基于转录组数据和短序列测序的系统发育研究方法

Phylogenomics based on transcriptomics and short-read sequences

蒙冠良¹, 周程冉², 刘山林³, 周欣^{3,*}

¹ Centre of Taxonomy and Evolutionary Research, Zoological Research Museum Alexander Koenig, D-53113, Bonn, Germany; ² 深圳华大生命科学研究院, 深圳 518083; ³ 昆虫学系, 中国农业大学, 北京 100193

*通讯作者邮箱: xinzhou@cau.edu.cn

引用格式: 蒙冠良, 周程冉, 刘山林, 周欣. (2021). 基于转录组数据和短序列测序的系统发育研究方法. *Bio-101* e1010649. Doi: 10.21769/BioProtoc.1010649.

How to cite: Meng, G. L., Zhou, C. R., Liu, S. L. and Zhou, X. (2021). Phylogenomics based on transcriptomics and short-read sequences. *Bio-101* e1010649. Doi: 10.21769/BioProtoc.1010649. (in Chinese)

摘要: 明确的系统发育关系 (phylogeny) 是分类学、进化事件的推断及时间校正、相关性状的祖先状态构建等的重要基础和前提。在明确了不同支系的系统发育关系的前提下, 通过分析比较性状在不同类群中的演化历史, 可以了解物种经受了哪些选择压力, 不同性状是否倾向于共同演化或只是对相同选择压力的适应。自高通量短序列测序技术出现以后, 通过广泛采样及使用大规模数据集来研究不同类群的系统发育关系的方法已经成为共识。但由于大规模全基因组测序成本高、分析难度大, 简化基因组测序仍具有广泛应用前景。其中, 利用转录组测序方法来获得大量适用于系统发育分析的直系同源基因具有可观的成本优势和较高的可行性。本文简要概括了基于转录组方法进行系统发育研究的主要步骤, 包括样本制备、转录组测序与组装、直系同源基因筛选与评估和系统发育分析等, 希望能为相关学者提供一定的参考。

关键词: 转录组, 系统发育, 进化, 组装, 短序列测序

研究背景

理解物种之间的系统发育关系是构建生物演化框架的关键, 也是进一步开展生物性状演化、生物地理变迁、分子遗传机制等分支学科的重要前提。漫长的生命演化历史以及多次出现的大规模生命灭绝和爆发事件 (Hoyal Cuthill *et al.*, 2020), 对重建符合/接近真

实演化历史的系统发育关系造成了重大挑战 (Misof *et al.*, 2014; Wickett *et al.*, 2014; Tank *et al.*, 2015)。近年来, 使用基因组级别的基因集来重建系统发育关系已成为演化生物学领域的重要手段 (例如 Amemiya *et al.*, 2013; Zhang *et al.*, 2014; Chaw *et al.*, 2019; Li *et al.*, 2021)。虽然基因组序列包含了所有具有系统发育信号的基因, 但受限于基因组测序、组装、注释等在技术或资源等方面的困难, 基因组序列在分子系统发育研究中的使用仍非常有限。在开展大规模多物种的演化研究中, 不同物种基因组之间存在显著的异质性, 包括基因组大小、杂合度、基因组加倍事件等方面的差异, 都会对从基因组中获取直系同源基因 (orthologous gene; 即不同物种中来自于同一祖先的属于同一个基因家族的基因, 而同一物种中由于基因复制等产生的同一家族的基因, 则称为旁系同源基因 paralogous genes) 造成很大的影响。相对而言, 不同生物类群在编码基因的总数上较为恒定, 而转录组序列中包含了绝大部分适用于系统演化分析的直系同源基因。因此, 与全基因组测序相比, 基于转录组测序的方法具有可观的成本优势和较高的可行性。基于转录组方法的基因组系统发育 (phylogenomics) 研究在近年来取得了突飞猛进的发展, 在昆虫 (例如 Misof *et al.*, 2014)、植物 (例如 Leebens-Mack *et al.*, 2019) 等诸多类群得以广泛应用, 对高级阶元的系统发育关系构建做出了重要的贡献。本文简要概括了基于转录组进行系统发育研究的主要步骤, 包括: 样本制备、转录组组装、直系同源基因筛选、多序列比对与评估和系统发育分析等 (图 1)。

图 1. 基于短序列测序和转录组数据进行系统发育分析的工作流程

一、样品制备和数据获取

1. 样品收集与保存

核糖核酸 (Ribonucleic acid, RNA) 是单链分子, 具有容易降解的生物特性。为了获得高质量 RNA 样品, 需要采集新鲜的组织样本以保证 RNA 分子的完整性。在野外或者实验室获取新鲜的生物样品后, 通常可以采用以下两种方式进行保存: (1) 液氮保存法: 将生物样品快速置于液氮 (常压下约为 -196°C) 中; (2) 存储液保存法: 将样品放入 RNA 保存液中 (如 RNA-later 试剂等) 进行保存, 可以有效减缓 RNA 酶的降解作用。RNA 保存液可以在常温下存放, 并且可以通过分装的形式提前放置于不同规格的离心管 (50 ml, 10 ml 等) 中, 比液氮更为便携, 因此更适合在野外采集工作中使用。在利用 RNA 保存液进行生物样品保存需注意以下事项: (1) 需要快速地将组织样本切割为一定大小的组织块 (或利用杵臼碾碎), 使保存液尽可能充分浸润组织; (2) 试剂与组织体积比例保持在 5:1 左右 (或根据实际储存液相关要求配置); (3) 短期野外采集可在室温操作, 长期保存应尽量置于 $\leq 4^{\circ}\text{C}$ 的环境中

以延长 RNA 寿命。实践操作证明，经 RNA-later 保存液常温保存和邮寄的样品，经 1 个月左右仍可提供高质量的 RNA 样品供高通量测序建库使用。

对于昆虫样品可把腹部剔除，以减少测序数据中来源于肠道微生物的信息。为了减少个体间基因组异质性对后续转录组组装造成的干扰，建议在满足 RNA 文库构建起始量的前提下，尽量利用单个体完成 RNA 的提取、文库构建和转录组数据的产生。最后，需要强调的是，在样本操作、运输过程中需要严格防止样品发生交叉污染，如：需要避免使用受污染的工具来对样本进行操作，可以使用封口膜对收集了样品的器具进行封口以防止保存液泄露等。

2. RNA 提取

针对真核生物已经有较为成熟的 RNA 提取对方法和相应的商业试剂盒产品。原则上，根据试剂厂商的标准操作流程进行即可。例如，利用 TRIzol 试剂盒 (Invitrogen) 进行总 RNA 提取的主要步骤为：匀浆处理、分层处理、沉淀、洗涤和溶解保存。其次，在建库之前需要检测 RNA 的纯度 (28S/18S)、浓度、完整性 (RNA Integrity Number, RIN) 和 RNA 总量是否满足建库要求，相关检测可以使用 Qubit、Agilent 2100 Bioanalyzer 或者 NanoDrop 等仪器完成。

3. 文库构建与测序

目前常用的 NGS (Next-Generation Sequencing) 测序平台 (或"短序列测序平台") 包括 Illumina HiSeq、MGI DNBSEQ 等。不同平台文库构建及测序流程会有一些的差异，需要根据产品说明书进行实际操作。动物转录组 RNA 的建库流程一般包括：(1) RNA 纯化：如利用 Oligo (dT) 磁珠对全转录组中的 mRNA 进行富集，减少其它类型的 RNA (如 lncRNA、rRNA 等) 的含量；(2) 片段化：使用片段化缓冲液将纯化后的转录组 RNA 消化为目的片段大小，如 250 bp；(3) 修复及扩增：利用目的片段为模板进行 cDNA 双链的合成，修复末端，添加接头与标签并扩增，从而得到插入片段长度为指定长度的测序文库 (见图 1 I. 样品制备和数据获取)。

每个样本的测序数据量依基因组大小和所使用的测序平台而异。对于 NGS 测序平台，小型基因组 (例如 500 Mb) 的样本建议每个样本测 3 Gb 数据量以上，较大的基因组 (例如鱼类) 建议每个样本测 8 Gb 数据量以上。测序策略可以采用双末端 150 bp (Paired-end 150 bp, 150 PE) 等。测序数据量越大，越有利于低丰度转录本的获取及组装。

二、转录组数据分析

1. 下机数据质控

NGS 测序平台的测序下机数据一般保存在 Fastq 格式 (Cock *et al.*, 2010) 的文件中, 称为 raw data。Raw data 中往往存在一些低质量的序列, 如包含测序接头序列的 reads、包含过多模糊碱基 ("Ns")、或者过多低质量的碱基等 (图 1 II. 转录组数据分析)。接头序列或者测序错误一方面会增加计算资源的消耗, 另一方面会导致组装错误或者 reads 无法正确比对到参考基因组上。因此, 有必要对 raw data 进行质控和过滤, 相关标准可参考 Misof *et al.*, (2014): 1) 去除含接头的序列; 2) 去除质量值在 Q20 以下的碱基占比超过 20% 的数据; 3) 去除 N 含量较高 (> 5%) 的数据。常见的用于完成相关分析的软件有: SOAPnuke (Chen *et al.*, 2018; <https://github.com/BGI-flexlab/SOAPnuke>)、FastQC (Andrews, 2010; <https://github.com/s-andrews/FastQC>)、Cutadapt (Martin, 2011; <https://cutadapt.readthedocs.io/en/stable/>)、NGS QC Toolkit (Patel and Jain, 2012)、Trimmomatic (Bolger *et al.*, 2014; <http://www.nipgr.ac.in/ngsqctoolkit.html>) 和 Skewer (Jiang *et al.*, 2014) 等。

