

基因组时代线粒体基因组拼装策略及软件应用现状

匡卫民, 于黎

云南大学生命科学学院, 省部共建生物资源保护与利用国家重点实验室, 昆明 650091

摘要: 随着测序技术的不断发展, 越来越多物种的全基因组数据被测定和广泛应用。在二代基因组数据爆发式增长的同时, 除了核基因组数据, 线粒体基因组数据也非常重要。高通量测序的全基因组序列中除了核基因组序列也包括线粒体基因组序列, 如何从海量的全基因组数据中提取和拼装线粒体基因组序列并加以应用成为线粒体基因组在分子生物学、遗传学和医学等方面的研究方向之一。基于此, 从全基因组数据中提取线粒体基因组序列的策略及相关的软件不断发展。根据从全基因组数据中锚定线粒体 reads 的方式和后续拼装策略的不同, 可以分为有参考序列拼装方法和从头拼装方法, 不同拼装策略及软件也表现出各自的优势和局限性。本文总结并比较了当前从全基因组数据中获得线粒体基因组数据的策略和软件应用, 并对使用者在使用不同策略和相关软件方面给予建议, 以期在线粒体基因组在生命科学的相关研究中提供方法上的参考。

关键词: 全基因组; 线粒体基因组; 有参考序列拼装方法; 从头拼装方法; 拼装软件

Mitogenome assembly strategies and software applications in the genome era

Weimin Kuang, Li Yu

State Key Laboratory for Conservation and Utilization of Bio-Resource in Yunnan, School of Life Sciences, Yunnan University, Kunming 650091, China

Abstract: With rapid advances in next-generation sequencing technologies, the genomes of many organisms have been sequenced and widely applied in different settings. Mitochondrial genome data is equally important and the high-throughput whole-genome data typically contain mitochondrial genome (mitogenome) sequences. How to extract and assemble the mitogenome from massive whole-genome sequencing (WGS) data remain a hot area in molecular biology, genetics and medicine. The cataloging and analysis of accumulating mitogenome data promotes the development of assembly strategies and corresponding software applications related to mitochondrial DNA from the WGS data. Mitogenome assembly strategies can be divided into mitogenome-reference strategy and *de novo* strategy. Each strategy has different advantages

收稿日期: 2019-08-07; 修回日期: 2019-09-25

基金项目: 国家自然科学基金项目(编号: 31872213), 云南省教育厅科学研究基金产业化培育项目(编号: 2016CYH02)和云南省研究生学术新
人奖资助项目[Supported by the National Natural Science Foundation of China (No.31872213), Industrialization Cultivation Project of
Scientific Research Fund of Yunnan Education Department (No. 2016CYH02) and the Academic Graduate Students Foundation of Yunnan
Province]

作者简介: 匡卫民, 博士, 专业方向: 遗传学。E-mail: kuangwm0714@sina.com

通讯作者: 于黎, 博士, 研究员, 研究方向: 动物遗传与进化。E-mail: yuli@ynu.edu.cn

DOI: 10.16288/j.yczs.19-227

网络出版时间: 2019/10/29 16:37:23

URI: <http://kns.cnki.net/kcms/detail/11.1913.R.20191029.1041.001.html>

and limitations with respect to the difference of bait mitogenome-linked short reads from the WGS data and corresponding assembly strategy. In this review, we summarize and compare current mitogenome assembly strategies and the software applications available. We also provide suggestions related to use different assembly strategies and software applications, and the expected benefits and limitations of methods references in life science.

Keywords: whole-genome sequencing; mitogenome; mitogenome-reference assembly; *de novo* assembly; assembly software

线粒体基因组(mitochondrial genome)作为一种特殊且容易获取的遗传标记,因具有高突变速率、无基因重组、高拷贝数和母系遗传等特点^[1],被广泛应用在系统发育和生物地理研究^[2~5]、群体遗传^[6~13]、医学^[14~17]和生态学研究^[18~20]等领域。在早期的研究阶段,线粒体基因组序列的获取是首先通过长链链式反应(long range PCR, LR-PCR)和克隆 PCR 扩增,然后再通过引物步移(primer walking)桑格(Sanger)测序。这种方法准确性高,但通量低、耗时耗力和花费高。随着测序技术的发展,特别是新一代测序技术(next-generation sequencing, NGS)的发展及测序成本的快速下降,使得线粒体基因组序列的获取变得更为容易。目前,NGS 及其衍生技术(如 LR-PCR 加 NGS、RNA 测序加缺口填补(gap filling)和直接鸟枪法测序^[21~23]等)使得高通量测序成为普遍现象。相比传统的 Sanger 测序技术,NGS 技术通量高、可以更快速且用更低的花费获得全基因组序列(whole-genome sequencing, WGS)、外显子序列和基因转录本^[24]。新一代测序技术的基本原理是:测序平台对样本总 DNA 或分离纯化后的线粒体 DNA 随机打断成 50~700 bp 的单链 DNA 文库(DNA 长短取决于文库构建平台),并将短片段的两端与测序接头序列连接起来,然后对产生的几百万条的 DNA 分子进行测序,高效、准确、快速地获得大量 DNA 序列,最后通过生物信息分析从海量的全基因组数据中获取线粒体基因组。近年来,以 Pacific Biosciences (PacBio) 和 Oxford Nanopore 单分子测序技术为代表的第三代测序技术飞速发展,其测序过程无需进行 DNA 随机打碎和 PCR 扩增,并且读长增加到几十 kb,甚至到 100 kb,拼装后得到更高质量的全基因组序列。基因组技术的发展也促使线粒体序列数据爆发式地增加。因此,越来越多的研究者尝试采用多个不同

的策略从 WGS 数据中获取线粒体基因组^[23,25~39]。

在 NGS 时代如何高效分离和富集线粒体 DNA 而避免核 DNA 的污染是线粒体基因组测序及后续分析的关键,目前主要包括两种分离策略:(1)在 NGS 测序前,从总 DNA 中物理分离纯化线粒体 DNA。这种策略先通过氯化铯密度梯度离心/差速离心或者试剂盒富集磁珠将核 DNA 和线粒体 DNA 分离^[40,41],然后将分离纯化后的线粒体 DNA 进行文库构建和高通量测序。这样,通过在 NGS 测序前就将核 DNA 和线粒体 DNA (或叶绿体 DNA)分离,以保证获得的数据是来自于线粒体(或叶绿体)。该方法的优势在于避免了核 DNA 的污染,即线粒体序列转移到核基因的序列(nuclear mitochondrial pseudogenes, Numts^[42])。但是,物理分离纯化的方法所用的试剂盒价格昂贵、操作比较繁琐和耗时耗力、对样品的质量和数量也都有一定的要求,因此目前仍然存在许多挑战^[43,44],特别是在珍稀野生保护动物和古 DNA (ancient DNA, aDNA)的研究领域则更为困难。(2)先进行 PCR 扩增,对扩增产物进行 NGS 测序。该策略是先用引物扩增出线粒体基因组目的片段,再将扩增产物直接上机进行 NGS 测序,无需构建 DNA 文库^[45]。该方法的优势在于需要的起始 DNA 样本量少,特别适合小型昆虫和环境 DNA 研究领域,关键在于模板 DNA 的质量和 PCR 引物的特异性。

NGS 数据被广泛应用在生命科学的很多领域,尤其是在进化生物学、群体遗传学等揭示物种的起源和扩散历史方面发挥了重要的作用。研究者们常常发现核基因数据和线粒体数据表现出不一致的谱系关系,特别是具有复杂的群体历史的类群(比如基因交流、遗传漂变、偏向性迁徙和祖先谱系分选等)。可见,在分析 NGS 数据时,除了核基因组数据外,线粒体基因组数据也非常重要。然而,目前通过 NGS

方法获得的全基因组数据中即包括了线粒体基因组数据和核基因组数据。在全基因组数据中, 虽然与核基因 reads 的测序深度相比, 线粒体 reads 的测序深度是核基因的 100~1000 倍(细胞中存在几十到数百个拷贝)^[46], 但是线粒体基因组总的 reads 数量只占总 WGS 的 reads 很少一部分, 而且常常受到核基因和叶绿体(绿色植物) reads 的污染。因此, 使用高效的生物信息工具和分析策略从海量的全基因组数据中快速准确地获得线粒体基因组 reads 并完整准确地进行后续线粒体基因组拼装就显得非常重要^[36]。本文将总结当前常用的从 WGS 数据中获取线粒体基因组序列的拼装策略及相关软件, 并对使用者在使用不同策略和相关软件方面给予建议。

