生物多样性 2017, **25** (6): 577–599 doi: 10.17520/biods.2017097
Biodiversity Science http://www.biodiversity-science.net

•综述•

结合系统发育与群体遗传学分析检验杂交是否存在的技术策略

毛建丰[1*] 马永鹏[2] 周仁超[3]

1 (北京林业大学生物科学与技术学院林木育种国家工程实验室和林木花卉育种教育部重点实验室, 北京 100083)
2 (中国科学院昆明植物研究所东亚植物多样性与生物地理学重点实验室, 昆明 650201)
3 (中山大学生命科学学院, 广州 510275)

摘要: 杂交通常指不同类群间(种间或种内)经有性途径的遗传交流。越来越多的研究表明, 作为一种遗传交换过程, 杂交是生物多样性形成、维持和丧失的重要机制, 它广泛参与了动物、植物、微生物等的类群分化。然而, 我们对杂交过程中遗传交换的普遍性、存在模式、产生机制的认识还非常有限。当前, 高通量测序技术的飞速发展和基因组学研究技术的普遍应用, 为深入评价遗传交换的普遍性和进化意义提供了前所未有的契机。如何选用恰当的研究技术与策略检验潜在的杂交并评价它的特征, 成为大家普遍面临的问题。本文试图综合来自系统发育和群体遗传等相互关联学科中不同的技术策略, 以当前流行的高通量测序技术为核心, 结合表型和细胞遗传学等多种数据获取和分析手段, 概括不同分析策略的特点, 联系必要的实例研究, 为生物多样性与进化领域的学者提供检测遗传交换的参考。

关键词: 遗传交换; 基因流; 生物多样性; 系统发育; 群体遗传

Approaches used to detect and test hybridization: combining phylogenetic and population genetic analyses

Jian-Feng Mao[1*], Yongpeng Ma[2], Renchao Zhou[3]

1 *National Engineering Laboratory for Tree Breeding, Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants of Ministry of Education, College of Biological Sciences and Technology, Beijing Forestry University, Beijing 100083*
2 *Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201*
3 *School of Life Sciences, Sun Yat-sen University, Guangzhou 510275*

Abstract: Hybridization among diverging (interspecific or intraspecific) groups involves gene flow and genetic recombination. Increasingly, studies have shown that hybridization, a process of genetic exchanges, occurs widely in the divergence and unity of animals, plants, and microorganisms, and acts as an important mechanism for the formation and maintenance of biological diversity. The rapid development of high-throughput sequencing technology and the widespread application of genome-level techniques provides an unprecedented opportunity for us to further evaluate the universality and evolutionary significance of hybridization. However, selecting appropriate research techniques and strategies to detect the potential hybridization and evaluate its characteristics becomes a common question. In this review, we attempt to synthesize methods from phylogenetics and population genetics of the genomic era to provide biodiversity and evolutionary researchers a practical reference for testing hybridization.

Key words: genetic exchange; gene flow; biodiversity; phylogenetics; population genetics

收稿日期: 2017-03-26; 接受日期: 2017-05-04
基金项目: 国家自然科学基金(31370255; 31670664)
* 通讯作者 Author for correspondence. jianfeng.mao@bjfu.edu.cn

杂交(hybridization)通常指真核生物不同类群间(种间或种内)经有性途径的遗传交换(genetic exchange)。杂交使不同亲本的遗传物质共存于杂交子代，而杂交子代减数分裂过程中的同源染色体遗传交换实现了遗传物质在不同类群间的交换和重新组合。除了有性过程的杂交，重组(recombination)经常被用来描述原核生物中不经有性过程实现的遗传物质重新组合。这种重组大量存在于病毒中。水平基因转移(lateral/horizontal gene transfer)指那些不同于杂交和重组中实现的基因由亲代垂直传递给子代的遗传交换过程。此外，基因交流(gene flow)、渐渗(introgression)、伴随基因流的分化(divergence/isolation with gene flow/migration)和网状进化(reticulate evolution)等概念也常被用于概括不同时空、进化过程或系统发育特征的遗传交换。这些概念的形成、应用范围和侧重有差异，但本质上都是遗传交换(Arnold, 2016)。不同于生命之树(tree of life)理论，生命之网(web of life)理论认为生物的演化过程是一个由遗传交换构成的网状模式，它在更普遍的意义上概括了遗传交换对生物多样性产生和维持的重要作用(Arnold, 2016; Mallet et al, 2016)。本文在认同不同遗传交换过程在进化生物学共性意义的基础上，将技术策略综述的视角限定在对杂交(或者说狭义的杂交)的检测上，以此避免大而不全带来的误导。

越来越多的证据表明，杂交普遍存在于生物类群分化和维持中，具有重要的进化生物学意义。在农林业上，杂交是一种重要的育种手段。通过人工杂交，可以快速获得杂种优势、实现有益变异在不同育种材料间的重新组合，甚至产生全新的表型。在自然类群中，杂交也有类似的作用：可以带来新的遗传变异，且速度比突变要快得多(Anderson & Hubricht, 1938; Martinsen et al, 2001)。因此，杂交增加了选择中性位点的等位基因数量；引入具有选择优势的等位基因，增加获得这些变异的类群的适合度，促成快速的适应性转变(Choler et al, 2004; Martin et al, 2006; Castric et al, 2008; Kim et al, 2008)。

杂交对遗传多样性分布格局、类群分化、物种形成也有重要影响。首先，杂交可以直接导致基因流从一个类群流入另一个类群；它可能致使一个类群替代另一个类群，或新生的杂种类群替代亲本类群，造成等位基因或者类群的灭亡(Ellstand & Elam, 1993)；也可导致分化类群的融合(Grant & Grant, 2014)。其次，杂交还可导致新物种的形成，杂交和后续的多倍化过程构成的异源多倍化是一种快速的物种形成途径，特别是在植物中(Hegarty & Hiscock, 2008; Soltis & Soltis, 2009)。不经由多倍化的同倍性杂交物种形成(homoploid hybrid speciation)也可以发生，在此物种形成模式中，杂种谱系通常和亲本类群在生态或者地域上彼此分隔开来(Gross & Rieseberg, 2005; Abbott et al, 2010)。杂交对物种形成的另一个作用是，它通过传递有利于适应性分化的等位基因，进而强化种间隔离的形成，并促进物种形成(Abbott et al, 2013)。

