系统发育基因组学方法研究进展

李佳璇, 梁丹, 张鹏*

中山大学生命科学学院, 广州 510006

* 联系人, E-mail: zhangp35@mail.sysu.edu.cn

收稿日期: 2018-12-20; 接受日期: 2019-01-31; 网络版发表日期: 2019-04-15

摘要 系统发育基因组学是系统发育学与基因组学融合产生的一个交叉学科, 主要通过分析基因组水平的大规模分子数据来解析生物之间的系统进化关系. 本文简述了5种常用的基于高通量测序技术的系统发育基因组学数据获取方法: 扩增子测序(targeted amplicon sequencing)、转录组测序(transcriptome sequencing)、简化基因组测序(reduced-representation sequencing)、目标序列捕获(target sequence capture)和低覆盖度的全基因组测序(low coverage whole-genome shotgun sequencing), 归纳比较了每种方法的特点、适用范围、成本花费等. 此外, 本文还简述了系统发育基因组学数据分析中的5个基本步骤: 数据预处理与组装、直系同源基因的鉴定、序列比对、数据质量控制和系统发育树的构建, 并对这5个步骤的分析要点与常用软件的特点做了归纳. 虽然系统发育基因组学分析是重建生命之树的一条必由之路, 当前它也面临着一系列的问题: 如大量数据的分析导致计算量与时间花费大幅增加、系统误差的影响会随着数据量增大而变得更为明显、不同数据集可能会给出高支持率但不一致的结果. 最后, 本文综合国内外学者的最新研究成果, 对这些问题进行了讨论并给出了相应的建议.

关键词 高通量测序, 系统发育基因组学, 系统发育分析, 系统误差, 系统发育信号

系统发育学(Phylogenetics)是一门追溯生物类群起源与进化, 并通过重建系统发育树来研究生命类群之间进化关系的学科, 是进化生物学研究的重要领域之一. 构建高置信度的系统发育关系不仅是生物分类的基础, 也是阐明物种起源、扩散、进化、揭示物种形成机制的前提^[1~3]. 由于技术条件的限制, 早期的分子系统发育学研究常常基于单个或少量基因. 但由于受到谱系不完全分选(incomplete lineage sorting)、旁系同源基因(paralog)、饱和替换(substitution saturation)、生物杂交(hybridization)等因素的影响, 基于单个基因构建的系统发育树有时并不能真实反映生物的

进化历史^[4~6]. 此外, 对于一些难以解析的进化节点, 少量基因的分析也难以提供足够的信息. 因此, 近年来学界逐渐出现一种将基因组学和系统发育学结合的研究范式, 旨在利用基因组级别的分子数据重建生物类群的系统进化关系、研究生物的起源与进化过程, 系统发育学研究由此也进入了一个崭新的阶段——系统发育基因组学(Phylogenomics)^[7~11]. 近年来, 高通量测序技术和基因组学研究的快速发展为系统发育基因组学带来了前所未有的机遇. 随着测序成本的快速下降, 越来越多物种的全基因组已经完成了测序, 但考虑到实验成本以及序列拼接、注释等一系列问题, 利用全基

引用格式: 李佳璇, 梁丹, 张鹏. 系统发育基因组学方法研究进展. 中国科学: 生命科学, 2019, 49: 456–471

Li J X, Liang D, Zhang P. Recent progress in phylogenomic methods (in Chinese). Sci Sin Vitae, 2019, 49: 456–471, doi: 10.1360/N052018-00222

基因组数据进行常规系统发育研究目前仍不太可行. 实际上, 对于大多数系统发育分析而言, 并不需要完整的全基因组序列, 只需要得到部分基因组数据即可满足研究需要. 因此针对部分基因组序列的高通量测序方法对于系统发育基因组学研究的重要性日益突显^[12,13]. 目前, 系统发育基因组学分析已在各大生物类群如真菌、动物、植物的系统发育重建中得到了成功应用^[14-23], 但伴随而来的海量数据也对现有的数据分析方法提出了新的挑战. 本文综合国内外学者的最新研究成果, 对近年来国际系统发育基因组学研究所发展起来的各种数据获取的实验方法与数据分析流程进行了介绍, 探讨了系统发育基因组学分析中需要注意的一些普遍性问题, 希望为国内各生物类群系统发育基因组学研究的开展提供借鉴.

1 基于高通量测序技术的系统发育基因组学数据获取方法

1.1 扩增子测序

近年来, 多基因分析被广泛应用于物种间系统发育关系重建^[4,13,24]. 传统方法是利用PCR技术扩增目的基因片段, 然后进行Sanger测序获得数据. 对于研究物种少(<50)和基因数目少(<10)的研究, 该方案的实现难度低, 普通的分子系统学实验室均可以开展工作. 近年来, 针对不同生物类群的通用基因标记集正不断被开发出来, 如102个脊椎动物通用核基因标记^[25]、96个蛇类通用基因标记集^[26]以及95个甲虫通用核基因标记集^[27]等, 这为系统学多基因分析提供了大量标记资源. 然而, 当DNA样本和/或基因数目增多时, 需要处理的PCR产物将达到成千上万个, 全部采用Sanger测序, 无论从劳力还是花费上来看都是难以承受的. 随着高通量测序技术的出现, 分子系统学家逐渐开始采用成本更低的高通量测序, 如454^[28,29], Ion Torrent^[30], Illumina^[31-33]等平台来进行大批量的PCR产物测序. 其中, 短读长的Illumina测序平台凭借其测序成本低(约\$0.01/MB)、通量高的优势, 目前已成为高通量扩增子测序的主要实现平台.

基于Illumina的扩增子测序的大致流程是: 将同一样本的不同PCR产物进行混合, 利用片段化酶或超声波仪器进行DNA打断后连接物种特异性标签(barcode), 最后将不同样本的标记片段混合并进行文库构

建测序; 在后期数据处理过程中, 首先利用特异性标签将不同物种的reads进行分选, 然后用基因组或转录组组装软件(如Trinity^[34], SPAdes^[35])进行序列拼接, 随后对获得的contigs进行直系同源基因鉴定, 最终获得目标扩增子序列(具体实验与数据分析流程可参考文献^[33]). 与Sanger扩增子测序相比, Illumina扩增子测序在测序成本与实验劳动强度上有着非常明显的优势. 例如, 一个100个物种×100个基因的系统发育基因组学项目, 需要测定10000个PCR产物的序列, 按双向Sanger测序, 每个测序反应12元计, 只是测序花费就达到24万, 这是大多数分子系统学实验室所难以负担的. 如果采用Illumina扩增子测序, 只需要10~20 GB的测序量便可以完成任务, 加上建库成本, 总花费不会超过3000元. 同时, Illumina扩增子测序的实验操作过程更为简单, 无需对每个PCR产物进行单独纯化, 对于有杂带、无可见条带的PCR产物可以直接混合建库, 只要测序通量高, 也可获得相应数据. 除此之外, Illumina扩增子测序对基因标记的长度没有限制, 可以直接测定长度在10000 bp以上的扩增子. 虽然Illumina测序平台的读长较短, 在序列组装过程中有可能会出现问题, 但可通过后续的生物信息学分析加以去除.

扩增子测序的不足之处在于: (i) 对DNA质量要求较高, 若DNA降解程度严重, PCR扩增成功率将大大降低, 这使得扩增子测序方案无法使用一些质量低的标本; (ii) DNA消耗量大, 按每个PCR反应消耗20 ng DNA模板算, 完成100个基因的扩增需要消耗2 μg基因组DNA, 这对于一些个体微小且珍贵的标本是不可行的; (iii) 当研究规模较大时, PCR的劳动强度会比较高. 以扩增96个分子标记为例, 假设一个实验室有3台PCR仪器, 一般每天可处理3个样本, 若总样本数为200, 则至少需要连续工作3个月来进行PCR扩增. 无论如何, 高通量扩增子测序的实验条件与数据分析流程都较为简单, 是传统Sanger测序分子系统学实验室在较短的时间里最容易掌握的一种高通量数据技术, 最终获得的数据也比较容易与公共数据库数据相整合, 在未来的系统发育基因组学研究中仍将会占有一席之地.