1 有参考序列拼装策略及软件应用

有参考序列拼装策略需要选择近缘物种的线粒体基因组或部分片段作为参考序列从研究类群的全基因组数据中捕获线粒体 reads。根据从 WGS 数据中捕获线粒体 reads 是否需要完整的线粒体基因组作为参考序列, 目前常用的策略可以分为: (1) 基于线粒体整个基因组的拼装策略; (2) 基于线粒体片段的拼装策略^[47,48](图 1)。在数据分析流程上, 首先使用全基因组比对工具(如 BWA^[49])将总 reads 映射

(mapping)到线粒体参考序列上, 根据序列的相似性捕获线粒体 reads, 然后再使用不同的序列延长策略对捕获到的线粒体 reads 进行序列延伸, 直到延长到完整的线粒体基因组长度。

1.1 基于线粒体基因组拼装策略及软件应用

基于线粒体基因组作为参考序列获取物种或群体的线粒体基因组序列的方法被广泛应用在系统发育和群体遗传学研究。如 Ko 等^[50]将现存大熊猫的线粒体基因组作为参考序列, 获取到一个 2.2 万年前大熊猫的线粒体基因组。其原理是根据同源比对的研究方法, 将 WGS 数据映射到近缘物种的线粒体基因组上, 再根据线粒体 reads 间相互重叠情况, 从而完成序列的延长(图 1)。这种方法较容易获取和参考基因组一致的序列(consensus sequence), 并且准确性高, 运算速度较快且不耗计算资源。

随着测序技术的发展, 对数据分析能力的需求也在增加, 特别是人类线粒体基因组研究领域, 包括人类进化历史、人类线粒体疾病等方面的研究^[51,52], 推动了人类线粒体基因组的拼装和注释相关软件的发展(表 1)。MIA 是较早用于人类线粒体基因组拼装的软件, 研究者对尼安德特古人类骨头提到的 DNA 进行高通量测序后, 用现代人的线粒体基因组作为参考序列, 使用该软件获取到尼安德特古人类的线

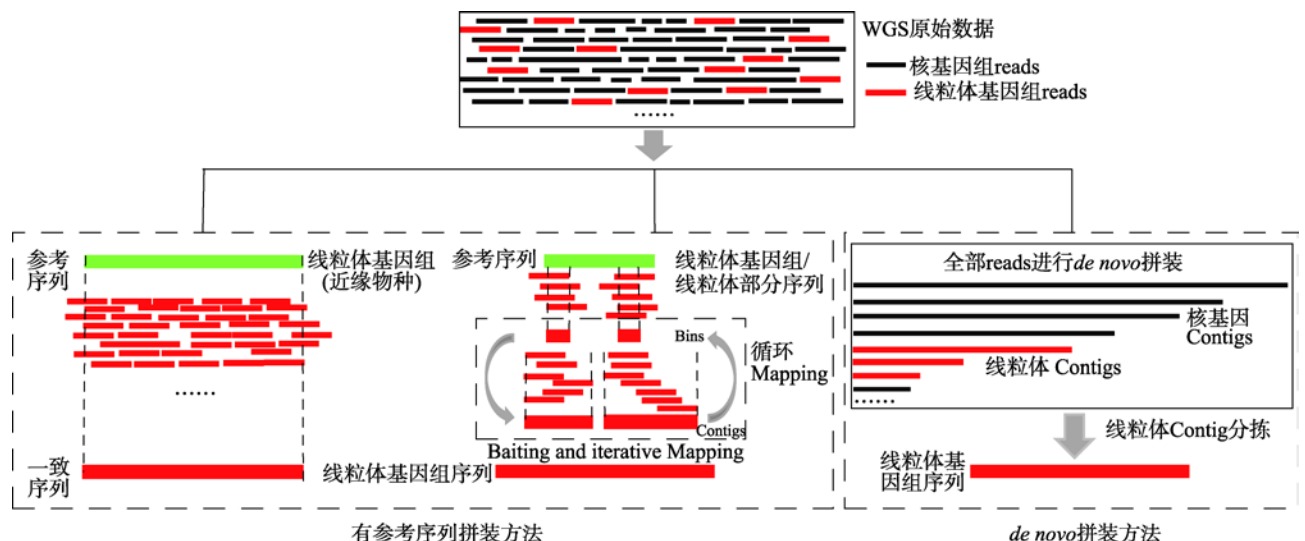


图 1 从全基因组测序数据中获得及拼装线粒体基因组策略

Fig. 1 Strategies of mitogenome assembly from whole-genome sequencing data

分析流程图根据参考文献[36,47,66]修改绘制。实线框代表全基因组短 reads 序列; 虚线框代表获取线粒体基因组序列的方法。

表 1 线粒体基因组拼装软件信息

Table 1 Mitogenome assembly software

软件名称	是否需要参考序列/ 参考序列类型	适用 物种	输入文件格式、 类型	变异 注释	结构可 视化	运行 环境	编程 语言	软件网址
MIA	是/自定义参考序列	任意 物种	Fastq、SE reads 和 PE reads	×	×	CUI	C/C++	https://github.com/mpieva/mapping-iterative-assembler
MitoBam- Annotator	是/rCRS	人	Bam	√	√	Web	Java	http://bioinfo.bgu.ac.il/bsu/software/MITO-BAM
MitoSeek	是/rCRS 和 hg19	人	Bam	√	×	GUI	Perl	https://github.com/riverlee/MitoSeek
mtDNA- profiler	是/rCRS	人	Fasta	×	√	Web	Java	http://mtprofiler.yonsei.ac.kr
MITObim	是/自定义参考序列	任意 物种	Bam	×	×	CUI	Perl	https://github.com/chrishah/MITObim
Mit-o-matic	是/rCRS	人	Fastq、SE reads 和 PE reads	√	√	Web/GUI	Java	http://genome.igib.res.in/mitomatic
MToolBox	是/rCRS 和 RSRS	人	Fastq/Bam/Sam、 SE reads 和 PE reads	√	×	Web/CUI	Python	https://sourceforge.net/projects/mtoolbox
ARC	是/自定义参考序列	任意 物种	Fastq、SE reads 和 PE reads	×	×	Web/CUI	Python	https://github.com/ibest/ARC
Phy-Mer	是/自定义参考序列	任意 物种	Fasta/fastq/Bam、 SE reads 和 PE reads	×	√	CUI	Python	https://github.com/danielnavarrogonzalez/phy-mer
mtDNA- Server	是/rCRS 和 RSRS	人	Fastq/Bam/VCF、 SE reads 和 PE reads	√	√	Web	Java	https://mtdna-server.uibk.ac.at
IOGA	是/自定义参考序列	任意 物种	Fastq、SE reads 和 PE reads	×	×	CUI	Python	https://github.com/holmrensen/IOGA
NOVOPlasty	是/自定义参考序列	任意 物种	Fastq/fasta、SE reads 和 PE reads	×	×	Web/CUI	Perl	https://github.com/ndierckx/NOVOPlasty
Norgal	否	任意 物种	Fastq、SE reads 和 PE reads	×	×	CUI	Python/ Java	https://bitbucket.org/kosaidtu/norgal
Organelle- PBA	是/自定义参考序列	任意 物种	PacBio reads	×	×	CUI	Perl	https://github.com/laubombarily/Organelle_PBA
MitoSuite	是/rCRS, RSRS, hg19, GRCh37 和 38	人	Bam/Sam	√	√	GUI	Python	https://mitosuite.com
ORG.Asm	是/自定义参考序列	任意 物种	Fastq、SE reads 和 PE reads	×	×	CUI	Python	https://git.metabarcoding.org/org-asm/org-asm
MitoZ	否	任意 物种	Fastq、SE reads 和 PE reads	√	√	CUI	Python	https://github.com/linzhi2013/MitoZ
GetOrganelle	是/自定义参考序列	任意 物种	Fastq、SE reads 和 PE reads	×	×	CUI	Python	https://github.com/Kinggerm/GetOrganelle
Trimitomics	是/自定义参考序列	任意 物种	RNA-seq reads、 PE reads	×	×	Unknown	Unknown	Unknown

按拼装软件发表时间先后顺序排列。“√”表示可以实现的功能；“×”表示不可以实现的功能；GUI：图形用户界面；CUI：命令行运行界面；Web (web server)：网络图形用户界面。

