物种树构建后的基因渗入检验方法

Methods of Detecting Introgression after Species-tree Construction

丁颖, 纪繁迪, 黄华腾*

生命科学学院, 陕西师范大学, 西安, 陕西 710119

*通讯作者邮箱: huanghuateng@snnu.edu.cn

引用格式: 丁颖, 纪繁迪, 黄华腾.(2021). 物种树构建后的基因渗入检验方法. *Bio-101* e1010678. Doi: 10.21769/BioProtoc.1010678.

How to cite: Ding, Y., Ji, F. D. and Huang, H. T. (2021). Methods of Detecting Introgression after Species-tree Construction. *Bio-101* e1010678. Doi: 10.21769/BioProtoc.1010678. (in Chinese)

摘要: 近年来越来越多的分子系统发育研究开始关注物种间基因渗入现象。一方面很多系统发育基因组学研究发现仅仅不完全谱系分选不能够完全解释数据集内部的异质性, 另一方面有大量种间基因渗入的实例被发现。因此, 很多系统发育基因组学分析会在物种树构建之后检验物种间是否存在基因渗入。目前该领域使用较为普遍的几类方法有: 基于位点模式统计量的方法、基于基因树信息的方法和 **PhyloNet**。本文简单介绍了各类方法的原理和运行所需数据, 并对它们的适用范围进行了讨论。示例分析步骤及可执行 R 脚本见文末在线文档链接。

关键词: 基因渗入, 系统发育, 不完全谱系分选, 物种树构建

研究背景

准确构建类群分化的历史是进化研究的基础。随着高通量测序技术的发展, 系统发育迎来了系统发育基因组学 (phylogenomics) 的时代 (于黎和张亚平, 2006)。在海量序列数据用于系统发育分析的同时, 由生物过程或非生物因素导致的数据异质性也越发明显 (Jeffroy *et al.*, 2006; Singhal *et al.*, 2021)。比如, 物种快速分化导致的不完全谱系分选 (incomplete lineage sorting, ILS) 可以使不同基因谱系分化历史不同 (Maddison, 1997); 碱基替换模型的错配可以导致基因树估计误差 (Richards *et al.*, 2018) 等。虽然

系统发育方法的发展解决了其中一些问题 (如用基于溯祖模型的物种树方法解决 ILS 问题; Mirarab *et al.*, 2014), 但仍有很多因素并未在通用的系统发育树构建方法中得到考虑。其中, 物种之间的基因渗入是近年来被越来越多的系统发育研究关注的一个因素 (Blair and Ané, 2020; Dasmahapatra *et al.*, 2012; Fontaine *et al.*, 2015; Mallet *et al.*, 2016)。

物种间基因渗入的形式可能是个体迁移导致的基因流、原有物种杂交形成新物种, 或者通过其他载体 (如病毒) 的水平基因转移 (Husnik and McCutcheon, 2018; Jiao *et al.*, 2021)。不论具体机制, 对系统发育分析来说, 基因渗入使物种分化历史不再是严格的二叉树结构, 而是网状结构 (network-like evolution)。这违反了大多数系统发育学方法的假设 (Baptiste *et al.*, 2013; Blair and Ané, 2020)。基因渗入和不完全谱系分选都会增加数据内部的异质性—不同基因有不同的谱系分化历史。当两者在物种分化历史中同时存在时, 基于溯祖模型的物种树方法可能会出现错误 (Blair and Ané, 2020)。Tian *et al.* (2016) 发现, 即使基因渗入发生在姐妹物种当中 (即基因渗入不改变基因树的拓扑结构), 也有可能使物种树方法出现异常 (gene flow anomaly zone; Long and Kubatko, 2018)。

近年来大量研究表明, 物种间基因渗入比以往认为的更加常见 (Mao *et al.*, 2018)。虽然水平基因转移被认为多见于原核生物, 但真核生物中种间杂交并不鲜见—有研究估计大概 25% 的开花植物和 10% 的动物物种存在杂交现象 (Grant and Grant, 1992; Mallet, 2005; Mallet, 2007; Mallet *et al.*, 2016)。有学者认为通过种间杂交带入新等位基因可能是一种重要的适应性进化机制 (Mao *et al.*, 2018; Oziolor *et al.*, 2019)。近期快速分化类群中的基因渗入很大程度上能够模糊该类群的系统发育关系 (如袖蝶属; Edelman *et al.*, 2019)。高级阶元间系统发育关系也会受古老的基因渗入事件的影响: 如 Suvorov *et al.* (2021) 对蜻蜓目的研究发现在三个主要亚目分化时期, 均翅亚目和间翅亚目之间存在的基因渗入。