2. 转录组组装

短序列测序平台的读长一般只有 150-300 bp。因此, 需要对测序得到的短片段序列进行拼接组装才能获得全长的转录本序列。主流的组装策略主要分为: 有参组装 (reference-based assembly)、从头组装 (*de novo* assembly) 及混合组装 (Martin and Wang, 2011; Li *et al.*, 2014; Huang *et al.*, 2016; Zhou *et al.*, 2020)。

2.1 有参组装

有参组装指依赖于目标物种或近缘物种的参考基因组的转录组组装策略, 其主要步骤包括: (1) 利用 TOPHAT2 (Kim *et al.*, 2013; <http://ccb.jhu.edu/software/tophat/manual.shtml>)、STAR (Dobin *et al.*, 2013; <https://github.com/alexdobin/STAR>)、HISAT2 (Kim *et al.*, 2019; <http://daehwankimlab.github.io/hisat2/>) 等软件把原始序列比对到参考基因组上。相比于其它基因组比对软件, 上述软件的主要特点是在比对过程中能够考虑不同的可变剪接形式。同时, 该步骤还能有效降低样品污染和测序错误对组装结

果的影响。(2) 使用 Cufflinks (Trapnell *et al.*, 2013)、StringTie (Pertea *et al.*, 2015) 等软件, 先将比对上同一个基因座 (locus) 的相关 reads 构建"图", 由所有可能形式的可变剪接体 (isoforms; 由于真核生物等生物中具有选择性剪接机制, 同一基因产生的 mRNA 前体会由于不同剪接形式进一步产生不同的 mRNA, 即 mRNA 剪接异构体) 所组成的; 然后通过"解图", 即穿越不同的路径, 从而获得各个独立的可变剪接。

有参组装对低表达的转录本具有更强的敏感性, 有助于发现新的转录本 (即该转录本不存在于参考基因组的注释结果中), 并且还具有较低的计算资源消耗, 降低了运行成本。但是其组装结果的质量水平依赖于参考基因组的质量以及参考物种与目标物种的亲缘关系, 因此不适用于参考基因组质量差 (例如基因组 N50 只有几 K bp 长, BUSCO (Manni *et al.* 2021; <https://busco.ezlab.org/>) 基因完整性小于 70%) 或参考物种过于远缘 (例如不同的目) 的情况。可以预见, 随着一系列大型基因组计划 (如 5000 种节肢动物基因组计划 (The 5000 arthropod genomes initiative, i5K; <http://i5k.github.io/>) (i5K Consortium, 2013)、万种鸟基因组计划 (The Bird 10,000 Genomes Project, B10K; <https://b10k.genomics.cn/>) (Zhang, 2015)、地球生物基因组计划 (The Earth BioGenome Project, EBP) (Lewin *et al.*, 2018)、脊椎动物基因组计划 (The Vertebrate Genomes Project, VGP; <https://vertebrategenomesproject.org/>) (Genome 10K Community of Scientists 2009; Rhie *et al.*, 2021) 的开展, 有参组装的分析策略将适用于越来越多的研究。

2.2 从头组装

在目标物种或其近缘物种的参考基因组缺失, 或者参考基因组的质量较差时 (例如相关基因缺失或不完整), 可以使用从头组装的分析方案。

从头组装常用的算法以 *de Bruijn* Graph (DBG) 理论为依据, 把测序序列拆分成特定长度的短片段 (称为 kmer) 来构建 DBG 图, 然后通过"解图"来获得基因序列以及其不同的转录本。DBG 也被广泛应用于基因组组装软件中, 但是基因组组装软件假设基因组的不同位置具有基本一致的测序深度。而转录本组装软件必须考虑到不同基因、甚至同一个基因的不同可变剪接体的表达量差异。广泛使用的转录本组装软件有 Trinity (Grabherr *et al.*, 2011)、Oases (Schulz *et*

al., 2012)、IDBA-Tran (Peng *et al.*, 2013)、SOAPdenovo-trans (Xie *et al.*, 2014) 等。**Kmer** 的大小决定了不同节点之间的重叠长度, 该重叠长度对组装结果的质量有重要影响。一般情况下, 推荐使用较大的 **kmer**, 从而获得高质量的转录本组装结果。但过大的 **kmer** 不利于低表达量转录本的组装。较小的 **kmer** 可以更有效地获取低表达量的转录本, 但同时会导致组装结果碎片化。因此, 一些研究 (例如 De Vries *et al.*, 2017; Robertson *et al.*, 2017) 也尝试使用多个 **kmer** 大小进行转录本组装, 最后把不同的结果进行整合, 相关工具有 Trans-ABYSS (Robertson *et al.*, 2010)、Multiple-k (Surget-Groba and Montoya-Burgos, 2010) 和 Rnnotator (Martin *et al.*, 2010) 等。在 1KITE 昆虫转录组演化 (The 1K Insect Transcriptome Evolution Project; <https://1kite.org/>) 研究中 (Misof *et al.*, 2014), 首先利用 SOAPdenovo-trans (Xie *et al.*, 2014) 进行转录组组装, 组装时参数 **kmer** 设置为 31, **-e** 参数设置为 3; 再利用补洞软件 GapCloser (Luo *et al.*, 2012) 使用默认参数进行补洞。

2.3 混合组装

在已有目标物种或其近缘物种的中等质量参考基因组 (例如基因组 N50 几百 K bp 左右, BUSCO (Manni *et al.*, 2021) 基因完整性 80% 左右) 的前提下, 为了兼顾组装结果完整性和计算资源消耗, 可以考虑使用混合组装的策略。混合组装指同时采用有参组装和从头组装进行转录本组装, 具体来说有两种策略: 先比对再组装和先组装再比对 (Martin and Wang, 2011)。先比对再组装, 指的是先把所有测序 **reads** 比对到参考基因组上 (即有参组装), 然后把未比对上的 **reads** 进行从头组装, 这种方法的优点是只需要较少的计算资源。而如果参考基因组的质量比较差 (比如组装错误、基因缺失或不完整), 那么可以考虑使用先组装再比对的策略, 一方面可以通过从头组装来检测新的转录本及新的可变剪接体, 另一方面充分利用了有参组装的高敏感性的优势, 从而可以获得更长、更完整的转录本 (Martin and Wang, 2011)。

3. 组装结果质控

对组装结果进行质控非常重要。该步骤主要检测样本是否被外源物种污染以及样本之间是否存在交叉污染。外源物种污染可能的来源包括: (1) 该样本自身的肠道内含物 (例如微生物、食物残渣等), 或者共生、寄生生物; (2) 实验过程中其它样本所导

致的污染，例如同期进行的其它项目样本。同一个项目的不同样本之间也可能产生交叉污染，原因包括：(1) 样本采集、保存过程中共用未经彻底清洗干净的工具；(2) 样本运输过程中发生样本管发生破裂、保存液泄露；(3) 实验过程中导致的污染，例如同时进行 RNA 提取、建库和测序等。此外，混合测序（尤其是不同样品混合在同一个 lane 测序）过程中也可能产生标签跳跃（index cross-talk）(Wright and Vetsigian, 2016)，导致序列被拆分（demultiplexing）到错误的样品。

检测污染的手段包括：(1) 利用线粒体或者叶绿体等来源的标记基因对样本进行鉴定，确认分类结果是否与目标物种一致；(2) 与 NT 库 (<https://www.ncbi.nlm.nih.gov/nucleotide/>) 等数据库进行比对，分析比对上的物种类群组成；(3) 样本之间（尤其是同一个 lane 测序的样本）进行相互比对，分析是否存在高度相似的序列。

此外，还可以使用 VecScreen 软件 (<http://www.ncbi.nlm.nih.gov/tools/vecscreen/>) 和 UniVec 数据库 (<https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/>) 鉴定载体 (vector)、linker 或者接头 (adapter) 序列等污染。需要注意的是，NGS 实验过程中通常不涉及克隆载体。因此，检测到的载体污染有可能是假阳性结果，或者是外源基因（例如病毒或者细菌的基因）整合到了基因组导致。后者对进化基因组分析具有重要的意义，但也会干扰后续的直系同源基因的查找。所以，如果一条序列的末端中度或者高度匹配到载体序列、linker DNA，我们建议删除该序列。而如果一条序列的中间部分匹配到载体序列或者 linker DNA，建议将匹配部分从组装结果中删除，并把该序列在中间打断成两个子片段。另外，在把组装结果上传 NCBI Transcriptome Shotgun Assembly (TSA) 数据库 (<https://www.ncbi.nlm.nih.gov/genbank/tsa/>) 过程中，TSA 数据库会进一步检查是否存在污染序列污染，我们建议把相应检测到的序列全部删除。

对组装结果过滤后，可以对组装数据进一步过滤与统计，如：删除长度低于 200 bp 的序列后进行组装结果的统计。主要统计指标包括：总长度、组装序列 (scaffold) 数量、最长序列长度、GC 含量、N50 长度与对应的序列数量；以上指标有助于了解样品转录组的数据基本情况，如在 1KITE 数据中昆虫转录组组装数据的 N50 大多在 1,000 bp 左右。