1.2 转录组测序

转录组测序(RNA-seq)是特定细胞在某一功能状态下所能转录出来的所有RNA总和(主要包括mRNA和非编码RNA)进行高通量测序的一种技术. 通过高通

量测序能够全面快速地获取某一物种特定组织或器官在某一状态下的全部转录本序列信息^[36]。在系统发育基因组学研究中, 该技术通常被用于不同分类尺度下的系统发育关系重建、生物适应性进化等分析上。获得转录组测序数据的一般流程是: 首先提取高质量的总RNA, 然后进行mRNA富集或者去除高丰度的rRNA, 随后合成双链cDNA, 构建测序文库进行高通量测序。获得测序数据后, 首先对原始数据进行质量控制和筛选, 随后利用Trinity, ABySS^[34,37]等转录组拼接软件对数据进行从头组装(*de novo* assembly), 接着进行去冗余、寻找可读框(open reading frames)等操作, 最后通过生物信息学比对, 鉴定出直系同源基因, 用于系统发育基因组学分析。转录组测序的优势是不需要参考基因组信息, 能够一次性获得大量编码序列信息, 非常适合非模式生物的研究。与扩增子测序相比, 转录组测序不需要利用参考序列进行分子标记的筛选、引物设计及PCR扩增, 省时、大大降低实验劳动强度, 且获得的基因数目远多于扩增子测序, 是目前系统发育基因组学研究中广泛采用的数据获取方式, 在真菌、节肢动物、陆生植物等不同生物类群中得到成功运用^[38-41]。例如, Misof等人^[42]测定了103种现存昆虫代表种类的转录组序列并结合其他已发表的基因组数据, 涵盖了昆虫纲30个目及节肢动物外类群共计144个代表物种, 利用筛选得到的1478个单拷贝蛋白质编码同源基因, 构建了至今最完整、可信度最高的昆虫高级阶元系统发育树。

然而利用转录组数据进行系统发育基因组学研究的主要缺点在于以下四个方面。(i) 由于要提取高质量的RNA, 转录组测序需要活体标本或液氮保存的新鲜组织样品, 这大大增加了样品材料收集与运输的难度。(ii) 无法分析博物馆藏标本: 对于一些难以获得活体标本的物种(如濒危生物), 博物馆藏标本几乎是材料来源的唯一途径。(iii) 实验成本仍然较高: 虽然近年来转录组测序的价格已经大降低, 但目前仍然在2000元/样品左右;(iv) 数据分析难度较高, 单个转录组测序的数据量一般在6~10 GB, 数据处理对于计算机的运算与存储有较高的硬件要求, 而且转录组测序一般会获得上千个同源基因的数据, 对这种体量的数据进行分析处理也需要较高的生物信息学分析能力。

1.3 简化基因组测序

简化基因组测序(reduced-representation genome sequencing)是指利用限制性内切酶消化基因组DNA, 回收特定长度的酶切片段并进行高通量测序。这类测序技术能够大幅度降低物种基因组复杂程度, 显著降低测序成本, 广泛应用于SNP标记开发、基因分型、群体遗传结构研究等方面。在系统发育基因组学中, 常被用来解决低阶元系统发育关系和谱系地理学中的问题, 代表方法有限制性酶切位点关联DNA测序(restriction-site associated DNA sequencing, RAD-seq)^[43,44]、简化多态序列复杂度测序(complexity reduction of polymorphic sequences, CROPS)^[45]、简化代表文库测序(reduced-representation libraries sequencing, RRLs)^[46]、基因分型测序(genotyping by sequencing, GBS等^[47]。其中应用最为广泛的标准RAD-seq是利用限制性核酸内切酶消化物种基因组DNA, 在酶切基因组片段两端添加P1测序接头, 随后对片段进行打断处理, 回收一定大小范围的片段后加上P2测序接头, PCR扩增富集同时具有P1和P2接头的DNA片段, 最后进行高通量测序。该技术实验操作简单、成本低、无需参考基因组, 能够高效地获取大量SNP标记用于系统发育关系构建和群体遗传结构分析。如Meier等人^[48]为了探究维多利亚湖流域丽鱼的爆炸性适应辐射过程, 对该流域的156个慈鲷科物种进行RAD-seq, 获得了436166个SNP位点, 建树结果表明所有来自维多利亚湖流域的丽鱼呈单系起源且绝大多数内部节点的Bootstrap支持率大于90%, 证明利用RAD-seq数据能够很好地解析辐射进化生物类群的系统关系。然而, RAD-seq方法对样本DNA质量要求较高^[49], 不能应用在DNA高度降解的样本中。除此之外, 当分析进化跨度大的生物类群时, 由于不同进化枝中的类群所共有的酶切位点会发生丢失, RAD-seq测序会存在直系同源核酸序列数据缺失率高(missing data)、遗传多样性估算存在偏差, 发现无效等位基因(null alleles)等问题^[50-52]。近年来基于RAD-seq的改进方法不断涌现, 包括2b-RAD^[53,54], ddRAD^[55], ezRAD^[56], Rapture^[57]等。本文从DNA降解的影响, 实验操作流程、实验成本等方面对这些RAD技术进行比较(表1), 不同的研究学者可结合研究类群的特点、研究目的等, 选择最佳的实验技术手段。

表 1 5种RAD-seq技术的比较

Table 1 Comparison of five RAD-seq technologies

方法	标准RAD	2bRAD	ezRAD	ddRAD	Rapture
限制性内切酶数目	1	1	≥1	2	1
限制性内切酶种类	稀有酶	II B型限制性内切酶	同裂酶	稀有酶和常见酶	稀有酶
DNA降解对实验的影响	大	小	大	大	大
起始DNA量	100 ng~1 μg	低至1 ng	50 ng~1 μg	50 ng~100 ng	50 ng
基因组二次打断	需要	不需要	不需要	不需要	需要
插入片段大小	≤300 bp	33~36 bp	≤300 bp	≤300 bp	≤300 bp
文库制备时间	~24 h	~4 h	~6 h	~8 h	~30 h
实验操作流程	复杂	简单	简单	简单	复杂
剪切位点选择	改变限制性内切酶	改变限制性内切酶	改变限制性内切酶和片段大小选择	改变限制性内切酶和片段大小选择	设计不同区域的捕获探针和改变限制性内切酶
实验成本	高	低	中等	很低	中等

1.4 目标序列捕获技术

目标序列捕获技术是基于DNA杂交原理,通过专门设计的探针与基因组序列进行杂交,富集基因组中特定区域并进行高通量测序的方法.该方法的一般流程是,首先通过参考基因组序列比对分析选择可以富集的保守序列区域,根据序列的保守性设计合成具有生物素标记的长度为80~120 bp的DNA或RNA单链探针,随后将探针与目标样本的基因组测序文库进行杂交并固定在生物素亲和磁珠上,最后洗脱未杂交上的文库片段,回收目标文库片段进行高通量测序^[58].序列捕获技术对DNA质量要求较低,适用于DNA降解较为严重的样本^[59,60],同时实验操作技术和设计捕获探针的生物信息学流程较为成熟,在捕获目标区域的长度和捕获基因的数量上有较高的灵活性,数据回收率和利用率较高,因而近年来已经成为了系统发育基因组学获取数据的主流方法.目前常用的目标序列捕获方案和特点如表2所示.其中锚定序列捕获(anchored enrichment, AHE)^[61]和超保守序列元件捕获技术(ultra-conserved element, UCE)^[62]仅通过一套通用探针就能够对进化跨度大的不同物种进行序列捕获^[69~73].如Prum等人^[74]应用AHE技术对198个鸟类物种进行序列捕获,成功构建了高解析度的鸟类的系统发育关系.目前商业化探针的主要供应商有Aglient SureSelect, Roche NimbleGen和MyBait等,价格都比较昂贵,因此,为了降低实验成本,许多学者将线粒体基因组或基因

组目标区域的PCR扩增产物作为捕获探针,进行目标序列捕获,也得到了不错的效果^[66~68].这类实验室自制探针的方法,实验成本低、灵活性强,可应用在不同生物类群当中.然而,由于基因片段数量的限制,自制探针的方法能捕获到的标记数目一般较少,而且由于分子标记引物设计需要相关的参考基因组序列,对非模式生物类群的应用会受到一定的限制^[75].