生物类群中杂交的普遍存在获得了越来越多的认可。然而，相对于复杂的生物多样性，清晰描述的杂交事件仍显得非常少。实际上，作为一个重要的进化过程和生物多样性维持机制，杂交受到的关注还远远不够。据估算，有25%的高等植物物种(Mallet, 2005)和10%的动物物种(Schwenk et al, 2008)间存在着种间杂交；此外，在真菌(Nelson, 1963; Depotter et al, 2016)、细菌(Duncan et al, 1989; Cohan & Kane, 2001; Cohan, 2002; Earl et al, 2008)、病毒(Worobey & Holmes, 1999; Bujarski, 2013; Lefeuvre & Moriones, 2015; Pérez-Losada et al, 2015; Su et al, 2016)的基因组进化中，杂交普遍存在。以植物为例，*Flora of China*共记载我国3万余种植物(Wu et al, 2014)，按25%计算，我们推测其中9,000种植物间存在杂交。同时，日益加剧的全球气候变化和人类活动将对生物多样性的分布特征产生显著影响，进而改变种间关系，导致原已形成的种间生殖隔离改变或降低，进一步增加杂交的机会(Stenz et al, 2015; Novikova et al, 2016; Vallejo-Marin & Hiscock, 2016)。相较于上面的估算，得到确认的植物杂交案例非常稀少，我们认为可能不足1%。因此，要全面澄清杂交产生的机制、杂交对多样性形成和维持的影响及其进化生物学意义还为时尚早。目前已经开展或正在进行的工作所覆盖的杂交事件，估计不超过总量的1‰。检测杂交及评估其发生的范围和机制，将成为未来生物多样性监测、管理和保护中不可缺少的组成部分。在后志书时代的今天，检测杂交等遗传交换事件的存在，进而揭示它对生物多样性形成和维持的重要价值，

是我们面临的重要挑战。

当前, 高通量测序和基因组技术的飞速发展提供了快捷、简便、廉价、高覆盖度的分子标记, 为在更大尺度上鉴定杂交, 甚至澄清杂交产生的机制和基因组效应提供了前所未有的契机。传统上, 杂交的检出往往借助表型特征(包括形态、生理、次生代谢物等)变异模式、细胞遗传学手段、系统发育分析、群体遗传分析等手段。常规的分析手段中, 表型分析未针对遗传物质本身的特征进行评估, 所以只能用于杂交的初步认定而不能作为确认的工具; 细胞遗传学分析受材料限制较大, 对实验技术要求高, 应用范围并不广。利用分子标记技术的分析中, 基于多位点的系统发育分析和群体遗传分析在基因组时代有了全新的发展, 系统发育基因组学和群体基因组学逐步替代了原有的基于少数位点的分析, 得到了越来越广的应用, 它们将是我们检测和评估杂交的重要手段。本文针对如何检测杂交的存在这一普遍面临的问题, 综述了已经出现的、来自于不同研究领域的分析技术和策略, 在此基础上, 我们把重点放在了结合基因组学方法的技术策略上。旨在为动物、植物、微生物等各个不同类群生物多样性相关领域的研究人员提供参考, 提高研究者对杂交在生物进化中意义的认识, 提高有关研究中识别遗传重组的能力, 为更深入地认识生物多样性的形成和维持机制, 建立高效的生物多样性监测、保护和管理策略奠定方法学基础。

1　表型变异模式可以给出杂交存在的依据

在分子标记技术出现前, 对杂交的描述大都依赖表型分析。表型指个体在包括形态、行为、物候、生殖、生理、生活史、次生代谢等各方面的表现。F_1代杂种往往体现出两个亲本都不具备的杂种优势; 不论低世代还是高世代杂种, 除了受到亲本遗传物质比例的影响之外, 它们往往在一些性状上表现出居中的表型, 同时都普遍在个别性状上体现出超亲分离(transgressive segregation)。在调查的171个研究实例中, 155个(91%)有超亲分离存在; 在分析的1,229个性状中, 有44%显示了超亲分离, 其中58%的植物性状和35%的动物性状显示了超亲分离(Rieseberg et al, 1999)。杂种显示超亲分离与杂交本身有直接关系(Rieseberg et al, 2003)。

大量的调查分析表明, 和生殖隔离密切相关的性状更多地表现出超亲分离, 比如植物的环境适应性和开花物候性状、动物的生活史和行为性状(Rieseberg et al, 1999)。总之, 杂种和亲本种间在表型的变异模式上有联系, 但这种联系还没有确定的模式可依。由此可见, 表型在揭示杂交是否存在或者确定亲本来源上有很多限制, 它无法提供确切的依据。但表型指标甚至仅仅是形态指标往往也能提供非常有价值的信息, 比如将它们与基因组水平的标记结合起来可以为物种的界定提供重要信息(Pyron et al, 2016)。笔者亲历了一个体现形态数据重要性的实例。在几十年前对松属(*Pinus*)的修订中, 我国的分类学家就依据形态特征猜测高山松(*Pinus densata*)要么来自云南松(*P. yunnanensis*)和油松(*P. tabuliformis*)的杂交, 要么作为祖先种分化为另外两种松树(吴中伦, 1956)。之后, 研究人员开展了形态、生殖、物候、解剖、幼苗萌发、人工杂交、多重分子标记的群体遗传等多方面的研究, 已有的证据支持了前一个推测(Wang XR et al, 1990, 2001; Wang & Szmidt, 1994; Song et al, 2002, 2003; Ma et al, 2006; Mao et al, 2009; Wang BS et al, 2011; Gao et al, 2012; 张立沙等, 2012; 梁冬等, 2013; Xing et al, 2014; Zhao et al, 2014)。

目前, 随着计算技术的蓬勃发展, 用于分析表型数据的统计方法已经非常成熟。描述统计(descriptive statistics)可以针对单一变量或多个变量。单一变量的描述统计中, 比较不同类群在特定性状上的平均值、中位数、变异幅度、方差、标准差等, 是统计分析的基本步骤; 多变量描述统计中, 可以考察不同类群在多个变量上的变异模式(作散点图)或者考察性状间的相关性在不同类群上的差异(性状间的相关模式受样本遗传背景的影响, 遗传背景不同, 性状间相关模式也会改变)。多重比较、回归分析、方差分析的目的都在于对类群间在单一或者多个性状上的差异显著性进行检验, 澄清它们的大小关系和相似程度。聚类分析和主成分分析则更侧重类群在多性状变异上的变异模式, 在这两类分析中也可以引入分化显著性的检验, 并且可以通过做图直观地观察类群间的相似模式和类群内的变异。比如通过对杜鹃花属(*Rhododendron*)内马缨杜鹃(*R. delavayi*)、露珠杜鹃(*R. irroratum*)两个种及其杂交后代大样本的形态测量(物种个体内部和不同个体的形态性状重复)和严格的统计分析(包括形态性状的

outlier检验与去除, 多因素分析和杂交指数的计算)发现: 不论亲本种还是杂交后代都可以准确鉴定; 同时发现中国植物志书对亲本种很多叶、花的特征描述实际上无法区分亲本, 因为这些特征的种内变异幅度已经远远超过种间变异; 更有意思的是研究发现对区分马缨杜鹃和露珠杜鹃贡献最大的特征是柱头直径和面积(Marczewski et al, 2016), 原因可能是柱头面积对于传粉者具有重要的选择作用。这一发现对于接下来研究这两个亲本种间的生殖隔离及其遗传基础有重要启示。