鉴于基因渗入对系统发育树构建的影响及其普遍程度, 目前很多系统发育基因组学分析在构建物种树之后, 都会进行基因渗入检验, 以期能够更全面地解释导致数据异质性的因素。比如 Cai *et al.* (2021) 构建金虎尾目系统发育关系时, 发现仅 ILS 和基因树估计误差不足以解释基因树间的不一致性问题, 基因渗入是可能的原因。程等人 (2020) 也系统探讨了检测估计种间基因渗入的方法, 包括基于统计量的检验方法和基于进化模

型的方法。本文将介绍目前使用较多的几类基因渗入检验方法：基于位点模式 (site pattern) 的 D 统计量 (Durand *et al.*, 2011; Green *et al.*, 2010)、 f 统计量 (Durand *et al.*, 2011; Martin *et al.*, 2015)、基于基因树信息的卡方检验 (拓扑频率; Allman *et al.*, 2020)、枝长检验 (BLT; Suvorov *et al.*, 2021) 和 QuIBL (Edelman *et al.*, 2019), 以及 PhyloNet (Than *et al.*, 2008)。需要说明的是, 本文介绍的前两类方法仅可以判断基因渗入是否存在, 量化基因渗入比例需要其他模型, 而 PhyloNet 可以检测出类群间是否存在基因渗入并能够量化其遗传比例。此外, 本文主要介绍了适用于一般系统发育基因组学数据的方法, 即数据中每个物种通常只有单个或者极少数个体的数据, 一般没有全基因组测序, 而是几百到几千个不连锁的基因位点。因此, 本文不包括基于群体遗传(每个物种需要多个个体)或基于位点间连锁不平衡信息的基因渗入检验方法。目前大部分系统发育分析采用物种树方法, 因此, 本文主要关注能够同时考虑 ILS 和基因渗入的检验方法。下面将介绍各类方法的原理和优缺点, 并列出了相关软件或可参考教程。

一、基于位点模式 (site pattern) 统计量的检验方法

基于位点模式统计量的检验方法一般运用于分化时间较短的类群。使用时, 先筛出序列数据中的信息位点, 统计不同位点模式的出现频率, 用这些频率值进行基因渗入检验。这类方法一般假设位点之间无连锁, 因此非常适合简化基因组测序数据——每个基因片段长度很短, 可变位点较少 (Davey and Blaxter, 2010)。使用者也可以从每个基因片段中随机抽取单个可变位点, 或者用经验数据得到的统计量标准误差做检验 (见下文)。这类方法使用简单, 计算速度快, 但每次检验对物种个数有限制 (包括外群在内只能有 4 个或 5 个物种)。当物种总数超过限制时, 一般由研究者根据其它已知信息挑选关注物种, 或者穷尽系统发育树上所有的四分类单元树 (4-taxon tree) 进行多次检验 (Eaton and Overcast, 2020)。

最早被开发的是 D 统计量 (D -statistic), 又被称为 ABBA-BABA 检验 (Durand *et al.*, 2011; Green *et al.*, 2010)。该检验适用于四分类单元树 (图 1 A 和 B), 包括三个物种 (S_1, S_2, S_3) 和它们的外群物种 O , 它们之间的系统发育关系需满足 $((S_1, S_2), S_3), O$ 。这里, ABBA 和 BABA 表示两种位点模式: 每个字母代表一个物种在特定位点的碱基状态 (即 AGTC 中一种), ABBA 模式是指在该位点上物种 S_2 和 S_3 的碱基相同 (状态 B), 而物种 S_1 和外群物种 O 的碱基相同 (默认为祖先状态 A); BABA 则是物种 S_1 和 S_3 碱基相

同, 物种 S_2 保留祖先状态。当只有 ILS 存在时, 在序列中这两个模式应以相同的频率出现 (图 1)。如果 S_3 和任一其它两个内群物种存在基因渗入, 两种模式的位点数量将出现差异, 即 ABBA-BABA 检验。标准化的两种模式位点数量差异即 D 统计量:

$$D(S_1, S_2, S_3, O) = \frac{N_{ABBA}(S_1, S_2, S_3, O) - N_{BABA}(S_1, S_2, S_3, O)}{N_{ABBA}(S_1, S_2, S_3, O) + N_{BABA}(S_1, S_2, S_3, O)} \quad (1)$$

其中, N 指序列中符合某种模式的位点个数。如果所有位点间都无连锁、相互独立, D 值是否显著偏离零可以用二项式分布检验。如果有连锁 (如全基因组数据), 可以将基因组划分成区块, 然后使用刀切法 (jackknife) 计算经验 D 值的分布标准方差 (Efron, 1981; Kunsch, 1989)。当 D 值显著小于零时, 则表明物种 S_1 与 S_3 之间可能存在基因渗入; 当 D 值显著大于零时, 则表明物种 S_2 与 S_3 之间可能存在基因渗入。