1.5 低覆盖度的全基因组测序

近年来,对于一些基因组较小的生物类群,如虱目,有些学者尝试利用低覆盖度的全基因组测序进行系统发育基因组学研究^[76].该方法直接对每个样品的基因组DNA随机打断、构建Illumina鸟枪测序文库并进行10倍左右覆盖度的测序,随后从测序数据中提取同源基因序列.针对这一实验策略,Allen等人^[77]开发了一种自动化限制目标序列组装方法(automated target restricted assembly method, aTRAM).该方法需要提供一个近缘或远缘物种的参考基因序列数据集,结合BLAST比对、从头组装和迭代策略,能够从一个全基因组双端(paired-end)测序文库中快速获取相对应的目标基因序列.2017年,Boyd等人^[78]对61种鸽羽虱进行全基因组测序,利用aTRAM软件获得的977个直系同源基因数据集,构建了至今可信用度最高的鸽羽虱属内的系统发育树.与UCE, AHE为代表的目标序列捕获方法相比,结合aTRAM的低覆盖度的全基因组测序技术无需进行探针设计、合成过程及杂交实验,只需在

表 2 4种目标序列捕获技术的比较

Table 2 Comparison of four target sequence capture approaches

方法	代表文献	研究对象	基因组捕获区域	串联数据集大小	探针来源	杂交方式
锚定序列捕获	[61]	动物	核基因序列	~123 KB	商业合成	液相
超保守序列元件捕获	[62]	动物	核基因序列	~352 KB	商业合成	液相
外显子捕获	[63]	动物	核基因序列	~25.3 MB	商业合成	液相
	[64]	动物	核基因序列、线粒体基因	~4 MB	商业合成	固相
	[65]	动物	核基因序列	~400 KB	商业合成	液相
基于PCR产物制备探针的捕获	[66]	动物	线粒体基因组	~17 KB	实验室自制	液相
	[67]	植物	叶绿体基因组	~146 KB	实验室自制	液相
	[68]	动物	核基因序列	~52 KB	实验室自制	液相

后期的数据分析过程中, 提供一个目标参考序列数据集, 就可达到类似目标序列捕获的结果. 该方法对样品DNA质量的要求低、操作简单、花费少, 但这一方法数据利用率很低(<5%), 仅适合研究小型基因组的生物类群, 对于基因组大于1 GB以上的物种, 这一方法就并不那么有效了.

1.6 总结

从当前系统发育基因组学分析在国内外的发展状况来看, 简化基因组测序和目标序列捕获技术是当前系统发育基因组学研究获取数据的两个主要技术手段. 相比而言, 目标序列捕获技术对DNA质量要求更低, 可对捕获目标区域、分子标记的数目及长度进行选择, 几乎适用所有的生物类群, 且捕获效率高、数据回收速度快、数据集的缺失率低, 适合不同阶元的系统发育分析, 具有更好的发展前景. 简化基因组测序在高阶元进化分析中数据缺失率较高, 但可以有效解决复合种及种下水平的系统发育关系和系统地理学问题. 研究者应该考虑不同方法的优缺点与适用范围(表3), 选择合适的数据获取方法来开展研究工作.

2 系统发育基因组学数据处理流程

相比于传统的基于Sanger测序的常规系统发育数据分析, 系统发育基因组学研究中产生的高通量测序数据增加了数据分析的难度和复杂度. 目前系统发育基因组学数据分析流程一般分为数据预处理与组装、直系同源基因的鉴定、序列比对、数据质量控制和系统发育树的构建(图1).

2.1 数据预处理与组装

为了提高数据质量, 对高通量测序数据分析之前要对它进行预处理, 去除低质量测序数据和污染序列(主要是人和细菌的DNA污染), 以免影响分析结果. 可使用Trimmomatic^[79], Bowtie2^[80]等软件对测序所得到的原始数据(raw data)进行质量控制和筛选, 去除测序序列中所有含N的、质量较低的序列(phred quality score>30)以及污染序列, 得到高质量的序列数据(clean reads). 如果是多样本混合高通量测序, 在这一步中还可以依据样品特异性标签(barcode或index)将clean reads分选到对应的实验样本中, 随后再利用转录组拼接软件(Trinity^[34], StringTie^[81], Cufflinks^[82], SOAPdenovo-Trans^[83]), 基因组拼接软件(SOAPdenovo 2^[84], SPAdes^[35], ABySS^[37], Velvet^[85])或细胞器拼接软件(NOVOpasty^[86], MITObim^[87])对clean reads进行序列从头拼接(*de novo assembly*) 或利用BWA^[88], Bowtie2等基于参考序列进行数据定位(mapping), 最终获得拼接好的DNA片段重叠群(contigs).

2.2 直系同源基因的鉴定

现有的系统发育分析都是在所研究的基因序列是直系同源的假设基础上进行的, 因此直系同源基因的鉴定是系统发育分析的重中之重. 在组装得到的contig群中快速鉴定出直系同源基因, 最为常用的策略是双向BLAST方法: 即如果1号物种的A序列在2号物种的第一个BLAST Hit为B序列, 反过来2号物种的B序列在1号物种的第一个BLAST Hit还是A序列, 则A序列与B序列为1号物种与2号物种的直系同源序列. 这

表 3 5种基于高通量测序技术的系统发育基因组学数据获取方法的比较

Table 3 Comparison of five methods of data collection in phylogenomics based on high-throughput sequencing

方法	扩增子测序	转录组测序	简化基因组测序	目标序列捕获	低覆盖度的全基因组测序
适用的分类阶元范围	种级至目级水平	种级至纲级水平或更高	种下、种间及属内	种级至纲级水平或更高	种级至纲级水平或更高
获得标记的长度	中等	长	短	短到中等	中等
数据集缺失率	低	中等	高	中等	中等
数据集信息量 (串联数据集大小或SNPs数目)	50~200 KB	1~20 MB	几万~几十万SNPs	1.5~26 KB	100 KB以上
核酸质量要求	DNA片段>5 KB	需新鲜样品提RNA	DNA片段>20 KB	DNA片段>200 bp	DNA片段>200 bp
实验操作难度	简单	简单	简单	中等	简单
数据分析难度	简单	中等	简单	中等	简单
实验成本	中等	高	低	中等	低

图 1 系统发育基因组学数据处理流程. 从获得高通量测序原始数据开始, 圆角方框与箭头代表数据分析的顺序和步骤, 灰色虚线和直角方框表示相应步骤所对应的常用软件及方法(网络版彩图)

Figure 1 The bioinformatic pipeline for analyzing high-throughput phylogenetic data. Beginning with high-throughput data, Fillet square boxes and arrows represent the order and steps of data analysis. Gray dotted lines and square boxes represent common used software and methods (color online)

一策略适用于任何生物类群, 在没有参考序列存在的情况下就能够简便有效地鉴定出直系同源序列, 特别适用于转录组分析, 但该方法计算需求较大. 此外, 研究者还可以采用 orthoMCL-DB^[89], orthoDB^[90], OMA^[91], Inparanoid^[92]等直系同源基因数据库鉴定直系同源基因, 这种分析策略对计算资源要求较低, 但鉴定精度依赖于库中是否有相近的直系同源基因. 研究者可根据所研究的生物类群、实验室对软件的熟悉程度、数据分析系统的硬件配置要求等方面来选择合

适的直系同源基因鉴定方法.