三、系统发育分析

1. 直系同源基因筛选

基因转录后普遍存在可变剪切等转录调控过程。因此，转录组测序和组装得到的结果中会包含基因的多种转录本，需要从中筛选出适用于系统关系重建的基因序列，特别是单拷贝直系同源基因 (single copy orthologs)，进行后续系统发育分析 (见图 1 III. 系统发育分析)。千种昆虫转录组演化项目 (Misof *et al.*, 2014) 采用了 HaMStRad 软件 (Misof *et al.*, 2014; <https://github.com/mptksen/HaMStRad>)，用于从转录组组装结果中筛选单拷贝直系同源基因，该软件基于 HaMStR v.8 (Ebersberger *et al.*, 2009) 进行了开发优化。其主要流程包括：(1) 首先从 OrthoDB (Kriventseva *et al.*, 2007) 数据库中下载特定类群的单拷贝直系同源基因的蛋白质序列，以及所对应的各个物种 (称为"参考物种" (reference species)) 的基因集 (蛋白质序列)；(2) 利用 MAFFT (Katoh *et al.*, 2002; Katoh and Toh, 2008) 软件对这些蛋白质序列进行多序列比对；(3) 基于步骤 2 中的比对结果，利用 HMMER 软件 (Eddy, 2011) 构建隐马尔可夫模型 (profile Hidden Markov models, pHMMs)；(4) 利用所构建的 pHMMs 模型和 HMMER 软件 (Eddy, 2011) 对转录组组装结果进行搜索，得到待选的转录本序列；(5) 接着把所有待选序列作为问询序列 ("query") 利用 BLAST 软件 (Camacho *et al.*, 2009) 与步骤 1 中参考物种的基因集 (即 "database") 进行比对，找出双向最优的比对结果 (reciprocal best hits)，即要求最佳比对上的至少一个参考物种的基因必须参与构建了该 query 序列所对应的 pHMM 模型。HaMStRad (Misof *et al.*, 2014; <https://github.com/mptksen/HaMStRad>) 也允许多个非重叠 (non-overlapping) 的转录本序列比对上同一个直系同源基因组 (Orthologous Group, OG)，并自动把它们连接起来；(6) 为了进一步考虑测序和组装错误导致的移码突变，使用 Exonerate (Slater and Birney, 2005) 软件以及单拷贝直系同源基因的蛋白质序列，对筛选出来的转录本序列进行开放阅读框预测 (open reading frames)；(7) 最后，如果同属于一个转录本的序列被分配到了多个不同 OGs，该待选转录本将被剔除。由于测序或组装的原因，样本中一些被预测出来的单拷贝直系同源基因仍然可能存在内部终止密码子，我们可以把它们替换为 "NNN"，并翻译成对应的未知氨基酸 "X"。类似功能的软件还有 OrthoGraph (Petersen *et al.*, 2017) 等。

2. 多序列比对

系统发育分析利用同源性状 (homologous characters; 指不同物种所拥有的同一个性状, 例如某个基因座或序列位点, 继承自这些物种的共同祖先。同源性状可进一步划分为祖征(apomorphy)和衍征(plesiomorphy), 见后文解释) 构建系统树。因此, 在得到单拷贝直系同源基因之后, 需要对每个基因分别进行多序列比对 (Multiple Sequence Alignment, MSA), 使得同源碱基位点相互对齐 (见图 1 III. 系统发育分析)。通常, 先将 CDS 序列翻译成蛋白序列, 然后对蛋白质序列进行多序列比对, 得到蛋白质的多序列比对结果 ("蛋白质-MSA")。由于密码子存在简并性, 即一个氨基酸可以由多个不同的密码子编码, 对于远缘物种, 蛋白质水平的序列对比 CDS 水平的序列比对具有更高的可靠性, 并且可以避免在比对过程中引入非真实的移码突变 (Suyama *et al.*, 2006)。常用的多序列比对工具包括 MAFFT (Nakamura *et al.*, 2018; <https://mafft.cbrc.jp/alignment/software/>)、ClustalW (Larkin *et al.*, 2007) 和 MUSCLE (Edgar, 2004) 等。接着, 利用 pal2nal.pl (Suyama *et al.*, 2006) 或 TrimAl (Capella-Gutiérrez *et al.*, 2009; <http://trimal.cgenomics.org/>) 软件, 依据"蛋白质-MSA"构建 CDS 水平的、基于密码子的多序列比对结果 ("CDS-MSA")。根据 CDS-MSA, 我们还可以提取由每个密码子第 1 和第 2 位碱基组成的集合, 以及四重简并位点 (four-folded sites, 4D 位点; 即密码子的某个位点上任何核苷酸都编码同样的氨基酸对应的位点) 组成的集合。此外, 一个蛋白编码基因往往编码了一个或多个蛋白质结构域 (protein domain), 而蛋白质结构域是蛋白质结构、功能和进化的基本单元, 因此在基于分区 (partition) 的系统发育分析方法中 (见后文"4. 进化模型的选择"章节), 编码蛋白质结构域的序列区间被认为比基因更适合作为一个进化单元 (Misof *et al.*, 2014)。利用 Pfam (<http://pfam.xfam.org/>) 和 pfam_scan.pl (Mistry *et al.*, 2021) 以及 HMMER (Eddy, 2011), 研究人员可以鉴定出相关蛋白质结构域并构建由蛋白质结构域组成蛋白质比对和 CDS 比对, 以用于系统发育分析。

在得到多序列比对结果之后, 还需要鉴定出可能的异常序列 (outlier), 然后将它们进行重新比对或者删除。异常序列的来源主要有四种 (Jehl *et al.*, 2015): (1) 该序列与其余序列不是同源序列, 其来源可能是直系同源基因筛选步骤存在缺陷; (2) 由于测序、转录本注释 (即对转录本的结构或功能进行注释) 或翻译错误, 导致该序列中的某些部分与其它序列不是同源关系; (3) 转录本不完整, 可能由实验处理、测

序数据量不足、转录本组装和注释不完善等因素导致；(4) 该基因在不同物种之间分化程度过大。其中，非同源序列来源的异常序列最容易检测，因为其空位 (gap) 和保守区间 (conserved blocks) 的模式与其它序列明显不同。对于第二种情况，其非同源序列部分的长度越长，越容易检测。尽管多序列比对流程已经成熟，当广泛存在长度差异大的序列或者序列分化程度较大的时候，比对结果会变得不可靠，因此我们需要尤其关注这个问题。Misof *et al.* (2014) 通过比较参考物种之间及样本-参考物种之间的 BLOSUM62 遗传距离 (Henikoff and Henikoff, 1992) 的大小关系，或者样本-参考物种的序列重叠长度的方式来识别异常序列，然后使用 MAFFT (Nakamura *et al.*, 2018) 软件对异常序列进行重新比对，接着按照同样的方法再次检测异常序列，如仍无法完成比对则将它们完全剔除。此外，常用的对 MSA 进行质控的工具还有 Gblocks (Talavera and Castresana, 2007) 和 TrimAl (Capella-Gutiérrez *et al.*, 2009) 等。

3. 组成异质性和缺失数据的评估

目前被广泛使用的各种进化模型大多数均假设分子进化符合组成同质性 (compositional homogeneity)，即：稳定性 (Stationary)、可逆性 (Reversible) 和均质性 (Homogeneous) (SRH) 的条件 (Ababneh *et al.*, 2006a and 2006b; Naser-Khdour *et al.*, 2019) (见图 1 III. 系统发育分析)。稳定性指的是 DNA 或者氨基酸的频率恒定，不随进化时间发生变化。可逆性指进化过程是稳定且没有特定方向的，即不同碱基或氨基酸之间的替换速率在正、反方向上是相等的 (Naser-Khdour *et al.*, 2019)。均质性是指分子瞬时替换速率恒定，且进化树的不同支系具有相同的速率。然而，已有研究表明不符合 SRH 条件的分子数据会严重影响系统发育树的拓扑结构，导致系统误差 (Naser-Khdour *et al.*, 2019)。一个解决方法是，在构建系统发育树之前，对多序列比对结果进行 SRH 条件检测，剔除不符合上述条件的序列和位点。更详细的方法步骤，可参考 Jermin *et al.* (2017) 和 Naser-Khdour *et al.* (2019)。此外，也可以使用一些对 SRH 条件要求相对宽松的进化模型和软件 (Foster, 2004; Lartillot and Philippe, 2004; Dutheil and Boussau, 2008; Zou *et al.*, 2012; Woodhams, *et al.*, 2015)。但相关软件通常需要消耗大量计算资源，因而应用范围受到一定限制。

此外，数据非随机缺失也可能会影响系统发育推断的结果 (Burleigh *et al.*, 2009; Cho *et al.*, 2011; Dell'Ampio *et al.*, 2014)。AliStat (Wong *et al.*, 2020) 可以通过基因比对结果 (Alignments) 进行两两比较，评估缺失数据的分布情况。把该信息与异常物种 (rogue species; 见后文) 信息结合，我们可以鉴定出一些受到缺失数据影响的样品。

