)采用长 PCR 和 Roche 454 技术测定了 2 种爪蟾(*Xenopus borealis* 和 *Xenopus victorianus*)的线粒体基因组。

3.2　无需预扩增 mtDNA 的高通量方法

这种策略不需要长 PCR 扩增 mtDNA,在获得高纯度的总 DNA 后即依据不同测序平台技术制备测序文库,进行高通量测序,然后对测序结果进

行拼接,得到的叠连群与参考基因组或 GenBank 数据库比对,得到大部分线粒体基因组序列信息,为了填补拼接后的缺口,在已知的序列信息基础上设计特异性引物进行 PCR 扩增和测序,获得完整的线粒体基因组序列。由于这种方法的测序文库是建立在总 DNA 基础上的,mtDNA 只占整个读序的一小部分,所以分析测序数据时一般需要一个参考线粒体序列(可以是近缘物种的)分选出线粒体序列进行拼接。

已经在 Roche FLX 和 Illumina Genome Analyzer 平台上应用这种策略测定了包括几种古 DNA 在内的线粒体基因组。对椎实螺(*Radix balthica*)总 DNA 进行的高通量测序(Roche/454)获得了平均长度为 318 bp 的 286 643 条高质量读序(reads),完成序列拼接后,在 SwissProt 数据库中进行 BLAST 比对,获得了 2 个独立的线粒体 DNA 片段,分别长 11 060 bp 和 1 937 bp,再设计 3 对引物进行 PCR 扩增以填补缺口,对扩增产物进行测序从而获得了完整的线粒体基因组(Feldmeyer *et al.*, 2010)。在 Roche FLX 测序平台分析了来源于冰川时期已灭绝的披毛犀(*Coelodonta antiquitatis*)标本的毛干总 DNA,从总计获得的 70 082 读序中拼接处完整的线粒体基因组(Willerslev *et al.*, 2009)。应用 Illumina Genome Analyzer 平台已经成功从已灭绝的野牛(*Bos primigenius*)肱骨 DNA 提取物中获得了高质量的线粒体基因组序列(Edwards *et al.*,2010)。

3.3 从转录组数据中获得线粒体基因组序列的方法

转录组测序已经成为快速获得基因组编码序列的成熟方法,而从总 RNA 出发获得的转录组数据含有大量的线粒体转录本,因此,只要从拼接好的 Unigene 或叠连群分选出线粒体基因,就可以获得几乎所有的线粒体编码序列。与无需预扩增 mtDNA 的高通量方法一样,分选线粒体序列也需要一个近缘物种的参考基因组。采用该策略从 Roche 454 测序平台获得的 4 种雀形目鸟类脑组织转录组数据成功提取了线粒体序列并进行了系统发生分析(Nabholz *et al.*,2010)。

4 总结与展望

以上总结了目前为止测定线粒体基因组序列的主要策略和方法。在实际工作中,选择哪种方法取决于很多因素,包括研究规模、样品大小和保存状态、经费情况等。总的来说,目前的高通量技术对于小规模研究费用仍然很高,所以引物步移法还是这类研究的首选方法。由于大多数昆虫身体微小,而测定线粒体基因组原则上只能采用一个个体的总 DNA,所以基于长 PCR 扩增产物的引物步移法尤其适合昆虫线粒体基因组研究。但对于大规模线粒体基因组研究来说,NGS 技术无疑是省时省力的最佳选择。

参考文献(References)

Edwards CJ, Magee DA, Park SD, McGettigan PA, Lohan AJ, Murphy A, Finlay EK, Shapiro B, Chamberlain AT, Richards MB, Bradley DG, Loftus BJ, MacHugh DE, 2010. A complete mitochondrial genome sequence from a mesolithic wild aurochs (*Bos primigenius*). *PLoS ONE*, 5(2):e9255.

Feldmeyer B, Hoffmeier K, Pfenninger M, 2010. The complete mitochondrial genome of *Radix balthica* (Pulmonata, Basommatophora), obtained by low coverage shot gun next generation sequencing. *Mol. Phylogenet. Evol.*, 57(3):1329 – 1333.

Lloyd RE, Foster PG, Guille M, Littlewood DT, 2012. Next generation sequencing and comparative analyses of *Xenopus* mitogenomes. *BMC Genomics*, 13(1):496.

Machida RJ, Miya MU, Yamauchi MM, Nishida M, Nishida S, 2004. Organization of the mitochondrial genome of Antarctic krill *Euphausia superba* (Crustacea: Malacostraca). *Mar. Biotechnol.* (*NY*), 6(3):238 – 250.

Mardis ER, 2008. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387 – 402.

Nabholz B, Jarvis ED, Ellegren H. 2010, Obtaining mtDNA genomes from next-generation transcriptome sequencing: a case study on the basal Passerida (Aves: Passeriformes) phylogeny. *Mol. Phylogenet. Evol.*, 57(1):466 – 470.

Simon C, Buckley TR, Frati F, Stewart JB, Beckenbach AT, 2006. Incorporating molecular evolution into phylogenetic analysis, and a new compilation of conserved polymerase chain reaction primers for animal mitochondrial DNA. *Annu. Rev. Ecol. Evol. Syst.*, 37:545 – 579.

Song N, Liang A, 2009. The complete mitochondrial genome sequence of *Geisha distinctissima* (Hemiptera: Flatidae) and comparison with other hemipteran insects. *Acta Biochim. Biophys Sin.* (Shanghai), 41(3):206 – 216.

Tamura K, Aotsuka T, 1988. Rapid isolation method of animal mitochondrial DNA by the alkaline lysis procedure. *Biochem. Genet.*, 26(11/12):815 – 819.

Tzeng CS, Hui CF, Shen SC, Huang PC, 1992. The complete nucleotide sequence of the *Crossostoma lacustre* mitochondrial genome: conservation and variations among vertebrates. *Nucleic Acids Res.*, 20(18):4853 – 4858.

Yamauchi MM, Miya MU, Nishida M, 2004. Use of a PCR-based approach for sequencing whole mitochondrial genomes of insects: two examples (cockroach and dragonfly) based on the method developed for decapod crustaceans. *Insect Mol. Biol.*, 13(4):435 – 442.

Willerslev E, Gilbert M, Binladen J, Ho SYW, Campos PF, Ratan A, Tomsho LP, da Fonseca RR, SherA, Kuznetsova TV, Nowak-Kemp M, Roth TL, MillerW, Schuster SC, 2009. Analysis of complete mitochondrial genomes from extinct and extant rhinoceroses reveals lack of phylogenetic resolution. *BMC Evol. Biol.*, 9:95.

刘念, 胡婧, 黄原, 2006. 应用长 PCR 扩增蝗虫线粒体全基因组. 动物学杂志, 41(2):61 – 65.

鲁成, 刘运强, 廖顺尧, 李斌, 向仲怀, 韩华, 王学刚, 2002. 家蚕线粒体基因组全序列测定与分析. 农业生物技术学报, 10(2):163 – 170.

叶维萍, 黄原, 2003. 长 PCR 技术及其在动物学研究中的应用. 动物学杂志, 38(3):105 – 109.