种间基因渐渗检测方法及其应用研究进展

程祥^{1,2}, 李玲玲^{1,2}, 肖钰^{1,2}, 陈晓阳^{1,2}, 胡新生^{1,2*}

1. 华南农业大学林学与风景园林学院, 广州 510642;

2. 广东省植物种质资源与利用重点实验室, 广州 510642

* 联系人, E-mail: xinsheng@scau.edu.cn

收稿日期: 2020-04-10; 接受日期: 2020-06-22; 网络版发表日期: 2020-09-01

中央财政林业改革发展资金项目(批准号: [2018]GDTK-08)和华南农业大学人才引进启动项目(批准号: 4400-K16013)资助

摘要 伴随着谱系分选进程, 从共同祖先群体分裂后形成的子代群体间的基因流逐渐衰减, 并可能延续至新物种形成. 种间基因组序列差异或种间群体基因频率的变异部分地记录了这些种群历史和当代动态过程, 可以根据种间基因组序列变异或基因频率变异推测这些发生过的事件. 本文系统探讨了检测种间基因渐渗的方法, 包括基于种间群体基因频率变异分析(F_{st} , LD, STRUCTURE, AFS, TreeMix)和基于种间DNA序列差异分析(ABBA-BABA, IM, IIM, ABC)方法, 分析了这些方法的原理和应用局限性, 拓展这些方法中的一些重要假设将是今后研究任务之一, 包括基因重组效应、区分配子与合子迁移的贡献、交配系统与基因流互作以及适应性基因迁移的阻碍作用. 澄清这些影响有助于深入理解物种形成过程.

关键词 基因渐渗, 物种形成, 隔离-迁移模型, 谱系分选, 基因树

基因流指基因在空间上从一个位置或群体移动到另一个位置或另一个群体的过程. 当群体间存在频繁的基因交换时, 基因流(尤其是适应性的功能基因流)可减少群体间的遗传分化, 维持同一物种的完整性^[1]. 当群体因歧化选择等进化过程作用而逐渐加大遗传分化时, 分化的群体终究趋于生殖隔离, 形成新物种^[2]. 起初群体间的基因流逐渐减弱, 基因流维持物种完整性的功能也就衰减, 这时遗传分化的群体或物种之间通过杂交及随后的回交来实现物种间基因交换, 这一过程称为种间基因渐渗. 在生物学概念上, 种内群体间的基因流和种间基因渐渗是两个不同层次上的群体基因交换, 有着不同的群体遗传和进化基础.

联系到物种形成与地理关系时, 物种形成可以在两个地理位置上完全隔离的群体间发生, 即异域物种形成, 也可以在地理上相邻的群体间发生, 即邻域物种形成, 甚至可在分布重叠的群体之间发生, 即同域物种形成. 在物种形成过程中, 地理隔离的作用就是直接阻碍群体间的基因流, 从而间接强化自然选择和漂变作用. 在无地理隔离条件下, 自然选择是促进分化群体生殖隔离的主要动力, 漂变的作用常常可以忽略^[3], 因此, 基因流和基因适应性歧化选择是物种形成的两个重要过程. 当群体间基因流大于一定的阈值, 迁移的等位基因可以掩埋接受群体的适应性等位基因, 使得群体遗传分化变小^[4], 维持物种的完整性, 且在相

引用格式: 程祥, 李玲玲, 肖钰, 等. 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, 2020, 50: 1388–1404

Cheng X, Li L L, Xiao Y, et al. Advances in the methods of detecting interspecific gene introgression and their applications (in Chinese). Sci Sin Vitae, 2020, 50: 1388–1404, doi: 10.1360/SSV-2020-0126

同的迁移率条件下, 选择系数大的有利等位基因有更快的速率来维持物种的完整性^[5]. 因此, 物种形成意味着控制生殖隔离的相关基因选择适应性强度大于群体间的基因流^[1], 尽管如此, 基因流, 尤其中性基因流, 在物种形成的过程中仍在进行.

上述分析指出, 基因流和自然选择是物种形成过程中的两个主要相互作用过程. 伴随着谱系分选过程, 起始的种内群体间基因流逐渐演变成种间基因渐渗. 许多因素参与改变种间基因渐渗, 例如, 基因流与重组互作, 可导致早晚期迁移的基因年龄或DNA片段长度存在差异, 同一条染色体上不同部位可存在不同程度的种间基因交换^[6]. 不同交配系统的物种可产生不相等的种间基因渐渗^[7,8]. 种间遗传背景的不亲和性, 如DMI(Dobzhansky-Muller-incompatibility)、物种所处的生态环境、生殖生态学特征、基因迁移载体、空间物理障碍等因素都可能影响种间基因渐渗. 检测和估计种间基因流有助于理解遗传保护单位、物种的系统分类地位、杂种和生物多样性等问题. 因此, 检测种间基因渐渗在进化生物学和保护遗传学上具有重要意义.

已有的相关文献显示, 许多直接和间接方法可用于推测种内群体间的基因流, 如利用无线电跟踪或者雷达跟踪, 应用全球定位技术和微型化技术, 使用卫星遥感技术来追踪动物等^[9], 这些方法的缺点是非常耗时, 且常常回报有限. 理论上, 部分直接方法也可应

用于检测种间基因渐渗, 但很少有报道, 更为普遍的是应用间接法估计群体间基因流. 表1简要概括了应用分子标记估计种内群体间基因流的方法^[28], 包括估计单世代平均迁移数($N_e m$)、相对迁移率(m)、迁移距离和扩散方差等. 这些方法的前提假设不同, 如用 F_{st} 估计迁移是在假定迁移-漂变平衡条件下, 估计历史上发生的群体间基因流^[10], 目前也有建议采用IM(isolation-with-migration)方法估计种内群体间基因流^[29]. 由于种内群体间基因流与种间基因渐渗的差异, 有些方法难以适合同时分析两个层面的基因流, 如很少有报道应用距离隔离和扩散方差估算等方法分析种间基因渐渗. 亲本分析用于估计当代花粉扩散也主要局限在种内群体间或种内群体内基因流^[30,31], 但基因溯祖理论框架可用于种内群体间基因流(基因树)^[22]和种间基因渐渗(物种系统树)^[32].

至今已有许多统计方法用于分析种间基因渐渗. Sousa和Hey^[6]系统地评价了应用基因组数据分析种间基因流的不同统计方法和模型. 王茜等人^[33]讨论了应用核质分子标记的连锁不平衡来检测种间基因渐渗. 本文将从检测方法的应用角度分析种间基因渐渗, 并注重分析基于基因组序列的方法原理、有关统计估计问题及近5年实际应用的研究进展: 首先讨论纯统计检测方法及有关软件应用, 其次讨论基于特定进化模型的检测方法, 评价每种方法的优缺点, 结合案例分析动植物种间基因渐渗, 为深入理解物种形成过程中基

表 1 种内群体间基因流估计方法

Table 1 Methods of estimating gene flow among populations of the same species

估计参数	理论模型	统计量	样本	参考文献
$N_e m$ 或 m	岛屿模型, 迁移-漂变平衡	F_{st}	空间上两个或更多群体抽样	[10,11]
m, N_e	岛屿模型	F_{st}	时间上两个或更多世代群体抽样	[12]
m	单群体	亲本基因来源分析	单个群体多位点基因型分型	[13]
m	分配检验法	混杂分析, 即单株基因组成源于不同群体概率分布	空间上两个或更多群体抽样	[14~17]
m	私有基因有限扩散	回归分析	空间上多群体抽样	[18,19]
m	岛屿模型	似然函数估计	空间上多群体抽样	[20,21]
m	基因树	F_{st}	空间上多群体抽样	[22]
M_p, θ_i	基因树	似然函数估计	两群体样本	[23]
距离隔离效应测定	距离隔离模型	遗传距离与对数地理距离的回归分析	空间上多群体抽样	[24,25]
花粉迁移距离	二世代方法	Φ_{FT}	母本和种子两个世代样本	[26]
基因扩散距离方差	杂交带(基因频率随距离渐变)	特征长度与扩散距离方差关系	空间上多群体抽样	[27]

因流的角色提供参考。

1 纯统计检测

1.1 F_{st} 统计量分布

F_{st} 指标准化的群体间基因频率的方差^[10], 其本身无任何进化意义。 F_{st} 可用于解释群体遗传分化程度, 所以又被称为群体遗传分化系数, 可反映群体内随机从两个个体各抽取一个等位基因, 它们来自同一祖先的概率, 取值范围在0和1之间($0 \leq F_{st} \leq 1$)。 当与进化动力连接后, F_{st} 可用来估计群体遗传参数^[34], 其中以前应用最广的就是估计群体间单世代的平均基因流数量, 即 $N_e m = (1/F_{st} - 1)/4$ 。 该理论是在岛屿模型假设下(没有选择、亚群体的数量无穷大并且有相同有效群体大小、每个群体迁移率相等), 群体处于迁移-漂变平衡等条件下获得的。 由于多数天然群体很少符合这些条件^[35], 也有学者提倡不用这方法估计迁移数^[29]。 尽管如此, 由于 F_{st} 在统计学上的计算简洁性以及有明确的生物学意义, 仍被广泛应用于比较种内群体间和种间群体的基因组分化研究。

由于是统计量, F_{st} 估计受样本容量影响, Weir^[34]建议的应用方差分析方法估计较好地考虑到了样本容量的影响, 可测验 F_{st} 是否显著。 实际比较两个物种基因组分化时, 可以通过沿染色体计算每个SNP(single nucleotide polymorphisms)位点 F_{st} 数值或计算一定大小的滑动窗口长度的 F_{st} 均值获得 F_{st} 沿染色体变异模式。 数值较大的DNA区域表示种间遗传分化大, 可能由于歧化选择导致, 数值小的基因组片段可能是由于基因渐渗或类似的选择所致, 但不直接估计 $N_e m$ 值。