4. 进化模型的选择

不同的基因或位点可能受到不同的选择压力。因此，在系统发育分析中，应充分考虑不同基因或位点在进化速率和进化模式方面的差异性 (Yang 1996; Luo *et al.*, 2010; Kumar *et al.*, 2012; Lanfear *et al.*, 2017)，例如蛋白编码基因的第 1、2、3 位密码子。研究者需要为所拥有的数据找到最佳的分区方案 (partition schemes)，并为每个分区找出最适合的进化模型 (图 1 III. 系统发育分析)。分区方案并不是分区越多越好，所使用的进化模型 (即核苷酸或氨基酸替代模型。一个进化模型通常包括 3 个要素即：不同碱基 (或氨基酸) 的频率，例如 A、T、G、C 每种碱基的频率均为 0.25；不同碱基 (或氨基酸) 之间相互替换的速率 (即替换矩阵)，如考虑转换和颠换的差异；相互替换速率在不同位点之间的分布情况，例如假设所有位点具有相同的替换速率，或存在一定比例的不变位点 (invariant sites)，或替换速率服从 Gamma 分布。不同进化模型之间的区别在于相关模型参数的值是使用固定的先验值还是将被从用户数据中进行估计以及估算方法的差别) 应该能够充分考虑分子数据的进化历程 (evolutionary process)，但同时应避免过度参数化 (overparameterization) (Lanfear *et al.*, 2012)。常用的工具有 PartitionFinder2 (Lanfear *et al.*, 2017) 及 ModelFinder (Kalyaanamoorthy *et al.*, 2017)。以 PartitionFinder2 (Lanfear *et al.*, 2017) 为例，主要分析流程包括：(1) 软件首先根据用户提供的数据构建一个系统发育树用于后续步骤；(2) 用户针对所有比对好的多序列比对结果 (MSAs) 自定义基本数据模块 (data blocks)，例如：把一个蛋白编码基因的密码子的第 2 位组成的位点作为一个基本数据模块，或者把一个内含子作为一个基本数据模块；(3) 依照用户指定的分区策略 (例如贪婪算法、聚类算法等) 形成不同的分区方案，即对基本数据模块单独进行分析或合并分析，形成不同的数据子集；(4) 基于步骤 1 的进化树结果，对每个数据子集进行计算，得出每个待选进化模型的对数似然值 (log likelihood)，并依据赤池信息准则 (Akaike Information

Criterion, AIC)、纠正后的赤池信息准则 (corrected AIC, AICc) 或贝叶斯信息准则 (Bayesian Information Criterion, BIC), 选出最适合于该数据子集的进化模型; (5) 针对每种分区方案包含的数据子集所对应的最佳模型, 将所得到的对数似然值进行累加, 得到各个分区方案的对数似然值; (6) 最后, 根据 AIC、AICc 或者 BIC 准则选出最优的分区方案。该分区方案将被采用来构建系统发育树。

5. 构建系统发育树

目前常用的分子系统发育树构建策略 (图 1 III. 系统发育分析) 有两种: (1) 串联 (concatenation) 基因构树。该方法首先将多个基因的比对结果首尾串联起来, 形成一个超长的序列比对结果 (concatenated alignment), 然后选取合适的进化模型进行建树。这种方法基于所有基因之间及基因与物种之间具有相同的演化历史的假设 (Shen *et al.*, 2021)。而导致该假设不成立的因素有隐性旁系同源基因 (hidden paralogy, 即一些基因由于其所属的基因家族的部分成员在进化过程中发生丢失或未被完整鉴定导致这些基因被错误地认为是直系同源基因, 而实际上它们属于旁系同源基因) (例如 Rasmussen and Kellis, 2012)、水平基因转移 (Horizontal Gene Transfer, HGT; 即基因信息在不同物种之间的横向转移 (相对于亲代-子代遗传信息的垂直传递而言)) (例如 Davidson *et al.*, 2015) 和不完全谱系筛选 (Incomplete Lineage Sorting, ILS; 即由于物种种化和祖先多态性的保留, 导致基因树与物种树有差异的情况) (例如 Scornavacca and Galtier, 2017) 等。一般串联建树获得的进化树枝长以单碱基替换数 (substitution per site) 表示。(2) 基于溯祖理论 (coalescence-based) 的构树方法, 或称为并联建树。该方法首先对每个基因分别比对 (alignment) 后进行构树, 得到每个基因的系统发育关系, 称为"基因树" (gene tree)。接着, 基于溯祖理论将所有基因树进行整合, 最终得到"物种树" (species tree, 即描述不同物种之间的进化关系的树)。这种方法可以减少基因数据缺失或者不完全谱系分选对构树结果造成的偏差。并联建树得到的进化树枝长以溯祖单位 (coalescent unit) 表示。基于转录组测序的系统发育研究在采样或者测序质量、组装效果等方面常存在差异, 物种间基因缺失的现象比较常见, 因此基于溯祖理论的构树方法得到越来越广泛的应用。这方面的代表软件有 ASTRAL-III (Zhang *et al.*, 2018 ; <https://github.com/smirarab/ASTRAL>)、MP-EST (Liu *et al.*, 2010 ; <http://faculty.franklin.uga.edu/liu/mp-est>)、STAR (Liu *et al.*, 2009) 等。但这种方法

受基因树的错误估计的影响较大 (Shen *et al.*, 2021)。导致基因树被错误估计的原因包括低质量的基因比对 (gene alignments) 等, 例如一些基因比对只有很少的信息位点 (informative sites, 一个位点至少存在 2 种不同的碱基 (或氨基酸) 并且每种碱基 (或氨基酸) 出现的次数至少为 2) 或者测序数据的非随机缺失 (例如 Springer and Gatesy, 2016; Simmons and Kessenich, 2020)。

利用基因比对数据构建基因树或物种树的常用软件包括 RAxML (Stamatakis 2014; 使用方法可参考教程: <http://evomics.org/learning/phylogenetics/raxml/>)、IQ-Tree (Nguyen *et al.*, 2015; <http://www.iqtree.org/>)、PhyML (Guindon *et al.*, 2010; <https://github.com/stephaneguindon/phyml>) 和 MrBayes (Ronquist *et al.*, 2012; <http://nbisweden.github.io/MrBayes/>) 等。前三者均使用最大似然估计算法, 在给定的基因序列的条件下找出最可能的拓扑结构和枝长, 即系统发育树。而 MrBayes (Ronquist *et al.*, 2012) 则使用贝叶斯马尔可夫链蒙特卡洛 (Bayesian Markov chain Monte Carlo (MCMC)) 方法进行系统演化关系分析。

6. 异常物种分析

异常物种 (rogue species) 是指在一系列系统发育树 (例如 bootstrap 树) 中可能出现在多个位置的物种。通常无法准确地推断出其与相关物种的进化关系。异常物种的存在往往会导致系统发育树上的其它节点具有较低的 bootstrap 支持值, 甚至出现多叉树 (polytomy, 即一个系统发育树上的某个祖先节点同时分化出 3 个或更多后代支系的情况)。RogueNaRok 软件 (Aberer *et al.*, 2013; <https://github.com/abrerer/RogueNaRok>) 可以通过剔除法鉴定异常物种, 即: 把若干个物种从 bootstrap 树中剔除之后, 一致树 (consensus tree) 的支持度将得到提升 (图 1 III. 系统发育分析)。

7. 物种树与基因树的差异分析

物种树与基因树的差异可能是由非同源相似 (homoplasy)、基因渐渗 (introgression)、快速演化 (rapid diversification) 等历史事件导致的 (图 1 III. 系统发育分析)。导致非同源相似的因素包括趋同演化 (convergent evolution)、平行演化 (parallel evolution) 和演化逆转 (evolutionary reversal)。水平基因转移、杂交 (hybridization) 等过程则会把某些基因从一个物种渗入到另一个物种。而相对短时间内发生的、连续的物种分化事件一方面导致新形成的物种没有足够时间来积累新

的突变 (mutation), 另一方面祖先的遗传变异在连续的分化事件中被保留了下来, 导致不完全谱系筛选, 这种现象在有效种群大小 (effective population size, 一个真实群体的有效群体大小等同于一个能够产生同样强度的遗传漂变的理想群体的大小) 越大的祖先物种中越明显 (Stiller and Zhang, 2019)。此外, 当一些支系发生了众多的突变的时候, 如果随机变化的数量比非随机的、具有系统发育信号的变化还多的时候, 会导致长枝相吸 (Long Branch Attraction, LBA) 现象, 其原因可能是不同类群具有不同的进化速率或者数据特异性 (Simpson, 2019)。

为了提高系统树的支持度和可靠性, 可考虑的方法有: (1) 对相关类群进行更密集的抽样, 获得更多物种的数据 (Stiller and Zhang, 2019)。(2) 使用合适的进化模型, 包括使用更好的位点分区方案和考虑数据异质性的影响 (见“4. 进化模型的选择”部分) (Stiller and Zhang, 2019)。(3) 由于不同类型的数据经受着不同的进化约束力 (evolutionary constraints), 具有不同的系统发育信号, 在进化分析中可使用不同的数据类型 (例如蛋白基因、非编码区域、线粒体序列、结构变异) (Stiller and Zhang, 2019)。(4) 选择合适的数据集合, 使用更多的有效数据位点, 以减少基因交流等事件的影响 (Zhang *et al.*, 2021)。(5) 剔除利用串联构树方法 (如 IQ-Tree (Nguyen *et al.*, 2015)) 和并联构树方法 (如 ASTRAL-III (Zhang *et al.*, 2018)) 得到的基因树的拓扑结构不一致的基因, 这可以通过计算基因树与物种树 (由串联或并联方法得到) 的拓扑距离得到。研究表明, 在较低的不完全谱系筛选水平及较低至中等比例的基因树被错误估计的情况下, 剔除其基因树被错误估计的基因可以有效提高系统发育树的准确性。但是, 在中等或高水平的不完全谱系筛选和高比例的基因树被错误估计的情况下, 剔除这种类型的基因虽然可以降低基因树之间的拓扑结构差异性, 但是得到的系统树与真实的物种树的拓扑结构并非始终一致。导致基因树被错误估计的因素包括: 没有足够的有效信息位点、基因座内部存在重组导致溯祖模型的不当使用、测序数据位点的非随机缺失等 (Shen *et al.*, 2021)。