在表型数据的统计分析中, 统计作图必不可少, 应用恰当的统计作图能直观反映类群内和类群间的变异模式。我们熟知的有: 体现单一变量数据分布的频率图、概率密度图、箱图, 体现双变量的二维散点图, 三变量的立体散点图(虽不受统计学者的推崇, 但仍有广泛应用)等。此外, 聚类分析的聚类图(可以是树状或网状)、主成分分析图、biplot (一种与主成分分析密切相关的作图方法)也是很有价值的统计作图技术, 值得参考。Adams (1982)对如何利用多元统计技术分析表型数据进而检验杂交进行了较为系统全面的综述, 可作为参考。

2 细胞遗传学水平的分析

核型分析和染色体减数分裂行为的观察, 配合染色体荧光原位杂交(fluorescence *in situ* hybridization, FISH)技术, 也可为检测杂种或者多倍体提供重要线索。利用这些技术揭示出的染色体数目、染色体构象、同源染色体配对和分离行为、减数分裂异常现象的频率等, 可以用于判断是否为异源多倍体、是否是杂种以及杂种的育性等。但由于取材和技术操作的限制, 现在这方面的应用在逐步减少。国内的一个研究实例是红花油茶(*Camellia reticulata*): 它同时含有二倍体($2n = 2x = 30$)、四倍体($2n = 4x = 60$)和六倍体($2n = 6x = 90$) 3种类型, 通过FISH技术对不同倍性的红花油茶及近缘种的分析, 确定了异源四倍体的红花油茶是二倍体的红花油茶与二倍体的西南红山茶(*C. pitartii*)杂交形成; 而异源六倍体的红花油茶是异源四倍体的红花油茶和二倍体的怒江山茶(*C. saluenensis*)杂交形成(Liu & Gu, 2011)。染色体荧光原位杂交技术在植物和进化研究中的应用可参考Chester等(2010)的综述。

3 系统发育与群体遗传学分析

与表型分析和细胞遗传学的间接分析不同, 在系统发育和群体遗传学研究中, 借助恰当的分子标记技术和分析策略可对杂交的存在做直接检验, 并有可能澄清杂交发生的历史、存在的模式、物种形成及多样性维持的机制等。当前, DNA测序技术飞速发展, 测序速度、数据通量、测序文库构建的灵活性、测序片段长度均有大幅提高, 同时测序成本、建库操作复杂性等在迅速降低。各种动植物、微生物基因组从头测序或是群体水平重测序技术应用越来越广泛。这些都为系统发育重建和在谱系地理学及群体遗传学水平检测杂交的存在做了极好的铺垫, 极大地拓展了我们认识杂交的能力。无论研究的目标物种是否有参考基因组, 都不需要太大投入(目前几百元人民币)就可以拿到一个包含上万甚至更多单碱基突变(single nucleotide polymorphism, SNP)位点的DNA数据样本。传统的分子标记正在被各种基因组水平的测序策略所取代。这里, 我们首先对现在流行的几个基于新一代测序技术(next generation sequencing, NGS)的分子标记技术原理和应用进行简要的综述。同时, 对重要的数据分析技术和策略进行总结, 指出它们的原理和适用特点。再次, 分别从系统发育和群体遗传两个不同领域对杂交检验的分析方法进行综述。目的在于将各类理论和技术的特点呈现出来, 为同行提供选择的依据。在分子标记方面, 我们列出了部分重要的技术; 在数据分析方面, 除了本文提供的相关信息, 最近Payseur和Rieseberg (2016)针对检测各不同类群中杂交的存在和澄清杂交模式的问题列举了多种分析策略和大量最新的应用实例, 可供参考。

3.1 基因组时代的分子标记技术

3.1.1 PCR扩增产物测序

该技术也被称为扩增子测序(amplicon sequencing), 是直接将PCR扩增得到的产物用NGS技术测序。它是Sanger测序的延续, 但其通量高, 可以同时对多个位点、多个样本进行测序。在测序模板的准备中, 可以对每个样本都单独进行多个PCR扩增, 得到每个样本多个不同特异位点的PCR产物, 然后把产物混合进行测序(即平行扩增测序, parallel tagged sequencing) (Meyer et al, 2008); 还可以利用多PCR引物同时扩增目标基因组的多个特定位点,

然后进行样本混合测序(即混合扩增测序, multiplex PCR method) (Mamanova et al, 2010)。在测序文库构建中, 可以对单一样本添加标签序列, 从而在样本混合测序后加以区分, 获得单一样本的序列。当然, 也可以不区分样本进行混池测序。平行扩增测序过程中, 对大量样本众多位点进行单独的PCR扩增是很费时的; 而混合扩增测序时, 虽然可以在单一PCR中对多个特异位点进行扩增, 但往往不同位点扩增效率不一致, 这又会带来新的问题。针对这个问题, 发展了微液滴PCR (microdroplet PCR), 它可以实现在极小液滴中独立开展数量庞大的PCR反应(Livak, 2003; Sims et al, 2009)。这些依赖PCR扩增的技术的一个突出优势是它们对待测的模板DNA要求比较低, 这也是PCR的一个重要特点。当样品量奇缺、DNA质量不高时, 这个技术的优势就凸显出来(Shen et al, 2013)。对DNA已有所降解的标本进行测序时, 这个策略是个不错的选择。该分子标记技术的特征参见图1。