然而 D 统计量有两个缺点: 无法定量反应基因渗入的程度和判断基因渗入的方向。后续研究在这两方面进行了拓展, 提出了 f 统计量, 包括 f_G (Durand *et al.*, 2011)、 f_{hom} 和 f_d (Martin *et al.*, 2015)。 f 统计量和 D 统计量在计算中分子相同 (公式 1), 只是标准化采用的分母不同。如果是近期的基因渗入, 这些统计量能够反应基因组上基因渗入的比例 (Zheng and Janke, 2018)。须说明的是, f_G 需要物种 S_3 中至少两个谱系的信息:

$$\hat{f}_G = \frac{S(S_1, S_2, S_3, O)}{S(S_1, S_{3a}, S_{3b}, O)} \quad (2)$$

其中, S_{3a} 和 S_{3b} 是来自 S_3 的两个谱系, 而 $S(S_1, S_2, S_3, O) = N_{ABBA}(S_1, S_2, S_3, O) - N_{BABA}(S_1, S_2, S_3, O)$ 。 $\hat{f}_{hom}$ 中用 S_3 替代 S_{3a} 和 S_{3b} 样本:

$$\hat{f}_{hom} = \frac{S(S_1, S_2, S_3, O)}{S(S_1, S_3, S_3, O)} \quad (3)$$

其中 $S(S_1, S_3, S_3, O) = N_{ABBA}(S_1, S_3, S_3, O)$, 即仅物种 O 和 S_1 有相同碱基的位点个数。值得一提的是, 如果有群体数据, 等位基因频率也可以用于计算 D 统计量和 f 统计量 (Martin *et al.*, 2013)。

D_{FOIL} 检验 (Pease and Hahn, 2015) 是将基于四分类单元树的 D 统计量扩展到五个物种 (图 1C 和 D), 进而推断基因渗入的方向。五个物种需满足的系统发育关系为 $((S_1, S_2), (S_3, S_4)), O$, O 为外群物种, 且物种 S_1 、 S_2 的分化时间晚于 S_3 、 S_4 (图 1C 和 D)。以五个物种中 S_2 和 S_3 之间存在基因渗入为例: 如果基因是从 S_2 渗入 S_3 , 那么基因树拓扑中 $((((S_2, S_3), S_1), S_4), O)$ 频率会增加, 即 (S_2, S_3) 因基因渗入而形成类群的姐妹群是“供

体”(物种 S_2) 的原有姐妹种 S_1 。反映到位点模式，BBBAA 的位点个数会增多 (图 1C)。反之，如基因从 S_3 渗入 S_2 ，则拓扑为 $((((S_2, S_3), S_4), S_1), O)$ 的基因树会增加， (S_2, S_3) 的姐妹群为 S_3 原有姐妹种 S_4 ，ABBBAA 模式的位点个数会增多 (图 1D)。考虑四个内群物种间所有可能的基因渗入，Pease *et al.* (2015) 设计了一套基于四个统计量的基因渗入方向检验 (详见原文)，统称为 D_{FOIL} 检验。

相关软件或参考教程：

1. D 统计量检验

基于 python 2，参考教程：<https://ipyrad.readthedocs.io/en/latest/API-analysis/cookbook-abba-baba.html>

基于 R，参考教程：https://github.com/simonhmartin/tutorials/blob/master/ABBA_BABA_whole_genome/README.md

2. f 统计量，可在 Dsuite 软件包 (Malinsky *et al.*, 2021) 中计算：

<https://github.com/millanek/Dsuite>

3. D_{FOIL} 统计量检验，参考：<http://www.github.com/jbpease/dfoil>

图 1. D 统计量和 D_{FOIL} 检验的位点模式 黑色粗线代表物种树，绿线为 ILS 导致的和物

种树拓扑不一致的基因树，蓝线为基因渗入产生的和物种树拓扑不一致的基因树，短红线指示基因树上碱基状态从 A 到 B 突变的谱系。(A) 和 (B) 展示在四分类单元树中，ILS 导致的两种不兼容基因树 (绿线)，分别产生 ABBA 和 BABA 两种模式的位点，而物种 S_2 和 S_3 间基因渗入 (A) 会增加 ABBA 模式位点的数量。(C) 和 (D) 展示五分类单元树不同基因渗入方向 (蓝色箭头) 会产生的基因树拓扑和对应的位点模式。

二、基于基因树信息的检验方法

基于位点模式统计量的方法往往假设一个位点上没有多次替换，并且不考虑碱基替换模型，因此适合分化时间较短的物种间基因渗入检验。相较而言，基于基因树信息的检验方法适用的时间尺度更长，可以检测更古老的基因渗入事件。因为这类方法仅使用基因树信息，不需要序列文件，使用者需要先用其它构树软件为数据集中每个基因片段单独建树。最简单的检验方法是基于三物种拓扑频率分布的卡方检验 (χ^2 test)。假设这三个物种的物种树为 $((S_1, S_2), S_3)$ ，ILS 可以产生两种和物种树拓扑不兼容的基因树： $((S_2, S_3), S_1)$ 和 $((S_1, S_3), S_2)$ ，且产生两种拓扑的概率相等 (图 2)。卡方检验是检验数据集中不兼容基因树拓扑的频率是否符合相等的理论预期。