粒体基因组^[53]。随着人类线粒体基因组数据的不断累积和研究领域的不断扩大，对数据分析能力和软

件的功能提出了新要求。一些网络或 windows 图形用户界面的软件被广泛使用，包括 MitoBamAnno-

tator^[54]、MitoSeek^[55]、mtDNA-profiler^[56]、mit-o-matic^[57]、MToolBox^[58]、Phy-Mer^[59]、mtDNA-Server^[60]和 MitoSuite^[61]等。这类软件支持多种输入文件格式,除了 mtDNA-profiler 和 mit-o-matic 外,其他软件都支持二进制的 Bam 格式文件。因此,这些软件可以直接读取不同软件的输出数据,加快了整个分析流程。值得注意的是,各种软件供用户选择的参考基因组数量有差异,如 MitoBamAnnotator、mtDNA-profiler 和 mit-o-matic 仅提供了 1 套人类基因组(rCRS), MitoSeek (rCRS, hg19)、mtDNA-Server (rCRS, RSRS) 和 MToolBox (rCRS, RSRS) 提供了 2 套基因组数据,而 MitoSuite 提供了 5 套人类参考基因组(rCRS、RSRS、hg19、GRCh37 和 38)。使用 Phy-Mer 软件,用户可以自定义参考基因组序列。此外,通过 MitoBamAnnotator、MitoSeek、MToolBox、mtDNA-Server、mit-o-matic 和 MitoSuite 软件,用户可以设置相应参数(比如最小等位基因频率,MAF)来检测线粒体基因组的变异位点和异质性位点(heteroplasmic sites,即线粒体基因组序列上同一个位置存在两种及两种以上的碱基类型,来源可能是外源污染,包括测序错误、特异性扩增,reads 匹配错误等,也可能是内源线粒体异质体)。MitoBamAnnotator 主要评估和预测线粒体异质性位点潜在的功能,但使用功能比较单一。MitoSeek 和 MToolBox 扩展了分析功能,包括线粒体拷贝数目、比对质量、结构变异检测等功能。MitoSeek 还可以借助 Circos^[62]软件对检测出的变异进行可视化,包括基因结构变异(structural variations, SVs)和单核苷酸变异(single nucleotide polymorphism, SNPs)。MToolBox 优势在于可以单次分析多个个体,并且将变异信息记录到 VCF 文件中,更容易被解析和注释。从用户操作运行方面比较, MitoSeek 和 MToolBox 是一款基于 Perl 编程语言的 Linux 运算环境,并且需要加载多个独立的 Perl 模块和比对软件(BWA)以及变异检测软件(GATK^[63]),对于非生物信息研究背景的用户安装和使用这类软件相对较困难。mtDNA-Server 和 mit-o-matic 软件是网络用户图形分析工具,用户不需要复杂的安装过程,仅通过注册的邮箱后上传数据并进行分析,操作和数据分析相对简单,缺点是受输入文件大小的限制,特别是高测序深度的个体上传数据较缓慢。

近期开发的 MitoSuite 软件扩展了更多实用功能,功能更强大,包括人类线粒体基因组的拼装、变异检测、疾病变异注释和功能预测、拷贝数目、质量检测和覆盖度的可视化等。MitoSuite 相比于其他早期的软件,不需要安装其他复杂的计算模块,是图形化操作系统且能本地运行的一款容易操作的软件,可以直接从 Bam 文件中自动建立一致性序列后进行系统发育或群体遗传学的研究^[61],所以对于人类线粒体基因组的研究领域,选择 MitoSuite 更具有优势。

综上所述,使用上述方法及相关软件从全基因组数据中获取线粒体基因组序列,首先借助全基因组比对软件,包括常用的 BWA 和 Bowtie/Bowtie2^[64],将从总 reads 中捕获到线粒体基因组 reads。这两种比对软件优势在于可以对 reads 错配或 reads 多处匹配进行筛选和过滤,通过后续的质控获取到纯净的线粒体 reads。但是,无法区分 Numts 和线粒体拷贝数,从而影响线粒体异质性的检测。另外,这些方法及相关软件需要选择近缘物种的线粒体基因组参考序列,如果选择进化关系较远的物种的线粒体基因组作为参考序列,在全基因组比对的过程中可能会发生 reads 错配或者因序列分歧大导致部分区域比对不上而出现缺失数据(gap),从而影响到后续线粒体基因组拼装的准确性和完整性^[38]。因此,选择合适物种的线粒体基因组作为参考序列是该方法和软件应用的关键。对于要研究的物种无法确定其近缘物种,或者是确定了其近缘物种但没有已有线粒体基因组数据的情况下,这个方法就有很大的局限性^[36,39]。

1.2 基于线粒体片段拼装策略及软件应用

上述借助近缘物种的线粒体全基因组作为参考序列的拼装策略及相关的软件多数适用于人的线粒体基因组拼装、变异检测和变异注释等。随着越来越多其他物种的研究,线粒体基因组分析也被广泛应用在非模式物种的研究中^[65]。仅用人的基因组作为参考序列的软件来获取和分析其他物种的线粒体基因组序列就表现出很大的局限性,因此迫切需要开发适用范围更广的线粒体基因组拼装软件。与总 reads 直接映射到线粒体基因组参考序列的拼装策略类似,但可以选择遗传关系较远或较近物种的线

粒体基因组,甚至线粒体部分序列,来进行其它物种的线粒体基因组序列获取和拼装。该方法首先借助全基因组比对软件将过滤后的 WGS 数据映射到参考序列上,高覆盖度且连续的线粒体 reads 组成序列块(bins),这些单独的 bins 或者根据 bins 重叠情况连接成 Contigs 替换原先的参考序列,并作为下次映射的靶序列(baiting sequencing),依次反复将 WGS 数据映射到新生成的靶序列上延长序列,最后延长到完整的线粒体基因组长度(图 1)。反复映射和替换靶序列可以避免参考序列和拼装方法的偏好性。拼装过程中需要调整 Kmer 值(拼装过程中 reads 打断成长度为 K 的一段固定核苷酸序列)大小,反复将 WGS 数据映射到靶序列上进行序列延长,因此需要消耗大量的计算资源,原始数据越大越消耗计算资源。如果选择遗传关系越远的物种或选择的靶序列越短,拼装时的序列延长则需要更多的循环次数,计算时间也就越长。

Hahn 等^[66]开发的 MITObim 软件可以直接从 WGS 数据中拼装非模式物种的线粒体基因组,这个软件嵌入了 MIRA 和 IMAGE 计算模块。相比 MIA, MITObim 的准确性可以达到 99.5% 以上,在重复区域可以有效的填补 gap,计算速度和内存消耗也占有优势,成为目前最广泛使用的线粒体基因组拼装软件。该软件不支持双端序列(paired-end reads, PE reads),支持 IonTorrent、454 和 PacBio 测序平台数据,而且建议原始数据 reads 数量不要超过 20~40 百万条。如果超出,建议从原始 reads 中随机抽取部分 reads,这样就减少 reads 的数量,不过这样可能会影响拼装结果的准确性和完整性。当然,MITObim 也无法解决线粒体基因组拼装中一些尤为复杂的问题,如 Numts、复杂的无脊椎动物和植物的线粒体拼装等^[67]。ARC^[47]软件的拼装过程类似于 MITObim 软件,两者都可以选择亲缘关系较远的物种的线粒体基因组或者线粒体部分序列就可以得到完整的线粒体基因组序列,主要的差异在于序列延长方式。ARC 是直接对 bins 进行拼装完成序列的延长,而 MITObim 则是反复将总 reads 往靶序列上映射完成延长序列。相比其他全基因组拼装软件,ARC 不是将总 reads 进行从头拼装,而是先通过映射的方式对 reads 重叠的 bins 进行拼装,优势在于不耗内存,运

行速度较快。此外,ARC 基本上不受降解严重的 DNA 质量和低质量的 reads 的影响,特别是 aDNA,而且运算速度比 MITObim 和传统的拼装方法快^[47]。Li 等^[68]使用 ARC 软件对 19 个隐杆线虫(*Caenorhabditis*)物种进行线粒体基因组拼装,测试了不同测序平台(Roche、454、Illumina 和 Ion Torrent)对线粒体基因组拼装的影响,结果发现 ARC 软件对 454 平台的数据进行分析时会崩溃,可能的原因是序列长度范围大导致数据分析需要较大的计算资源。但是 ARC 拼装的完整性都要比 MITObim 好。然而,Dierckxsens 等^[47]用 ARC 软件对角胫叶甲属(*Gonioctena Intermedia*)进行线粒体基因组拼装,结果发现尽管 ARC 准确性高(99.99%),但不能将线粒体拼装到一条 Contig 上,完整性较差(覆盖到线粒体基因组的 85.39%)。