针对不同数据集或方法可能获得的不一致的系统树, 可以对不同拓扑结构、枝长、支持度等进行统计和比较。分析过程可以参考: (1) 系统树支持度: 在系统树构建的时候软件会计算出系统树的对数最大似然值 (log likelihood) (Goldman *et al.*, 2000), 该树似然值越大说明系统树支持度越高; 进化枝上的 bootstrap 分值或贝叶斯方法获得的后验概率也能在一定程度上代表相关拓扑结构的支持度。(2) 基因树

支持度：计算以全部核基因树为背景时的拓扑结构支持度，可以利用 **ASTRAL** (Mirarab *et al.*, 2014) 计算 **normalized quartet score**，分数越高，说明支持度越高；如果分数较低，则表明该物种树中可能存在较高的不完全谱系分选现象或其它可能导致基因树与物种树不一致的现象存在。或利用 **DiscoVista** (Sayyari *et al.*, 2018) 进行基因树支持度计算。(3) 系统树拓扑结构差异：利用 **Robinson-Foulds distance** (RF 距离, Topological distance) (Kuhner and Felsenstein 1994) 表示，RF 值越小证明两棵系统树之间的拓扑结构相似性越高，该值可以利用 **R** (Team, 2013) 的 **ape** (Paradis *et al.*, 2004) 包计算。(4) 系统发育假设检测：检测方法包括 **Approximately Unbiased 检验** (AU test, 近似无偏检验) (Shimodaira, 2002)、**Shimodaira-Hasegawa 检验** (SH test) (Shimodaira and Hasegawa, 1999) 和 **Kishino-Hasegawa 检验** (KH test) (Kishino and Hasegawa, 1989) 等。

获得研究对象的物种树后，可以进一步检测物种树和基因树不一致背后的进化事件，即非同源相似、基因渐渗、快速演化等 (Stiller and Zhang, 2019; Zhang *et al.*, 2021) (图 1 III. 系统发育分析)。我们可以利用 **Densitree** (Bouckaert, 2010) 对随机抽取的多个基因树进行可视化，亦可利用 **DiscoVista** (Sayyari *et al.*, 2018; <https://github.com/esayyari/DiscoVista>) 等方法进行基因树拓扑结构分布统计。同时，目前已经有多种工具可以实现对基因渐渗或不完全谱系分选事件的鉴定，如基于基因组的窗口序列信息的 **ABBA-BABA test** (D statistics) (Kulathinal *et al.*, 2009; Green *et al.*, 2010; Durand *et al.*, 2011; Martin *et al.*, 2015)、**DFOIL test** (Pease and Hahn, 2015; <https://github.com/jbpease/dfoil>) 等、基于基因树拓扑结构与枝长信息的 **QUIBL** (Edelman *et al.*, 2019; <https://github.com/miriammiyaqi/QuIBL>)、基于基因树网状结构的 **Phylonet** (Than *et al.*, 2008 ; <http://old-bioinfo.cs.rice.edu/phylonet/>)、**Phylonetworks** (Solís-Lemus *et al.*, 2017 ; <https://phylonetwork.readthedocs.io/en/latest/>) 等。其中，基于窗口序列的方法更适用于全基因组数据，基于基因树的方法则可以尝试在转录组数据结果中进行分析，综合进行判断。

最后，研究表明，除了序列比对、软件差异、进化模型、搜索的树的数量及随机起始种子的选择等因素，低系统发育信号、计算机的处理器类型和多线程均会影响数据的分析可重复性 (Shen *et al.*, 2020)。

8. 物种分化时间估计

分子钟理论认为，分子进化速率在物种形成过程中保持大致恒定，可以通过化石时间对得到的系统发育树进行时间校正，推断出相关支系分化事件发生的时间 (Bromham and Penny, 2003) (图 1 III. 系统发育分析)。其中，化石证据的使用涉及到类群和地质学的不确定性，如何选取合适的化石证据是一个重要的问题。Parham *et al.* (2012) 提出了以下选取原则：(1) 针对所研究类群的内部及最近缘的外群 (内群 (ingroup) 是指我们所研究的、感兴趣的相关物种，外群(outgroup) 则是在内群发生物种分化之前就已经分化出来的物种或类群。利用外群物种，我们可以明确所构建出来的系统发育树的演化方向)，考虑所有获得良好鉴定结果的化石证据；(2) 总结相关基于化石的衍征 (apomorphy, 即后代继承自其最近共同祖先的、相对于其余类群所新获得的性状或性状状态，可用于鉴定某个类群是否是单系群(monophyly)；而如果除了后代及其最近共同祖先的其它类群也拥有该性状或性状状态，则把该性状或性状状态称为祖征 (plesiomorphy)) 诊断方法得到的系统发育关系或者使用了这些化石样本的最新的研究成果；(3) 总结已经发表的化石的地质时间及其它地层范围，并且与最新版的国际年代地层表的时间进行同步。

广泛使用的、利用贝叶斯 (Bayesian) 方法来推断物种分歧时间的软件有 MCMCtree (Yang, 2007) 、BEAST (Drummond and Rambaut, 2007) 和 BEAST2 (Bouckaert *et al.*, 2014)。详细的使用教程可参考 Dos Reis and Yang (2019) 和 <https://beast2-dev.github.io/beast-docs/beast2/DivergenceDating/DivergenceDatingTutorial.html> 和 <https://taming-the-beast.org/tutorials/> (Barido-Sottani *et al.*, 2018)。

在 Misof *et al.* (2014) 中，作者依据 Parham *et al.* (2012) 的标准选取了 37 个化石证据，利用 BEAST (Drummond and Rambaut, 2007) 软件及所预测出来的单拷贝直系同源基因的蛋白质序列对系统发育树进行时间校正。由于数据量过大，BEAST 软件 (Drummond and Rambaut, 2007) 无法一次运行所有序列数据，作者根据 PartitionFinder (Lanfear *et al.*, 2012) 给出的分区方案，对每个分区分别进行了单独 BEAST (Drummond and Rambaut, 2007) 分析，并要求每个分区覆盖了所有的研究物种 (内群和外群共计 144 个物种)，每个分区的长度至少为 500 个氨基酸位点，且任一分区内每个物种的序列至少有部分位点不是未知氨基酸 (即"X")。对于

进化树上的每个校正节点，作者分别设置了一个先验分布：使用"uniform"分布并指定了最小及最大时间范围，或者使用"log-normal" ($\mu = 2$; $d = 0.5$) 分布并指定了一个最小的时间。使用 log-normal 先验分布是由于作者认为对某些节点使用 uniform 先验分布过于保守 (关于如何设置合适的先验分布，请读者参考 <https://taming-the-beast.org/tutorials/FBD-tutorial/FBD-tutorial.pdf> 和 Ho and Phillips (2009) 以及 Heath (2012))。同时，作者设置其系统树的根节点时间范围是 411.5 – 580 百万年之前。对于每个分区，作者使用 PartitionFinder (Lanfear *et al.*, 2012) 给出的最佳进化模型。但由于 BEAST (Drummond and Rambaut, 2007) 只实现了 RAxML (Stamatakis, 2014) 或 ExaML (Kozlov *et al.*, 2015) 软件中的部分进化模型，因此作者只能在使用 BEAST (Drummond and Rambaut, 2007) 分析时把最佳进化模型以最相似的可用模型进行代替 (注：RAxML (Stamatakis, 2014) 或 ExaML (Kozlov *et al.*, 2015) 软件中已实现的进化模型见相应软件的官方文档关于-m 参数的介绍，例如 <https://cme.h-its.org/exelixis/resource/download/NewManual.pdf>。BEAST (Drummond and Rambaut, 2007) 和 BEAST2 (Bouckaert *et al.*, 2014) 已实现的进化模型可参考 <https://www.beast2.org/features/>)。同时，作者使用的分子钟模型是 "uncorrelated lognormal relaxed clock"。通常我们需要运行 MCMCTree (Yang, 2007) 或者 BEAST (Drummond and Rambaut, 2007) 至少两次，并检测相关参数的估计值是否可以达到收敛 (convergence)，即趋于稳定。如果通过 Tracer 软件 (Rambaut *et al.*, 2018) 来辅助判断，判定的标准通常要求有效样本大小 (Effective Sample Size, ESS) 大于 200。如果发现结果未收敛，则需要增加迭代数 (iterations) 或 (和) burn-in 的大小来重新运行。接着，对于每个分区，在删除每个 run 的 burn-in trees 后，把不同的 run 的 sampled trees 进行合并，然后随机抽取 10,000 个树 (Misof *et al.*, 2014)，最后把不同分区的 sampled trees 利用 BEAST (Drummond and Rambaut, 2007) 中的 LogCombiner 进行合并。下一步，利用 TreeAnnotator 1.8 (Rambaut and Drummond, 2013) 来计算及注释一致树 (consensus tree)，得到每个节点的中位数及置信区间。最后，使用 FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>) 来对结果进行可视化。