根据种内群体间基因组数据分析群体分化 F_{st} 分布, 近几年在动植物研究方面都有许多报道, 如海岛*Pristionchus pacificus*线虫^[36]、珊瑚礁鱼鲷科(Serranidae)低纹鲷属内亚种(*Hypoplectrus* spp)^[37]、树袋熊(*Phascolarctos cinereus*)^[38]等种内群体间基因组分化。 但在种间基因组分化方面, F_{st} 分布应用于不同的研究目的, 例如, 利用基因组SNP分析短叶丝兰(*Yucca brevifolia*)和*Y. jaegeriana*种间遗传分化($F_{st}=0.25$)^[39], 不同生态型的红色交嘴雀(*Loxia curvirostra*)和白翅交嘴雀(*L. leucoptera*)基因组差异^[40]; 利用斑唧鹀(*Pipilo maculatus*)和领唧鹀(*P. ocai*)种间具有极大 F_{st} 值的特定位点, 在两个杂交带找到了类似于种间基因渐渗的基因

组区段^[41]。 Walsh等人^[42]分析尖尾雀(*Ammospiza caudacuta*)和*A. nelsoni*雀种间基因渐渗, 发现异域物种基因组间分化接近2倍的同域物种间基因组分化。 Murray等人^[43]分析白厚皮桉(*Eucalyptus albens*)和铁木桉(*E. sideroxylon*)种间基因组分化 $F_{st}=0.15$, 比种内群体间基因组分化 $F_{st}=0.01 \sim 0.02$ 大, 但认为基因组分化源于距离隔离和环境隔离。 近期, Battey^[44]发现, 棕煌蜂鸟(*Selasphorus rufus*)、亚伦的蜂鸟(*S. sasin*)及星蜂鸟(*S. calliope*)三个种间Z染色体的 F_{st} 要比常染色体的大, 认为与种间生殖隔离有关。 值得注意的是, 在DNA特定的区域, 不同的进化过程可能会产生类似的 F_{st} 模式, 如种间群体有相同的自然选择类型和频繁的种间基因渐渗都可导致小的基因组分化, 因此仅从基因组 F_{st} 的变化模式推测基因渐渗过程可能会带来偏差。

1.2 LD统计量分布

LD(linkage disequilibrium)通常指两个基因位点的配子连锁不平衡, 也可包含高阶基因型连锁不平衡^[34], 反映了两个位点的基因频率或基因型频率相关性。 同 F_{st} 统计量类似, 四种基本进化动力(突变、选择、漂变及迁移)及交配系统可影响LD, 重组减弱LD。 理论和模拟研究证明, 只有当存在一定的选择上位性时, 最大的基因型LD可以大于低阶配子LD, 其他的进化过程(包括线性的自然选择加性模型)都有相反的结论^[45], 因此, 这种关系可用于全基因组筛选具有相互作用的SNP位点^[46,47]。 根据群体遗传学有关迁移对LD产生性质及迁移与重组互作的结果, 以下特点可用来推测种间基因渐渗发生情况: (i) LD区段长度较大, 全基因组LD片段长度并不呈反“J”分布, 如趋于正态分布, 有可能近期发生种间基因渐渗, 类似于群体混杂的结果; (ii) LD片段很短, 全基因组LD长度呈“J”分布, 多数LD片段很短, 少数LD片段较长, 如人类基因组Hap-Map项目调查的CHB(Han Chinese Beijing)群体^[48], 说明群体经历了很长的历史, 重组将早期LD片段打断变小, 若种间存在基因渐渗, 可能发生在早期, 如在谱系分选过程中的并系(paraphyly)或更早的复系(polyphyly)阶段; (iii) 当种间基因相互渐渗程度不相等时, 种间全基因组LD片段平均长度及片段长度的分布应存在差异。

除了早期的有关人类群体基因组研究报道外, 应用上述特征分析动植物种间基因渐渗的研究报道很

少. 近期多数报道基因组局部选择与LD的互作, 鉴定/筛选与物种分选有关的基因位点, 这些位点呈现出高度的种间遗传分化. 如Reddy等人^[49]比较陆地棉(*Gossypium hirsutum*)和海岛棉(*G. barbadense*)栽培种间基因组的遗传多样性, 发现两物种存在保守的LD片段及单倍型, 证明它们在驯化过程中存在一致的进化模式. Stankowski等人^[50]分析*Mimulus aurantiacus*种不同授粉生态型的杂交带, 分子标记间存在低程度的LD, 认为对生态型形成有贡献的位点分散在基因组不同的位置. Schilling等人^[51]理论模拟显示, 在物种形成过程中, 由单位点到多位点耦合增加了LD及对基因流的障碍. 这些结果主要反映了选择与迁移的互作效应, 种内选择位点区域LD形成增强了对种间基因渐渗的生物学障碍.

1.3 非先验识别种间群体结构分析

当不同种间群体被视为高度遗传分化的群体时, 也可应用种内群体结构分析来分析种间群体分化. 对要分析的种间群体的每一个体, 事先不假定其来源, 根据传统群体概念来界定群体以及参数的贝叶斯先验假设分析多位点基因型数据, 在不同群体数假设下, 计算合理的群体间遗传结构和群体内遗传组成, 之后再确定每个个体的来源, 判断可能的种间杂交个体. 类似于 F_{st} 和LD分析, 这方面群体结构分析没有涉及到进化动力过程. 目前已有一些软件用于这类分析, 如基于Bayes途径的STRUCTURE^[14]和基于极大似然估计分析的Admixture^[52]及Frappe^[53]等, 本文只简要介绍最常用的STRUCTURE分析原理及其应用.

事先不假定每个个体来源, 对于给定的群体数(K), 每个群体用一组位点上的等位基因频率来描述, 将每个个体的基因型随机分配到各个假定的群体, 同时计算群体的等位基因频率, 依据群体概念(Hardy-Weinberg平衡, HWE)及位点连锁平衡($LD=0$)确定群体组成. 具体计算是根据给定的群体数 K 和二倍体基因型数据 X , 所有群体的等位基因频率 P , 建立似然函数 $Pr(X|K, P)$, 再根据先验分布假设 $Pr(K)$ (均匀分布)和 $Pr(P)$ (Dirichlet分布), 建立后验分布函数 $P(K, P|X)$, 再根据所得到的后验分布抽样数值计算 K 和 P . 在假定每个个体的基因位点可能存在来自不同群体时(个体混合模型), 其参数为 Q 向量, 从后验分布 $P(K, P, Q|X)$ 也可以同时估计参数 Q . 计算的最后目的是通过分配个体

每个基因到给定的群体内, 获得贝叶斯估计, 使得尽量达到群体内位点间连锁平衡及HWE^[13], 合理的群体数 K 采用 $\log(Pr(X|K))$ 对 K 的二阶导数除以 $\log(Pr(X|K))$ 的标准差 $sd(\log(Pr(X|K)))$ 得到 ΔK 的极大值确定, $\log(Pr(X|K))$ 的标准差 $sd(\log(Pr(X|K)))$ 由重复计算获得^[54].

STRUCTURE被广泛地应用于种内群体遗传结构研究中, 理论上也可用于种间群体结构分析. 在谱系分选过程中, 最后形成的单一物种可以被看作是一个遗传分化的群体, 当对两个或多个物种(群体)分析时, 用STRUCTURE可以估计每个个体的遗传组成在不同种间(群体)的分布比率 Q , 这些比率 Q 估值可以判断是否存在种间基因渐渗现象.

目前已有许多报道应用STRUCTURE分析动植物种间杂交, 例如, Miralles等人^[55]分析南非无须鳕(*Merluccius capensis*)和游鱼(*M. paradoxus*)的杂交, Fogelqvist等人^[56]分析种柳树种间杂交(*Salix dasyclados*, *S. schwerinii*及*S. viminalis*), Jordan等人^[57]分析了异交种紫萼路边青(*Geum rivale*)和自交种欧亚路边青(*G. urbanum*)的种间杂交, López-Caamal等人^[58]分析了*Tithonia tubaeformis*和圆叶肿柄菊(*T. rotundifolia*)杂交现象等. 同样, 这类分析提供了是否存在种间基因渐渗的证据, 但仍不能估计种间基因渐渗程度.

1.4 ABBA-BABA检验

ABBA-BABA模型提供了一种新的途径来分析种间基因渐渗, 该方法起初是由Green等人^[59]提出的, 用于分析现代人和尼安德特人在核基因组水平上的遗传混杂, 采用Patterson D 统计量来测验三个遗传关系密切的群体. Durand等人^[60]进一步推导出在祖先群体为非随机交配系统下, 用 D 统计量检验群体混杂的不同表达式. 该统计检验的基本原理是假设4个物种存在系统进化树(((P1, P2), P3), O), 进化树中P1和P2为两个关注的物种, 它们从共同祖先群体或物种P3分化而来, O为遗传关系更远的外类群或物种(outgroup), 从每个物种获得基因组序列数据并进行比对分析, 考虑每个SNP位点仅存在两个等位基因, 对于任一个SNP位点, 假定来自O群体DNA的等位基因为野生型“ A ”, 由突变衍生出来的等位基因为“ B ”, 因此对给定的系统树(((P1, P2), P3), O), 单个SNP位点基因树可以是ABBA模式, 也可以是BABA模式(图1), 在P1和P2的共同祖先P3群体为随机交配条件下, 两种基因树模式发

图1 在同一系统树((P1, P2), P3), O)下的两种基因树模式(ABBA, BABA)^[60]. P1和P2为遗传关联的物种, P3为P1和P2共同祖先群体或物种, O为外类群, 用于确定野生型等位基因, 基因A为野生型等位基因, B为突变型等位基因

Figure 1 Two distinct gene trees (ABBA, BABA) under a common species tree ((P1, P2), P3), O^[60]. P1 and P2 are genetically related species. P3 is a common ancestral species or population of species P1 and P2. O stands for an outgroup that is used to determine a wild allele type at individual polymorphic sites. Allele A stands for a wild allele type while B for the mutant allele

生期望频率应该相等. D 统计量的计算公式为:

$$D(P_1, P_2, P_3, O) = \frac{\sum_{i=1}^n (C_{ABBA}(i) - C_{BABA}(i))}{\sum_{i=1}^n (C_{ABBA}(i) + C_{BABA}(i))}, \quad (1)$$

式中 $C_{ABBA}(i)$ 和 $C_{BABA}(i)$ 分别为在第 i 位点 ABBA 和 BABA 模式数值(0或1). 当群体 P3 与 P1 或者与 P2 发生种间基因渐渗且 P3 群体为非随机交配系统时, 那么一种基因树模式的频率将会高于另一种基因树模式的频率, 产生显著的 D 值.