Dierckxsens 等^[38]开发了 NOVOPlasty 软件,类似于 SSAKE^[69]和 VCAKE^[70]算法,将排序后的 reads 存放在哈希表中,以便 reads 的快速读取,因此运算速度较快。NOVOPlasty 软件需要提供一条靶序列,可以是一条短 read、一段编码基因序列,甚至是完整的线粒体基因组序列。值得注意的是,NOVOPlasty 与 ARC 拼装策略不同的是,NOVOPlasty 借助提供的靶序列从 WGS 数据中获取线粒体基因组的一条 read,然后再对捕获到的 read 进行双向延伸。作者将 NOVOPlasty 与当前主流的拼装软件相比较,包括 MITObim、MIRA、ARC、SOAPdenovo2 和 CLCbio,结果发现:除了 ARC 外,其余软件都将线粒体拼装在一 Contig。通过对 NOVOPlasty 拼装到的序列进行质量评估,没有发现缺失位点和不确定的碱基位点,表明准确性和完整性高。NOVOPlasty 的计算速度最快、基因组覆盖度最高,CLCbio 准确性同样也达到了 100%,但是基因组的覆盖度不高(89.96%)。MIRA 和 ARC 都体现最高的基因组覆盖度,但是准确性最低。增加测序覆盖度和 reads 的长度可以提高 NOVOPlasty 的完整性和准确性,特别是高重复和 AT 含量高的区域。NOVOPlasty 运行不需要载入其他软件和模块,对于用户来说安装和操作比较简单^[38]。

目前用于叶绿体基因组拼装软件同样适合线粒体基因组的拼装,包括 IOGA^[71]、GetOrganelle^[72]和 ORG.Asm^[73]等。IOGA 和 GetOrganelle 类似于

MITObim 中的“Baiting and iterative 映射”分析流程。IOGA 分析过程需要 Bowtie2、SOAPdenovo2、SPAdes 3.0^[37]和其他程序来捕获线粒体 reads, 拼装过程还需要调整拼装参数 Kmer 大小(范围为 37~97), 最后通过拼装似然评估(assembly likelihood estimation, ALE)从候选的 Contigs 序列里确定线粒体基因组^[74]。这种方法适合降解程度较大的样品的线粒体基因组或叶绿体基因组拼装, 比如博物馆样品等。与其他拼装软件比较, IOGA 使用 ALE 检验来筛选拼装好的 Contigs, 最后通过最大似然值来判断最优的拼装序列。GetOrganelle 和 IOGA 数据分析流程非常相似。GetOrganelle 嵌入了独立的 Bowtie2、BLAST^[75]和 SPAdes 3.0 分析模块, 双端 reads 和单端 reads (single-end reads, SE reads)均可以作为 GetOrganelle 的输入文件。GetOrganelle 可以直接在 SPAdes 拼装的过程中进行 reads 错误校正和错配过滤, 保留高质量的 reads 作为后续分析, 而 IOGA 和 MITObim 则需要用其他过滤软件提前进行低质量 reads 的过滤。IOGA 和 GetOrganelle 拼装软件均嵌入 SPAdes 程序计算模块, 在拼装过程中需要反复调试 Kmer 值的大小。选择合适的 Kmer 不仅能够保证线粒体 Scaffolds 或 Contigs 的完整性和准确性, 还可以减少计算时间和运行内存^[72]。

最近, 随着单分子测序 PacBio 和 Nanopore 长片段测序技术的发展, 一些复杂物种的全基因组序列被测序和应用, 特别是多倍体物种和高重复的物种, 显示了长片段测序技术的优势^[27,76~80]。同时, 已经开发出了一些适用于拼装 PacBio 和 Nanopore 长 reads 的软件, 比如 HGAP^[81]、Falcon (<https://github.com/PacificBiosciences/falcon>)、Canu^[82]和 Sprai^[83]等, 而从这些平台测序得到的长 reads 进行线粒体和叶绿体基因组拼装的方法和算法还很缺乏。目前已经有一些研究者直接使用 PacBio 和 Nanopore 平台进行线粒体基因组测序并进行拼装^[25~29]。Soorni 等^[84]基于 Perl 编程语言开发的 Organelle-PBA 直接对 PacBio 平台测序到的全基因组长片段进行线粒体或叶绿体基因组的拼装。Organelle-PBA 安装和使用需要安装多种 Perl 模块和多种软件, 包括 BlasR^[85]、Samtools^[86]、Blast^[87]、SSPACE-LongRead^[88]、Sprai 和 BEDTools^[89]等。虽然 PacBio 和 Nanopore 测序平

台可以得到更长的 reads, 但是仍然存在一定的碱基错误率, 因此需要使用碱基校正软件进行碱基校正, 比如 Sprai。因 PacBio 和 Nanopore 测序平台不需要在建库的过程中进行 DNA 随机打断和扩增并且具有读长长特点, 所以可以完整得将线粒体基因组一次性测通, 有效避免了 Numts 的污染。但同时因为 PacBio 和 Nanopore 测序平台对样品 DNA 质量有极其严格的要求, 要保证 DNA 的完整性, 所以 Organelle-PBA 的使用也有局限性。

2 从头(*de novo*)拼装策略及软件应用

目前, 世界上越来越多的物种的全基因组数据和线粒体基因组数据被公布, 但也有绝大多数物种的基因组信息还未被测定, 针对没有参考基因组序列的物种, 从头拼装是一种快速和准确地获取遗传信息的策略, 这种方法被广泛应用在 DNA 和 RNA 序列拼装。线粒体基因组的从头拼装与核基因组的拼装过程相似, 首先从海量的全基因组数据中找到短 reads 的一致性序列, 然后再根据不同长度的大片段文库进行 Contigs 的排序和连接, 最后延长到 Scaffolds 水平。根据线粒体 reads 的来源不同, 可以分为从全基因组数据中从头拼装线粒体基因组策略和从转录组数据中从头拼装线粒体基因组策略(图 1)。

2.1 从全基因组数据中从头拼装线粒体基因组策略及软件应用

从头拼装线粒体基因组方法不需要提供完整的线粒体基因组或线粒体部分序列作为参考序列。从头拼装首先将 WGS 的全部 reads 进行从头拼装^[47,48], 即将核基因和线粒体基因 reads 都分别拼装为长片段序列, 然后依据线粒体基因组序列长度和高测序深度进行严格的 Contigs 过滤得到候选线粒体 Contigs, 最后反复将 WGS 数据映射到候选线粒体 Contigs 上, 不断延长 Contigs, 直到延长到完整线粒体基因组长度(图 1)。现有的软件有 Norgal^[36]和 MitoZ^[39]等。对于一些没有近缘物种线粒体基因组的物种, 或者 DNA 降解严重的样品(比如 aDNA 序列), 用有参考

序列拼装方法就有很大的局限性。所以,对 aDNA 或者环境 DNA 首先进行 NGS 测序,再进行线粒体基因组从头拼装即是一个行之有效的策略。但是,这种方法常常要借助于全基因组或转录组拼装的软件和计算模块(包括 SOAPdenovo2^[90]、SPAdes^[37]、Velvet^[91]、BIBRat^[92]、CLCbio (<https://www.qiagenbioinformatics.com/products/clc-assembly-cell>)、SOAPdenovo-Trans^[93]和 Trinity^[94]等)对整个基因组数据进行拼装,而且需要反复调整 Kmer 值的范围以达到最佳的拼装质量,所以耗费计算资源,计算速度较慢。

传统的从头拼装软件,包括 SOAPdenovo2、Newbler、SPAdes、Velvet、CLCbio、ALLPATHS^[95]和 Platanus^[96]等,在全基因组序列拼装过程中,其线粒体 Scaffolds 或 Contigs 常常被过滤掉。从头拼装线粒体基因组则借助传统的从头拼装软件,在分析过程中考虑线粒体 reads 的高测序深度,而不是将其删除。目前已经有许多动植物的线粒体基因组用从头的拼装方法获得了完整的线粒体基因组序列。Lee 等^[97]对桔梗科的桔梗(*Platycodon grandiflorus*)和党参(*Codonopsis lanceolata*)进行了低覆盖度基因组测序并对线粒体基因组进行拼装。他们首先使用 Celera、SOAPdenovo、SPAdes 和 CLCbio 等 4 种全基因组拼装软件对全部 reads 进行从头拼装,得到由核基因和线粒体组成的 Contigs 库,其次根据线粒体的 Contigs 和核基因组的 Contigs 平均测序深度的差异确定候选线粒体 Contigs,再将 WGS 数据比对到候选线粒体 Contigs 上,如此循环完成 Contig 的延长,最后得到完整的线粒体基因组^[97]。类似于这种拼装策略,Al-Nakeeb 等^[36]开发的 Norgal 软件,先使用 MEGAHIT^[98]拼装软件对 NGS 数据进行从头拼装,然后再将 NGS 数据重新映射到拼装好的 Contig 上,通过线粒体和核基因组的 reads 覆盖度来判断线粒体 Contig(s)。他们通过与其他不同策略的线粒体基因组拼装软件比较发现,Norgal 软件的准确性和 NOVOPlasty 软件相似,但是从运算速度上来比较,NOVOPlasty 远比 Norgal 和 MITObim 要快,原因是 Norgal 需要调整不同 Kmer 大小对整个基因组进行拼装,然后再比对 reads 和计算核基因组 reads 的测序深度来判断拼装的可靠性^[36]。