Suvorov *et al.* (2021) 开发的枝长检验 (Branch Length Test, BLT) 更进一步利用了不兼容基因树中的枝长信息 (Suvorov *et al.*, 2021)。仅有 ILS 作用时，两种不兼容基因树上两个姐妹物种间遗传距离的理论预期是相同的：具有 $((S_2, S_3), S_1)$ 拓扑的基因树上，物种 S_2 和 S_3 的遗传距离，应该和具有 $((S_1, S_3), S_2)$ 拓扑的基因树上物种 S_1 和 S_3 的遗传距离相等。这里，姐妹物种间遗传距离就是基因树上两个物种所对应的分支枝长之和 (图 2 红线长度)，所以我们可以用 Wilcoxon 秩和检验 (Rank sum test) 来检验基因树的枝长是否符合 ILS 的理论预期。基因渗入会导致一种不兼容基因树有显著更小的平均外部枝长。

QuIBL (Quantifying Introgression via Branch Lengths) 也是基于基因树枝长分布的检验方法 (Edelman *et al.*, 2019)。与 BLT 利用外部枝长不同，QuIBL 检验的是内部分支枝长 (图 2 黄线)。根据溯祖理论，不兼容拓扑基因树上的两个内部节点对应着所有物种共同祖先群体内的溯祖事件 (图 2B 和图 2C 中黑色实心圆点)。因此，两个内部节点的距离应该符合溯祖理论推导出来的指数分布，且该分布只有一个决定性参数，即祖先群体的大小。而对于和物种树拓扑一致的基因树，第一个内部节点有可能对应着两个

姐妹物种在共同祖先群体内的溯祖事件 (图 2A 空心圆点)。因此, 两个内部节点的距离分布不再符合单参数指数分布, 而是两个概率分布的混合。如果存在基因渗入, 那么就存在由于基因渗入导致的不兼容拓扑基因树, 这些内部枝长也不符合单个参数的指数分布。QuIBL 对于三种拓扑结构的基因树分别进行内部枝长的检验来判断是否有基因渗入的存在。

可见, 以上基于基因树的基因渗入检验方法对物种个数的限制仍然存在, 即每次检验局限于三、四个物种的集合, 或者物种树上一个特定的内部分支。当物种个数超出限制时, 已有研究一般穷尽所有物种组合, 或者依次检验物种树所有内部分支, 并对结果进行多重检验校准。由于这类检验方法使用的基因树由单个基因片段构建, 因此, 当基因片段较短、信息位点较少时, 该方法的灵敏度可能会因基因树估计误差而降低, 进而导致检验结果偏向于保守。

相关软件或参考教程:

1. 卡方检验: 通过 MSCquartets R package 可视化 (Rhodes *et al.*, 2021), R 包参考: <https://CRAN.R-project.org/package=MSCquartets>。
2. QuIBL, 参考教程: <https://github.com/miriammiyaqi/QuIBL>

图 2. 基于基因树信息的检验方法 黑色粗线代表 3 分类单元的物种树: $((S_1, S_2), S_3)$, 内嵌细线条为基因树, 共三种可能拓扑, (A) 为兼容拓扑, (B)、(C) 为两种不兼容拓扑。内部节点 (圆点) 对应基因树上姐妹物种溯祖事件。(B) 和 (C) 中两个溯祖事件都发生在三个物种的共同祖先群体中 (实心圆点), (A) 中一个溯祖事件可能发生 S_1 和 S_2 的共同祖先群体 (空心圆点)。红色分枝长度为 BLT 检验针对的姐妹种遗传距离, 而黄线是 QuIBL 方法检验的内部分支枝长。

三、PhyloNet

PhyloNet 软件包于 2008 年首次发表 (Than *et al.*, 2008), 包括多种构建基因渗入历史的方法, 在系统发育基因组学中得到广泛应用。需要说明, PhyloNet 中的一些方法同样依赖于基因树信息, 考虑到 PhyloNet 和上述两类检验方法区别较大, 因此这里作单独介绍。其中一个重要区别是在基础模型方面。其它已介绍的检验方法不涉及具体的基因渗入模型。它们本质上是检验零假设 (仅有 ILS) 是否成立, 如果能够推翻零假设则意味着基因渗入的可能。基因渗入的具体模型可以是群体遗传学的经典基因流模型: 一个群体的基因拷贝在每一代中有一定比例迁移到另一个群体中, 这种迁移将持续一段时间 (Hey and Nielsen, 2004; Nielsen and Wakeley, 2001); 也可以是杂交成种模型: 两个祖先谱系各自贡献一定比例的遗传背景, 形成一个新的谱系 (Jiao *et al.*, 2021; Yu *et al.*, 2012)。PhyloNet 明确基于杂交成种的基因渗入模型, 而且将这个过程视为进化上的一次性事件, 没有经典基因流模型中的持续时间 (但是有模拟显示 PhyloNet 在群体遗传经典基因流模型下仍有较好表现; Wen and Nakhleh, 2018)。