Martin 等人^[61]评价了使用 D 统计量定位种间基因渐渗位点, 并估计 3 种 P1 和 P2 由于基因渐渗导致的相同基因组的比率:

$$f_G = \frac{S(P_1, P_2, P_3, O)}{S(P_1, P_3, P_3, O)}, \quad (2)$$

式中分子 S 计算公式为 $S(P_1, P_2, P_3, O) = \sum_{i=1}^n (C_{ABBA}(i) - C_{BABA}(i))$ 和分母计算表达式类似, P3a 和 P3b 为从 P3 谱系抽取的两个类群, 所以 f_G 计算的是从 P3 到 P2 的单一方向基因流. 更为保守的 f_{hom} 值计算是假设 P2 完全被 P3 替代($P_2 = P_3$), 以确保 f_{hom} 值在 0 到 1 之间:

$$f_{hom} = \frac{S(P_1, P_2, P_3, O)}{S(P_1, P_3, P_3, O)}. \quad (3)$$

同时考虑 P2 和 P3 的双向基因流, 定义 P_D 为衍生等位基因 B 频率最高的群体(P2 或 P3), 基因流比率计算为:

$$f_d = \frac{S(P_1, P_2, P_3, O)}{S(P_1, P_D, P_D, O)}. \quad (4)$$

Martin 等人^[61]通过模拟及应用袖蝶属(*Heliconius*)

种的基因组数据调查 D 统计量的特征, 在有效群体小的情况下, D 估计值偏高, 不可靠, 但统计量 f_d 不存在类似的偏差, 可更好地应用于鉴别种间基因渐渗位点. Zheng 和 Janke^[62]用模拟的方法评价 D , f_G , f_{hom} 及 f_d 统计量, 认为用 D 检测基因流在遗传距离大(群体遗传分化时间)范围内表现稳健, 但对群体大小敏感, 建议应用 D 检测大群体的物种间基因流. 对于 f_G , f_{hom} 及 f_d 统计量, Zheng 和 Janke^[62]认为, 由于缺乏何时基因流发生的信息, 这些统计量难于应用到实际生物学问题研究上.

类似于上述分析, Eaton 等人^[63,64]进一步分解 ABBA-BABA 检验, 将 P3 物种分解成可以包含多个不同且有独立渐渗历史的亚谱系, 测验 5 基因模式[(((P1, P2), (P31, P32)), O)], 通过检验 3 对不同的等位基因数 D 统计量, 一次比对两个 P3 亚谱系(ABBBA/BABBA, ABBAA/BABAA 和 ABABA/BAABA). D_1 检测基因渐渗从 P31 到 P2 或 P1, D_2 检测基因渐渗从 P32 到 P2 或 P1, D_{12} 检测基因渐渗从 P31 和 P32 同时到 P2 或 P1:

$$D_1(P_1, P_2, P_{31}, P_{32}, O) = \frac{\sum_{i=1}^n (C_{ABBAA}(i) - C_{BABAA}(i))}{\sum_{i=1}^n (C_{ABBAA}(i) + C_{BABAA}(i))}, \quad (5)$$

$$D_2(P_1, P_2, P_{31}, P_{32}, O) = \frac{\sum_{i=1}^n (C_{ABABA}(i) - C_{BAABA}(i))}{\sum_{i=1}^n (C_{ABABA}(i) + C_{BAABA}(i))}, \quad (6)$$

$$D_{12}(P_1, P_2, P_{31}, P_{32}, O) = \frac{\sum_{i=1}^n (C_{ABBBA}(i) - C_{BABBA}(i))}{\sum_{i=1}^n (C_{ABBBA}(i) + C_{BABBA}(i))}. \quad (7)$$

这些 D 统计量计算可以应用 bootstrapping 或 Jackknifing 方法进行显著性检验.

在探讨上述方法时, 包含了在人(*Homo sapiens*)^[59]、蝴蝶(*Heliconius*)^[61]、栎树(*Quercus series virentes*)^[64]等物种的应用. 近期应用 ABBA-BABA 检验其他的动植物种间基因渐渗已有一些报道. 研究内容包括: 证实种间基因渐渗是物种丰富热点形成原因^[65], 证实栽培种向日葵(*Helianthus annuus*)与近缘种间基因流^[66], 马来西亚激流蛙复合体(*Amolops larutensis complex*)种间存在基因交换^[67], 云杉属(*Picea*)种间基因渐渗^[68], 揭示袖蝶属(*Heliconius*)蝴蝶种间基因渐渗与溯祖共存^[69], 芸香科(Rutaceae)的夏威夷蜜茱萸属(*Melicope*)种间存在祖先和现代杂交^[70], 鉴别山羊草

(*Aegilops Triticum*)的4个亚系种间系统发育关系来源于祖先群体结构分化而非杂交^[71], 西藏藏獒(*Tibetan Mastiff*)部分基因组区域源于藏灰狼(*Canis lupus laniger*)^[72], 鉴定沿海地区的齿栎(*Quercus dentata*)基因组区域源于内陆生长的蒙古栎(*Q. mongolica* var. *crispula*)基因组^[73]等。

与 F_{st} 和LD统计测验比较, ABBA-BABA方法测验更为具体的种间基因渐渗信息。由于 D 计算是基于等位基因状态IIS(identity in state)确定基因树拓扑结构的, 因此受祖先群体的交配系统、遗传分化、漂变等因素影响。P3群体近交、自交、漂变或选择等非随机交配条件会增加单个等位基因的频率, 从而增加等位基因野生型 A 或突变等位基因 B 的概率, 导致偏向ABBA或BABA模式, 产生显著的 D 值, 如何区分这些过程与真正的基因渐渗过程仍需要进一步研究。

1.5 TreeMix分析

TreeMix^[74]是利用全基因组范围内的等位基因频率数据进行分析, 从而推断多个群体分裂和混合事件。考虑一个祖先群体分裂成多个子代群体, 这些群体的基因频率因多世代漂变逐渐偏离祖先群体基因频率 x_A 。按照Cavalli-Sforza和Edwards^[75]模型计算频率的方差和群体样本间基因频率协方差 $\bar{w}_{ij}$, 实际分析时, 考虑SNP连锁影响, 将全基因组分解成许多相互独立的片段进行估计 $\bar{w}_{ij}$ 。为估计并进行检验群体间迁移权数 w 显著性, 采用Jackknife的办法估计标准误。

该方法的基本思路按以下主要步骤进行: (i) 以共同祖先群体基因频率方差为参考, 从多个子代群体的基因组序列样本数据, 计算成对群体样本基因频率的协方差 $\bar{w}_{ij}(i, j = 1, 2, \dots, m)$, 获得方差-协方差矩阵 V 估计值, 多群体样本基因频率向量 X 服从多元正态分布, 即 $X \sim MVN(x_A, V)$ 。若某群体与其他群体间存在迁移, 除了增加来自不同群体迁移比率权数 w 外, 该群体基因频率方差及与其他群体协方差采用类似方法计算。(ii) 寻找各群体间最大似然树, 似然函数 $L(\bar{W}|W) = \prod_{i=1}^m \prod_{j=i}^m N(\bar{w}_{ij} | G, \sigma_{ij}^2)$, G 为一个有根和有方向的无循环且具有分支长度和混杂权数的进化图; σ_{ij}^2 为 $\bar{w}_{ij}$ 估值的标准误。(iii) 计算残差协方差矩阵及按照一定方向添加群体间迁移分支边, 若似然函数显

著增加, 需调整分支图构相并重新计算分支长度和似然函数, 重复上述过程直到似然函数没有显著增加为止。

类似于STRUCTURE分析应用, 当把物种视为高度遗传分化的群体时, TreeMix分析也被用于分析物种间基因渐渗, 例如, Gagnaire等人^[76]分析了巨牡蛎(*Crassostrea gigas*)和角牡蛎(*Crassostrea angulata*)种间的全基因组分化, 证实它们存在种间基因渐渗。

除上述统计检验方法外, 还有其他一些方法, 如主成分PCA(principle component analysis)聚类分析^[77]和非参数群体结构分析AWclust^[78]等, 可以作为初步分析种间基因渐渗是否发生, 解决有和无的问题, 但要准确估计迁移率, 需要深入分析有关种间分化与引入基因流参数, 将基因流作为单独的进化过程来建立模型及似然函数, 估计并严格测验基因流。这就需要物种形成进行模型假设, 估计和测验基因流。

2 基于种间进化模型检验

2.1 IM模型

Felsenstein^[79]从统计学角度评价分子序列构建系统发育树的可靠性, 提出一个基于系统树框架估计群体参数的似然函数:

$$L(\Theta|X) \propto Pr(X|\Theta) = \sum_{G \in \Psi} Pr(X|G)P(G|\Theta), \quad (8)$$

式中 X 为序列数据, G 为系统/基因树, Ψ 为所有可能系统/基因树的集合, Θ 为要估计的参数向量。 $P(G|\Theta)$ 为先验分布, $Pr(X|G)$ 为给定基因树 G 条件下获得数据 X 的概率。由于系统树集 Ψ 随着参与的序列数增加而显著增大, 上述加和解析式难以获得, 该计算的难点在于穷尽所有的基因树。随后, Felsenstein^[80]应用Monte Carlo积分近似计算和自助样(bootstrap sampling)方法估计有效群体大小 $\Theta=Ne$ 或 $4N_e\mu$ 。Kuhner等人^[81]提出另一种近似计算MCMC(Markov Chain Monte Carlo)估计 $\Theta=4N_e\mu$ 。Beerli和Felsenstein^[23,82]应用近似计算MCMC估计两群体间的迁移率及有效群体数。这些近似计算估计没有得到广泛应用, 但为后来的IM模型探索奠定了基础。

Nielsen和Wakeley^[83]首次提出IM模型(图2), 该模型包含5个参数 $\Theta = \{\theta_1, \theta_2, \theta_a, M_1, M_2, T\}$, 3个群体, 即2个后裔群体(N_1, N_2)在 T 世代前来自1个共同的祖先群

图2 IM模型^[83]. 1个祖先群体(N_a, θ_a)在 T 世代前分裂成两个不完全分离的子代群体($N_1, \theta_1; N_2, \theta_2$), 分裂后子代群体间存在基因流(M_1, M_2)

Figure 2 IM model^[83]. One common ancestral population with parameters (N_a, θ_a) split in two progeny populations with parameters ($N_1, \theta_1; N_2, \theta_2$) at time T in the past. Two directional gene flow between progeny populations (M_1, M_2) occurred since the splitting time

体(N_a 或 θ_a), 用 M_1 和 M_2 度量种间基因渐渗, 用Metropolis-Hastings MCMC获得参数的后验分布. Hey和Nielsen^[84]进一步分析多位点情况下后验分布及MCMC计算方法, 并估计拟暗果蝇(*Drosophila pseudoobscura*)与波斯果蝇(*D. persimilis*)间的分裂时间 T , $P(G|\Theta)$ 的分布仍然采用数值计算.