随着用户对数据分析的需求越来越大,要求简

化及高效率的数据分析流程、功能全面和良好的用户体验的软件越来越成为迫切的需要。Meng 等^[39]开发的 MitoZ 软件可以“一键式”地对线粒体基因组进行拼装、注释和可视化。该软件包括了多种计算模块,包括原始数据的预处理、从头拼装、候选线粒体序列的富集和线粒体基因组的注释和可视化等功能。相比于其他软件,该软件能对低质量的 reads、碱基大量缺失的 reads 和建库中 PCR 冗余的 reads 进行过滤,以保证后续分析数据的可靠性。MitoZ 整合了 SOAPdenovo-Trans 的算法,从核基因组中的 reads 进行线粒体基因组的从头拼装,其原理是:根据线粒体基因组 reads 的平均测序深度远比核基因组的高,设置不同的 Kmer 参数来达到最佳的拼装效果。这个软件提供了两种拼装方式:快捷模式(quick model)和多 Kmer 模式。根据作者的建议尽可能使用多 Kmer 模式调整不同 Kmer 参数,以保证复杂线粒体基因组拼装的完整性和准确性。从拼装的基因数量和序列的总长度方面进行比较,MitoZ 比有参考序列的拼装策略更具有优势,特别是对于物种间相似度很低的基因。此外,除了各类软件算法的差异,重复序列、AT 含量和异质性率(异质性位点占总变异位点的数量)等也是影响线粒体基因组的拼装完整性和准确性的关键因素^[39]。MitoZ 对线粒体基因组的注释(Blast、Genewise、MiTFi 和 Infernal)以及可视化(Circos)功能集成了其他成熟的软件模块,因此间接地扩展了拼装软件的功能,也极大地简化了数据的分析过程。

2.2 从转录组数据中从头拼装线粒体基因组策略及软件应用

新一代测序技术的发展同时推动了转录组水平的研究,从转录组数据中获得基因组编码序列已经很成熟,而总的 RNA 转录本中包含大量的线粒体编码基因转录本,于是研究者开发了可以高效地从转录组数据中富集线粒体编码基因序列的一些软件。这些方法的原理是根据线粒体在细胞内多拷贝数的特征,线粒体编码基因 mRNA 的 reads 测序深度远比核基因组的编码基因 reads 高,具有高水平的基因表达量。Plese 等^[99]开发了 Trimitomics 软件能快速有效地从转录本 reads 里面对线粒体编码基因序列

进行拼装。该软件的分析流程包括了 NOVOPlasty、Bowtie2/Trinity 和 Velvet 等 3 个独立拼装过程: (1) 首先使用 NOVOPlasty 软件将全部的 RNA reads 进行从头拼装, 根据 Kmer 大小范围(25、39、45 和 51)确定线粒体编码序列的完整性; (2) 如果没有拼装到完整的线粒体编码序列或者拼装到部分序列, 则先使用 Trimmomatic 0.33^[100]对原始 RNA reads 进行过滤, 再用 Bowtie2^[64]软件将过滤后的 reads 比对到近缘物种的线粒体基因组上, 用 Trinity^[94,101]对 mapped-read 进行从头拼装; (3) 使用 Velvet 软件对全部的转录本进行从头拼装, 接着用 BlastN 软件^[102]确定得到的线粒体 Contigs。如果以上 3 种方法都没有拼装到完整的线粒体编码序列, 那么再使用 Geneious 软件整合以上 3 种方法拼装的结果, 再将整合的结果在 NCBI 数据库中进行同源性鉴定。作者通过对 6 个无脊椎动物进行线粒体编码基因的拼装, 结果发现 3 种拼装过程都能够覆盖到 97% 以上的线粒体编码基因序列。从拼装完整性和准确性来评估 NOVOPlasty、Bowtie2/Trinity 和 Velvet 拼装过程的可靠性, 结果发现 3 种拼装方法因物种差异而差异, 如 *A. valida* 和 *P. dumerilii* 这两种纽形动物, Bowtie2/Trinity 拼装流程得到的线粒体编码序列的质量更好。而从运行时间、运行内存上比较, NOVOPlasty 拼装流程更具有优势。值得注意的是, Trimitomics 软件提供 3 种拼装流程, 通过判断拼装结果的完整性来判断是否进行其他拼装流程。同时对于复杂物种的线粒体基因组, 还可以整合 3 种拼装流程的结果, 增加了可靠性。

3 拼装策略及软件使用建议

当使用者在使用不同的线粒体基因组拼装软件时, 首先要区分选择有参考线粒体序列拼装方法的软件还是从头拼装方法的软件。如果使用者要拼装的物种的遗传信息很清楚, 可以选择有参考拼装方法的软件。如果要拼装的物种缺乏相关的遗传背景, 特别是 aDNA, 建议选择从头拼装的策略。此外, 用户选择不同的软件还需要注意以下几点: (1) 了解各类软件的原理及适用性, 特别是一些软件对基因组上高重复区有偏好性; (2) 适用的物种, 人或者非

模式物种; (3) 不同的软件依赖于不同的数据类型, 首先需要区分数据是核基因组数据还是转录本数据, 长片段还是短片段序列, 单端 reads 还是双端 reads 等; (4) 不同的软件对输入的文件格式有不同的要求; (5) 根据使用者实际需要评估计算资源和操作系统选择不同的软件。影响线粒体基因组拼装的完整性和准确性的因素很多, 包括基因组序列特征(比如重复元件, 异质性)、测序深度和测序技术(reads 长度和碱基错误率)都给序列拼装带来了挑战。此外, 尽管基因组拼装算法和软件在不断地发展和优化, 但在 WGS 数据中很难区分线粒体和核基因相似的 reads, 以及 Numts 污染^[103]等问题, 都会造成不同拼装软件在拼装结果上的冲突和后续研究分析结果的推断^[104]。值得注意的是, 有研究报道发现, 不同的物种采用不同的拼装软件, 拼装到的线粒体基因组的完整性(比如蛋白质编码区、rRNA 和 tRNA 的数量)和准确性均有差异^[105]。如果计算资源允许的情况下, 应当选择多种拼装策略的软件进行线粒体基因组的拼装, 而对于低覆盖区域或不同拼装软件间导致结果不一致的区域或 gap, 还需要 Sanger 测序进行验证^[105]。

本文共列举了 19 个从 WGS 数据中拼装线粒体基因组的软件(表 1), 多数软件的代码和软件包存储在 GitHub, 优势在于它是基于网站和云的服务, 可以开源软件的代码, 以及跟踪和控制对代码的更改。这些软件中有 12 个软件是命令行运行的方式(CUI), 即可在 Linux 操作系统下完成, 用户可以在参数设置文本文件或者命令行参数中设置软件运行参数。命令行运行方式的优点是可以跨平台进行大数据的计算, 比如可以将任务提交到大型计算集群上进行计算, 缺点是使用者必须要熟悉大量的计算机命令, 而不是用鼠标操作就能实现。另外一种运行方式是网络(web server, Web)或 windows 图形用户界面运行(GUI), 用户通过简单的鼠标操作就可以完成参数设置, 非常适合对软件不熟悉或者生物信息研究的初学者。

此外, 本文列举的 19 个软件中, 共有 9 个是用 Python 和 Perl 语言编写的(表 1)。其他软件, 如 MIA 使用的则是 C/C++, 而 Norgal 使用面向对象编程语言 Java 编写。这些编程语言具有可移植性、可扩展性和可嵌入性、具有丰富的库等特点。

4 结语与展望

新一代测序技术的不断发展使得越来越多物种的全基因组数据信息被公开和应用,这些数据包含线粒体 DNA 和核 DNA。此外,即使在基因组时代,对线粒体基因组的研究仍然是不可缺少的,比如对于有复杂社会结构和与性别相关的扩散行为的物种的研究^[13,106]等。这些研究都促进了线粒体基因组数据爆发式增长和拼装策略及相关软件的发展。

线粒体基因组的拼装是非常复杂和快速发展的领域,包括获取线粒体基因组的技术和方法等都需要持续地改进和提高,好的拼装策略依赖于 WGS 数据集、计算能力和可获得的参考基因组。此外,成功获得一个高质量的线粒体基因组取决于许多因素,包括建库测序平台、基因组的结构特点(重复序列含量、GC 含量等)^[107]。数据类型也决定线粒体拼装的质量,如 aDNA。最近测序技术和提取 aDNA 的发展推动了古基因组的研究,并利用生物信息学的手段从 WGS 数据中拼装古线粒体基因组序列。aDNA 因长时间保存在土壤中或在博物馆中而导致 DNA 被降解成小的 DNA 片段,又加上发掘的 aDNA 的近缘物种的不确定性,因此为古线粒体基因组的拼装带来许多挑战。正如 Meng 等^[39]指出,开发一款灵活性和高效率的软件,具有良好的用户体验的软件,使得用户能够把更多的时间和精力集中在生物学问题研究上,而不是如何获取线粒体基因组。