最后，把相关信息如物种树、物种分化时间、生物地理历史等信息进行综合分析，对生物演化历程进行合理的阐释。

9. 经典系统进化课程推荐

线上有学者分享了一些经典的在线课程课程或者流程，对于读者学习了解系统进化有很大的帮助。如：https://github.com/simonhmartin/genomics_general 和 <https://github.com/simjoly/CourseComparativeMethods> 以及 <https://github.com/ForBioPhylogenomics/tutorials>

四、展望

转录组数据在动、植物等研究中具有广泛的应用，目前这些数据主要是由短序列测序平台产生。但是，短读长测序存在一些固有的缺陷，包括实验过程中的 GC 偏向性、短序列难以准确地比对到参考序列上的重复区域（真假基因、多拷贝基因）等 (Alkan *et al.*, 2011; Ozsolak and Milos, 2011)。随着长读长测序技术的推出，如 PacBio 公司的 Single-Molecule Real Time (SMRT) (Chin *et al.*, 2013)、Oxford Nanopore Technologies (ONT) 公司的纳米孔测序技术 (Laver *et al.*, 2015)，相关问题可以得到有效解决。长读长测序技术可以直接把 RNA 分子（或 cDNA 分子）测通，无需组装就可以获得转录本全长序列，从而能够更准确地鉴定不同类型的转录本 (Jain *et al.*, 2018)。通过不断地进行技术改进，这两种测序技术的数据准确度得到不断提升。而随着更大通量的仪器的推出，例如 PacBio 公司的 Sequel II System 测序仪和 ONT 公司的 PromethION 测序仪，单碱基测序成本进一步下降，这为研究人员开展更大规模、更高质量系统发育研究提供了新机遇。与此同时，更大规模的数据的产生也对数据的存储和传输、系统发育相关算法与软件的开发研究提出了新的挑战。

致谢

感谢审稿人和编辑对本文初稿提出的修改建议。本研究由国家自然科学基金面上项目 (31772493) 资助。

参考文献

1. Ababneh, F., Jermiin, L. S., Ma, C. and Robinson, J. (2006a). [Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences](#). *Bioinformatics* 22(10): 1225-1231.
2. Ababneh, F., Jermiin, L. S. and Robinson, J. (2006b). [Generation of the exact](#)

- [distribution and simulation of matched nucleotide sequences on a phylogenetic tree](#). *J Math Model Algor* 5: 291-308.
3. Aberer, A. J., Krompass, D. and Stamatakis A. (2013). [Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice](#). *Syst Biol* 62(1): 162-166.
 4. Alkan, C., Sajjadian, S. and Eichler, E. E. (2011). [Limitations of next-generation genome sequence assembly](#). *Nat Methods* 8: 61-65.
 5. Andrews, S. (2010). [FastQC: a quality control tool for high throughput sequence data](#).
 6. Bolger, A. M., Lohse, M. and Usadel, B. (2014). [Trimmomatic: a flexible trimmer for Illumina sequence data](#). *Bioinformatics* 30(15): 2114-2120.
 7. Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C. H., Xie, D., Suchard, M. A., Rambaut, A. and Drummond, A. J. (2014). [BEAST 2: a software platform for Bayesian evolutionary analysis](#). *PLoS Comput Biol* 10(4): e1003537.
 8. Bouckaert, R. R. (2010). [DensiTree: making sense of sets of phylogenetic trees](#). *Bioinformatics* 26(10): 1372-1373.
 9. Bromham, L. and Penny, D. (2003). [The modern molecular clock](#). *Nat Rev Genet* 4(3): 216-224.
 10. Burleigh, J. G., Hilu, K. W. and Soltis, D. E. (2009). [Inferring phylogenies with incomplete data sets: a 5-gene, 567-taxon analysis of angiosperms](#). *BMC Evol Biol* 9: 1-11.
 11. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. L. (2009). [BLAST+: architecture and applications](#). *BMC Bioinformatics* 10: 1-9.
 12. Capella-Gutiérrez, S., Silla-Martínez, J. M. and Gabaldón, T. (2009). [trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses](#). *Bioinformatics* 25(15): 1972-1973.
 13. Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., Li, Y., Ye, J., Yu, C., Li, Z., Zhang, X., Wang, J., Yang, H., Fang, L. and Chen, Q. (2018). [SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data](#). *Gigascience* 7(1): gix120.
 14. Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W. and Korlach, A.

- J. (2013). [Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data](#). *Nat Methods* 10(6): 563.
15. Cho, S., Zwick, A., Regier, J. C., Mitter, C., Cummings, M. P., Yao, J., Du, Z., Zhao, H., Kawahara, A. Y., Weller, S., Davis, D. R., Baixeras, J., Brown, J. W. and Parr, C. (2011). [Can deliberately incomplete gene sample augmentation improve a phylogeny estimate for the advanced moths and butterflies \(Hexapoda: Lepidoptera\)?](#) *Syst Biol* 60(6): 782-796.
16. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. and Rice, P. M. (2010). [The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants](#). *Nucleic Acids Res* 38(6): 1767-1771.
17. Davidson, R., Vachaspati, P., Mirarab, S. and Warnow, T. (2015). [Phylogenomic species tree estimation in the presence of incomplete lineage sorting and horizontal gene transfer](#). *BMC Genomics* 16: 1-12.
18. Dell'Ampio, E., Meusemann, K., Szucsich, N. U., Peters, R. S., Meyer, B., Borner, J., Petersen, M., Aberer, A. J., Stamatakis, A., Walz, M. G., Minh, B. Q., von Haeseler, A., Ebersberger, I., Pass, G. and Misof, B. (2014). [Decisive data sets in phylogenomics: lessons from studies on the phylogenetic relationships of primarily wingless insects](#). *Mol Biol Evol* 31(1): 239-249.
19. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T. R. (2013). [STAR: ultrafast universal RNA-seq aligner](#). *Bioinformatics* 29(1): 15-21.
20. Drummond, A. J. and Rambaut, A. (2007). [BEAST: Bayesian evolutionary analysis by sampling trees](#). *BMC Evol Biol* 7: 1-8.
21. Durand, E. Y., Patterson, N., Reich, D. and Slatkin, M. (2011). [Testing for ancient admixture between closely related populations](#). *Mol Biol Evol* 28(8): 2239-2252.
22. Dutheil, J. and Boussau, B. (2008). [Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs](#). *BMC Evol Biol* 8: 1-12.
23. Ebersberger, I., Strauss, S. and von Haeseler, A. (2009). [HaMStR: profile hidden markov model based search for orthologs in ESTs](#). *BMC Evol Biol* 9: 1-9.
24. Eddy, S. R. (2011). [Accelerated profile HMM searches](#). *PLoS Comput Biol* 7: e1002195.
25. Edelman, N. B., Frandsen, P. B., Miyagi, M., Clavijo, B., Davey, J., Dikow, R. B., Garcia-Accinelli, G., Van Belleghem, S. M., Patterson, N., Neafsey, D. E., Challis,

- R., Kumar, S., Moreira, G. R. P., Salazar, C., Chouteau, M., Counterman, B. A., Papa, R., Blaxter, M., Reed, R. D., Dasmahapatra, K. K., Kronforst, M., Joron, M., Jiggins, C. D., McMillan, W. O., Di Palma, F., Blumberg, A. J., Wakeley, J., Jaffe, D. and Mallet, J. (2019). [Genomic architecture and introgression shape a butterfly radiation](#). *Science* 366(6465): 594-599.
26. Edgar, R. C. (2004). [MUSCLE: multiple sequence alignment with high accuracy and high throughput](#). *Nucleic Acids Res* 32(5): 1792-1797.
27. Feng, S., Stiller, J., Deng, Y., Armstrong, J., Fang, Q., Reeve, A. H., Xie, D., Chen, G., Guo, C., Faircloth, B. C., Petersen, B., Wang, Z., Zhou, Q., Diekhans, M., Chen, W., Andreu-Sánchez, S., Margaryan, A., Howard, J. T., Parent, C., Pacheco, G., Sinding, M. H. S., Puetz, L., Cavill, E., Ribeiro, Â. M., Eckhart, L., Fjeldså, J., Hosner, P. A., Brumfield, R. T., Christidis, L., Bertelsen, M. F., Sicheritz-Ponten, T., Tietze, D. T., Robertson, B. C., Song, G., Borgia, G., Claramunt, S., Lovette, I. J., Cowen, S. J., Njoroge, P., Dumbacher, J. P., Ryder, O. A., Fuchs, J., Bunce, M., Burt, D. W., Cracraft, J., Meng, G., Hackett, S. J., Ryan, P. G., Jönsson, K. A., Jamieson, I. G., da Fonseca, R. R., Braun, E. L., Houde, P., Mirarab, S., Suh, A., Hansson, B., Ponnikas, S., Sigeman, H., Stervander, M., Frandsen, P. B., van der Zwan, H., van der Sluis, R., Visser, C., Balakrishnan, C. N., Clark, A. G., Fitzpatrick, J. W., Bowman, R., Chen, N., Cloutier, A., Sackton, T. B., Edwards, S. V., Foote, D. J., Shakya, S. B., Sheldon, F. H., Vignal, A., Soares, A. E., R., Shapiro, B., González-Solís, J., Ferrer-Obiol, J., Rozas, J., Riutort, M., Tigano, A., Friesen, V., Dalén, L., Urrutia, A. O., Székely, T., Liu, Y., Campana, M. G., Corvelo, A., Fleischer, R. C., Rutherford, K. M., Gemmell, N. J., Dussex, N., Mouritsen, H., Thiele, N., Delmore, K., Liedvogel, M., Franke, A., Hoepfner, M. P., Krone, O., Fudickar, A. M., Milá, B., Ketterson, E. D., Fidler, A. E., Friis, G., Parody-Merino, Á. M., Battley, P. F., Cox, M. P., Lima, N. C., B., Prosdocimi, F., Parchman, T. L., Schlinger, B. A., Loiselle, B. A., Blake, J. G., Lim, H. C., Day, L. B., Fuxjager, M. J., Baldwin, M. W., Braun, M. J., Wirthlin, M., Dikow, R. B., Ryder, T. B., Camenisch, G., Keller, L. F., DaCosta, J. M., Hauber, M. E., Louder, M. I., M., Witt, C., C., McGuire, J. A., Mudge, J., Megna, L. C., Carling, M. D., Wang, B., Taylor, S. A., Del-Rio, G., Aleixo, A., Vasconcelos, A. T. R., Mello, C. V., Weir, J. T., Haussler, D., Li, Q., Yang, H., Wang, J., Lei, F., Rahbek, C., Gilbert, M. T., P., Graves, G. R., Jarvis, E. D., Paten, B. and Zhang, G. (2020). [Dense sampling of](#)