围绕着 Θ 参数估计, 已有一系列理论方面的研究. Hey和Nielsen^[32]将公式(8)中后验分布计算转化为:

$$P(\Theta|X) = \int \psi P(\Theta|G)p(G|X)dG \quad (9)$$

$$\approx \frac{1}{k} \sum_{i=1}^k P(\Theta|G_i),$$

式中 $p(\Theta|G_i)$ 的解析式可以推导出来, 这样简化了计算后验分布及参数检验. 需要指出的是, 对于一个给定的 G , 溯祖过程与迁移过程是独立的, 为推导 $P(G|\Theta)$ 分布密度的解析式提供了方便. Hey团队已开发出IM系列程序包, 分析以公式(9)为框架的参数估计, 尽管如此, 实际计算仍然缓慢. Wang和Hey^[85]进一步推导出(8)式中的 $P(G|\Theta)$ 解析式, 直接通过函数数值积分来获得极大似然参数估计, 无需要通过MCMC计算, 这样也缩短了计算时间. 这些计算的关键限速步骤仍然是计算不同基因数树的概率. 基因树的似然估计常用Felsenstein的剪枝算法和局部更新进行, 但未穷尽所有可能的基因树^[86], 似然估计的计算量与DNA序列的长度相关, 增加内存和运算时间. 周纯葆等人^[87]针对HKY突变模型进行了IM模型的数据并行运算, 从而提高运算效率. Lohse等人^[88]基于Mathematica研制出一

种分支长度发生函数, 自动计算单个位点上任意突变样本的基因树概率, 快速计算极大似然估计方法, 其优点是分析少量样本多基因序列, 缺点是该方法只适用于相同的有效群体大小 $N_1=N_2$ 及对称基因渐渗 $M_1=M_2$ 情况. Yang^[89]提出了通过比较异域物种形成模式与邻域或同域三物种基因树的似然函数的比值LRT检验来间接判断是否存在种间基因渐渗. 在考虑已知三物种的系统发育树条件下, Zhu和Yang^[90]分析了对称基因流下的IM模型.

至今已有许多应用IM模型分析动植物种间基因渐渗的案例, 例如, Jiménez和Ornelas^[91]应用mtDNA和卫星DNA位点分析美洲绿蜂鸟(*Amazilia beryllina*)、蓝尾蜂鸟(*Amazilia cyanura*)及灰腹蜂鸟(*Amazilia saucerottii*)种间基因流, 证实物种间双向基因渐渗率不等; Pereira等人^[92]根据mtDNA和核分子标记证实火蜥蜴(*Salamandra salamandra*)在高纬度和低纬度生态种间杂交带中发生基因渐渗等; Walter等人^[93]研究发现, 第四纪出现的格陵兰鲨鱼(*Somniosus microcephalus*)与太平洋睡鲨(*Somniosus pacificus*)在加拿大北极及亚北极区存在遗传组成混杂的个体, 表明在最初物种分化后的二次接触上进行了渗透杂交; Koblmüller等人^[94]应用IM模型检测2种金枪鱼种金头类珍珠虎(*Altolamprologus compressiceps*)和喷点类珍珠虎(*A. calvus*)分化时间短, 且种间几乎没有基因流; Jordan等人^[57]证实紫萼路边青(*Geum rivale*)和自交种欧亚路边青(*G. urbanum*)的种间几乎没有基因流; Díaz等人^[95]证明3种无角苍蝇(*Anastrepha fraterculus*)种间存在连续和不对称的基因流动模式; Wang等人^[96]应用IM模型检测中国斑嘴鸭(*Anas zonorhyncha*)和绿头鸭(*A. platyrhynchos*)存在不完全分选和基因流, 而非单系.

IM模型有以下三个假设: (i) 处理多DNA序列但少位点, 计算时间随着位点增加而计算耗时; (ii) 进化过程相对简单, 没有提供更多基因流信息; (iii) 子代群体及祖先群体的有效群体大小为常数, 基因位点间相互独立等. 违反这些假设可能会为分析推断带来偏差^[97]. 虽然对于多位点基因组序列数据, 重组影响可能不占主要部分, 分析结果仍可提供有价值参考, 但要分析这些假设影响, 需要借助其他模型.

2.2 IIM模型

与IM模型不同, IIM(isolation-with-initial-migra-

tion)模型考虑一个祖先群体分裂为两个子代群体, 子代群体间起初存在基因流但随后完全隔离(图3), 模型假设可能更接近实际物种的谱系分选过程. 在IIM模型正式形成前, 已有一些相关的理论研究^[98~100], 直到Wilkinson-Herbots^[101]正式建立IIM理论模型, 但模型只考虑了对称基因流和有效群体大小相等时的情况.

Costa和Wilkinson-Herbots^[102]进一步拓展原有的假设, 提出一种高效快速的极大似然方法估计参数, 同时可利用似然比LRT检测: (i) 分裂开始直到现在仍存在潜在不对称的基因流的情况; (ii) 分裂开始直到过去的某个时间点仍存在潜在不对称的基因流, 之后完全隔离的情况; (iii) 分裂开始完全隔离的情况. 该模型适用于从大量独立位点的DNA序列数据.

给定两个物种随机抽取的DNA序列样本, 应用Costa和Wilkinson-Herbots^[102]的IIM模型计算涉及到三个独立时间上连续的马尔科夫链, 起始状态分别为两个谱系中的种群一、亚种群二、状态三(每个谱系分别在两个亚种群)或状态四(谱系已经合并). 用以状态一、二、三为起始状态的三个马尔科夫链来建立似然函数, 将多位点的似然函数乘积获得最后的似然函数, 来求得极大似然参数估计. 模型假设每个阶段有效群体大小固定, 种间有效群体大小可以不等, 各基因位点独立, 且无重组效应.

目前应用IIM模型分析动植物种间分化早期时的基因流的研究较少, 例如, 应用IIM模型分析欧洲刺泥鳅属(*Cobitis*)种间早期频繁基因流, 到后来演化成无性繁殖而产生杂合子形成障碍, 并形成新物种过程^[103]; 分析生态群落组装机制^[104]; 分析拟暗果蝇(*Drosophila pseudoobscura*)与波斯果蝇(*D. persimilis*)分化早期基因流^[105]等. 该模型更接近实际物种分选过程及高效运算, 因此可以预期该模型有广阔的应用前景.

2.3 AFS模型分析

类似于IM模型框架, 对于任意两个群体/物种, 等位基因频率谱AFS(allele frequency spectrum)模型假定它们的共同祖先群体分裂成两个子代群体/物种, 子代群体/物种间存在基因流(图4), 所涉及到的参数 Θ 有祖先(N_A)和子代群体的有效群体大小(N_1 , N_2)、分裂时间(T)、基因流(M)等. 依据Gutenkunst等人^[106]提出的模型, AFS与IM模型不同之处有: (i) AFS模型可同时考虑多个群体/物种间的基因流; (ii) 处理非中性SNP

图3 IIM模型^[101]. 1个祖先群体($2aN$)在 τ_0 世代以前遵循Wright-Fisher生长模型, 然后在 τ_0 世代前分裂成两个子代群体, 每个群体遵循Wright-Fisher生长模型, 在过去 τ_1 和 τ_0 世代之间两后裔群体存在基因流(m_1 , $2N$; m_2 , $2bN$), 但在0到 τ_1 世代前两后裔群体($2c_1N$, $2c_2N$)基因流停止

Figure 3 IIM model^[101]. One ancestral population ($2aN$) followed the Wright-Fisher demographic model before τ_0 generations in the past, and split into two progeny populations. Each progeny population followed Wright-Fisher demographic model since time τ_0 . Gene flow occurred between progeny populations from time τ_1 to τ_0 in the past, but did not occur since τ_1 in the past

位点; (iii) 依据突变的等位基因频率矩阵来计算, 突变等位基因是根据所研究的群体DNA序列与外类群/种群DNA序列比对确定的, 矩阵行与列数等于两群体的样本数(若是二倍体的, 各样本数乘以2), 矩阵中的元素对应两群体样本DNA序列在所在位置对应的各自样本DNA位点多态性出现的次数/频次, 如[1,0]表示在第一个群体样本DNA序列中某SNP位点出现1个突变等位基因和第二个群体样本DNA序列中对位点没有突变等位基因的个数. AFS矩阵包含了所有种间序列差异信息, 但仅隐含种间基因组进化关系. (iv) 采用扩散模型建立多群体突变等位基因数相对频率的联合密度 ϕ , 计算期望突变等位基因数; (v) 根据AFS矩阵每个元素位点间相互独立及Poisson分布假设, 建立似然函数 $L(\Theta|S)$, 采用有限元差分法近似计算和估计参数, 比基于基因树空间抽样的IM和IIM模型计算更高效.