参考文献(References):

- [1] Brown WM, George M Jr., Wilson AC. Rapid evolution of animal mitochondrial DNA. *Proc Natl Acad Sci USA*, 1979, 76(4): 1967–1971. [DOI]
- [2] Lei R, Frasier CL, Hawkins MT, Engberg SE, Bailey CA, Johnson SE, McLain AT, Groves CP, Perry GH, Nash SD, Mittermeier RA, Louis EE. Phylogenomic reconstruction of Sportive Lemurs (genus *Lepilemur*) recovered from mitogenomes with inferences for Madagascar biogeography. *J Hered*, 2017, 108(2): 107–119. [DOI]
- [3] Mueller RL, Macey JR, Jaekel M, Wake DB, Boore JL. Morphological homoplasy, life history evolution, and historical biogeography of plethodontid salamanders inferred from complete mitochondrial genomes. *Proc Natl Acad Sci USA*, 2004, 101(38): 13820–13825. [DOI]
- [4] Zhang P, Chen YQ, Zhou H, Liu YF, Wang XL, Papenfuss TJ, Wake DB, Qu LH. Phylogeny, evolution, and biogeography of Asiatic Salamanders (Hynobiidae). *Proc Natl Acad Sci USA*, 2006, 103(19): 7360–7365. [DOI]
- [5] Zhang P, Papenfuss TJ, Wake MH, Qu LH, Wake DB. Phylogeny and biogeography of the family Salamandridae (Amphibia: Caudata) inferred from complete mitochondrial genomes. *Mol Phylogenet Evol*, 2008, 49(2): 586–597. [DOI]
- [6] Cerný V, Fernandes V, Costa MD, Hájek M, Mulligan CJ, Pereira L. Migration of Chad speaking pastoralists within Africa based on population structure of Chad Basin and phylogeography of mitochondrial L3f haplogroup. *BMC Evol Biol*, 2009, 9: 63. [DOI]
- [7] Klimova A, Phillips CD, Fietz K, Olsen MT, Harwood J, Amos W, Hoffman JJ. Global population structure and demographic history of the grey seal. *Mol Ecol*, 2014, 23(16): 3999–4017. [DOI]
- [8] Lin LH, Ji X, Diong CH, Du Y, Lin CX. Phylogeography and population structure of the Reeves's Butterfly Lizard (*Leiolepis reevesii*) inferred from mitochondrial DNA sequences. *Mol Phylogenet Evol*, 2010, 56(2): 601–607. [DOI]
- [9] Miller W, Hayes VM, Ratan A, Petersen DC, Wittekindt NE, Miller J, Walenz B, Knight J, Qi J, Zhao F, Wang Q, Bedoya-Reina OC, Katiyar N, Tomsho LP, Kasson LM, Hardie RA, Woodbridge P, Tindall EA, Bertelsen MF, Dixon D, Pyecroft S, Helgen KM, Lesk AM, Pringle TH, Patterson N, Zhang Y, Kreiss A, Woods GM, Jones ME, Schuster SC. Genetic diversity and population structure of the endangered marsupial *Sarcophilus harrisii* (Tasmanian devil). *Proc Natl Acad Sci USA*, 2011, 108(30): 12348–12353. [DOI]
- [10] Roslin T. Spatial population structure in a patchily distributed beetle. *Mol Ecol*, 2001, 10(4): 823–837. [DOI]
- [11] Teacher AG, André C, Merilä J, Wheat CW. Whole mitochondrial genome scan for population structure and selection in the Atlantic herring. *BMC Evol Biol*, 2012, 12: 248. [DOI]
- [12] Uren C, Kim M, Martin AR, Bobo D, Gignoux CR, Van Helden PD, Möller M, Hoal EG, Henn BM. Fine-scale human population structure in southern Africa reflects ecogeographic boundaries. *Genetics*, 2016, 204(1): 303–314. [DOI]

- [13] Kuang WM, Ming C, Li HP, Wu H, Frantz L, Roos C, Zhang YP, Zhang CL, Jia T, Yang JY, Yu L. The origin and population history of the endangered golden snub-nosed monkey (*Rhinopithecus roxellana*). *Mol Biol Evol*, 2019, 36(3): 487–499. [DOI]
- [14] Haberman Y, Karns R, Dexheimer PJ, Schirmer M, Somekh J, Jurickova I, Braun T, Novak E, Bauman L, Collins MH, Mo A, Rosen MJ, Bonkowski E, Gotman N, Marquis A, Nistel M, Rufo PA, Baker SS, Sauer CG, Markowitz J, Pfefferkorn MD, Rosh JR, Boyle BM, Mack DR, Baldassano RN, Shah S, Leleiko NS, Heyman MB, Griffiths AM, Patel AS, Noe JD, Aronow BJ, Kugathasan S, Walters TD, Gibson G, Thomas SD, Mollen K, Shen-Orr S, Huttenhower C, Xavier RJ, Hyams JS, Denson LA. Ulcerative colitis mucosal transcriptomes reveal mitochondriopathy and personalized mechanisms underlying disease severity and treatment response. *Nat Commun*, 2019, 10(1): 38. [DOI]
- [15] Inak G, Lorenz C, Lisowski P, Zink A, Mlody B, Prigione A. Concise review: induced pluripotent stem cell-based drug discovery for mitochondrial disease. *Stem Cells*, 2017, 35(7): 1655–1662. [DOI]
- [16] Suomalainen A. Mitochondrial DNA and disease. *Ann Med*, 1997, 29(3): 235–246. [DOI]
- [17] Toda T. Molecular genetics of Parkinson's disease. *Brain Nerve*, 2007, 59(8): 815–823. [DOI]
- [18] Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, Brzezinski MA, Chaal BK, Chiovitti A, Davis AK, Demarest MS, Detter JC, Glavina T, Goodstein D, Hadi MZ, Hellsten U, Hildebrand M, Jenkins BD, Jurka J, Kapitonov VV, Kröger N, Lau WW, Lane TW, Larimer FW, Lippmeier JC, Lucas S, Medina M, Montsant A, Obornik M, Parker MS, Palenik B, Pazour GJ, Richardson PM, Rynearson TA, Saito MA, Schwartz DC, Thamtrakoln K, Valentin K, Vardi A, Wilkerson FP, Rokhsar DS. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science*, 2004, 306(5693): 79–86. [DOI]
- [19] Janzen DH, Burns JM, Cong Q, Hallwachs W, Dapkey T, Manjunath R, Hajibabaei M, Hebert PDN, Grishin NV. Nuclear genomes distinguish cryptic species suggested by their DNA barcodes and ecology. *Proc Natl Acad Sci USA*, 2017, 114(31): 8313–8318. [DOI]
- [20] Zarowiecki MZ, Huyse T, Littlewood DT. Making the most of mitochondrial genomes--markers for phylogeny, molecular ecology and barcodes in *Schistosoma* (Platyhelminthes: Digenea). *Int J Parasitol*, 2007, 37(12): 1401–1418. [DOI]
- [21] Hu M, Jex AR, Campbell BE, Gasser RB. Long PCR amplification of the entire mitochondrial genome from individual helminths for direct sequencing. *Nat Protoc*, 2007, 2(10): 2339–2344. [DOI]
- [22] Nabholz B, Jarvis ED, Ellegren H. Obtaining mtDNA genomes from next-generation transcriptome sequencing: a case study on the basal Passerida (Aves: Passeriformes) phylogeny. *Mol Phylogenet Evol*, 2010, 57(1): 466–470. [DOI]
- [23] Timmermans MJ, Dodsworth S, Culverwell CL, Bocak L, Ahrens D, Littlewood DT, Pons J, Vogler AP. Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Res*, 2010, 38(21): e197. [DOI]
- [24] Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet*, 2010, 11(1): 31–46. [DOI]
- [25] Lounsberry ZT, Brown SK, Collins PW, Henry RW, Newsome SD, Sacks BN. Next-generation sequencing workflow for assembly of nonmodel mitogenomes exemplified with North Pacific albatrosses (*Phoebastria* spp.). *Mol Ecol Resour*, 2015, 15(4): 893–902. [DOI]
- [26] Shearman JR, Sonthirod C, Naktang C, Pootakham W, Yoocha T, Sangsrakru D, Jomchai N, Tragoonrun S, Tangphatsornruang S. The two chromosomes of the mitochondrial genome of a sugarcane cultivar: assembly and recombination analysis using long PacBio reads. *Sci Rep*, 2016, 6: 31533. [DOI]
- [27] Kovar L, Nageswara-Rao M, Ortega-Rodriguez S, Dugas DV, Straub S, Cronn R, Strickler SR, Hughes CE, Hanley KA, Rodriguez DN, Langhorst BW, Dimalanta ET, Bailey CD. PacBio-based mitochondrial genome assembly of *Leucaena trichandra* (Leguminosae) and an intragenomic assessment of mitochondrial RNA editing. *Genome Biol Evol*, 2018, 10(9): 2501–2517. [DOI]
- [28] Wang SB, Song QW, Li SS, Hu ZG, Dong GQ, Song C, Huang HW, Liu YF. Assembly of a complete mitogenome of chrysanthemum *nankingense* using Oxford Nanopore long reads and the diversity and evolution of Asteraceae mitogenomes. *Genes*, 2018, 9(11): 547. [DOI]
- [29] Gan HM, Linton SM, Austin CM. Two reads to rule them all: Nanopore long read-guided assembly of the iconic Christmas Island red crab, *Gecarcoidea natalis* (Pocock, 1888), mitochondrial genome and the challenges of AT-rich mitogenomes. *Mar Genom*, 2019, 45: 64–71. [DOI]