特别要提出注意的是, 在不同生物类群中, 鉴别直系同源基因的难度是不一样的. 相对于脊椎动物而言, 植物在进化历史过程中多倍体化事件发生较为普遍, 无脊椎动物中基因复制和丢失比较容易发生^[93,94]. 在这两类生物中, 容易出现将旁系同源序列鉴定为直系同源序列的情况, 需要格外注意. 为降低这种错误的风险, 建议为每个基因构建单基因树, 然后判断一些单系性非常明显的类群(如被子植物、两侧对称动物)

是否在单基因树中出现不单系的情况。如果出现则表明在这个基因数据集中可能存在直系同源序列和旁系同源序列混合的问题, 需要将旁系同源序列从该基因数据集中移除。

2.3 序列比对

完成直系同源基因鉴定后, 下一步需要进行多序列比对, 对于蛋白质编码序列, 可以使用常规的MAFFT^[95], ClustalW^[96], MUSCLE^[97]等程序进行批量序列比对, 对于进化速度更快的非编码序列(如intron), 为了提高比对准确性, 建议使用针对高变序列专门开发的序列比对程序如SATé^[98], PASTA^[99]。

2.4 数据质量控制

特别要指出的是, 在系统发育基因组学数据分析过程中, 由高通量测序、序列组装、直系同源基因鉴定等引入的错误是难以完全避免的, 所以在序列比对后, 数据集中大概率会存在一些对后续分析造成影响的错误序列。为了进一步提高数据质量, 确保鉴定的直系同源基因数据集的准确性, 删除数据集中不可靠的序列或去除旁系同源序列是至关重要的。如何鉴别出错误的序列目前并没有一套大家普遍公认的处理流程。较为经典的例子是Simion等人^[100]关于多细胞动物系统进化关系研究。他们采用的数据质量控制策略较有代表性, 主要包括: (i) 通过BLAST公共数据库删除污染序列(主要是人或细菌的DNA污染); (ii) 去除由测序错误或体内转录降解的重复转录本; (iii) 删除不符合物种数目要求和缺少关键物种的alignment; (iv) 构建单基因树, 通过物种的聚类情况鉴别出可能的旁系同源序列; (v) 去除单基因树中枝长异常的序列(例如某一枝长大于全树总枝长的50%, 则该枝被删除); (vi) 删除数据集缺失率(missing data)较高的物种。根据本实验室的经验, 在这些步骤中, 删除污染序列、构建单基因树去除旁系同源序列、去除枝长异常的序列这三个步骤对数据质量的提升是最为关键的。

近期, Shen等人^[101]分析了植物、动物、真菌生命树上17个困难进化节点上的系统发育信号分布情况。他们发现在利用系统发育基因组数据解析某些困难节点中, 有时候信号分布具有明显的偏向性, 集中在单个或者少量几个基因中, 从数据集中删除这些基因会改变最终的分析结果。他们提出了一种基于似然率对每

个基因的系统发育信号进行检测的参量 Δ GLS, 可以计算出数据集中的每一个基因的信号贡献强度。基于这一分析, 研究者可以快速从大量的基因中定位出那些有决定性影响的基因, 判明它们是否是由人为错误引入(如存在旁系同源序列), 这样能够大大提高最终数据集的可靠性。

2.5 系统发育树的构建

超级矩阵法(supermatrix approach)和基于多物种溯祖模型(multi-species coalescent model)的物种树分析(species-tree approach)是系统发育分析进行进化树重建的两类主要方法。超级矩阵法是将所有基因的比对矩阵进行串联处理, 生成一个合并的大型数据矩阵集, 等同于把许多个基因合并成一个超级基因, 因此也被称为串联分析(concatenation analysis)。串联分析将不同基因中的信号进行混合, 其基本假定(assumption)是所有基因都遵循同一进化过程。串联分析中一个很重要的步骤是确定最优的数据分区模型。这是因为不同基因由于功能上的差异(如编码区和非编码区), 其进化速度是非常不同的, 在编码序列中, 由于简并密码子的存在, 第一、二、三位密码子间的进化速度有明显的差异。将这些进化速度差异很大的数据不加区别地合并分析, 无法真实反映生物进化的过程, 会对系统发育树推断的稳定性产生较大影响, 因此设置数据分区模型就显得十分必要。数据分区模型选择最常用的软件是PartitionFinder^[102](最新版本: PartitionFinder2^[103], 可处理形态学和基因组级别数据), 它能够基于贝叶斯信息标准(Bayesian information criterion, BIC)给出最优数据分区方案, 同时还会给出每个分区模块的最优进化模型, 在分析速度与使用便利性上已大大超越老牌的Modeltest^[104], jModeltest^[105]以及ProtTest^[106]等软件。

目前在系统发育基因组学分析中应用最为广泛的串联分析均是基于似然率计算的, 主要分为两大类, 第一类是基于最大似然法(maximum likelihood, ML)的建树软件, 主要代表有RAxML^[107]和IQ-TREE^[108]。两者均可以处理包括上千个基因或上千个物种的大型数据集, 且支持多线程运行。其中IQ-TREE具有超快自展(ultrafast bootstrap approximation)的特性, 如果采用相同的进化模型, 计算速度是RAxML的10~40倍^[109]。第二类是基于贝叶斯法(Bayesian inference, BI)的建树软

件, 主要代表程序有MrBayes^[110], PhyloBayes^[111]和ExaBayes^[112]. 其中, MrBayes软件的使用频率最高, 针对大型数据集开发的ExaBayes软件计算速度相对较快; PhyloBayes软件中的CAT模型能够降低长枝吸引的影响, 但计算速度相对于其他软件来说会慢得多.

与串联法不同, 物种树方法承认不同基因有不同的进化历史, 这些分析会为数据集中的每个基因构建基因树, 最后从这些基因树中统计出最可能的物种树. 相对于串联法, 物种树方法在不完全分选发生的情况下表现更为稳健, 更有可能给出正确的结果. 有研究表明, 在解析胎盘哺乳动物进化历史中的辐射事件时, 物种树方法要优于串联法^[113]. 由于是统计分析, 物种树方法需要分析相当数量的基因树, 且假定的每个基因的基因树都应当是准确的. 当这一假定不被满足的时候(如基因数量太少、大多数基因序列过短, 无法准确推断出准确的基因树), 物种树分析的表现可能就会不理想.

基于多物种溯祖模型的物种树建树软件主要分为两大类, 第一类是基于频率统计的方法(summary method), 以基因树为输入进行统计分析, 基于基因树拓扑结构的概率分布计算得出物种树, 常用的软件有MP-EST^[114]和ASTRAL^[115]. 两者均能够利用成百上千的基因重建物种树, 耗时短、计算量少, 并且有研究表明, 在谱系不完全分选存在的情况下也能得到一致的结果^[114,116,117]. 相比而言, ASTRAL的计算速度更快, 被广泛地应用于基因组级别的数据分析中. 另外一类是基于全参数溯祖模型(fully parametric coalescent model)的方法, 同时推断基因树与物种树, 最为常用的软件是BEAST^[118], 该方法得到的物种树准确度高, 但计算量非常大, 很少能用于处理包含几百个基因或者几百个物种的大数据集.

3 当前系统发育基因组学研究面临的挑战

3.1 计算量与时间花费

利用基因组级别的数据进行系统发育分析对于计算机的计算能力来说是一个巨大的挑战. 随着数据量的不断增大, 计算花费的时间也会大幅增加. 常规的基于十几个基因的单次建树分析通常在几个小时之内即可获得结果, 但基于基因组级数据的单次建树分析常常需要几个星期甚至几个月才能完成. 为提高分析

效率, 目前的分析程序多采用多线程或并行化计算以提高运算速度. 如RAxML程序目前共有四种版本: Sequential, Pthreads, MPI, Hybrid Pthreads/MPI. 其中Pthreads, MPI, Hybrid Pthreads/MPI版本能够利用多线程计算或并行化计算来加速最大似然法分析, 可用于处理包括上千个基因或上千个物种的大型数据集. Pthreads版本可以利用单颗处理器的多个线程的合并计算能力来加速似然值的计算, 这种策略消耗的内存较少, 适合于高主频、内存小的单台计算服务器; MPI版本则将多个ML分析(如自展)并行计算于不同的处理器核心, 这种策略消耗内存量大, 适合用于具有多核处理器的集群系统.

此外, 一般计算机显卡中的图形处理器(graphics processing unit, GPU)的处理核心数量要远远高于CPU中的处理核心数量, 在并行计算方面可提供数十倍乃至上百倍于CPU的性能, 一些研究者据此提出了利用计算机显卡GPU强大的并行计算能力来提高运算速度的思路. 如Bao等人^[119]开发了GPU版的MrBayes程序, 他们提出了一种高性能的优化算法aMC³(adaptive MC³), 该算法大大提高了马尔科夫链的并发度, 能够充分利用GPU的并行计算资源, 大大加快运算速度. 与标准MrBayes程序相比, GPU版的MrBayes程序在一台装有一块显卡普通个人计算机上取得了63倍的提速, 在四块显卡的个人电脑上取得170倍的提速. 然而, 调用GPU进行并行计算会受限于显示内存的大小, 因而目前还无法处理太大的数据.

总的来说, 在系统发育基因组学研究中, 处理大规模数据集必然会带来计算量增大、时间成本增加的问题, 解决这一问题需要从算法优化、多核计算、并行计算等多方面进行更深入的研究, 开发更多适合处理大数据的软件版本, 提高计算速度、缩短计算时间.