理论模拟研究显示在不同参数设置下, 由AFS矩阵计算得到的成对群体概率分布图不一致, 可以区分IM模型与其他模型过程^[6,106], 但分辨程度至今未见报道. 由于选择效应(γ)被纳入参数模型, AFS也可分析非中性DNA序列区域, 如非同义密码位点的SNP. 目前应用 $\partial a \partial i$ 程序包来实现AFS模型分析^[106], 已有广泛的实际研究报道. 例如, Le Moan等人^[107]应用AFS模型分析欧洲凤尾鱼(*Engraulis encrasicolus*)沿海海洋生态型基因组差异时发现, 冰期后二次接触及随后的异

图4 左边为AFS模型^[106], 1祖先群体(N_A , θ)在 τ 世代前分裂成两个子代群体(N_1 , N_2), 子代群体间存在基因流(M); 右边为AFS矩阵, 从群体1和2中分别随机抽取 n_1 和 n_2 样本, DNA测序, 与外类群对应的DNA比对确定两样本中SNP的突变等位基因型, 然后应用群体基因组比对来计算AFS矩阵各元素观测值

Figure 4 Species tree in the left panel is the AFS model^[106] where one ancestral population split into two progeny populations at time τ in the past and gene flow occurred between the two populations since then. AFS matrix in the right panel is for model analysis. Two samples of sizes n_1 and n_2 are derived from two progeny populations and sequenced. Wild and mutant alleles at individual sites are determined in comparison with outgroup DNA sequences. Each entry value in the matrix is calculated based on the alignment comparison of DNA sample sequences from two progeny populations

域隔离造成基因组分化, 但二次接触时基因流破坏早期的基因组分化。Hsieh等人^[108]应用AFS模型分析4个中非西部侏儒狩猎采集者, 发现了3个来自喀麦隆的巴卡俾格米人(Pygmies)和9个约鲁巴人(Yorubas)的基因组差异, 该结果更好地支持了约鲁巴人和侏儒之间的双向不对称基因流, 而非以前的从约鲁巴人到侏儒的单向基因流认识。Rougeux等人^[109]应用全基因组SNP位点分析了历史上的鲱形白鲑种(*Coregonus clupeaformis*)间二次接触时异质基因渐渗引起基因组不同区域变异; Wang等人^[110]筛选4000个在心叶华葱芥(*Sinallaria limprichtiana*)和大叶葱芥(*S. grandifolia*)有差异表达的基因, 应用AFS模型分析证实它们存在持续的种间基因流。Nikolic等人^[111]应用AFS模型分析欧洲鳗鲡(*Anguilla Anguilla*)和美洲鳗鲡(*A. rostrata*)种形成历史, 证实两种鳗鲡在150000~160000世代前分化, 但在87000~92000世代前存在二次接触和杂交, 64%的基因组存在有限的基因交换。这些研究有些除了推测种间基因渐渗(M)外, 也估计物种分裂时间(T)^[106]。

AFS模型从基因频率途径分析物种分化、种间基因渐渗, 去除了复杂的从基因溯祖途径或基因树空间 G 抽样分析, 目前已有广泛应用, 但以下问题值得注意: (i) 多群体等位基因频率估计或AFS矩阵的观测值均受样本大小影响, 样本对AFS模型的检测功效、对参数估计影响目前还没有报道。(ii) AFS模型假设, 如位点相同的突变率、无重组、位点间无连锁不平衡, 可能会影响参数估计, 尤其是当SNP位点数多的情

况下。位点间突变率存在变异, 如常用的 Γ -分布来描述位点间突变率分布等。Sousa和Hey^[6]对AFS模型弱点有一些评价, 如AFS模型对检测群体间基因流的程度敏感性, 这些需要统计检验等。(iii) AFS分析时, 祖先群体与子代群体的有效群体大小是假定固定的, 实际群体的大小有可能变化, 如冰期后群体扩张或收缩等, 这与IM、IIM模型假设类似。以上限制条件对于解释实际结果可能会有影响。

2.4 ABC分析

近似贝叶斯计算(approximate Bayesian computation, ABC)提供了一种基于似然函数或贝叶斯公式, 采用不同与传统极大似然或贝叶斯估计的途径去估计参数, 避免复杂的似然函数或后验分布函数计算, 有关分析原理已有许多文献报道^[112,113], 这里简要概括如下: 假定观测数据 D , 所要估计参数为 Θ , 参数的先验分布为 $\pi(\Theta)$, 依据贝叶斯定理可得参数得后验分布 $P(\Theta|D)$ 为:

$$P(\Theta|D) = \frac{P(D|\Theta)\pi(\Theta)}{P(D)}, \quad (10)$$

式中分母 $P(D)$ 为试验数据概率或常数。应用ABC算法估计参数 Θ 步骤: (i) 从先验分布抽取参数 Θ^* ; (ii) 在一定的统计模型 M 下(依据具体的假设模型定)模拟并产生相同样本大小的模拟数据 D^* ; (iii) 选用一个或多个简略且尽量为参数的充分统计量 $S(\cdot)$, 计算模拟数据

$S(D^*)$ 与实际观测数据 $S(D)$ 的距离 $\rho(S(D^*), S(D))$; 假定一个小数 ϵ , 当 $\rho(S(D^*), S(D)) < \epsilon$ 时, 接受 Θ^* , 否则拒绝 Θ^* ; (iv) 重复步骤(i)~(iii)获得系列参数 Θ 值, 这些接受的参数值服从或趋于后验分布, $P(\Theta|D) = P(\Theta|S(D)) \approx P(\Theta|\rho(S(D^*), S(D)) < \epsilon)$. 在步骤(iii)中, 有一些不同的算法来调整接受的参数抽样值使得随后的抽样参数值分布更接近后验分布, 包括基于线性(regression-based conditional density estimation)或非线性(feed-forward neural network model)回归的密度估计、序贯蒙特卡罗抽样(sequential Monte Carlo sampling, SMC)、ABC群体蒙特卡罗(population Monte Carlo, PMC)等方法^[112~114].

自从ABC方法在群体遗传学应用后^[112,115,116], 该方法广泛地应用于生物学不同研究方面, 如生态学和谱系地理学^[117,118]. 目前研究人员已开发许多ABC应用分析程序, 如DIYABC^[119]和R-abc程序^[120](wikipedia网站很好地概括了一些主要的ABC分析程序). 在应用ABC分析种间基因渐渗时, 常用的步骤为: (i) 获取两物种观测的同源基因组序列数据或物种多位点的卫星DNA分子标记数据(D); (ii) 根据参数先验分布假设及各种物种形成模型 M , 如图2中的IM, 图3中的IIM以及图5中的三种模型等参考模型、应用模型程序, 如ms^[121]或msnsam^[122], 生成模拟数据^[123]; (iii) 应用ABC程序分析, 计算参数概括统计量 $S(D^*)$, 比较每一种模型的模拟数据与实际观测数据 $S(D)$ 差异, 选择最好的

模拟数据和模型, 获得参数后验分布, 推测参数及生物学意义.

近几年, 应用ABC方法分析动植物种或高度分化群体的分化时间、有效群体大小、基因渐渗等方面有许多报道. 例如, Roux等人^[124]应用ABC方法分析61对动物分化群体/种分化问题, 估计种间基因流, 澄清了大多数模棱两可的案例, 并发现了一些隐型种; Portnoy等人^[125]应用16个核编码微卫星和线粒体DNA控制区1077个碱基对, 分析美国东南海岸和墨西哥湾北部湾鲨叉鲨(*Carcharhinus isodon*)分化群体间的基因流; Piñeiro等人^[126]应用DIYABC分析中非森林番荔枝科的绿野仙人掌(*Greenwayodendron suaveolens* subsp. *Suaveolens*)和Detarioideae科的树种(*Scorodophloeus zenkeri*), 根据核SSR数据和ABC模型分析种间分化时间(<2 Ma), 认为两物种分化是检验冰期对树木种群干旱化效应的适宜尺度; Gladstein和Hammer^[127]应用ABC方法和基因组SNP数据分析了阿什凯纳兹犹太人(Ashkenazi Jew)与塞法迪奇/米斯拉希犹太人(Sephardic/Mizrahi Jews)分化群体的分化时间、有效群体大小及基因流. 注意ABC分析是推测后验分布来估计物种或高度分化的群体间基因流等参数, 有的研究会结合其他方法进行综合分析, 如上述阿什凯纳兹犹太人群体历史研究采用AFS作为统计量 $S(D)$ 与ABC联合分析.

除上述模型分析外, 还有一些应用较少的模型, 如

图5 与IM及IIM模型不同的三种物种形成模型^[6]. 隔离模型指祖先群体在 τ_0 世代前分裂成两个物种, 随后种间无基因渐渗; 随机交配模型指祖先群体无分裂, 所研究的两个“物种”为同一种内的两个群体; 二次接触模型指祖先群体在 τ_0 世代前分裂成两个物种, 在 τ_1 至 τ_0 世代前物种间无基因流, 到 τ_2 世代后发生种间基因渐渗

Figure 5 Three species tree models that are distinct from IM and IIM models^[6]. The first one is the isolation model where one ancestral population split into two progeny populations at time τ_0 in the past and no gene introgression occurred between progeny populations since then. The second model is the panmictic model where the ancestral population did not split, and the two focal species are considered as two populations of the same species. The last one is the secondary contact model where one ancestral population split into two progeny populations at time τ_0 in the past. Gene flow among progeny populations did not occur in the time interval from τ_1 to τ_0 in the past, but occurred since τ_1 in the past

根据迁移基因组IBD(identity by descents)片段长度密度分布估计近期种间基因渐渗. Pool和Nielsen^[128]提出似然函数, 根据观察到的迁移DNA在接受群体的IBD片段长度分布, 估计不同时间段从一个种间到另一个物种的迁移率, 该方法的应用前提是可以根据全基因组分析, 能够鉴别迁移DNA片段的长度及分布. 近期, Schrider等人^[129]提出FILET(finding introgressed loci via extra-trees)方法分析种间基因渐渗等.

3 问题与展望

澄清谱系分选与基因流的关系是进化生物学研究的重要内容之一. 表2概括了上述讨论的种间基因渐渗主要检测方法, 包括基于基因频率和基于DNA序列分析两种途径进行的分析, 前者需要适当的样本来估计基因频率, 后者一般只需要少量样本测序分析且在基因溯祖理论框架下分析物种间基因渐渗. 当获得群体基因组数据时, 可同时从两途径分析, 如AFS分析. 两类数据是相互关联的, 群体内个体DNA序列差异是群体内基因频率变异的基础, 都不同程度地揭示物种间经历的生态和进化过程. 这些分析都面临以下共同

问题.

一是在多基因位点情况下, 重组效应常被忽略不计, 或假设位点间相互独立, 实际数据分析或设计也是尽量避开重组影响. 重组可导致同一DNA片段不同位置的IBD不同, 或位点间LD存在, 重组与基因流互作导致不同位置片段年龄不同. 然而, 增加重组效应这一层信息会使得似然函数的建立变得更为复杂, 且难以计算^[6], 也许对于DNA序列数据, 重组效应对于同时应用大量基因位点分析作用可能是微小的, 但在方法上考虑重组效应有待进一步突破.