- [30] Maughan PJ, Chaney L, Lightfoot DJ, Cox BJ, Tester M, Jellen EN, Jarvis DE. Mitochondrial and chloroplast genomes provide insights into the evolutionary origins of quinoa (*Chenopodium quinoa* Willd.). *Sci Rep*, 2019, 9(1): 185. [DOI]
- [31] Mofiz E, Seemann T, Bahlo M, Holt D, Currie BJ, Fischer K, Papenfuss AT. Mitochondrial genome sequence of the Scabies Mite provides insight into the genetic diversity of individual scabies infections. *PLoS Negl Trop Dis*, 2016, 10(2): e0004384. [DOI]
- [32] Ni P, Bhuiyan AA, Chen JH, Li J, Zhang C, Zhao S, Du X, Li H, Yu H, Liu X, Li K. De novo assembly of mitochondrial genomes provides insights into genetic diversity and molecular evolution in wild boars and domestic pigs. *Genetica*, 2018, 146(3): 277–285. [DOI]
- [33] Niu WT, Yu SG, Tian P, Xiao JG. Complete mitochondrial genome of *Echinophyllia aspera* (Scleractinia, Lobophylliidae): mitogenome characterization and phylogenetic positioning. *Zookeys*, 2018, 793: 1–14. [DOI]
- [34] Sahoo PK, Singh L, Sharma L, Kumar R, Singh VK, Ali S, Singh AK, Barat A. The complete mitogenome of brown trout (*Salmo trutta fario*) and its phylogeny. *Mitochondrial DNA A DNA Mapp Seq Anal*, 2016, 27(6): 4563–4565. [DOI]
- [35] Shi YC, Liu Y, Zhang SZ, Zou R, Tang JM, Mu WX, Peng Y, Dong SS. Assembly and comparative analysis of the complete mitochondrial genome sequence of *Sophora japonica* 'JinhuaiJ2'. *PLoS One*, 2018, 13(8): e0202485. [DOI]
- [36] Al-Nakeeb K, Petersen TN, Sicheritz-Pontén T. Norgal: extraction and de novo assembly of mitochondrial DNA from whole-genome sequencing data. *BMC Bioinformatics*, 2017, 18(1): 510. [DOI]
- [37] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*, 2012, 19(5): 455–477. [DOI]
- [38] Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res*, 2017, 45(4): e18. [DOI]
- [39] Meng GL, Li YY, Yang CT, Liu SL. MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. *Nucleic Acids Res*, 2019, 47(11): 63. [DOI]
- [40] Bignell GR, Miller AR, Evans IH. Isolation of mitochondrial DNA. *Methods Mol Biol*, 1996, 53: 109–116. [DOI]
- [41] Li G, Davis BW, Eizirik E, Murphy WJ. Phylogenomic evidence for ancient hybridization in the genomes of living cats (Felidae). *Genome Res*, 2016, 26(1): 1–11. [DOI]
- [42] Yang QQ, Li ZH, Liu LJ. Advance and application of mtDNA COI barcodes on insects. *Chin Bull Entomol*, 2012, 49(06): 1687–1695.
杨倩倩, 李志红, 伍祎, 柳丽君. 线粒体 COI基因在昆虫 DNA 条形码中的研究与应用. 应用昆虫学报, 2012, 49(06): 1687–1695. [DOI]
- [43] Sha M, Lin LL, Li XJ, Huang Y. Strategy and methods for sequencing mitochondrial genome. *Chin Bull Entomol*, 2013, 50(01): 293–297.
沙淼, 林立亮, 李雪娟, 黄原. 线粒体基因组测序策略和方法. 应用昆虫学报, 2013, 50(01): 293–297. [DOI]
- [44] Li TJ, Cao YX, Zhao HC, Yu Y, Qiao J. Research progress of sequencing method for animal mitochondrial genome. *Tianjin Med J*, 2016, 44(06): 796–800.
李天杰, 曹延祥, 赵红翠, 于洋, 乔杰. 动物线粒体基因组测序方法的研究进展. 天津医药, 2016, 44(06): 796–800. [DOI]
- [45] Groenenberg DSJ, Harl J, Duijm E, Gittenberger E. The complete mitogenome of *Orcula dolium* (Draparnaud, 1801); ultra-deep sequencing from a single long-range PCR using the Ion-Torrent PGM. *Hereditas*, 2017, 154: 7. [DOI]
- [46] King JL, Larue BL, Novroski NM, Stoljarova M, Seo SB, Zeng X, Warshauer DH, Davis CP, Parson W, Sajantila A, Budowle B. High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. *Forensic Sci Int Genet*, 2014, 12: 128–135. [DOI]
- [47] Hunter SS, Lyon RT, Sarver BAJ, Hardwick K, Forney LJ, Settles ML. Assembly by Reduced Complexity (ARC): a hybrid approach for targeted assembly of homologous sequences. *bioRxiv*, 2015: 014662. [DOI]
- [48] Machado DJ, Lyra ML, Grant T. Mitogenome assembly from genomic multiplex libraries: comparison of strategies and novel mitogenomes for five species of frogs. *Mol Ecol Resour*, 2016, 16(3): 686–693. [DOI]
- [49] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 2009, 25(14): 1754–1760. [DOI]
- [50] Min-Shan Ko A, Zhang YQ, Yang MA, Hu YB, Cao P, Feng XT, Zhang LZ, Wei FW, Fu QM. Mitochondrial genome of a 22,000-year-old giant panda from southern