3.2 系统误差问题

系统发育分析中的误差来源有两类. 第一类为随机误差(stochastic error), 这类误差是由样本量或基因片段所包含的系统发育信号不足导致的, 当数据量逐步增加, 随机误差对结果的影响会逐渐减少直至消除^[120-122]. 系统发育基因组学分析都会有大量的数据, 因此随机误差基本上不会对系统发育基因组学分析的最终结果造成影响. 第二类为系统误差(systematic error), 这类误差主要是由构建系统发育树时所应用的数

据进化模型所基于的一些假定与分子数据实际的进化过程不一致所导致的。与随机误差不同的是, 系统误差会随数据量的增加而不断增大, 最终导致得到了支持率极高的拓扑结构, 但却未能正确地反映真实的生物进化历史。所以, 在系统发育基因组学时代, 应该谨慎解读高支持率的系统发育关系, 排除系统误差造成的影响。

比较常见的系统误差包括三种。第一种是序列组成异质性(compositional heterogeneity)所造成的系统误差。目前常用的建树模型都假设DNA序列在不同生物类群的进化过程中, 其碱基的使用频率是不会发生改变的, 即进化过程中序列组分是同质的。然而在真实的情况下, 不同生物的序列组成成分存在一定的差异, 如果分析的数据集在不同物种间存在较大的序列组成异质性, 则在构树过程中往往会将序列组成相似的物种错误地聚为一支, 与真实进化关系存在偏差^[123]。在无脊椎动物(如节肢动物门)、植物、真菌等生物类群中, 序列组成异质性现象较为常见^[124-126]。一般来说, 相对于第一、二位密码子, 第三位密码子比较容易出现较强的序列组成异质性; 相对于核酸序列, 蛋白质序列的组成异质性会更低。因此, 在系统发育基因组学分析中, 可以尝试对DNA序列进行简并化处理或在蛋白质水平上进行系统发育分析, 以降低序列组成异质性可能带来的系统误差影响。

第二种是速率异质性(rate heterogeneity)所造成的系统误差。不同生物类群经历了不同的进化历史, 因而具有不同的进化速率^[127]。当类群间进化速率高度不一致时, 某些进化快的类群可能由于多重突变随机地在某些位点上获得相同的碱基, 而现有的进化模型会把这种足以掩盖真实进化历史信息的非同源相似(homoplasy)当作真实的“信号”, 从而将进化速度快的类群进行错误聚类, 造成“长枝吸引”现象(long-branch attraction, LBA)^[128]。降低长枝吸引影响的方法很多, 在这里只列举几个常用的策略: (i) 因为使用单个远缘生物作为外类群会在系统树中出现明显的长枝, 增加了长枝吸引发生的概率, 所以研究者应当采用亲缘关系近多个物种作为分析的复合外类群; (ii) 避免进化树中出现仅包含一个物种的长枝情况, 对于进化快的类群, 尽量多选择一些物种加入分析, 可以截断长枝^[129]; (iii) 去除进化太快的序列位点或者基因, 可以降低长枝吸引的影响; (iv) 采用对LBA不敏感的速率异质性

进化模型, 如Phylobayes程序的CAT模型^[130]。长枝吸引是系统发育基因组学分析中最容易发生的误差来源, 在各种生物类群与进化尺度下均可能发生, 需要在分析中格外注意。

第三种是碱基突变速率变化所造成的系统误差(heterotachy), 是指对于特定的碱基位点而言, 突变速率不是恒定不变的, 它会在某个时间段急剧增加, 也会在某个时间段下降, 但是这种随时间产生的速率转变在序列水平上无法检测, 而大多数现有模型的假设认为位点突变速率是恒定的, 当模型假设与现实出现严重不拟合的情况时, 就会得到一个错误的结果^[131,132]。对于碱基突变速率变化这种系统误差的特性目前研究较少, 也没有较为可行的检测与分析算法, 所以在系统发育基因组学分析中一般较少考虑。

事实上, 这三种系统误差是分子系统发育研究中的共性问题, 也存在于基于少量基因的分子系统发育研究中, 但在系统发育基因组学分析中, 系统误差的影响会随数据量的增大而变得更为明显, 需要格外注意。

3.3 结论不一致问题

在利用大量数据解决系统发育的困难问题时, 经常会出现用不同数据集和不同分析方法时所得的结果不一致的现象。造成这一问题的一个重要因素在于, 在海量的数据中, “信号”(signal)与“噪音”(noise)总是并存的, 当“噪音”在数据中所占比例超过“信号”所占比例, 则很可能导致不同研究得到高支持率但拓扑树型不一致的后果^[133]。为了从海量数据中有效提取真实“信号”, 降低“噪音”干扰, 学者们通常会根据数据质量(如数据集完整度)、数据特性(如基因进化速度)或预测数据信号强度等不同标准, 筛选一个合适的数据集来进行最后的建树分析。如Salichos和Rokas^[134]在利用23个酵母基因组的1070个直系同源基因解决酵母的系统发育关系时发现, 串联建树方法构建的系统发育树中所有节点的Bootstrap支持率均为100%, 但构建的1070个单基因树与串联树的拓扑结构均存在不一致性。他们研究了这种不一致性可能造成的后果, 发现基于常规数据筛选标准(如完整度、进化速率)过滤得到的不同数据集会发生显著的结果不一致问题, 而选择具有较强系统发育信号的基因时(基因信号强弱以单基因树的平均自展值average bootstrap support为指标),

不同数据集间结果的不一致性能得到有效地消除。

近期, Chen等人^[135]以有颌类脊椎动物的系统发育关系为例, 比较评估了不同数据筛选方案在提高数据信噪比、解决进化节点上的表现。他们发现, 不同基因在解析一棵进化树上的不同节点的表现是非常不同的, 同样一个基因, 在解析A节点时表现很好, 但在解析B节点上可能表现较差, 甚至给出错误的答案。因此, 不加分辨地将各种基因混合使用来同时解决多个困难节点并不合理。因此, 他们提出两种新的数据筛选方案, 称之为问题特异性数据筛选方案(question-specific approaches)。第一种筛选方法称为“假说控制法”(hypothesis-control), 是指在解析三个类群的系统发育关系时, 有且仅可能有三种可能的拓扑结构。当某基因树支持三种假说中的一种, 则代表该基因对解决该问题已明确表态, 反之, 若某基因树不支持三种假说中任意一种(无匹配基因, “non-matching” gene), 这说明该基因对解决该问题是没有贡献的, 甚至会产生“噪音”, 因此将这些基因从数据集中删除。第二种筛选方法称为“节点控制法”(node-control), 是指只保留基因树能支持给定进化节点(the control node)的基因。给定的节点应该是已经被广泛认同(如胎盘哺乳动物的单系性节点)并与所要研究的进化节点靠近, 如果某个基因不能重现给定节点, 则说明该基因可能是信号不足(stochastic error)或存在一定的系统误差, 将这些基因从数据集中删除可以达到将数据信号集中到特定问题的效果。最终结果表明, 问题特异性数据筛选策略能够有效地解决有颌类脊椎动物系统关系中具有争议的困难问题, 如支持海鲢总目(Elopomorpha)与骨舌总目(Osteoglossomorpha)聚类共同构成真骨鱼类(Euteleost)基出支、核心陆生鸟类(land birds)单系起源等。Wang等人^[136]近期在利用转录组数据研究异翅亚目昆虫的系统发育关系时也发现, 问题特异性数据筛选策略过滤出的基因数据要优于不加过滤的原始数据集。

总的来说, 虽然人们经常认为拓扑结构的支持率高意味着进化关系得到了很好的解析, 但在系统发育基因组学时代, 基于大量数据的分析常常能够获得支

持率极高的拓扑结构, 然而, 高支持率的结果可能并不是真实的进化关系, 而是由于系统误差造成的假象, 这就要求研究者在分析数据与解释结果的时候要更加谨慎。系统发育关系不一致现象在系统发育基因组学时代是普遍存在的, 数据量的大小与结果的正确与否并无关系。研究者应该通过各种数据筛选方法优化自己的数据, 通过对不同数据集、不同分析方法得出结果进行比较, 检查数据分析中是否出现不一致现象, 探讨不一致产生的可能原因, 来评估最终结果的可靠性。