另一个区别在于对物种数量的限制。上述两类方法都是针对 3-4 个物种集合的检验方法, 而 PhyloNet 是一种同时考虑 ILS 和基因渗入的“物种树”构建方法 (Wen *et al.*, 2018)。因此理论上, PhyloNet 对物种的个数没有限制, 仅仅是实际中的计算复杂度限制了可以应用 PhyloNet 数据集的大小 (见下文)。由于存在杂交成种事件, PhyloNet 进行系统发育重建的结果是一个网络树。网络树也可用 Newick 形式表示, 与严格的二叉树不同之处在于杂交节点有两个亲本节点 (Cao *et al.*, 2019)。以下面的网络树为例 (图 3):

$((((S_1:1.0)\#H1:1.0::0.7,S_3:1.0):5,(\#H1:1.0::0.3,S_2:1.0):4):1.0,S_4:1.0);$

其中“#H1”代表杂交节点, 节点编号后跟分枝长度, 双冒号后为遗传比例。图 3 显示了该网络可视化结果: 即物种 S_1 是杂交物种, 基因谱系有 70% 的可能来源于 S_3 , 30% 来源于 S_2 。

图 3. 网络树的可视化 蓝色分枝上数字代表杂交物种谱系来源比例。(A) Dendroscope (Huson and Scornavacca, 2012) 对拓扑结构的可视化。因为该程序不处理枝长信息，图中线条长度不对应 Newick 中枝长。(B) R 语言包 APE (Cardona *et al.*, 2008; Paradis and Schliep, 2019) 对同一网络树的可视化。Newick 格式网络树可用 `read.evonet` 函数读入，`plot.evonet` 函数绘图，然后结合 `nodeLabels` 函数显示杂交节点 “#H1”。

构建网络树比传统的系统发育树构建（即构建物种分化的二叉树）更为复杂。搜寻最优网络树不仅要优化二叉树的拓扑结构和枝长，还考虑杂交物种组合的各种可能。因此，构建网络树需要耗费更多的计算资源。计算资源的限制使得很多 PhyloNet 方法在超过 10 个类群和 3 个杂交节点的数据集上无法运行 (Yu and Nakhleh, 2015)，而且多数方法都需使用者指定最大杂交次数来限制搜索空间。然而，如何指定最大杂交次数在 PhyloNet 实际运用中仍较有争议 (Ferreira *et al.*, 2021)。另外，在传统的系统发育构建中，不同的拓扑结构对应的参数个数是相同的（如分支个数相同），而更多的杂交节点的网络树具有更多的参数。因此需要防止模型的过拟合（即推断出不存在的杂交事件）。在 PhyloNet 包中，基于最大似然法的方法通过信息准则来选择杂交节点的个数（如 `InferNetwork_ML`; Yu *et al.*, 2014），而基于贝叶斯框架的方法（如 `MCMC_GT`; Wen *et al.*, 2016）通过杂交个数的先验分布来控制模型复杂度。

PhyloNet 是一个非常灵活的软件包。针对不同输入数据类型，有基于已构建的基因树的方法，也有直接从比对序列开始的方法（表 1）。物种只有一个谱系或者有多个谱系都可以运行，而且允许有缺失数据（即一些位点没有包含所有的谱系）。为降低基因树构建误差的影响，PhyloNet 基于基因树的方法都可以对基因树上支持度低于指定阈值的节点进行压缩（-b 选项），以及对基因树分别设置权重。根据设置的返回网络数量（-n 选

项), PhyloNet 可以提供搜索过程中排名靠前的不同网络拓扑/树。

表 1. PhyloNet 软件包中各种推断网络树的方法

PhyloNet 命令	原理	输入	输出
InferNetwork_MP	简约法	基因树拓扑	网络拓扑及谱系来源比例
InferNetwork_ML (_bootstrap/_cv)	最大似然法	基因树	网络树
InferNetwork_MPL	最大伪似然法	基因树拓扑	网络树
MCMC_GT	贝叶斯	基因树	网络树
MCMC_SEQ	贝叶斯	序列	网络树、基因树及碱基替换相关参数
MCMC_BiMarkers	贝叶斯	双等位基因序列	网络树及突变率
MLE_BiMarkers	最大似然法	双等位基因序列	网络树及突变率

注：网络树包括网络拓扑、枝长和杂交节点谱系来源比例。

InferNetwork_MP 方法 (Yu *et al.*, 2013) 是在 MDC (Minimizing the number of deep coalescences; Maddison, 1997) 准则下, 根据基因树的拓扑结构推断网状树。这种方法计算速度较快, 即使处理多达 40 个类群和 5 个杂交节点的数据集, 大部分情况也可以在 24 h 内完成计算。缺点是估计的网络树只有拓扑结构和杂交节点的谱系来源比例, 没有枝长信息, 而且对于 ILS 比较多的网络树, 基于 MDC 准则的最大简约法在统计上会有不一致性问题(consistency issue; Than and Rosenberg, 2011; Yu *et al.*, 2013)。