- China reveals a new panda lineage. *Curr Biol*, 2018, 28(12): R693–R694. [DOI]
- [51] Taylor RW, Turnbull DM. Mitochondrial DNA mutations in human disease. *Nat Rev Genet*, 2005, 6(5): 389–402. [DOI]
- [52] Torroni A, Achilli A, Macaulay V, Richards M, Bandelt HJ. Harvesting the fruit of the human mtDNA tree. *Trends Genet*, 2006, 22(6): 339–345. [DOI]
- [53] Green RE, Malaspina AS, Krause J, Briggs AW, Johnson PLF, Uhler C, Meyer M, Good JM, Maricic T, Stenzel U, Prüfer K, Siebauer M, Burbano HA, Ronan M, Rothberg JM, Egholm M, Rudan P, Brajković D, Kućan Z, Gušić I, Wikström M, Laakkonen L, Kelso J, Slatkin M, Pääbo S. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell*, 2008, 134(3): 416–426. [DOI]
- [54] Zhidkov I, Nagar T, Mishmar D, Rubin E. MitoBam-Annotator: A web-based tool for detecting and annotating heteroplasmy in human mitochondrial DNA sequences. *Mitochondrion*, 2011, 11(6): 924–928. [DOI]
- [55] Guo Y, Li J, Li CI, Shyr Y, Samuels DC. MitoSeek: extracting mitochondria information and performing high-throughput mitochondria sequencing analysis. *Bioinformatics*, 2013, 29(9): 1210–1211. [DOI]
- [56] Yang IS, Lee HY, Yang WI, Shin KJ. mtDNAprofiler: a Web application for the nomenclature and comparison of human mitochondrial DNA sequences. *J Forensic Sci*, 2013, 58(4): 972–980. [DOI]
- [57] Vellarikall SK, Dhiman H, Joshi K, Hasija Y, Sivasubbu S, Scaria V. mit-o-matic: a comprehensive computational pipeline for clinical evaluation of mitochondrial variations from next-generation sequencing datasets. *Hum Mutat*, 2015, 36(4): 419–424. [DOI]
- [58] Calabrese C, Simone D, Diroma MA, Santorsola M, Guttà C, Gasparre G, Picardi E, Pesole G, Attimonelli M. MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. *Bioinformatics*, 2014, 30(21): 3115–3117. [DOI]
- [59] Navarro-Gomez D, Leipzig J, Shen L, Lott M, Stassen AP, Wallace DC, Wiggs JL, Falk MJ, Van Oven M, Gai X. Phy-Mer: a novel alignment-free and reference-independent mitochondrial haplogroup classifier. *Bioinformatics*, 2015, 31(8): 1310–1312. [DOI]
- [60] Weissensteiner H, Forer L, Fuchsberger C, Schöpf B, Kloss-Brandstätter A, Specht G, Kronenberg F, Schönherr S. mtDNA-Server: next-generation sequencing data analysis of human mitochondrial DNA in the cloud. *Nucleic Acids Res*, 2016, 44(W1): W64–69. [DOI]
- [61] Ishiya K, Ueda S. MitoSuite: a graphical tool for human mitochondrial genome profiling in massive parallel sequencing. *PeerJ*, 2017, 5: e3406. [DOI]
- [62] Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. *Genome Res*, 2009, 19(9): 1639–1645. [DOI]
- [63] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 2010, 20(9): 1297–1303. [DOI]
- [64] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 2012, 9(4): 357–359. [DOI]
- [65] Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends Genet*, 2008, 24(3): 133–141. [DOI]
- [66] Hahn C, Bachmann L, Chevreux B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Res*, 2013, 41(13): e129. [DOI]
- [67] Hahn C. Assembly of ancient mitochondrial genomes without a closely related reference sequence. *Methods Mol Biol*, 2019, 1963: 195–213. [DOI]
- [68] Li R, Ren X, Bi Y, Ding Q, Ho VWS, Zhao Z. Comparative mitochondrial genomics reveals a possible role of a recent duplication of NADH dehydrogenase subunit 5 in gene regulation. *DNA Res*, 2018, 25(6): 577–586. [DOI]
- [69] Warren RL, Sutton GG, Jones SJ, Holt RA. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, 2007, 23(4): 500–501. [DOI]
- [70] Jeck WR, Reinhardt JA, Baltrus DA, Hickenbotham MT, Magrini V, Mardis ER, Dangel JL, Jones CD. Extending assembly of short DNA sequences to handle error. *Bioinformatics*, 2007, 23(21): 2942–2944. [DOI]
- [71] Bakker FT, Lei D, Yu JY, Mohammadin S, Wei Z, Van De Kerke S, Gravendeel B, Nieuwenhuis M, Staats M, Alquezar-Planas DE, Holmer R. Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an iterative organelle genome assembly pipeline. *Biol J Linn Soc*, 2016, 117(1): 33–43. [DOI]
- [72] Jin JJ, Yu WB, Yang JB, Song Y, dePamphilis CW, Yi

- TS, Li DZ. GetOrganelle: a simple and fast pipeline for de novo assembly of a complete circular chloroplast genome using genome skimming data. *bioRxiv*, 2018: 256479. [DOI]
- [73] Coissac E, Hollingsworth PM, Laverigne S, Taberlet P. From barcodes to genomes: extending the concept of DNA barcoding. *Mol Ecol*, 2016, 25(7): 1423–1428. [DOI]
- [74] Clark SC, Egan R, Frazier PI, Wang Z. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics*, 2013, 29(4): 435–443. [DOI]
- [75] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinformatics*, 2009, 10: 421. [DOI]
- [76] Bayliss SC, Hunt VL, Yokoyama M, Thorpe HA, Feil EJ. The use of Oxford Nanopore native barcoding for complete genome assembly. *Gigascience*, 2017, 6(3): 1–6. [DOI]
- [77] Cao MD, Nguyen SH, Ganesamoorthy D, Elliott AG, Cooper MA, Coin LJ. Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *Nat Commun*, 2017, 8: 14515. [DOI]
- [78] Deschamps S, Zhang Y, Llaca V, Ye L, Sanyal A, King M, May G, Lin HN. A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nat Commun*, 2018, 9(1): 4844. [DOI]
- [79] Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res*, 2015, 25(11): 1750–1756. [DOI]
- [80] Lin MM, Qi XJ, Chen JY, Sun LM, Zhong YP, Fang JB, Hu CG. The complete chloroplast genome sequence of *Actinidia arguta* using the PacBio RS II platform. *PLoS One*, 2018, 13(5): e0197393. [DOI]
- [81] Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods*, 2013, 10(6): 563–569. [DOI]
- [82] Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*, 2017, 27(5): 722–736. [DOI]
- [83] Miyamoto M, Motooka D, Gotoh K, Imai T, Yoshitake K, Goto N, Iida T, Yasunaga T, Horii T, Arakawa K, Kasahara M, Nakamura S. Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. *BMC Genomics*, 2014, 15(1): 699. [DOI]
- [84] Soorni A, Haak D, Zaitlin D, Bombarely A. Organelle_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing data. *BMC Genomics*, 2017, 18(1): 49. [DOI]
- [85] Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, 2012, 13: 238. [DOI]
- [86] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. The sequence alignment/map format and SAMtools. *Bioinformatics*, 2009, 25(16): 2078–2079. [DOI]
- [87] McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*, 2004, 32(Web Server issue): W20–25. [DOI]
- [88] Boetzer M, Pirovano W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics*, 2014, 15: 211. [DOI]
- [89] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 2010, 26(6): 841–842. [DOI]
- [90] Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, 2012, 1(1): 18. [DOI]
- [91] Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*, 2008, 18(5): 821–829. [DOI]
- [92] Zhang TW, Luo YF, Chen YP, Li XN, Yu J. BIGrat: a repeat resolver for pyrosequencing-based re-sequencing with Newbler. *BMC Res Notes*, 2012, 5: 567. [DOI]
- [93] Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S, Zhou X, Lam TW, Li Y, Xu X, Wong GK, Wang J. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, 2014, 30(12): 1660–1666. [DOI]
- [94] Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B,

- Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*, 2013, 8(8): 1494–1512. [DOI]
- [95] Maccallum I, Przybylski D, Gnerre S, Burton J, Shlyakhter I, Gnirke A, Malek J, Mckernan K, Ranade S, Shea TP, Williams L, Young S, Nusbaum C, Jaffe DB. ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. *Genome Biol*, 2009, 10(10): R103. [DOI]
- [96] Kajitani R, Yoshimura D, Okuno M, Minakuchi Y, Kagoshima H, Fujiyama A, Kubokawa K, Kohara Y, Toyoda A, Itoh T. Platanus-alley is a de novo haplotype assembler enabling a comprehensive access to divergent heterozygous regions. *Nat Commun*, 2019, 10(1): 1702. [DOI]
- [97] Lee HO, Choi JW, Baek JH, Oh JH, Lee SC, Kim CK. Assembly of the mitochondrial genome in the campanulaceae family using Illumina low-coverage sequencing. *Genes*, 2018, 9(8): 383. [DOI]
- [98] Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 2015, 31(10): 1674–1676. [DOI]
- [99] Plese B, Rossi ME, Kenny NJ, Taboada S, Koutsouveli V, Riesgo A. Trimitomics: an efficient pipeline for mitochondrial assembly from transcriptomic reads in non-model species. *bioRxiv*, 2018, 19(5): 1230–1239. [DOI]
- [100] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 2014, 30(15): 2114–2120. [DOI]
- [101] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, Di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*, 2011, 29(7): 644–652. [DOI]
- [102] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*, 1990, 215(3): 403–410. [DOI]
- [103] Li M, Schroeder R, Ko A, Stoneking M. Fidelity of capture-enrichment for mtDNA genome sequencing: influence of NUMTs. *Nucleic Acids Res*, 2012, 40(18): e137. [DOI]
- [104] Li Y, Li X, Chen Y. Research summary of mitochondria pseudogene. *J Mianyang Norm Univ*, 2012, 31(05): 68–75.
李艳, 黎霞, 陈艳. 线粒体假基因研究综述. 绵阳师范学院学报, 2012, 31(05): 68–75. [DOI]
- [105] Velozo Timbó R, Coiti Togawa R, Costa MMC, Andow DA, Paula DP. Mitogenome sequence accuracy using different elucidation methods. *PLoS One*, 2017, 12(6): e0179971. [DOI]
- [106] Peters JL, Bolender KA, Pearce JM. Behavioural vs. molecular sources of conflict between nuclear and mitochondrial DNA: the role of male-biased dispersal in a Holarctic sea duck. *Mol Ecol*, 2012, 21(14): 3562–3575. [DOI]
- [107] Ekblom R, Smeds L, Ellegren H. Patterns of sequencing coverage bias revealed by ultra-deep sequencing of vertebrate mitochondria. *BMC Genomics*, 2014, 15: 467. [DOI]

(责任编辑: 吴东东)