二是不同倍性基因载体的基因流贡献. 已知动物群体的基因流可以通过配子或杂合子迁移形式(如鱼类), 而植物群体则以单倍体花粉流或二倍体种子流实现, 早期植物群体结构的研究是通过不同遗传方式的核质分子标记来估计花粉流(m_p)/种子流(m_s)的相对贡献^[131,132]. 类似地, 对于DNA序列数据, 也可采用核质基因序列IBD系数来估计^[133]. 容易理解通过花粉传播和授粉过程实现种间基因渐渗, 那么种子扩散是怎样作用于种间基因渐渗呢? 对于异域物种形成, 可能的途径是种子迁移, 随后亲本杂交或与花粉流形成的杂种进行回交等方式贡献种间基因渐渗, 与直接的花粉

表 2 物种间基因渐渗检测方法与理论模型

Table 2 Methods of detecting gene introgression between species and their theoretical models

数据类型	统计变量	模型与检测	参考文献
基因频率	F_{st}	无模型假设: 由 F_{st} 沿着基因组上的分布模式推测群体间基因流或群体分化	[47,130]
	LD	无模型假设: 由基因组上LD片段长度及其分布的分布模式推测群体间基因流	[60]
	Q	纯统计模型: 给定群体数 K 以及先验分布 $Pr(K)$, $Pr(P)$ 和 $Pr(Q)$, 建立后验分布, 通过MCMC计算后验分布, 估计 Q , 推测每个样本来自 K 群体得遗传组成比率 Q	[14]
基因频率和全基因组序列	AFS	采用扩散过程建立两个或多个群体等位基因频率谱(AFS)密度函数, 根据Poisson分布假设, 建立似然函数, 应用有限元差分数值计算估计群体大小、迁移参数、群体分化时间等	[106]
	迁移比率 w 检验(TreeMix)	应用各群体基因频率计算方差-协方差矩阵, 建立群体有根进化分支图, 根据似然函数是否显著增加来确定群体间迁移边界存在性	[74]
全基因组序列	ABBA-BABA分析 D	无模型假设: 当ABBA与BABA基因树拓扑相发生的频次存在显著差异时, 种间基因渐渗发生	[59]
	迁移IBD片段长度密度分布, 似然估计 m	模型: 迁移-漂变, 依据迁移IBD片段长度的密度分布, 建立似然函数估计两群体间迁移率	[128]
	IM后验分布函数估计	模型: 隔离-迁移模型(IM), 基于所有潜在的基因树构相, 建立参数后验分布, 采用MC计算及LRT测验	[83]
	IIM似然函数估计	模型: 分裂-起初迁移-隔离模型(IIM), 基于所有潜在的基因树拓扑结构, 建立似然函数, 采用数值计算估计及LRT测验	[101]
	ABC模拟估计	各种物种形成模型, 采用模型模拟数据与观测数据比对来确定最佳模型, 推测后验分布	[112,115,116]

流相差一个世代. 已知杂交带理论或冰期后二次接触产生的基因频率随空间距离渐变模式反映出种子流的贡献^[134,135]. 对于邻域物种形成, 近距离的种子流作用可能会增大, 但同时近距离的花粉流的作用也会大, 两者对于种间基因渐渗贡献有可能差异不大; 对于同域物种形成, 种子流的贡献可能微小, 主要通过授粉实现种间杂交和基因渐渗. 至今, 已知的种间基因渐渗检测方法(表2)还没有区分不同倍性基因载体的基因流, 如何估计它们的相对贡献对于认识物种分选过程中基因渐渗的作用有重要意义.

三是物种交配系统影响. 在基于DNA序列分析(基因树)的方法中, 物种的交配系统是假设随机的(Wright-Fisher群体增长模式). 已知一些动物种维持近交系统或选型交配(*assortative mating*), 一些植物表现自交, 多数植物种表现出混合交配系统. 交配系统与基因流的互作会影响现有的种间基因渐渗检测方法(表2), 主要表现在两个方面: (i) 改变基因树拓扑结构, 若祖先群体存在非随机交配, 增加子代群体基因之间共祖概率, 或增加基因共祖与基因流的混杂, 影响基因检测分析, 如ABAB-BABA分析和DNA序列分析在

位点间独立假设(无LD)等; (ii) 增加子代物种间基因流不相等^[7], 当子代群体演化成不同类型的交配系统时, 如*Geum*属中的异交种紫萼路边青(*Geum rivale*)和自交种欧亚路边青(*G. urbanum*)的基因流^[57], 模型中 $M_1 \neq M_2$. 长期以来, 交配系统与物种形成是两个相互独立发展的研究领域, 但近交系统作为杂合子形成前基因流障碍有助于物种形成是容易理解的^[7,8], 因此, 如何在检测种间基因渐渗同时估计交配系统具有重要意义.

四是物种间非中性基因渐渗. 除了AFS等少数分析考虑了选择性位点外, 大部分DNA序列分析方法(表2)都是建立在中性过程假设上的. 与物种生殖隔离的基因通常是选择适应的, 谱系分选过程导致与生殖隔离关联的基因趋于不同方向上的固定, 形成显著的差异表达, 这些基因的种间基因渐渗可能性很小. 其他适应性基因会有不同程度的种间选择差异, 不排除种间基因渐渗, 这些适应性基因流(即使很少发生)对物种形成起较大的阻碍作用^[135,136], 估计这部分基因流有助于深入理解物种形成所需要长或短时间的的原因.

参考文献

- 1 Morjan C L, Rieseberg L H. How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. *Mol Ecol*, 2004, 13: 1341–1356
- 2 Wright S. *Evolution and The Genetics of Populations, Volume 3: Experimental Results and Evolutionary Deductions*. Chicago: The University of Chicago Press, 1977
- 3 Coyne J A, Orr H A. *Speciation*. Sunderland: Sinauer Associates, Inc, 2004
- 4 Hu X S. Mating system and the critical migration rate for swamping selection. *Genet Res*, 2011, 93: 233–254
- 5 Zhang X X, Cheng X, Li L L, et al. The wave of gene advance under diverse systems of mating. *Heredity*, 2020, doi: 10.1038/s41437-020-0333-9
- 6 Sousa V, Hey J. Understanding the origin of species with genome-scale data: modelling gene flow. *Nat Rev Genet*, 2013, 14: 404–414
- 7 Hu X S. Mating system as a barrier to gene flow. *Evolution*, 2015, 69: 1158–1177
- 8 Pickup M, Brandvain Y, Fraïsse C, et al. Mating system variation in hybrid zones: facilitation, barriers and asymmetries to gene flow. *New Phytol*, 2019, 224: 1035–1047
- 9 Freeland J R, Petersen S D, Kirk H. *Molecular Ecology*, 2nd ed. New Jersey: Wiley-Blackwell, 2011
- 10 Wright S. The genetical structure of populations. *Ann Eugen*, 1949, 15: 323–354
- 11 Vitalis R, Couvet D. Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics*, 2001, 157: 911–925
- 12 Wang J, Whitlock M C. Estimating effective population size and migration rates from genetic samples over space and time. *Genetics*, 2003, 163: 429–446
- 13 Ellstrand N C, Marshall D L. Interpopulation gene flow by pollen in wild radish, *Raphanus sativus*. *Am Nat*, 1985, 126: 606–616
- 14 Pritchard J K, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*, 2000, 155: 945–959

- 15 Wilson G A, Rannala B. Bayesian inference of recent migration rates using multilocus genotypes. *Genetics*, 2003, 163: 1177–1191
- 16 Faubet P, Gaggiotti O E. A new Bayesian method to identify the environmental factors that influence recent migration. *Genetics*, 2008, 178: 1491–1504
- 17 Broquet T, Yearsley J, Hirzel A H, et al. Inferring recent migration rates from individual genotypes. *Mol Ecol*, 2009, 18: 1048–1060
- 18 Slatkin M. Rare alleles as indicators of gene flow. *Evolution*, 1985, 39: 53–65
- 19 Slatkin M, Barton N H. A comparison of three indirect methods for estimating average levels of gene flow. *Evolution*, 1989, 43: 1349–1368
- 20 Rannala B, Hartigan J A. Estimating gene flow in island populations. *Genet Res*, 1996, 67: 147–158
- 21 Tufto J, Engen S, Hinder K. Inferring patterns of migration from gene frequencies under equilibrium conditions. *Genetics*, 1996, 144: 1911–1921
- 22 Hudson R R, Slatkin M, Maddison W P. Estimation of levels of gene flow from DNA sequence data. *Genetics*, 1992, 132: 583–589
- 23 Beerli P, Felsenstein J. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, 2001, 152: 763–773
- 24 Wright S. Isolation by distance. *Genetics*, 1943, 28: 114–138
- 25 Rousset F. Genetics differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, 1997, 145: 1219–1228
- 26 Smouse P E, Dyer R J, Westfall R D, et al. Two-generation analysis of pollen flow across a landscape. I. Male gamete heterogeneity among females. *Evolution*, 2001, 55: 260–271
- 27 Barton N H. The structure of the hybrid zone in *Uroderma bilobatum* (Chiroptera: Phyllostomatidae). *Evolution*, 1982, 36: 863–866
- 28 Broquet T, Petit E J. Molecular estimation of dispersal for ecology and population genetics. *Annu Rev Ecol Evol Syst*, 2009, 40: 193–216
- 29 Marko P B, Hart M W. The complex analytical landscape of gene flow inference. *Trends Ecol Evol*, 2011, 26: 448–456
- 30 Ismail S A, Ghazoul J, Ravikanth G, et al. Does long-distance pollen dispersal preclude inbreeding in tropical trees? Fragmentation genetics of *Dysoxylum malabaricum* in an agro-forest landscape. *Mol Ecol*, 2012, 21: 5484–5496
- 31 Kassa A, Konrad H, Geburek T. Mating pattern and pollen dispersal in the wild olive tree (*Olea europaea* subsp. *cuspidata*). *Tree Genets Genomes*, 2018, 14: 1
- 32 Hey J, Nielsen R. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc Natl Acad Sci USA*, 2007, 104: 2785–2790
- 33 Wang X, Cheng X, Zhou W, et al. Assessing the ecological and evolutionary processes underlying cytonuclear interactions (in Chinese). *Sci Sin Vitae*, 2019, 49: 951–964 [王茜, 程祥, 周玮, 等. 细胞核质互作形成的生态与进化过程分析. 中国科学: 生命科学, 2019, 49: 951–964]
- 34 Weir B S. Genetic Data Analysis II. Sunderland: Sinauer Associates, 1996
- 35 Hu X S. A review on understanding the genetic structure of population (in Chinese). *Sci Silv Sin*, 2002, 38: 119–128 [胡新生. 群体遗传结构的理解. 林业科学, 2002, 38: 119–128]
- 36 McGaughan A, Rödelserperger C, Grimm D G, et al. Genomic profiles of diversification and genotype-phenotype association in island nematode lineages. *Mol Biol Evol*, 2016, 33: 2257–2272
- 37 Picq S, McMillan W O, Puebla O. Population genomics of local adaptation versus speciation in coral reef fishes (*Hypoplectrus* spp, Serranidae). *Ecol Evol*, 2016, 6: 2109–2124
- 38 Kjeldsen S R, Zenger K R, Leigh K, et al. Genome-wide SNP loci reveal novel insights into koala (*Phascolarctos cinereus*) population variability across its range. *Conserv Genet*, 2016, 17: 337–353
- 39 Royer A M, Streisfeld M A, Smith C I. Population genomics of divergence within an obligate pollination mutualism: Selection maintains differences between Joshua tree species. *Am J Bot*, 2016, 103: 1730–1741
- 40 Parchman T L, Buerkle C A, Soria-Carrasco V, et al. Genome divergence and diversification within a geographic mosaic of coevolution. *Mol Ecol*, 2016, 25: 5705–5718
- 41 Kingston S E, Parchman T L, Gompert Z, et al. Heterogeneity and concordance in locus-specific differentiation and introgression between species of towhees. *J Evol Biol*, 2017, 30: 474–485
- 42 Walsh J, Kovach A I, Olsen B J, et al. Bidirectional adaptive introgression between two ecologically divergent sparrow species. *Evolution*, 2018, 72: 2076–2089
- 43 Murray K D, Janes J K, Jones A, et al. Landscape drivers of genomic diversity and divergence in woodland *Eucalyptus*. *Mol Ecol*, 2019, 28: 5232–5247