- [bird diversity increases power of comparative genomics](#). *Nature* 587: 252-257.
28. Foster, P. G. (2004). [Modeling compositional heterogeneity](#). *Syst Biol* 53(3): 485-495.
29. Goldman, N., Anderson, J. P. and Rodrigo, A. G. (2000). [Likelihood-based tests of topologies in phylogenetics](#). *Syst Biol* 49(4): 652-670.
30. Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N. and Regev, A. (2011). [Full-length transcriptome assembly from RNA-Seq data without a reference genome](#). *Nat Biotechnol* 29: 644.
31. Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H. Y., Hansen, N. F., Durand, E. Y., Malaspina, A., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušić, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L. F., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D. and Pääbo, S. (2010). [A draft sequence of the Neandertal genome](#). *Science* 328(5979): 710-722.
32. Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O. (2010). [New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0](#). *Syst Biol* 59(3): 307-321.
33. Henikoff, S. and Henikoff, J. G. (1992). [Amino acid substitution matrices from protein blocks](#). *Proc Natl Acad Sci* 89: 10915-10919.
34. Huang, X., Chen, X. G. and Armbruster, P. A. (2016). [Comparative performance of transcriptome assembly methods for non-model organisms](#). *BMC Genomics* 17: 1-14.
35. Consortium, I. (2013). [The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment](#). *J Hered* 104(5): 595-600.
36. Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R.,

- Beggs, A. D., Dillthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., Snutch, T. P., Tee, L., Paten, B., Phillippy, A. M., Simpson, J. T., Loman, N. J. and Loose, M. (2018). [Nanopore sequencing and assembly of a human genome with ultra-long reads](#). *Nat Biotechnol* 36: 338-345.
37. Jehl, P., Sievers, F. and Higgins, D. G. (2015). [OD-seq: outlier detection in multiple sequence alignments](#). *BMC Bioinformatics* 16: 269.
38. Jermin, L. S., Jayaswal, V., Ababneh, F. M. and Robinson, J. (2017). [Identifying optimal models of evolution](#). *Bioinformatics* 379-420.
39. Jiang, H., Lei, R., Ding, S. W. and Zhu, S. (2014). [Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads](#). *BMC Bioinformatics* 15: 1-12.
40. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. and Jermin, L. S. (2017). [ModelFinder: fast model selection for accurate phylogenetic estimates](#). *Nat Methods* 14: 587-589.
41. Katoh, K., Misawa, K., Kuma, K. and Miyata, T. (2002). [MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform](#). *Nucleic Acids Res* 30(14): 3059-3066.
42. Katoh, K. and Toh, H. (2008). [Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework](#). *BMC Bioinformatics* 9: 1-13.
43. Kim, D., Paggi, J. M., Park, C., Bennett, C. and Salzberg, S. L. (2019). [Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype](#). *Nat Biotechnol* 37: 907-915.
44. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S. L. (2013). [TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions](#). *Genome Biol* 14: 1-13.
45. Kishino, H. and Hasegawa, M. (1989). [Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea](#). *J Mol Evol* 29: 170-179.
46. Kozlov, A. M., Aberer, A. J. and Stamatakis, A. (2015). [ExaML version 3: a tool for phylogenomic analyses on supercomputers](#). *Bioinformatics* 31: 2577-2579.
47. Kriventseva, E. V., Rahman, N., Espinosa, O. and Zdobnov, E. M. (2007).

- [OrthoDB: the hierarchical catalog of eukaryotic orthologs](#). *Nucleic Acids Res* 36: D271-D275.
48. Kuhner, M. K. and Felsenstein, J. (1994). [A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates](#). *Mol Biol Evol* 11(3): 459-468.
 49. Kulathinal, R. J., Stevison, L. S. and Noor, M. A. F. (2009). [The genomics of speciation in *Drosophila*: diversity, divergence, and introgression estimated using low-coverage genome sequencing](#). *PLoS Genet* 5: e1000550.
 50. Kumar, S., Filipski, A. J., Battistuzzi, F. U., Kosakovsky Pond, S. L. and Tamura, K. (2012). [Statistics and truth in phylogenomics](#). *Mol Biol Evol* 29(2): 457-472.
 51. Lanfear, R., Calcott, B., Ho, S. Y. W. and Guindon S. (2012). [PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses](#). *Mol Biol Evol* 29(6): 1695-1701.
 52. Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T. and Calcott, B. (2017). [PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses](#). *Mol Biol Evol* 34(3): 772-773.
 53. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson J. D., Gibson, T. J. and Higgins, D. G. (2007). [Clustal W and Clustal X version 2.0](#). *Bioinformatics* 23: 2947-2948.
 54. Lartillot, N. and Philippe, H. (2004). [A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process](#). *Mol Biol Evol* 21(6): 1095-1109.
 55. Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., and Studholme, D. J. (2015). [Assessing the performance of the oxford nanopore technologies minion](#). *Biomol Detect Quantif* 3: 1-8.
 56. Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., van Sluys, M., Soltis, P. S., Xu, X., Yang, H. and Zhang, G. (2018). [Earth BioGenome Project: sequencing life for the future of life](#). *Proc Natl Acad Sci* 115: 4325-4333.
 57. Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J. A., Stewart, R. and Dewey, C.

- N. (2014). [Evaluation of de novo transcriptome assemblies from RNA-Seq data](#). *Genome Biol* 15: 1-21.
58. Liu, L., Yu, L. and Edwards, S. V. (2010). [A maximum pseudo-likelihood approach for estimating species trees under the coalescent model](#). *BMC Evol Biol* 10: 302.
59. Liu, L., Yu, L., Pearl, D. K. and Edwards, S. V. (2009). [Estimating species phylogenies using coalescence times among sequences](#). *Syst Biol* 58(5): 468-477.
60. 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, D. W., Yiu, S. M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T. W. and Wang, J. (2012). [SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler](#). *Gigascience* 1(1): 1-18.
61. Martin, J., Bruno, V. M., Fang, Z., Meng, X., Blow, M., Zhang, T., Sherlock, G., Snyder, M. and Wang, Z. (2010). [Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads](#). *BMC Genomics* 11: 1-8.
62. Martin, J. A. and Wang, Z. (2011). [Next-generation transcriptome assembly](#). *Nat Rev Genet* 12: 671-682.
63. Martin, M. (2011). [Cutadapt removes adapter sequences from high-throughput sequencing reads](#). *EMBnet J* 17: 10-12.
64. Martin, S. H., Davey, J. W. and Jiggins, C. D. (2015). [Evaluating the use of ABBA-BABA statistics to locate introgressed loci](#). *Mol Biol Evol* 32(1): 244-257.
65. Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S. and Warnow, T. (2014). [ASTRAL: genome-scale coalescent-based species tree estimation](#). *Bioinformatics* 30(17): i541-i548.
66. Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., Frandsen, P. B., Ware, J., Flouri, T., Beutel, R. G., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A. J., Aspöck, U., Aspöck, H., Bartel, D., Blanke, A., Berger, S., Böhm, A., Buckley, T. R., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y., Jermiin, L. S., Kawahara, A. Y., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Y., Li, Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., McKenna, D. D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J. L., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., von Reumont, B. M., Schütte, K., Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N. U., Tan, M., Tan, X., Tang, M.,

- Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walz, M. G., Wiegmann, B. M., Wilbrandt, J., Wipfler, B., Wong, T. K. F., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D. K., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y., Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Y., Xu, X., Zhang, Y., Yang, H., Wang, J., Wang, J., Kjer, K. M. and Zhou, X. (2014). [Phylogenomics resolves the timing and pattern of insect evolution](#). *Science* 346(6210): 763-767.
67. Nakamura, T., Yamada, K. D., Tomii, K. and Katoh, K. (2018). [Parallelization of MAFFT for large-scale multiple sequence alignments](#). *Bioinformatics* 34(14): 2490-2492.
 68. Naser-Khdour, S., Minh, B. Q., Zhang, W., Stone, E. A. and Lanfear, R. (2019). [The prevalence and impact of model violations in phylogenetic analysis](#). *Genome Biol Evol* 11(12): 3341-3352.
 69. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. and Minh, B. Q. (2015). [IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies](#). *Mol Biol Evol* 32(1): 268-274.
 70. Ozsolak, F. and Milos, P. M. (2011). [RNA sequencing: advances, challenges and opportunities](#). *Nat Rev Genet* 12: 87-98.
 71. Paradis, E., Claude, J. and Strimmer, K. (2004). [APE: analyses of phylogenetics and evolution in R language](#). *Bioinformatics* 20(2): 289-290.
 72. Parham J.F., Donoghue P.C.J., Bell C.J., Calway T.D., Head J.J., Holroyd P.A., Inoue J.G., Irmis R.B., Joyce W.G., Ksepka D.T., Patané, J. S. L., Smith, N. D., Tarver, J. E., van Tuinen, M., Yang, Z., Angielczyk, K. D., Greenwood, J. M., Hipsley, C. A., Jacobs, L., Makovicky, P. J., Müller, J., Smith, K. T., Theodor, J. M., Warnock, R. C. M. and Benton, M. J. (2012). [Best practices for justifying fossil calibrations](#). *Syst Biol* 61(2): 346-359.
 73. Patel, R. K. and Jain, M. (2012). [NGS QC Toolkit: a toolkit for quality control of next generation sequencing data](#). *PLoS One* 7: e30619.
 74. Pease, J. B. and Hahn, M. W. (2015). [Detection and polarization of introgression in a five-taxon phylogeny](#). *Syst Biol* 64(4): 651-662.
 75. Peng, Y., Leung, H. C. M., Yiu, S. M., Lv, M. J., Zhu, X. G. and Chin, F. Y. L. (2013). [IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels](#). *Bioinformatics* 29(13): i326-i334.

76. Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T. and Salzberg, S. L. (2015). [StringTie enables improved reconstruction of a transcriptome from RNA-seq reads](#). *Nat Biotechnol* 33: 290-295.
77. Petersen, M., Meusemann, K., Donath, A., Dowling, D., Liu, S., Peters, R. S., Podsiadlowski, L., Vasilakopoulos, A., Zhou, X., Misof, B. and Niehuis, O. (2017). [Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes](#). *BMC Bioinformatics* 18: 1-10.
78. Rambaut, A. and Drummond, A. J. (2013). TreeAnnotator (Version 1.8). <http://beast.bio.ed.ac.uk>.
79. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. and Suchard, M. A. (2018). [Posterior summarization in Bayesian phylogenetics using tracer 1.7](#). *Syst Biol* 67(5): 901-904.
80. Rasmussen, M. D. and Kellis, M. (2012). [Unified modeling of gene duplication, loss, and coalescence using a locus tree](#). *Genome Res* 22: 755-765.
81. dos Reis, M. and Yang, Z. (2019). [Bayesian molecular clock dating using genome-scale datasets](#). In: Anisimova, M. (Ed.). In: *Evolutionary Genomics*. Humana, New York, NY. ISBN: 978-1-4939-9074-0.
82. Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y. S., Newsome, R., Chan, S. K., She, R., Varhol, R., Kamoh, B., Prabhu, A., Tam, A., Zhao, Y., Moore, R. A., Hirst, M., Marra, M. A., Jones, S. J. M., Hoodless, P. A. and Birol, I. (2010). [De novo assembly and analysis of RNA-seq data](#). *Nat Methods* 7: 909-912.
83. Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A. and Huelsenbeck, J. P. (2012). [MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space](#). *Syst Biol* 61(3): 539-542.
84. Sayyari, E., Whitfield, J. B. and Mirarab, S. (2018). [DiscoVista: interpretable visualizations of gene tree discordance](#). *Mol Phylogenet Evol* 122: 110-115.
85. Schulz, M. H., Zerbino, D. R., Vingron, M. and Birney, E. (2012). [Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels](#). *Bioinformatics* 28(8): 1086-1092.
86. Genome 10K Community of Scientists. (2009). [Genome 10K: a proposal to obtain](#)

- [whole-genome sequence for 10 000 vertebrate species](#). *J Hered* 100(6): 659-674.
87. Scornavacca, C. and Galtier, N. (2017). [Incomplete lineage sorting in mammalian phylogenomics](#). *Syst Biol* 66(1): 112-120.
 88. Shen, X. X., Li, Y., Hittinger, C. T., Chen, X. X. and Rokas, A. (2020). [An investigation of irreproducibility in maximum likelihood phylogenetic inference](#). *Nat Commun* 11: 1-14.
 89. Shen, X. X., Steenwyk, J. L. and Rokas, A. (2021). [Dissecting incongruence between concatenation-and quartet-based approaches in phylogenomic data](#). *Syst Biol* 0(0): 1-18.
 90. Shimodaira, H. (2002). [An approximately unbiased test of phylogenetic tree selection](#). *Syst Biol* 51(3): 492-508.
 91. Shimodaira, H. and Hasegawa, M. (1999). [Multiple comparisons of log-likelihoods with applications to phylogenetic inference](#). *Mol Biol Evol* 16(8): 1114.
 92. Simmons, M. P. and Kessenich, J. (2020). [Divergence and support among slightly suboptimal likelihood gene trees](#). *Cladistics* 36: 322-340.
 93. Simpson, M.G. (2019). *Plant systematics*. Academic press. 34-35.
 94. Slater, G. S. C. and Birney, E. (2005). [Automated generation of heuristics for biological sequence comparison](#). *BMC Bioinformatics* 6: 1-11.
 95. Solís-Lemus, C., Bastide, P. and Ané, C. (2017). [PhyloNetworks: a package for phylogenetic networks](#). *Mol Biol Evol* 34(12): 3292-3298.
 96. Springer, M. S. and Gatesy, J. (2016). [The gene tree delusion](#). *Mol Phylogenet Evol* 94: 1-33.
 97. Stamatakis, A. (2014). [RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies](#). *Bioinformatics* 30(9): 1312-1313.
 98. Stiller, J. and Zhang, G. (2019). [Comparative phylogenomics, a stepping stone for bird biodiversity studies](#). *Diversity* 11(7):115.
 99. Surget-Groba, Y. and Montoya-Burgos, J. I. (2010). [Optimization of de novo transcriptome assembly from next-generation sequencing data](#). *Genome Res* 20: 1432-1440.
 100. Suyama, M., Torrents, D. and Bork, P. (2006). [PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments](#). *Nucleic Acids Res* 34: W609-W612.
 101. Talavera, G. and Castresana, J. (2007). [Improvement of phylogenies after](#)

- [removing divergent and ambiguously aligned blocks from protein sequence alignments](#). *Syst Biol* 56(4): 564-577.
102. Team, R. C. (2013). R: A language and environment for statistical computing. Available from <https://www.r-project.org/>.
 103. Than, C., Ruths, D. and Nakhleh, L. (2008). [PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships](#). *BMC Bioinformatics* 9: 1-16.
 104. Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L. and Pachter, L. (2013). [Differential analysis of gene regulation at transcript resolution with RNA-seq](#). *Nat Biotechnol* 31: 46-53.
 105. Wong, T. K. F., Kalyaanamoorthy, S., Meusemann, K., Yeates, D. K., Misof, B. and Jermiin, L. S. (2020). [A minimum reporting standard for multiple sequence alignments](#). *NAR Genomics Bioinforma* 2: lqaa024.
 106. Woodhams, M. D., Fernández-Sánchez, J. and Sumner, J. G. (2015). [A new hierarchy of phylogenetic models consistent with heterogeneous substitution rates](#). *Syst Biol* 64(4): 638-650.
 107. Wright, E. S. and Vetsigian, K. H. (2016). [Quality filtering of Illumina index reads mitigates sample cross-talk](#). *BMC Genomics* 17: 876.
 108. Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T. W., Li, Y., Xu, X., Wong, G. and Wang, J. (2014). [SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads](#). *Bioinformatics* 30(12): 1660-1666.
 109. Yang, Z. (1996). [Among-site rate variation and its impact on phylogenetic analyses](#). *Trends Ecol Evol* 11: 367-372.
 110. Yang, Z. (2007). [PAML 4: phylogenetic analysis by maximum likelihood](#). *Mol Biol Evol* 24(8): 1586-1591.
 111. Zhang, C., Rabiee, M., Sayyari, E. and Mirarab, S. (2018). [ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees](#). *BMC Bioinformatics* 19: 15-30.
 112. Zhang, D., Rheindt, F. E., She, H., Cheng, Y., Song, G., Jia, C., Qu, Y., Alström, P. and Lei, F. (2021). [Most genomic loci misrepresent the phylogeny of an avian radiation because of ancient gene flow](#). *Syst Biol* 70(5): 961-975.
 113. Zhang G. (2015). [Bird sequencing project takes off](#). *Nature*. 522: 34.

114. Zhou, C., Meng, G. and Liu, S. (2020). [From reads to genes](#). In: Shelomi, M. (Ed.). In: *Transcriptomics in entomological research*. CABI, 24-44.
115. Zou, L., Susko, E., Field, C. and Roger, A. J. (2012). [Fitting nonstationary general-time-reversible models to obtain edge-lengths and frequencies for the Barry--Hartigan model](#). *Syst Biol* 61(6): 927-940.