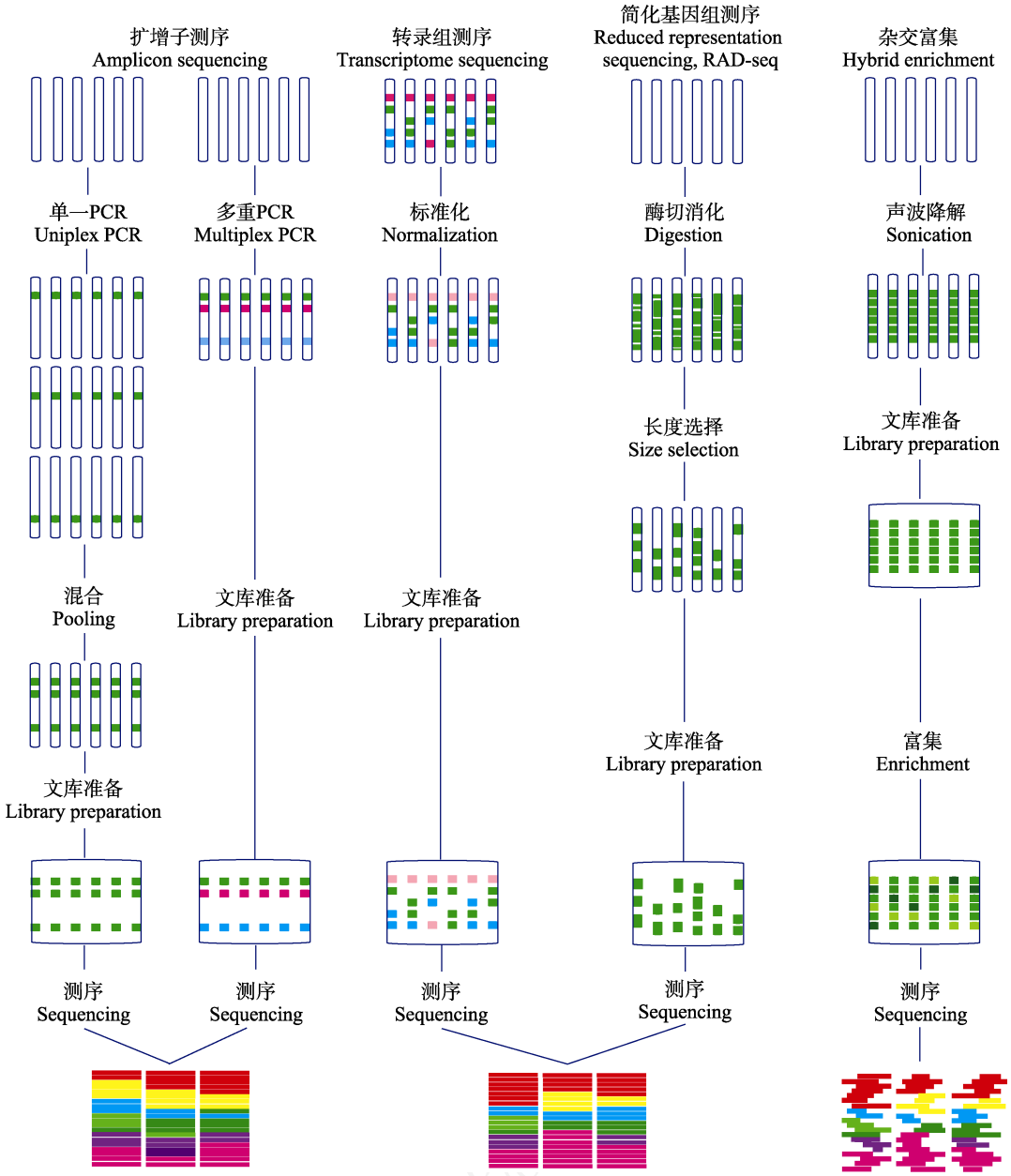

图1　基因组时代的分子标记技术(改编自Lemmon & Lemmon, 2013)。图中简要列出了各测序技术的特点、操作流程和数据特征。
Fig. 1　Molecular genotyping technologies in the genomic era (adapted from Lemmon & Lemmon, 2013). The figure lists the features, operating procedures and data processing of each sequencing technique.

同时, 对只关注少量位点但包含多个个体的项目, 扩增子测序有很大的吸引力(Griffin et al, 2011)。在检测杂交存在方面, 一个典型应用实例是鲨鱼的研究。研究人员通过扩增子测序建立的分子标记技术, 检测到帝氏真鲨(*Carcharchinus tilstoni*)和黑边鲭真鲨(*C. limbatus*)间的杂交(Morgan et al, 2012)。目前, 这项技术策略还被用于获取来自全线粒体基因组(Chan et al, 2010; Morin et al, 2010; Gunnarsdóttir et al, 2011)、全叶绿体基因组(Parks, 2009)、宏基因组(metagenomics)遗传变异的研究中。

3.1.2 基于限制性酶切的简化基因组测序技术

这类技术都包含一个重要步骤, 即对目标样本基因组进行限制性内切酶酶切, 有些还包含对酶切产物片段大小的筛选(Baird et al, 2008; van Tassell et al, 2008; Kim et al, 2016)。酶切过程中, 可用单一的内切酶, 也可用两个内切酶的组合(Peterson et al, 2012); 酶切的特征可能是甲基化敏感的、甲基化不敏感的或IIB型限制性核酸内切酶(Wang et al, 2012)。酶的组合和酶切片段筛选给了这类技术极大的灵活性, 加上样本标签后可以实现各种组合的混样测序, 这类技术的潜力巨大。基于类似的简化原理, 已有大量特定技术衍生出来(如reduced-representation library sequencing (RRL), restriction-site-associated DNA sequencing (RAD), genotyping by sequencing (GBS)), 有不少综述性论文对它们进行了详细介绍(Davey et al, 2011; Andrews et al, 2016)。该分子标记技术的特征参见图1。

但这类技术存在一些问题。第一, 人工筛选片段长度时, 由于人为操作误差可能造成个别样本上直系同源位点缺失值增多。尽管现在有了自动切胶仪器(比如PippinPrep^{TM}), 但售价相对较高(近两万美元)。第二, 无效等位基因的存在, 是由于限制性酶切位点的突变造成酶切片段丢失所致。错误地识别无效等位基因会影响对直系同源基因的评估, 进而导致系统发育分析的误差。第三, 这类基于限制性酶切的技术可能无法用于包含分化时间较长类群的系统发育分析, 因为酶切位点位置的变异会造成同源序列的减少。最后, 由于受测序技术的限制, 这类技术得到的位点多为200–300 bp长的序列, 这样短的片段往往无法提供足够的信息用来构建位点特异的基因树。不过, 基于Illumina的双末端测序已经可以把单一位点序列长度增加到500 bp以上(Etter et al, 2011)。