- 44 Battey C J. Evidence of linked selection on the Z chromosome of hybridizing hummingbirds. *Evolution*, 2019, 74: 725–739
- 45 Hu X S. Evolution of zygotic linkage disequilibrium in a finite local population. *PLoS ONE*, 2013, 8: e80538
- 46 Hu X S, Yeh F C. Assessing postzygotic isolation using zygotic disequilibria in natural hybrid zones. *PLoS ONE*, 2014, 9: e100568
- 47 Hu X S, Yeh F C, Hu Y, et al. High mutation rates explain low population genetic divergence at copy-number-variable loci in *Homo sapiens*. *Sci Rep*, 2017, 7: 43178
- 48 Hu X S, Yeh F C, Wang Z. Structural genomics: Correlation blocks, population structure, and genome architecture. *Curr Genomics*, 2011, 12: 55–70
- 49 Reddy U K, Nimmakayala P, Abburi V L, et al. Genome-wide divergence, haplotype distribution and population demographic histories for *Gossypium hirsutum* and *Gossypium barbadense* as revealed by genome-anchored SNPs. *Sci Rep*, 2017, 7: 41285
- 50 Stankowski S, Sobel J M, Streisfeld M A. Geographic cline analysis as a tool for studying genome-wide variation: a case study of pollinator-mediated divergence in a monkeyflower. *Mol Ecol*, 2017, 26: 107–122
- 51 Schilling M P, Mullen S P, Kronforst M, et al. Transitions from single- to multi-locus processes during speciation with gene flow. *Genes*, 2018, 9: 274
- 52 Alexander D H, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, 2009, 19: 1655–1664
- 53 Tang H, Peng J, Wang P, et al. Estimation of individual admixture: Analytical and study design considerations. *Genet Epidemiol*, 2005, 28: 289–301
- 54 Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol*, 2005, 14: 2611–2620
- 55 Miralles L, Machado-Schiaffino G, Garcia-Vazquez E. Genetic markers reveal a gradient of hybridization between cape hakes (*Merluccius capensis* and *Merluccius paradoxus*) in their sympatric geographic distribution. *J Sea Res*, 2014, 86: 69–75
- 56 Fogelqvist J, Verkhozina A V, Katyshev A I, et al. Genetic and morphological evidence for introgression between three species of willows. *BMC Evol Biol*, 2015, 15: 193
- 57 Jordan C Y, Lohse K, Turner F, et al. Maintaining their genetic distance: Little evidence for introgression between widely hybridizing species of *Geum* with contrasting mating systems. *Mol Ecol*, 2017, 27: 1214–1228
- 58 López-Caamal A, Reyes-Chilpa R, Tovar-Sánchez E. Hybridization between *Tithonia tubaeformis* and *T. rotundifolia* (Asteraceae) evidenced by nSSR and secondary metabolites. *Plant Syst Evol*, 2018, 304: 313–326
- 59 Green R E, Krause J, Briggs A W, et al. A draft sequence of the Neandertal genome. *Science*, 2010, 328: 710–722
- 60 Durand E Y, Patterson N, Reich D, et al. Testing for ancient admixture between closely related populations. *Mol Biol Evol*, 2011, 28: 2239–2252
- 61 Martin S H, Davey J W, Jiggins C D. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol Biol Evol*, 2014, 32: 244–257
- 62 Zheng Y, Janke A. Gene flow analysis method, the D-statistic, is robust in a wide parameter space. *BMC Bioinf*, 2018, 19: 10
- 63 Eaton D A R, Ree R H. Inferring phylogeny and introgression using RADseq data: An example from flowering plants (*Pedicularis*: Orobanchaceae). *Syst Biol*, 2013, 62: 689–706
- 64 Eaton D A R, Hipp A L, González-Rodríguez A, et al. Historical introgression among the American live oaks and the comparative nature of tests for introgression. *Evolution*, 2015, 69: 2587–2601
- 65 Zarza E, Faircloth B C, Tsai W L E, et al. Hidden histories of gene flow in highland birds revealed with genomic markers. *Mol Ecol*, 2016, 25: 5144–5157
- 66 Baute G J, Owens G L, Bock D G, et al. Genome-wide genotyping-by-sequencing data provide a high-resolution view of wild *Helianthus* diversity, genetic structure, and interspecies gene flow. *Am J Bot*, 2016, 103: 2170–2177
- 67 Chan K O, Alexander A M, Grismer L L, et al. Species delimitation with gene flow: A methodological comparison and population genomics approach to elucidate cryptic species boundaries in Malaysian torrent frogs. *Mol Ecol*, 2017, 26: 5435–5450
- 68 Feng S, Ru D, Sun Y, et al. Trans-lineage polymorphism and nonbifurcating diversification of the genus *Picea*. *New Phytol*, 2019, 222: 576–587
- 69 Kubatko L S, Chifman J. An invariants-based method for efficient identification of hybrid species from large-scale genomic data. *BMC Evol Biol*, 2019, 19: 1
- 70 Paetzold C, Wood K R, Eaton D A R, et al. Phylogeny of Hawaiian *Melicope* (Rutaceae): RAD-seq resolves species relationships and reveals

- ancient introgression. *Front Plant Sci*, 2019, 10: 1074
- 71 Huynh S, Marcussen T, Felber F, et al. Hybridization preceded radiation in diploid wheats. *Mol Phylogenet Evol*, 2019, 139: 106554
- 72 Miao B, Wang Z, Li Y. Genomic analysis reveals hypoxia adaptation in the Tibetan mastiff by introgression of the grey wolf from the Tibetan Plateau. *Mol Biol Evol*, 2017, 34: 734–743
- 73 Nagamitsu T, Uchiyama K, Izuno A, et al. Environment-dependent introgression from *Quercus dentata* to a coastal ecotype of *Quercus mongolica* var. *crispula* in northern Japan. *New Phytol*, 2020, 226: 1018–1028
- 74 Pickrell J K, Pritchard J K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet*, 2012, 8: e1002967
- 75 Cavalli-Sforza L L, Edwards A W. Phylogenetic analysis: Models and estimation procedures. *Am J Hum Genet*, 1967, 19: 233–257
- 76 Gagnaire P A, Lamy J B, Cornette F, et al. Analysis of genome-wide differentiation between native and introduced populations of the cupped oysters *Crassostrea gigas* and *Crassostrea angulata*. *Genome Biol Evol*, 2018, 10: 2518–2534
- 77 Patterson N, Price A L, Reich D. Population structure and eigenanalysis. *PLoS Genet*, 2006, 2: e190
- 78 Gao X, Starmer J D. AWclust: point-and-click software for non-parametric population structure analysis. *BMC Bioinf*, 2008, 9: 77
- 79 Felsenstein J. Phylogenies from molecular sequences: inference and reliability. *Annu Rev Genet*, 1988, 22: 521–565
- 80 Felsenstein J. Estimating effective population size from samples of sequences: a bootstrap Monte Carlo integration method. *Genet Res*, 1992, 60: 209–220
- 81 Kuhner M K, Yamato J, Felsenstein J. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, 1995, 140: 1421–1430
- 82 Beerli P, Felsenstein J. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc Natl Acad Sci USA*, 2001, 98: 4563–4568
- 83 Nielsen R, Wakeley J. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, 2001, 158: 885–896
- 84 Hey J, Nielsen R. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, 2004, 167: 747–760
- 85 Wang Y, Hey J. Estimating divergence parameters with small samples from a large number of loci. *Genetics*, 2010, 184: 363–379
- 86 Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 1981, 17: 368–376
- 87 Zhou C B, Lang X Y, Wang Y G, et al. A parallel implementation of the isolation with migration model (in Chinese). *e-Sci Technol Appl*, 2012, 3: 24–29 [周纯葆, 郎显宇, 王彦桐, 等. 隔离迁移 (Isolation with Migration) 模型数值计算的并行实现. *科研信息化技术与应用*, 2012, 3: 24–29]
- 88 Lohse K, Harrison R J, Barton N H. A general method for calculating likelihoods under the coalescent process. *Genetics*, 2011, 189: 977–987
- 89 Yang Z. A likelihood ratio test of speciation with gene flow using genomic sequence data. *Genome Biol Evol*, 2010, 2: 200–211
- 90 Zhu T, Yang Z. Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow. *Mol Biol Evol*, 2012, 29: 3131–3142
- 91 Jiménez R A, Ornelas J F. Historical and current introgression in a Mesoamerican hummingbird species complex: a biogeographic perspective. *PeerJ*, 2015, 4: e1556
- 92 Pereira R J, Martínez-Solano I, Buckley D. Hybridization during altitudinal range shifts: nuclear introgression leads to extensive cyto-nuclear discordance in the fire salamander. *Mol Ecol*, 2016, 25: 1551–1565
- 93 Walter R P, Roy D, Hussey N E, et al. Origins of the Greenland shark (*Somniosus microcephalus*): Impacts of ice-olation and introgression. *Ecol Evol*, 2017, 7: 8113–8125
- 94 Koblmüller S, Nevado B, Makasa L, et al. Phylogeny and phylogeography of *Altolamprologus*: ancient introgression and recent divergence in a rock-dwelling Lake Tanganyika cichlid genus. *Hydrobiologia*, 2017, 791: 35–50
- 95 Díaz F, Luís A, Lima A, Nakamura A M, et al. Evidence for introgression among three species of the *Anastrepha fraterculus* group, a radiating species complex of fruit flies. *Front Genet*, 2018, 9: 359
- 96 Wang W, Wang Y, Lei F, et al. Incomplete lineage sorting and introgression in the diversification of Chinese spot-billed ducks and mallards. *Curr Zool*, 2019, 65: 589–597
- 97 Strasburg J L, Rieseberg L H. How robust are “isolation with migration” analyses to violations of the IM model? A simulation study. *Mol Biol Evol*, 2010, 27: 297–310
- 98 Teshima K M, Tajima F. The effect of migration during the divergence. *Theor Popul Biol*, 2002, 62: 81–95