4 总结与展望

近年来, 随着高通量测序技术的快速发展, 许多生物类群的全基因组测序计划已被提上日程: 如10000个脊椎动物代表物种的全基因组测序(Genome 10K Project; <https://genome10k.soe.ucsc.edu/>)、5000种节肢动物的全基因组测序(i5k initiative; <http://arthropodgenomes.org/wiki/i5K>)^[137]、1000种酵母全基因组测序(1000 Fungal Genomes Project; <http://1000.fungalgenomes.org/home/>)等。2017年, 国际生物基因组学会议(BioGenomics 2017)更是提出了“地球生物基因组计划(Earth Bio Genome Project)”, 规划在未来10年对地球上所有已知的真核生物(包括单细胞生物、真菌、植物及动物)进行基因组测序。有理由期待, 在不久的将来, 大部分物种的全基因组数据将会陆续公布, 系统发育基因组学将真正进入全基因组时代, 数据将不再是限制我们开展研究的瓶颈。然而, 越来越多的基因组数据也会带来新的问题。如何从体量惊人的基因组数据中可靠鉴别出直系同源基因并进行准确的序列比对? 如何提出更符合分子进化历史且适用基因组级别数据的数据进化模型与算法? 如何开发计算通量更高的分析程序, 使其能够同时处理更多的样本与更多的基因? 如何对数据中的真实进化信号进行甄别与富集, 以降低结果不一致发生的概率? 种种这些问题都将对未来的系统发育基因组学研究提出新的挑战, 也将是系统发育基因组学方法学上未来的研究热点。

参考文献

- 1 Soltis D E, Soltis P S. Contributions of plant molecular systematics to studies of molecular evolution. *Plant Mol Biol*, 2000, 42: 45–75
- 2 Soltis D E, Albert V A, Savolainen V, et al. Genome-scale data, angiosperm relationships, and “ending incongruence”: A cautionary tale in

- phylogenetics. *Trends Plant Sci*, 2004, 9: 477–483
- 3 Nei M, Kumar S. Molecular Evolution and Phylogenetics. New York: Oxford University Press, 2000
- 4 Degnan J H, Rosenberg N A. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol*, 2009, 24: 332–340
- 5 Zou X H, Ge S. Conflicting gene trees and phylogenomics (in Chinese). *J Syst Evol*, 2008, 46: 795–807 [邹新慧, 葛颂. 基因树冲突与系统发育基因组学研究. 植物分类报, 2008, 46: 795–807]
- 6 Nakhleh L. Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends Ecol Evol*, 2013, 28: 719–728
- 7 Delsuc F, Brinkmann H, Philippe H. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet*, 2005, 6: 361–375
- 8 Posada D. Phylogenomics for systematic biology. *Syst Biol*, 2016, 65: 353–356
- 9 Spatafora J W, Bushley K E. Phylogenomics and evolution of secondary metabolism in plant-associated fungi. *Curr Opin Plant Biol*, 2015, 26: 37–44
- 10 Yu L, Zhang Y P. Phylogenomics—an attractive avenue to reconstruct “Tree of Life” (in Chinese). *Hereditas*, 2006, 28: 1445–1450 [于黎, 张亚平. 系统发育基因组学—重建生命之树的一条迷人途径. 遗传, 2006, 28: 1445–1450]
- 11 Wang Z Q, Xie Z Y, Cai Y F, et al. Advances in phylogenomics (in Chinese). *Hereditas*, 2014, 36: 669–678 [王章群, 解增言, 蔡应繁, 等. 系统发育基因组学研究进展. 遗传, 2014, 36: 669–678]
- 12 Lemmon E M, Lemmon A R. High-throughput genomic data in systematics and phylogenetics. *Annu Rev Ecol Evol Syst*, 2013, 44: 99–121
- 13 McCormack J E, Hird S M, Zellmer A J, et al. Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol Phylogenetics Evol*, 2013, 66: 526–538
- 14 Hackett S J, Kimball R T, Reddy S, et al. A phylogenomic study of birds reveals their evolutionary history. *Science*, 2008, 320: 1763–1768
- 15 Dunn C W, Hejnol A, Matus D Q, et al. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 2008, 452: 745–749
- 16 Jung K H, Cao P, Seo Y S, et al. The rice kinase phylogenomics database: A guide for systematic analysis of the rice kinase super-family. *Trends Plant Sci*, 2010, 15: 595–599
- 17 Chen M Y, Liang D, Zhang P. Phylogenomic resolution of the phylogeny of laurasiatherian mammals: Exploring phylogenetic signals within coding and noncoding sequences. *Genome Biol Evol*, 2017, 9: 1998–2012
- 18 Feng Y J, Blackburn D C, Liang D, et al. Phylogenomics reveals rapid, simultaneous diversification of three major clades of Gondwanan frogs at the Cretaceous-Paleogene boundary. *Proc Natl Acad Sci USA*, 2017, 114: E5864–E5870
- 19 Zhang S Q, Che L H, Li Y, et al. Evolutionary history of Coleoptera revealed by extensive sampling of genes and species. *Nat Commun*, 2018, 9: 205
- 20 Torruella G, de Mendoza A, Grau-Bové X, et al. Phylogenomics reveals convergent evolution of lifestyles in close relatives of animals and fungi. *Curr Biol*, 2015, 25: 2404–2410
- 21 Rujirawat T, Patumcharoenpol P, Lohnoo T, et al. Probing the phylogenomics and putative pathogenicity genes of *Pythium insidiosum* by Oomycete genome analyses. *Sci Rep*, 2018, 8: 1–14
- 22 Brown M W, Heiss A A, Kamikawa R, et al. Phylogenomics places orphan protistan lineages in a novel eukaryotic super-group. *Genome Biol Evol*, 2018, 10: 427–433
- 23 Fernández R, Kallal R J, Dimitrov D, et al. Phylogenomics, diversification dynamics, and comparative transcriptomics across the spider tree of life. *Curr Biol*, 2018, 28: 1489–1497.e5
- 24 Regier J C, Shultz J W, Zwick A, et al. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature*, 2010, 463: 1079–1083
- 25 Shen X X, Liang D, Feng Y J, et al. A versatile and highly efficient toolkit including 102 nuclear markers for vertebrate phylogenomics, tested by resolving the higher level relationships of the Caudata. *Mol Biol Evol*, 2013, 30: 2235–2248
- 26 Li J N, He C, Guo P, et al. A workflow of massive identification and application of intron markers using snakes as a model. *Ecol Evol*, 2017, 7: 10042–10055
- 27 Che L H, Zhang S Q, Li Y, et al. Genome-wide survey of nuclear protein-coding markers for beetle phylogenetics and their application in resolving both deep and shallow-level divergences. *Mol Ecol Resour*, 2017, 17: 1342–1358
- 28 O’Neill E M, Schwartz R, Bullock C T, et al. Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the