3.1.3 目标富集测序

目标富集测序(targeted enrichment sequencing)也称为序列捕获测序(sequence capture), 是通过高通量办法对基因组中特定序列进行富集, 进而进行高通量测序的一类技术(Mamanova et al, 2010)。在目标富集测序中, 先将基因组DNA打碎, 然后通过固相(Albert et al, 2007; Hodges et al, 2007)或液相(Gnirke et al, 2009; Maricic et al, 2010)的DNA探针杂交, 将非目的片段洗脱, 捕获目的序列, 再进行样本混合建库和高通量测序。与基于限制性酶切的简化策略相比, 这个技术是有目的地对基因组中特定序列进行筛选、富集、测序。与PCR扩增产物测序相比, 目标富集测序的富集是通过探针杂交实现的, 它的捕获通量可以更高。这个技术需要提前知道要富集的序列信息(可以是基因组、转录组或其他来源的序列), 以此设计捕获探针, 实现捕获。在捕获过程中, 用于区分样本的标签可以在捕获前加入, 也可在捕获后加入(Kenny, 2011)。由于捕获试剂往往较为昂贵, 捕获前加样本标签更为节省。该分子标记技术的特征参见图1。

尽管目标富集测序技术相对成熟, 但目前的应用较多地针对人类疾病研究, 在系统发育分析中的应用仍在发展。探针开发是一个重要环节, 从转录组或基因组开发探针的工具和流程已有不少报道(Mayer et al, 2016; Pavy et al, 2016; Schmickl et al, 2016)。超保守序列(ultraconserved elements)也是开发捕获探针的重要来源, 目前哺乳动物(Bejerano et al, 2004; Reneker et al, 2012)、鸟类(Mccormack & Al, 2011)、两栖爬行类(Crawford et al, 2012)、昆虫(Branstetter et al, 2017)、高等植物(Freeling et al, 2009; Reneker et al, 2012)上的超保守序列已有报道。该技术还有一个特点, 就是通过对长目标序列多个不同区域设计探针, 可以实现对整条序列的捕获测序。这突破了目前流行的新一代测序的读长较短的限制。但这对数据分析也造成了问题, 因为目前流行的系统发育分析软件对大量长片段进行分析时计算速度往往很慢, 比如BEST (Edwards et al, 2007)和*BEST (Heled & Drummond, 2010)。

3.1.4 转录组测序

转录组测序(transcriptome sequencing)或者RNA测序(RNA sequencing)是针对基因组中的表达

序列进行测序, 本质上也是一种简化基因组测序技术(Marioni et al, 2008; Morin et al, 2008; Wang et al, 2009)。尽管这项技术通常被用于适应性进化和生态基因组学研究, 但已有的研究实例表明, 该技术获得的转录组序列可以用于重建系统发育和检测杂交的存在(Nabholz et al, 2011; Pease et al, 2016)。此外, 该技术对发掘单碱基突变、进行群体遗传研究也很有吸引力, 因为它不仅可实现对复杂基因组的简化, 而且提供了来自直接与功能相关转录序列的信息。从转录组组装序列中寻找直系同源基因的计算工具已有报道, HaMStR (Ebersberger et al, 2009)是比较不错的一个。该分子标记技术的特征参见图1。

转录组测序的一大特点就是它可以直接地获得大量来自外显子区和附近非转录区(untranslated region, UTR)的序列, 这些信息适合各种具有不同分化程度的类群。目前, 公开或半公开的转录组序列数据已大量积累, 植物方面如1,000种植物转录组测序项目(1KP Project, http://www.onekp.com/)、北京林业大学和上海生物信息技术研究中心的植物转录组数据库(http://lifecenter.sgst.cn/plantransdb/index.do/)、药用植物转录组测序项目(Medicinal Plant Transcriptome Project, http://www.uic.edu/pharmacy/MedPITranscriptome/); 昆虫方面有1KITE项目(1KITE Project, http://1kite.org/); 真核微生物方面有海洋真核微生物转录组测序项目(Marine Microbial Eukaryote Transcriptome Project, http://www.marinemicroeukaryotes.org/)。基于转录组测序的系统发育研究也在不断积累, 比如节肢动物、陆生植物类群。植物方面的突出实例来自复旦大学, 研究人员利用转录组测序技术或以其为辅助解析了被子植物大类群(Zeng et al, 2014)、十字花科(Huang et al, 2015)、蔷薇科(Xiang et al, 2017)的系统发育, 但未见对种间杂交是否存在进行探讨。

转录组测序的限制也很明显(Ozsolak, 2011)。首先, 由于样本测序针对RNA开展, 而高质量的RNA需要从新鲜组织得到, 所以该技术受限于材料, 特别是对于那些材料难以获得的类群; 其次, 由于基因表达有组织特异性并受到环境因素的影响, 需要尽量保证用于测序的RNA来自相同培养条件下不同样本相同发育时期的相同组织, 这样才能最大限度地保证得到的相似序列更多地来自直系同源谱系; 第三, 不同基因表达量差异也很大, 如果想得到低表达基因则需要较高的测序深度; 最后, 可变剪接的存在会给转录组组装及后续数据分析带来挑战(Martin & Wang, 2011; Godden et al, 2012)。

3.1.5　全基因组测序

这个技术的实现是显而易见的。对于已有参考基因组、基因组大小合适的类群, 全基因组测序(更多的是基因组重测序, genome resequencing)应用已经很常见了, 尤其是对近缘种间或种内群体分化的研究中, 其优势更加明显。人类的1,000 Genomes计划(http://www.internationalgenome.org/)(The Genomes Project Consortium et al, 2010)和植物中拟南芥(*Arabidopsis thaliana*)的1,001 Genomes计划(http://1001genomes.org/) (Weigel & Mott, 2009)都为我们作了很好的示范。值得一提的是, 基因组水平的证据表明种间遗传交换参与了人(Patterson et al, 2012; Hellenthal et al, 2014; Lazaridis et al, 2014; Sankararaman et al, 2014; Ackermann et al, 2016)和拟南芥(Stenz et al, 2015; Novikova et al, 2016)的起源和分化。

与其他几项技术相比, 该技术对测序量的要求相对较高, 但建库的成本相对较低。随着测序成本的降低, 该技术的应用会更为普遍。如果测序深度足够大, 该技术可以得到序列变化的详细信息, 如除了碱基替换, 插入、缺失、重组、易位、倒位、基因重复、基因组重复、基因组共线性等都可获得。实际上, 除了碱基替换, 其他信息还无法用基于溯祖模型(coalescent-based)的策略进行分析, 在已有的系统发育及群体遗传分析中, 这些信息也往往被忽略。这方面的理论和分析技术的开发应该是未来发展的方向。

3.1.6　数据处理流程

相对传统技术来说, 这些分子标记技术的优势十分明显, 尽管数据处理中的某些具体环节还在发展, 但成熟的技术流程和相应的软件并不难获得。测序仪输出的原始数据转换、短读(short read)长数据质控、短读长数据的比对、转录组或基因组的组装、单碱基突变或单倍型获取、各步骤质量评估和数据筛选等重要环节的分析对一般的系统发育和群体遗传学的研究团队并不难实现, 有大量的高质量技术指南可以参考(DePristo et al, 2011; Nielsen et al, 2011)。对某一种特定分子标记技术的分析策略