InferNetwork_ML (_bootstrap/_CV; Yu *et al.*, 2014) 是基于最大似然法的网络树构建方法。输入的基因树可以仅有拓扑结构, 也可以包含枝长信息 (需要以有效群体大小为单位的等距基因树)。这个方法优点在于没有统计不一致性的问题, 而且能够输出带有枝长的网络树, 但缺点是计算瓶颈—物种越多, 可能的网络结构越多, 所以不建议在超过 10 个类群和 3 个杂交节点的数据集上运行 (Yu and Nakhleh, 2015)。ML 方法有两个拓展: InferNetwork_ML_bootstrap 通过输入每棵基因树的自展重复得到网络树每个节点的支持度; InferNetwork_ML_CV 通过分割数据进行交叉验证来确定最优的杂交节点个数。InferNetwork_MPL 方法 (Yu and Nakhleh, 2015) 是使用基因树拓扑结构信息的伪似然估计 (maximum pseudo-likelihood), 原理上和物种树构建方法 MP-EST 类似 (Liu *et al.*, 2010)。MPL 极大的解决了最大似然法的计算瓶颈问题, 因此也是目前系统发育基因组学中运用得最普遍的方法。

MCMC_GT (Wen *et al.*, 2016)、MCMC_SEQ (Wen and Nakhleh, 2018) 和

MCMC_BiMarkers (Zhu *et al.*, 2018) 是三种基于贝叶斯框架的方法。三种方法都使用了可逆跳跃马尔可夫链蒙特卡罗法 (reversible-jump Markov chain Monte Carlo; RJMCMC; Green, 2003) 来处理不同网络结构参数个数不一致的问题。从命名可以看出这三种方法的输入数据类型分别为: GT 对应基因树 (仅需拓扑结构), SEQ 对应序列, BiMarkers 对应双等位基因序列 (如 SNPs, AFLPs 等)。后面两种从序列数据同时推断基因树和网络树的方法在理论上优势较大, 避免了基因树估计过程带来的误差 (Wen and Nakhleh, 2018)。但是因为计算量的原因, 如果网络结构中有大于 5-6 个类群和 3-4 个杂交节点, MCMC_SEQ 和 MCMC_BiMarkers 将运行困难 (Zhu and Nakhleh, 2018)。MCMC_GT 中提供了利用伪似然值 (-pseudo) 的选项来加速计算。对于双等位基因序列, 另外一个 MLE_BiMarkers (Zhu and Nakhleh, 2018) 方法中也有利用伪似然值加速运行的选项, 可以用于物种数目较多 (> 100) 的数据集。

最后, 需要说明, 由于网络树巨大的搜索空间, 一般 PhyloNet 分析都需要多次独立运行, 确保结果一致。

相关软件或参考教程:

PhyloNet 软件包下载及使用: <https://bioinfocsc.rice.edu/PhyloNet>

总结

大量研究发现物种间基因渗入普遍存在, 并且会影响系统发育关系的准确推断。因此, 本文简要讨论了基于位点模式统计量、基于基因树信息以及 PhyloNet 等几类基因渗入检验方法, 并列出了相关软件的参考教程。

数据获取

针对如何在 R 语言环境中使用这些分析方法 (具体步骤及展示结果), 我们提供了一个在线示例文档: <https://github.com/YingDings/Introgression-Detection-Methods/blob/main/example.pdf>