- North American tiger salamander (*Ambystoma tigrinum*) species complex. *Mol Ecol*, 2013, 22: 111–129
- 29 Bybee S M, Bracken-Grissom H, Haynes B D, et al. Targeted amplicon sequencing (TAS): A scalable next-gen approach to multilocus, multitaxa phylogenetics. *Genome Biol Evol*, 2011, 3: 1312–1323
 - 30 Wielstra B, Duijij E, Lagler P, et al. Parallel tagged amplicon sequencing of transcriptome-based genetic markers for *Triturus* newts with the Ion Torrent next-generation sequencing platform. *Mol Ecol Resour*, 2014, 160: 1080–1089
 - 31 Barrow L N, Ralicki H F, Emme S A, et al. Species tree estimation of North American chorus frogs (Hylidae: *Pseudacris*) with parallel tagged amplicon sequencing. *Mol Phylogenets Evol*, 2014, 75: 78–90
 - 32 Zieliński P, Stuglik M T, Dudek K, et al. Development, validation and high-throughput analysis of sequence markers in nonmodel species. *Mol Ecol Resour*, 2014, 14: 352–360
 - 33 Feng Y J, Liu Q F, Chen M Y, et al. Parallel tagged amplicon sequencing of relatively long PCR products using the Illumina HiSeq platform and transcriptome assembly. *Mol Ecol Resour*, 2016, 16: 91–102
 - 34 Grabherr M G, Haas B J, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotech*, 2011, 29: 644–652
 - 35 Bankevich A, Nurk S, Antipov D, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*, 2012, 19: 455–477
 - 36 Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet*, 2009, 10: 57–63
 - 37 Simpson J T, Wong K, Jackman S D, et al. ABySS: A parallel assembler for short read sequence data. *Genome Res*, 2009, 19: 1117–1123
 - 38 Oakley T H, Wolfe J M, Lindgren A R, et al. Phylotranscriptomics to bring the understudied into the fold: Monophyletic Ostracoda, fossil placement, and pancrustacean phylogeny. *Mol Biol Evol*, 2013, 30: 215–233
 - 39 Wickett N J, Mirarab S, Nguyen N, et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc Natl Acad Sci USA*, 2014, 111: E4859–E4868
 - 40 Unruh S A, McKain M R, Lee Y I, et al. Phylotranscriptomic analysis and genome evolution of the Cypridioideae (Orchidaceae). *Am J Bot*, 2018, 105: 631–640
 - 41 Irisarri I, Baurain D, Brinkmann H, et al. Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nat Ecol Evol*, 2017, 1: 1370–1378
 - 42 Misof B, Liu S, Meusemann K, et al. Phylogenomics resolves the timing and pattern of insect evolution. *Science*, 2014, 346: 763–767
 - 43 Baird N A, Etter P D, Atwood T S, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, 2008, 3: e3376
 - 44 Miller M R, Dunham J P, Amores A, et al. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res*, 2007, 17: 240–248
 - 45 van Orsouw N J, Hogers R C J, Janssen A, et al. Complexity reduction of polymorphic sequences (CRoPSTTM): A novel approach for large-scale polymorphism discovery in complex genomes. *PLoS ONE*, 2007, 2: e1172
 - 46 Van Tassell C P, Smith T P L, Matukumalli L K, et al. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Methods*, 2008, 5: 247–252
 - 47 Elshire R J, Glaubitz J C, Sun Q, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, 2011, 6: e19379
 - 48 Meier J I, Marques D A, Mwaiko S, et al. Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nat Commun*, 2017, 8: 14363
 - 49 Puritz J B, Matz M V, Toonen R J, et al. Demystifying the RAD fad. *Mol Ecol*, 2015, 23: 5937–5942
 - 50 Andrews K R, Good J M, Miller M R, et al. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat Rev Genet*, 2016, 17: 81–92
 - 51 Gautier M, Gharbi K, Cezard T, et al. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Mol Ecol*, 2013, 22: 3165–3178
 - 52 Davey J W, Cezard T, Fuentes-Utrilla P, et al. Special features of RAD sequencing data: Implications for genotyping. *Mol Ecol*, 2013, 22: 3151–3164
 - 53 Wang S, Meyer E, McKay J K, et al. 2b-RAD: A simple and flexible method for genome-wide genotyping. *Nat Methods*, 2012, 9: 808–810
 - 54 Guo Y, Yuan H, Fang D, et al. An improved 2b-RAD approach (I2b-RAD) offering genotyping tested by a rice (*Oryza sativa* L.) F₂ population. *BMC Genom*, 2014, 15: 956
 - 55 Peterson B K, Weber J N, Kay E H, et al. Double digest RADseq: An inexpensive method for *de novo* SNP discovery and genotyping in model

- and non-model species. *PLoS ONE*, 2012, 7: e37135
- 56 Toonen R J, Puritz J B, Forsman Z H, et al. ezRAD: A simplified method for genomic genotyping in non-model organisms. *PeerJ*, 2013, 1: e203
- 57 Ali O A, O'Rourke S M, Amish S J, et al. RAD capture (Rapture): Flexible and efficient sequence-based genotyping. *Genetics*, 2016, 202: 389–400
- 58 Gnirke A, Melnikov A, Maguire J, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotech*, 2009, 27: 182–189
- 59 Burbano H A, Hodges E, Green R E, et al. Targeted investigation of the neandertal genome by array-based sequence capture. *Science*, 2010, 328: 723–725
- 60 Briggs A W, Good J M, Green R E, et al. Primer extension capture: Targeted sequence retrieval from heavily degraded DNA sources. *J Vis Exp*, 2009, 3: 1573
- 61 Lemmon A R, Emme S A, Lemmon E M. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst Biol*, 2012, 61: 727–744
- 62 Faircloth B C, McCormack J E, Crawford N G, et al. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst Biol*, 2012, 61: 717–726
- 63 George R D, McVicker G, Diederich R, et al. Trans genomic capture and sequencing of primate exomes reveals new targets of positive selection. *Genome Res*, 2011, 21: 1686–1694
- 64 Bi K, Vanderpool D, Singhal S, et al. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genom*, 2012, 13: 403
- 65 Li C, Hofreiter M, Straube N, et al. Capturing protein-coding genes across highly divergent species. *Biotechniques*, 2013, 54: 321–326
- 66 Maricic T, Whitten M, Pääbo S. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS ONE*, 2010, 5: e14004
- 67 Mariac C, Searcelli N, Pouzadou J, et al. Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. *Mol Ecol Resour*, 2014, 14: 1103–1113
- 68 Peñalba J V, Smith L L, Tonione M A, et al. Sequence capture using PCR-generated probes: A cost-effective method of targeted high-throughput sequencing for nonmodel organisms. *Mol Ecol Resour*, 2014, 496: 1000–1010
- 69 McCormack J E, Tsai W L E, Faircloth B C. Sequence capture of ultraconserved elements from bird museum specimens. *Mol Ecol Resour*, 2016, 16: 1189–1203
- 70 Blaimer B B, Lloyd M W, Guillery W X, et al. Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens. *PLoS ONE*, 2016, 11: e0161531
- 71 Dornburg A, Townsend J P, Brooks W, et al. New insights on the sister lineage of percomorph fishes with an anchored hybrid enrichment dataset. *Mol Phylogenets Evol*, 2017, 110: 27–38
- 72 Alfaro M E, Faircloth B C, Harrington R C, et al. Explosive diversification of marine fishes at the Cretaceous-Palaeogene boundary. *Nat Ecol Evol*, 2018, 2: 688–696
- 73 Espeland M, Breinholt J, Willmott K R, et al. A comprehensive and dated phylogenomic analysis of butterflies. *Curr Biol*, 2018, 28: 770–778.e5
- 74 Prum R O, Berv J S, Dornburg A, et al. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature*, 2015, 526: 569–573
- 75 Jones M R, Good J M. Targeted capture in evolutionary and ecological genomics. *Mol Ecol*, 2016, 25: 185–202
- 76 Allen J M, Boyd B, Nguyen N P, et al. Phylogenomics from whole genome sequences using aTRAM. *Syst Biol*, 2017, 105: 786–798
- 77 Allen J M, Huang D I, Cronk Q C, et al. aTRAM-automated target restricted assembly method: A fast method for assembling loci across divergent taxa from next-generation sequencing data. *BMC Bioinf*, 2015, 16: 98
- 78 Boyd B M, Allen J M, Nguyen N P, et al. Phylogenomics using target-restricted assembly resolves intra-generic relationships of parasitic lice (Phthiraptera: *Columbicola*). *Syst Biol*, 2017, 66: 896–911
- 79 Bolger A M, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics*, 2014, 30: 2114–2120
- 80 Langmead B, Salzberg S L. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 2012, 9: 357–359
- 81 Perteau M, Perteau G M, Antonescu C M, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotech*, 2015, 33: 290–295