也有不少参考，比如RAD测序方面(Hapke & Thiele, 2016; Kim et al, 2016; McKinney et al, 2016; Shafer et al, 2016; Torkamaneh et al, 2016, 2017)。

3.1.7 数据质量控制和筛选

基因组技术带来了海量数据，同时也伴随着一些问题。有专家认为，随着谱系基因组学的数据量增大，潜在的误差来源也会增加(Philippe et al, 2005)。误差来源主要有两个：随机误差(stochastic error)和系统误差(systematic error) (Swofford et al, 1996)。随机误差是由数据信息量不足造成的，应该会随着数据量的增加而降低；系统误差主要来自于模型的选配，会随着数据量的增加而增大(Philippe et al, 2005; Kumar, 2012)。如果没能很好地控制系统误差，我们可能得到看似支持率很好但错误的系统发育重建(O'Neill et al, 2013)。当仅利用一个基因位点构建基因树时，增加序列长度可以降低随机误差；但如果没有考虑增加序列长度带来的位点内重组的风险，那么会相应地增加系统误差。当试图把不同位点合并成一个位点进而构建基因树时，同样的问题也会发生。

为了降低系统误差，需要注意以下几个要点。(1)要注意质量控制和筛选。新一代测序数据分析涉及多个步骤，每一步骤的质量控制都很重要(Rusk, 2009)。(2)要注意直系同源基因的选取。大多数系统发育和谱系地理研究的分析是针对直系同源基因的。在Sanger测序时代，人们投入了大量精力开发单拷贝基因位点，以此避免旁系基因的掺入带来的系统误差。但实际上，即便是使用单拷贝基因，风险仍然存在。大量的全基因组分析表明，基因的丢失和重复是普遍存在的现象(Dehal & Boore, 2005)。这类现象的存在会给直系同源基因的识别带来极大的困难。应对这个困难的办法有几个。比如，有人通过比较来自大量位点的基因树(Rasmussen & Kellis, 2012; Boussau et al, 2013)，进而评判直系同源关系；也有人通过严格的序列相似性检验(例如用reciprocal BLAST评判)构建直系同源数据集；还有人借助基因组的共线性分析(Zheng et al, 2005)。相比之下，最后一种可能最为可靠。(3)利用等位基因序列(单倍型序列，allele)信息而不是位点的一致性序列进行系统发育分析。基于溯祖模型的分析往往假定分析针对的是基因的拷贝(单倍型)。但现实中，我们往往选用的是多个单倍型的一致性序列(consensus sequence)，以此代表某个位点。尽管目前还没看到模拟分析评估用一致性序列的风险，但这样做明显与模型假设不符，可能会带来潜在系统误差。选用单倍型序列数据至少会使对群体大小的估算更准确。在Sanger时代要得到单倍型数据不太容易，但新一代测序技术中完全可以将单倍型重建出来(Bauer, 2011; Menelaou, 2013)。(4)提高比对(alignment)的准确性来降低系统误差(Felsenstein, 2004; Susko et al, 2005)。频繁出现的插入/缺失以及高的碱基替换率会直接影响比对的准确性(Gatesy et al, 1993)，从而导致错配和系统误差(Misof & Misof, 2009)。借助蛋白质序列比对及去除高风险区域能在一定程度上保证数据的质量。(5)选用恰当的位点。一般需要数千个位点才可以将分化历史较短的谱系重建出来(Liu et al, 2010)；而对于分化历史较长的系统发育事件可能不需要这么多。群体大小、物种分化时间、位点的特性是帮助我们确定位点数量的重要因素(Leaché & Rannala, 2010; Liu & Yu, 2011; Morrison, 2011)。新一代测序可以获得海量数据，依据研究目的对位点进行筛选是必要的(Townsend, 2007)。

实际上，我们可能也无法恰当地处理所有数据。对于旨在检验杂交是否存在的研究来说，数据的筛选会面临一系列问题。群体遗传学就是研究等位基因数量(allele dosage)的问题，不同等位基因可能来自不同亲本，数据筛选可能造成等位基因的丢失，进而丢失重要杂交信号。当然，对于系统发育研究而言，类似的问题同样存在。数据的有效利用非常重要，对含有"杂交"的样本进行数据获取时面临一定的风险，需要引起大家的注意。

3.1.8 模型的选取

恰当的数据分析模型的选择是重中之重，一直被认为是准确重建系统发育的核心(Felsenstein, 1978)。有两个显而易见的模型选取问题需要注意。首先，要确认序列进化模型的选择是否恰当(Lemmon & Moriarty, 2004; Sullivan & Joyce, 2005)。这方面，混合模型(mixture models) (Pagel & Meade, 2004)、数据分区(data parititioning) (Yang, 1996)和针对大数据的最优分区策略(optimal partitioning strategies) (Lanfear et al, 2012)等技术可以给我们很多帮助。其次，要注意是否恰当地对重组(比杂交内涵更广，但实质类似)进行了处理。重组会导致基因树间

的不一致(Rannala & Yang, 2008)。把来自分散在基因组不同区域的不同位点的序列(比如来自RAD测序的序列)进行合并, 再作系统进化分析, 这样的做法有很大风险, 它会带来错误的系统发育信息(Degnan & Rosenberg, 2006; Edwards et al, 2007; Kubatko & Degnan, 2007; Leaché & Rannala, 2010)。来自转录组测序的数据, 即便是同一个基因中, 不同外显子间的重组也不容忽视。实际上, 本文的目的在于检测重组的存在, 不论它是在系统发育尺度上还是群体水平上。我们推荐使用可以同时检验杂交和不完全谱系筛选的算法来重建系统发育。

模型初选后, 应该用恰当的检验来评判选定模型的可靠性。这方面有一系列的工具可以利用, 比如位点内或位点间重组(Mcguire et al, 1997; Kosakovsky et al, 2006; Martin et al, 2015)、各种选择(Delport et al, 2010)、杂交(Yu et al, 2011)、时序性进化速率差异(heterotachy) (Pagel & Meade, 2008)和模型总体上的配合性(Albert & Schluter, 2005; Abby et al, 2012; Ackermann et al, 2016)等。如果初选模型不适合, 那么应该及时调整选用更恰当的模型。如果位点间基因树冲突, 那么可考虑选用应对杂交或者不完全谱系分离的模型, 而不是仍然将序列合并使用, 或者把不相配合的部分数据删除。当然还是要删除那些体现强烈选择印记的、变异饱和的(saturated) (Castresana, 2000; Rodríguezezpeleta et al, 2007; Gnirke et al, 2009)和包含有缺失值(missing)或有缺失变异(deletion)的位点(Lemmon et al, 2009)。