致谢

本研究得到国家自然科学基金面上项目 (编号 31970407) 资助。

参考文献

1. Allman, E. S., Mitchell, J. D. and Rhodes, J. A. (2020). [Gene tree discord, simplex plots, and statistical tests under the coalescent](#). *Syst Biol*
2. Baptiste, E., van Iersel, L., Janke, A., Kelchner, S., Kelk, S., McInerney, J. O., Morrison, D. A., Nakhleh, L., Steel, M. and Stougie, L. (2013). [Networks: expanding evolutionary thinking](#). *Trends Genet* 29(8): 439-441.
3. Blair, C. and Ané, C. (2020). [Phylogenetic trees and networks can serve as powerful and complementary approaches for analysis of genomic data](#). *Syst Biol* 69(3): 593-601.
4. Cai, L., Xi, Z., Lemmon, E. M., Lemmon, A. R., Mast, A., Buddenhagen, C. E., Liu L. and Davis, C. C. (2021). [The perfect storm: Gene tree estimation error, incomplete lineage sorting, and ancient gene flow explain the most recalcitrant ancient angiosperm clade, Malpighiales](#). *Syst Biol* 70(3): 491-507.
5. Cao, Z., Liu, X., Ogilvie, H. A., Yan, Z. and Nakhleh, L. (2019). [Practical aspects of phylogenetic network analysis using phylonet](#). *bioRxiv* 746362.
6. Cardona, G., Rosselló, F. and Valiente, G. (2008). [Extended Newick: it is time for a standard representation of phylogenetic networks](#). *BMC bioinformatics* 9(1): 1-8.
7. Dasmahapatra, K. K., Walters, J. R., Briscoe, A. D., Davey, J. W., Whibley, A., Nadeau, N. J., Zimin, A. V., Hughes, D. S., Ferguson, L. C. and Martin, S. H. (2012). [Butterfly genome reveals promiscuous exchange of mimicry adaptations among species](#). *Nature* 487(7405): 94.
8. Davey, J. W. and Blaxter, M. L. (2010). [RADSeq: next-generation population genetics](#). *Brief Funct Genomics* 9(5-6): 416-423.
9. 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.
10. Eaton, D. A. and Overcast, I. (2020). [ipyrad: Interactive assembly and analysis of RADseq datasets](#). *Bioinformatics* 36(8): 2592-2594.
11. Edelman, N. B., Frandsen, P. B., Miyagi, M., Clavijo, B., Davey, J., Dikow, R. B., García-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.
12. Efron, B. (1981). [Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods](#). *Biometrika* 68(3): 589-599.
13. Ferreira, M. S., Jones, M. R., Callahan, C. M., Farelo, L., Tolesa, Z., Suchentrunk, F., Boursot, P., Mills, L. S., Alves, P. C. and Good, J. M. (2021). [The legacy of recurrent introgression during the radiation of hares](#). *Syst Biol* 70(3): 593-607.
14. Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, R. M., Neafsey, D. E., Sharakhov, I. V., Jiang, X., Hall, A. B., Catteruccia, F. and Kakani, E. (2015). [Extensive introgression in a malaria vector species complex revealed by phylogenomics](#). *Science* 347(6217).
15. Grant, P. and Grant, B. (1992). [Hybridization of bird species](#). *Science* 256(5054): 193-197.
16. Green, P. J. (2003). [Trans-dimensional markov chain monte carlo](#). *Highly Structured Stochastic Systems* 27(2003):179-198.
17. Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H., Hansen, N. F., Durand, E. Y., Malaspinas, A. S., 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.
18. Hey, J. and Nielsen, R. (2004). [Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*](#). *Genetics* 167(2): 747-760.
19. Hey, J. and Nielsen, R. (2007). [Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics](#). *Proc Natl Acad Sci U S A* 104(8): 2785-2790.

20. Husnik, F. and McCutcheon, J. P. (2018). [Functional horizontal gene transfer from bacteria to eukaryotes](#). *Nat Rev Microbiol* 16(2): 67-79.
21. Huson, D. H. and C. Scornavacca (2012). [Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks](#). *Syst Biol* 61(6): 1061-1067.
22. Jeffroy, O., Brinkmann, H., Delsuc, F. and Philippe, H. (2006). [Phylogenomics: the beginning of incongruence?](#) *Trends Genet* 22(4): 225-231.
23. Jiao, X., Flouris, T. and Yang, Z. (2021). [Multispecies coalescent and its applications to infer species phylogenies and cross-species gene flow](#). *Natl Sci Rev*
24. Kunsch, H. R. (1989). [The jackknife and the bootstrap for general stationary observations](#). *Ann Stat* 1217-1241.
25. 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(1): 1-18.
26. Long, C. and Kubatko, L. (2018). [The effect of gene flow on coalescent-based species-tree inference](#). *Syst Biol* 67(5): 770-785.
27. Maddison, W. P. (1997). [Gene trees in species trees](#). *Syst Biol* 46(3): 523-536.
28. Malinsky, M., Matschiner, M. and Svardal, H. (2021). [Dsuite - Fast D - statistics and related admixture evidence from VCF files](#). *Mol Ecol Res* 21(2): 584-595.
29. Mallet, J. (2005). [Hybridization as an invasion of the genome](#). *Trends Ecol Evol* 20(5): 229-237.
30. Mallet, J. (2007). [Hybrid speciation](#). *Nature* 446(7133): 279-283.
31. Mallet, J., Besansky, N. and Hahn, M. W. (2016). [How reticulated are species?](#) *Bioessays* 38(2): 140-149.
32. Mao, Y., Economo, E. P. and Satoh, N. (2018). [The roles of introgression and climate change in the rise to dominance of *Acropora* corals](#). *Curr Biol* 28(21): 3373-3382. e3375.
33. 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.
34. 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.