- 82 Trapnell C, Williams B A, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech*, 2010, 28: 511–515
- 83 Xie Y, Wu G, Tang J, et al. SOAPdenovo-Trans: *De novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, 2014, 30: 1660–1666
- 84 Luo R, Liu B, Xie Y, et al. SOAPdenovo2: An empirically improved memory-efficient short-read *de novo* assembler. *Gigascience*, 2012, 1: 18
- 85 Zerbino D R, Birney E. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res*, 2008, 18: 821–829
- 86 Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: *De novo* assembly of organelle genomes from whole genome data. *Nucl Acids Res*, 2016, 53: 955
- 87 Hahn C, Bachmann L, Chevreux B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucl Acids Res*, 2013, 41: e129
- 88 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 2009, 25: 1754–1760
- 89 Chen F, Mackey A J, Stoeckert C J, et al. OrthoMCL-DB: Querying a comprehensive multi-species collection of ortholog groups. *Nucl Acids Res*, 2006, 34: D363–D368
- 90 Waterhouse R M, Tegenfeldt F, Li J, et al. OrthoDB: A hierarchical catalog of animal, fungal and bacterial orthologs. *Nucl Acids Res*, 2013, 41: D358–D365
- 91 Altenhoff A M, Schneider A, Gonnet G H, et al. OMA 2011: Orthology inference among 1000 complete genomes. *Nucl Acids Res*, 2011, 39: D289–D294
- 92 Sonnhammer E L L, Östlund G. InParanoid 8: Orthology analysis between 273 proteomes, mostly eukaryotic. *Nucl Acids Res*, 2015, 43: D234–D239
- 93 Masterson J. Stomatal size in fossil plants: Evidence for polyploidy in majority of angiosperms. *Science*, 1994, 264: 421–424
- 94 Wendel J F. Genome evolution in polyploids. *Plant Mol Biol*, 2000, 42: 225–249
- 95 Katoh K, Standley D M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol*, 2013, 30: 772–780
- 96 Larkin M A, Blackshields G, Brown N P, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*, 2007, 23: 2947–2948
- 97 Edgar R C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucl Acids Res*, 2004, 32: 1792–1797
- 98 Liu K, Raghavan S, Nelesen S, et al. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science*, 2009, 324: 1561–1564
- 99 Mirarab S, Nguyen N, Guo S, et al. PASTA: Ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *J Comput Biol*, 2015, 22: 377–386
- 100 Simion P, Philippe H, Baurain D, et al. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Curr Biol*, 2017, 27: 958–967
- 101 Shen X X, Hittinger C T, Rokas A. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat Ecol Evol*, 2017, 1: 0126
- 102 Lanfear R, Calcott B, Ho S Y W, et al. PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol Biol Evol*, 2012, 29: 1695–1701
- 103 Lanfear R, Frandsen P B, Wright A M, et al. PartitionFinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol Biol Evol*, 2016, 34: 772–773
- 104 Posada D, Crandall K A. Modeltest: Testing the model of DNA substitution. *Bioinformatics*, 1998, 14: 817–818
- 105 Posada D. jModelTest: Phylogenetic model averaging. *Mol Biol Evol*, 2008, 25: 1253–1256
- 106 Abascal F, Zardoya R, Posada D. ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics*, 2005, 21: 2104–2105
- 107 Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 2014, 30: 1312–1313
- 108 Nguyen L T, Schmidt H A, von Haeseler A, et al. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*, 2015, 32: 268–274
- 109 Minh B Q, Nguyen M A T, von Haeseler A. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol*, 2013, 30: 1188–1195
- 110 Ronquist F, Teslenko M, van der Mark P, et al. MrBayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*, 2012, 61: 539–542

- 111 Lartillot N, Lepage T, Blanquart S. PhyloBayes 3: A Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, 2009, 25: 2286–2288
- 112 Aberer A J, Kobert K, Stamatakis A. ExaBayes: Massively parallel bayesian tree inference for the whole-genome era. *Mol Biol Evol*, 2014, 31: 2553–2556
- 113 Song S, Liu L, Edwards S V, et al. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc Natl Acad Sci USA*, 2012, 109: 14942–14947
- 114 Liu L, Yu L, Edwards S V. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol Biol*, 2010, 10: 302
- 115 Mirarab S, Reaz R, Bayzid M S, et al. ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics*, 2014, 30: i541–i548
- 116 Larget B R, Kotha S K, Dewey C N, et al. BUCKY: Gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics*, 2010, 26: 2910–2911
- 117 Ane C, Larget B, Baum D A, et al. Bayesian estimation of concordance among gene trees. *Mol Biol Evol*, 2007, 24: 1575
- 118 Heled J, Drummond A J. Bayesian inference of species trees from multilocus data. *Mol Biol Evol*, 2010, 27: 570–580
- 119 Bao J, Xia H, Zhou J, et al. Efficient implementation of MrBayes on multi-GPU. *Mol Biol Evol*, 2013, 30: 1471–1479
- 120 Baptiste E, Brinkmann H, Lee J A, et al. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. *Proc Natl Acad Sci USA*, 2002, 99: 1414–1419
- 121 Madsen O, Scally M, Douady C J, et al. Parallel adaptive radiations in two major clades of placental mammals. *Nature*, 2001, 409: 610–614
- 122 Rokas A, Williams B L, King N, et al. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 2003, 425: 798–804
- 123 Lockhart P J, Steel M A, Hendy M D, et al. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol Biol Evol*, 1994, 11: 605–612
- 124 Liu Q, Feng Y, Xue Q. Analysis of factors shaping codon usage in the mitochondrion genome of. *Mitochondrion*, 2004, 4: 313–320
- 125 Behura S K, Severson D W. Codon usage bias: Causative factors, quantification methods and genome-wide patterns: With emphasis on insect genomes. *Biol Rev*, 2013, 88: 49–61
- 126 Plotkin J B, Kudla G. Synonymous but not the same: The causes and consequences of codon bias. *Nat Rev Genet*, 2011, 12: 32–42
- 127 Andreassen K, Baldwin B G. Unequal evolutionary rates between annual and perennial lineages of checker mallows (*Sidalcea*, Malvaceae): Evidence from 18S–26S rDNA internal and external transcribed spacers. *Mol Biol Evol*, 2001, 18: 936–944
- 128 Philippe H, Zhou Y, Brinkmann H, et al. Heterotachy and long-branch attraction in phylogenetics. *BMC Evol Biol*, 2005, 5: 50
- 129 Wiens J J. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? *Syst Biol*, 2005, 54: 731–742
- 130 Lartillot N, Philippe H. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*, 2004, 21: 1095–1109
- 131 Kolaczkowski B, Thornton J W. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, 2004, 431: 980–984
- 132 Philippe H, Germot A. Phylogeny of eukaryotes based on ribosomal RNA: Long-branch attraction and models of sequence evolution. *Mol Biol Evol*, 2000, 17: 830–834
- 133 Philippe H, Brinkmann H, Lavrov D V, et al. Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biol*, 2011, 9: e1000602
- 134 Salichos L, Rokas A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, 2013, 497: 327–331
- 135 Chen M Y, Liang D, Zhang P. Selecting question-specific genes to reduce incongruence in phylogenomics: A case study of jawed vertebrate backbone phylogeny. *Syst Biol*, 2015, 64: 1104–1120
- 136 Wang Y H, Wu H Y, Rédei D, et al. When did the ancestor of true bugs become stinky? Disentangling the phylogenomics of Hemiptera-Heteroptera. *Cladistics*, 2019, 35: 42–66
- 137 i5K Consortium. The i5K Initiative: Advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J Hered*, 2013, 104: 595–600

Recent progress in phylogenomic methods

LI JiaXuan, LIANG Dan & ZHANG Peng

School of Life Sciences, Sun Yat-Sen University, Guangzhou 510006, China

Phylogenomics is the intersection of the fields of phylogenetics and genomics, aiming to use genome-scale molecular data to resolve the phylogenetic relationships among organisms. In this review, we briefly introduced five methods of data collection in phylogenomics by high-throughput sequencing: targeted amplicon sequencing, transcriptome sequencing, reduced-genome-representation sequencing, target sequence capture, and low coverage whole-genome shotgun sequencing. We summarized and compared the characteristics, applications and costs of each method. In addition, we reviewed the current work flow of data processing of a typical phylogenomic analysis, including five steps: data preprocessing and assembly, orthologous sequences identification, sequence alignment, data quality control and phylogenetic tree reconstruction. We also summarized the key points of each step and introduced software that are used in these steps. Phylogenomics is a promising way to reconstruct the Tree of Life. However, it also faces a series of problems at present. For example, the analysis of genome-scale data will lead to a significant increase in computation time, the impact of systematic error will become more serious when the data size increases and different data sets may yield strongly supported but inconsistent results. In the end of this review, we summarized and discussed the latest studies concerning these issues and made some suggestions to deal with these problems in the future.

high-throughput sequencing, phylogenomics, phylogeny estimation, systematic error, phylogenetic signal

doi: [10.1360/N052018-00222](https://engine.scichina.com/doi/10.1360/N052018-00222)