3.2　系统发育分析策略

不论是二倍体还是多倍体杂种, 其基因组中不同位点的基因可能来自于不同亲本种, 它们反映着不同的进化历程, 这正是利用系统发育分析检测杂交的依据(Yu et al, 2011)。系统发育分析中早就开始评估“网状进化”的存在(Sneath, 1975)。水平基因转移、杂交、重组、不完全谱系分选(incomplete lineage sorting)及基因重复和丢失(gene duplication and loss)都会让我们看到基因树间拓扑结构的不一致性(phylogenetic incongruence), 这种不一致性不能被表示成二歧分枝的树, 而更像一个网状结构。我们用网状进化来描述这个现象。在排除其他机制存在的情况下, 基因树间的不一致性往往正是我们用来检测杂交事件存在的系统发育方法。但不能说系统发育的不一致性一定意味着杂交, 它只是意味着杂交可能存在。清晰地排除不完全谱系分选和基因重复丢失带来的影响是检测杂交的必需步骤。尽管这样在方法上并不限于系统发育网络(phylogenetic network), 比如可以通过评判不同DNA片段建树后的不同树型(拓扑结构)的比例, 以及不同树型对应的片段在基因组的分布来区分杂交和不完全谱系分选。但相比之下, 系统发育网络同时考虑基因组水平大量位点信息, 有对杂交和不完全谱系分选同时和综合检测的潜力, 所以本文的重点放在了系统发育网络上。较为全面的系统发育分析流程见图2。

对比单一位点构建系统发育树, 进而检测网状进化的存在, 是一种初步评估遗传交换存在的方式, 这里不多论述。多物种溯祖(multispecies coalescent)是通过多位点信息构建物种树的策略(Degnan & Rosenberg, 2009; Edwards, 2016), 最大似然(maximum-likelihood) (Kubatko et al, 2009; Wu, 2012)和贝叶斯(Bayesian)算法(Liu, 2008; Heled & Drummond, 2010)都被成功地应用于这种技术, 对这个领域计算技术的发展也有系统的综述文章发表(Nakhleh, 2013)。然而, 一个重要问题是这个领域大多数已有模型都只把不完全谱系分选作为导致系统发育不一致性的唯一原因。事实上, 如果杂交等遗传交换参与到了类群分化中, 这种技术不但无法检测到遗传交换的存在, 甚至可能会给出相当不可靠的结果, 比如奇怪的枝长、特别的群体大小等。

系统发育网络允许节点(node, 实际上是系统进化中的祖先种)间的网状连接, 它可以替代二歧分枝的系统发育树, 从而体现杂交或其他网状进化的存在, 也是呈现杂交存在的重要方式。目前, 系统发育网络分为直接的(implicit network)和间接的(explicit network)两大类(Solis-Lemus & Ane, 2016)。直接网络简单地将基因树间的不一致性呈现在网络上(Than et al, 2008; Huson & Scornavacca, 2011; Grunewald et al, 2013), 它们往往运算速度快。但这种网络中的内部节点并不对应祖先类群, 导致很难对它们进行深入解读。与直接网络相对应, 间接网络中可呈现网状进化并且内部节点与祖先类群相对应, 对于系统发育分析来说, 它们的价值更大。组合法(combinatorial method)和基于模型法(model-based method)是两类推算间接网络的策略。组合法往往无法把不完全谱系分选与杂交区分开, 也没有