35. Nielsen, R. and Wakeley, J. (2001). [Distinguishing migration from isolation: a Markov chain Monte Carlo approach](#). *Genetics* 158(2): 885-896.
36. Oziolor, E. M., Reid, N. M., Yair, S., Lee, K. M., VerPloeg, S. G., Bruns, P. C., Shaw, J. R., Whitehead, A. and Matson, C. W. (2019). [Adaptive introgression enables evolutionary rescue from extreme environmental pollution](#). *Science* 364(6439): 455-457.
37. Paradis, E. and Schliep, K. (2019). [ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R](#). *Bioinformatics* 35(3): 526-528.
38. Pease, J. B. and Hahn, M. W. (2015). [Detection and polarization of introgression in a five-taxon phylogeny](#). *Syst Biol* 64(4): 651-662.
39. Rhodes, J. A., Baños, H., Mitchell, J. D. and Allman, E. S. (2021). [MSCquartets 1.0: quartet methods for species trees and networks under the multispecies coalescent model in R](#). *Bioinformatics* 37(12): 1766-1768.
40. Richards, E. J., Brown, J. M., Barley, A. J., Chong, R. A. and Thomson, R. C. (2018). [Variation across mitochondrial gene trees provides evidence for systematic error: how much gene tree variation is biological?](#) *Syst Biol* 67(5): 847-860.
41. Singhal, S., Colston, T. J., Grundler, M. R., Smith, S. A., Costa, G. C., Colli, G. R., Moritz, C., Pyron, R. A. and Rabosky, D. L. (2021). [Congruence and conflict in the higher-level phylogenetics of squamate reptiles: an expanded phylogenomic perspective](#). *Syst Biol* 70(3): 542-557.
42. Suvorov, A., Kim, B. Y., Wang, J., Armstrong, E. E., Peede, D., D'Agostino, E. R., Price, D. K., Wadell, P., Lang, M. and Courtier-Orgogozo, V. (2021). [Widespread introgression across a phylogeny of 155 *Drosophila* genomes](#). *BioRxiv* 2020.2012.2014.422758.
43. Suvorov, A., Scornavacca, C., Fujimoto, M. S., Bodily, P., Clement, M., Crandall, K. A., Whiting, M. F., Schrider, D. R. and Bybee, S. M. (2021). [Deep ancestral introgression shapes evolutionary history of dragonflies and damselflies](#). *bioRxiv* 2020.2006.2025.172619.
44. Than, C., Ruths, D. and Nakhleh, L. (2008). [PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships](#). *BMC Bioinformatics* 9(1): 1-16.

45. Than, C. V. and Rosenberg, N. A. (2011). [Consistency properties of species tree inference by minimizing deep coalescences](#). *J Comput Biol* 18(1): 1-15.
46. Tian, Y. and Kubatko, L. S. (2016). [Distribution of coalescent histories under the coalescent model with gene flow](#). *Mol Phylogenet Evol* 105: 177-192.
47. Wen, D. and Nakhleh, L. (2018). [Coestimating reticulate phylogenies and gene trees from multilocus sequence data](#). *Syst Biol* 67(3): 439-457.
48. Wen, D., Yu, Y. and Nakhleh, L. (2016). [Bayesian inference of reticulate phylogenies under the multispecies network coalescent](#). *PLoS Genet* 12(5): e1006006.
49. Wen, D., Yu, Y., Zhu, J. and Nakhleh, L. (2018). [Inferring phylogenetic networks using PhyloNet](#). *Syst Biol* 67(4): 735-740.
50. Yu, Y., Barnett, R. M. and Nakhleh, L. (2013). [Parsimonious inference of hybridization in the presence of incomplete lineage sorting](#). *Syst Biol* 62(5): 738-751.
51. Yu, Y., Degnan, J. H. and Nakhleh, L. (2012). [The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection](#). *PLoS Genet* 8(4): e1002660.
52. Yu, Y., Dong, J., Liu, K. J. and Nakhleh, L. (2014). [Maximum likelihood inference of reticulate evolutionary histories](#). *Proc Natl Acad Sci U S A* 111(46): 16448-16453.
53. Yu, Y. and Nakhleh, L. (2015). [A maximum pseudo-likelihood approach for phylogenetic networks](#). *BMC Genomics* 16(10): 1-10.
54. Zheng, Y. and Janke, A. (2018). [Gene flow analysis method, the D-statistic, is robust in a wide parameter space](#). *BMC Bioinformatics* 19(1): 10.
55. Zhu, J. and Nakhleh, L. (2018). [Inference of species phylogenies from bi-allelic markers using pseudo-likelihood](#). *Bioinformatics* 34(13): i376-i385.
56. Zhu, J., Wen, D., Yu, Y., Meudt, H. M. and Nakhleh, L. (2018). [Bayesian inference of phylogenetic networks from bi-allelic genetic markers](#). *PLoS Comput Biol* 14(1): e1005932.
57. 程祥, 李玲玲, 肖钰, 陈晓阳 and 胡新生 (2020). "种间基因渐渗检测方法及其应用研究进展." [中国科学:生命科学](#) 50(12): 1388-1404.

58. 于黎, 张亚平. (2006). 系统发育基因组学——重建生命之树的一条迷人途径. *遗传* 28(11): 1445-1450.