- 99 Innan H, Watanabe H. The effect of gene flow on the coalescent time in the human-chimpanzee ancestral population. *Mol Biol Evol*, 2006, 23: 1040–1047
- 100 Becquet C, Przeworski M. Learning about modes of speciation by computational approaches. *Evolution*, 2009, 63: 2547–2562
- 101 Wilkinson-Herbots H M. A fast method to estimate speciation parameters in a model of isolation with an initial period of gene flow and to test alternative evolutionary scenarios. 2015, Available from: URL: <https://arxiv.org/abs/1511.05478>
- 102 Costa R J, Wilkinson-Herbots H. Inference of gene flow in the process of speciation: An efficient maximum-likelihood method for the isolation-with-initial-migration model. *Genetics*, 2017, 205: 1597–1618
- 103 Janko K, Pačes J, Wilkinson-Herbots H, et al. Hybrid asexuality as a primary postzygotic barrier between nascent species: On the interconnection between asexuality, hybridization and speciation. *Mol Ecol*, 2018, 27: 248–263
- 104 Bunnefeld L, Hearn J, Stone G N, et al. Whole-genome data reveal the complex history of a diverse ecological community. *Proc Natl Acad Sci USA*, 2018, 115: E6507–E6515
- 105 Korunes K L, Machado C A, Noor M A F. Inversions shape the divergence of *Drosophila pseudoobscura* and *D. persimilis* on multiple timescales. *BroRxiv*, 2019
- 106 Gutenkunst R N, Hernandez R D, Williamson S H, et al. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet*, 2009, 5: e1000695
- 107 Le Moan A, Gagnaire P A, Bonhomme F. Parallel genetic divergence among coastal-marine ecotype pairs of European anchovy explained by differential introgression after secondary contact. *Mol Ecol*, 2016, 25: 3187–3202
- 108 Hsieh P H, Veeramah K R, Lachance J, et al. Whole-genome sequence analyses of western central African pygmy hunter-gatherers reveal a complex demographic history and identify candidate genes under positive natural selection. *Genome Res*, 2016, 26: 279–290
- 109 Rougeux C, Bernatchez L, Gagnaire P A. Modeling the multiple facets of speciation-with-gene-flow toward inferring the divergence history of lake whitefish species pairs (*Coregonus clupeaformis*). *Genome Biol Evol*, 2017, 9: 2057–2074
- 110 Wang T J, Ru D F, Zhang D, et al. Analyses of genome-scale variation reveal divergence of two *Sinallaria* species (Brassicaceae) with continuous but limited gene flow. *J Sys Evol*, 2019, 57: 268–277
- 111 Nikolic N, Liu S, Jacobsen M W, et al. Speciation history of European (*Anguilla anguilla*) and American eel (*A. rostrata*), analysed using genomic data. *Mol Ecol*, 2020, 29: 565–577
- 112 Beaumont M A. Approximate Bayesian computation in evolution and ecology. *Annu Rev Ecol Evol Syst*, 2010, 41: 379–406
- 113 Turner B M, Van Zandt T. A tutorial on approximate Bayesian computation. *J Math Psychol*, 2012, 56: 69–85
- 114 Blum M G B, François O. Non-linear regression models for approximate Bayesian computation. *Stat Comput*, 2010, 20: 63–73
- 115 Tavaré S, Balding D J, Griffiths R C, et al. Inferring coalescence times from DNA sequence data. *Genetics*, 1997, 145: 505–518
- 116 Pritchard J K, Seielstad M T, Perez-Lezaun A, et al. Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Mol Biol Evol*, 1999, 16: 1791–1798
- 117 Csilléry K, Blum M G B, Gaggiotti O E, et al. Approximate Bayesian computation (ABC) in practice. *Trends Ecol Evol*, 2010, 25: 410–418
- 118 Hoban S, Bertorelle G, Gaggiotti O E. Computer simulations: tools for population and evolutionary genetics. *Nat Rev Genet*, 2012, 13: 110–122
- 119 Cornuet J M, Santos F, Beaumont M A, et al. Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics*, 2008, 24: 2713–2719
- 120 Csilléry K, François O, Blum M G B. Abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol Evol*, 2012, 3: 475–479
- 121 Hudson R R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 2002, 18: 337–338
- 122 Ross-Ibarra J, Wright S I, Foxe J P, et al. Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLoS ONE*, 2008, 3: e2411
- 123 Li S, Jakobsson M. Estimating demographic parameters from large-scale population genomic data using approximate Bayesian computation. *BMC Genet*, 2012, 13: 22
- 124 Roux C, Fraïsse C, Romiguier J, et al. Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS Biol*, 2016, 14: e2000234
- 125 Portnoy D S, Hollenbeck C M, Bethea D M, et al. Population structure, gene flow, and historical demography of a small coastal shark (*Carcharhinus isodon*) in US waters of the western Atlantic Ocean. *ICES J Mar Sci*, 2016, 73: 2322–2332
- 126 Piñeiro R, Dauby G, Kaymak E, et al. Pleistocene population expansions of shade-tolerant trees indicate fragmentation of the African rainforest

- during the ice ages. *Proc R Soc B*, 2017, 284: 20171800
- 127 Gladstein A L, Hammer M F. Substructured population growth in the Ashkenazi Jews inferred with approximate Bayesian computation. *Mol Biol Evol*, 2019, 36: 1162–1171
- 128 Pool J E, Nielsen R. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics*, 2009, 181: 711–719
- 129 Schrider D R, Ayroles J, Matute D R, et al. Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. *PLoS Genet*, 2018, 14: e1007341
- 130 Holsinger K E, Weir B S. Genetics in geographically structured populations: defining, estimating and interpreting F_{st} . *Nat Rev Genet*, 2009, 10: 639–650
- 131 Ennos R A. Estimating the relative rates of pollen and seed migration among plant populations. *Heredity*, 1994, 72: 250–259
- 132 Hu X S, Ennos R A. Impacts of seed and pollen flow on population differentiation for plant genomes with three contrasting modes of inheritance. *Genetics*, 1999, 152: 441–450
- 133 Hu X S, Ennos R A. On estimation of the ratio of pollen to seed flow among plant populations. *Heredity*, 1997, 79: 541–552
- 134 Hu X S, Li B. Assessment of the ratio of pollen to seed flow in a cline for genetic variation in a quantitative trait. *Heredity*, 2001, 87: 400–409
- 135 Hu X S, Li B. Seed and pollen flow and cline discordance among genes with different modes of inheritance. *Heredity*, 2002, 88: 212–217
- 136 Lowe W H, Allendorf F W. What can genetics tell us about population connectivity? *Mol Ecol*, 2010, 19: 3038–3051

Advances in the methods of detecting interspecific gene introgression and their applications

CHENG Xiang^{1,2}, LI LingLing^{1,2}, XIAO Yu^{1,2}, CHEN XiaoYang^{1,2} & HU XinSheng^{1,2}

1 College of Forestry and Landscape Architecture, South China Agricultural University, Guangzhou 510642, China;

2 Guangdong Key Laboratory for Innovative Development and Utilization of Forest Plant Germplasm, Guangzhou 510642, China

With the proceeding of lineage sorting, gene flow gradually decays among progeny populations that are derived from splitting of a common ancestral population, and may probably continue till the phase of speciation. The differences between interspecific genome sequences or between interspecific population gene frequencies partially record these dynamics of historical and contemporary demography and occurrences, and hence inferences on these events can be attained from analyses of interspecific genomic sequences or gene frequencies. Here, we systematically discussed the methods of detecting interspecific gene introgression, including the approaches based on the differences among interspecific gene frequencies (F_{st} , LD, STRUCTURE, AFS, TreeMix) and the differences of interspecific DNA sequences (ABBA-BABA, IM, IIM, ABC). We discussed the principles and limits of these methods. Relaxations of some important assumptions underlying these methods remain open questions in the future, including the effects of recombination, the separate contribution of gamete vs. zygote dispersal to gene introgression, the interaction of mating system with gene flow, and the impediment to speciation arising from introgression of adaptive genes. Clarifications of these effects would aid in our deep understanding of the process of speciation.

gene introgression, speciation, isolation-with-migration, lineage sorting, gene tree

doi: [10.1360/SSV-2020-0126](https://doi.org/10.1360/SSV-2020-0126)