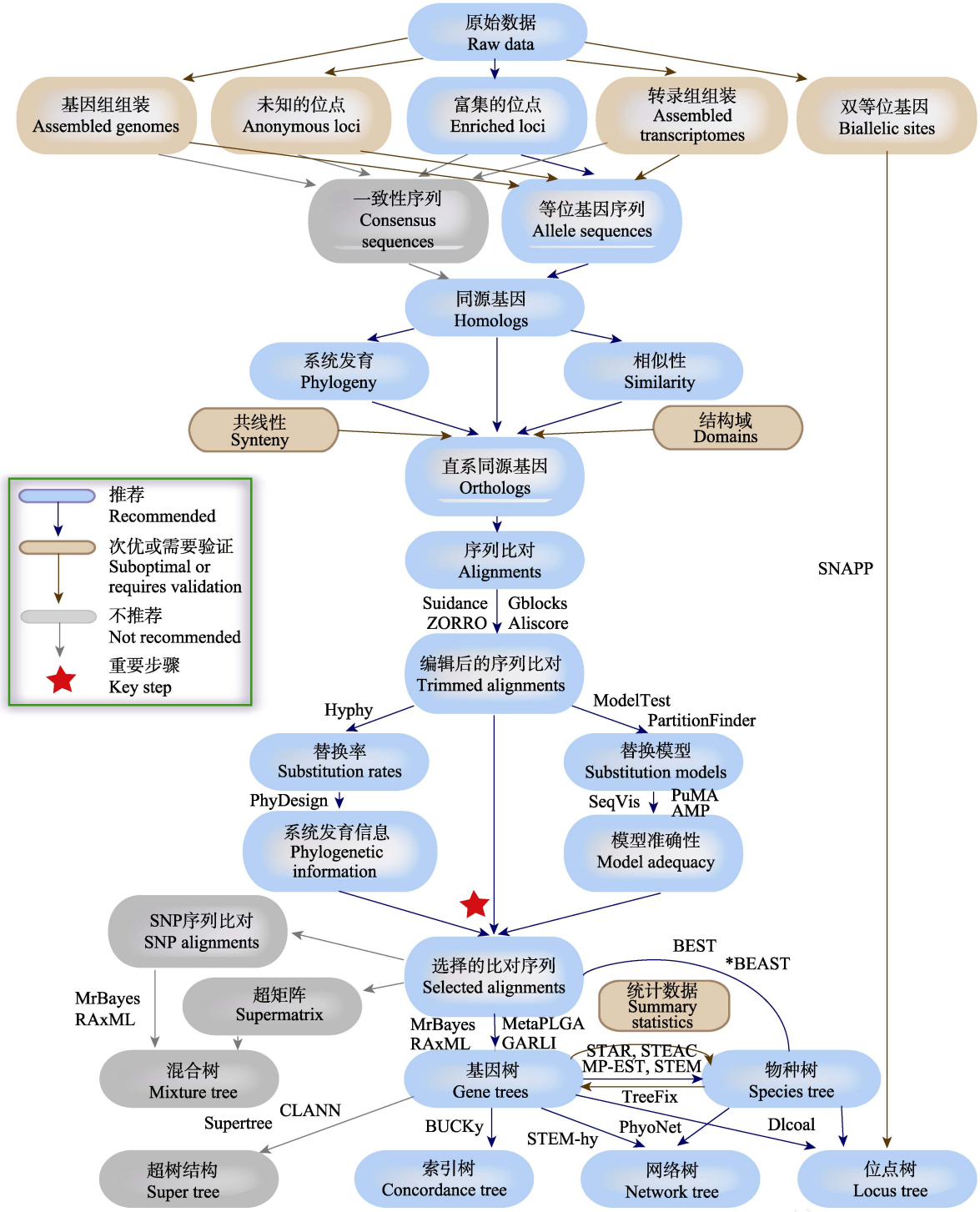

图2 系统发育策略检验杂交存在的基本流程(改编自**Lemmon & Lemmon, 2013**)。图中给出了利用基因组水平测序技术，通过数据处理、数据质量评估、数据筛选、直系同源基因识别、序列分析、碱基替代模型选取、系统发育树及系统发育网络重建等多个步骤的简要流程和相关的分析软件。依据操作的难易程度、数据的可靠性、是否能区分杂交和不完全谱系分选等因素，我们将各技术环节区分为“推荐”、“次优”和“不推荐”等等级，并列举了各环节部分可利用的软件。

Fig. 2 The brief work-flow of phylogenetic strategy used to test hybridization (adapted from Lemmon & Lemmon, 2013). The work-flow contains multiple steps including data processing, data quality evaluation, data screening, orthologous gene identification, sequence analysis, selection of base substitution model, phylogenetic tree and phylogenetic network reconstruction. We sort out the hierarchy of recommended, suboptimal or requires validation and not recommended, in consideration of the degree of operation, the reliability of the data, whether can distinguish hybridization, incomplete lineage sorting and other factors. And we list the available softwares for each step.

考虑基因不同进化历程(比如基因的重复和丢失)带来的误差(Gambette et al, 2012)。但这类方法的运算速度相对较快。基于模型的方法在利用多位点数据构建系统发育网络时同时考虑了杂交和不完全谱系分选(Strimmer & Moulton, 2000; Meng & Kubatko, 2009; Yu et al, 2012), 相比于组合法更为准确。这类基于模型的系统发育网络方法也被称为多物种网络溯祖(multispecies network coalescent, MSNC) (Yu et al, 2012, 2013, 2014)。PhyloNet (http://bioinfo.cs.rice.edu/phylonet/)是这类方法的一个重要软件工具, 它利用最大似然算法, 基于一系列基因位点信息估算系统发育网络(Yu et al, 2014)。PhyloNet可成功地应对由于不完全谱系分选造成的位点间的不一致性, 并且通过校验过程控制网络的复杂性。但它的计算量很大, 对超过10个分类群或4次潜在杂交事件的分析工作就不适用了。针对这种计算上的挑战, SNaQ (species networks applying quartets)算法通过计算假似然性(pseudolikelihood)极大提高了计算效能, 可以应用于有数十分类单位、数千位点的项目, 并且可以估算出大量的潜在杂交事件, 同时应对不完全谱系分选的存在(Solis-Lemus & Ane, 2016)。该方法对应的软件包叫做PhyloNetworks (https://github.com/crsl4/PhyloNetworks.jl/)。感兴趣的读者可以在网上找到非常完整的教程(http://crsl4.github.io/PhyloNetworks.jl/latest/)。事实上, PhyloNet的作者也通过计算假似然性提高了原有的计算效能(Yu & Nakhleh, 2015)。

不同于上面的系统发育网络技术, 如果研究体系中只包含2个二倍体类群和它们1次杂交事件产生的1个异源四倍体类群, 并且异源四倍体中2个二倍体基因组间没有重组, 那么Allopolyploids软件中的算法可以用来在排除不完全谱系分选的情况下估算这种杂交成种的历史(https://sites.google.com/site/touchingthedata/software/allopolyploids/)(Jones et al, 2013)。

3.3　群体遗传学分析策略

群体遗传学手段是用来探讨近缘种间或种内类群间分化、检测杂交的存在、澄清杂交特征的重要手段。实际上, 群体遗传学手段实现的第一步是从认识群体结构开始的, 有了群体结构的划分, 才有后面检测杂交存在、澄清杂交模式的可能。相应群体遗传分析流程的概括见图3。

3.3.1　主成分分析

主成分分析(principal component analysis, PCA)是一类经典的多元统计技术, 很早就开始被应用在遗传数据分析中(Menozzi et al, 1978)。主成分分析可以将样本间的关系呈现在主成分形成的二维或三维空间上, 但它不能提供杂交存在与否的检验(Patterson et al, 2012), 介于类群间的样本不一定就来自于类群间的杂交(Yang et al, 2012)。因此, EIGENSTRAT等软件基于PCA来估算杂交(Price et al, 2006)可能会得出错误的结果。除了EIGENSTRAT, adegenet也是一个主成分分析的工具(Jombart, 2008)。

3.3.2　聚类方法

这里提到的聚类方法(clustering methods)是指按等位基因频率分布模式, 将个体的来源归入相应祖先群体(ancestral populations)中。简单地说, 聚类的基本过程是: 先假定有*K*个可能存在的祖先群体; 然后依据个体基因型计算它们被判别为各群体的概率; 在计算时, 假定各位点群体上的频率分布服从哈代–温伯格平衡(Hardy-Weinberg equilibrium)。利用structure (Pritchard et al, 2000; Falush et al, 2003)和admixture (Alexander et al, 2009)等基于模型的聚类方法对每一个参试个体估算其来自于*K*个祖先群体中某个特定群体的遗传比例。祖先群体的数量*K*是个先验值, 可以依据一些已知的研究背景确定, 也可以通过最小*K*值的策略估算出来。旨在揭示杂交的群体遗传研究中, 此类聚类方法的应用非常普遍。InStruct是structure软件的拓展, 它没有哈代–温伯格平衡的前提假设, 可以同时对群体结构和近交率进行估算(Gao et al, 2007), 这对近交或自交比较频繁的植物研究有利用价值。类似的聚类运算的软件还有不少, 例如FRAPPE (Tang et al, 2007)、sNMF (Frichot et al, 2014)、Geneland (Guillot et al, 2004, 2005)等。它们在操作上类似, 运算速度相差不少, 有些可以考虑样本群体间的环境或地理信息。这类聚类方法可以应用于基因型数据、二倍化的单倍型数据(pseudo-homozygous genotype)或者基因型的似然值数据(genotype likelihood)。ANGSD软件包中的NGSAdmix软件(Skotte et al, 2013)可以基于基因组水平基因型的似然值数据而不是基因型数据做聚类分析。

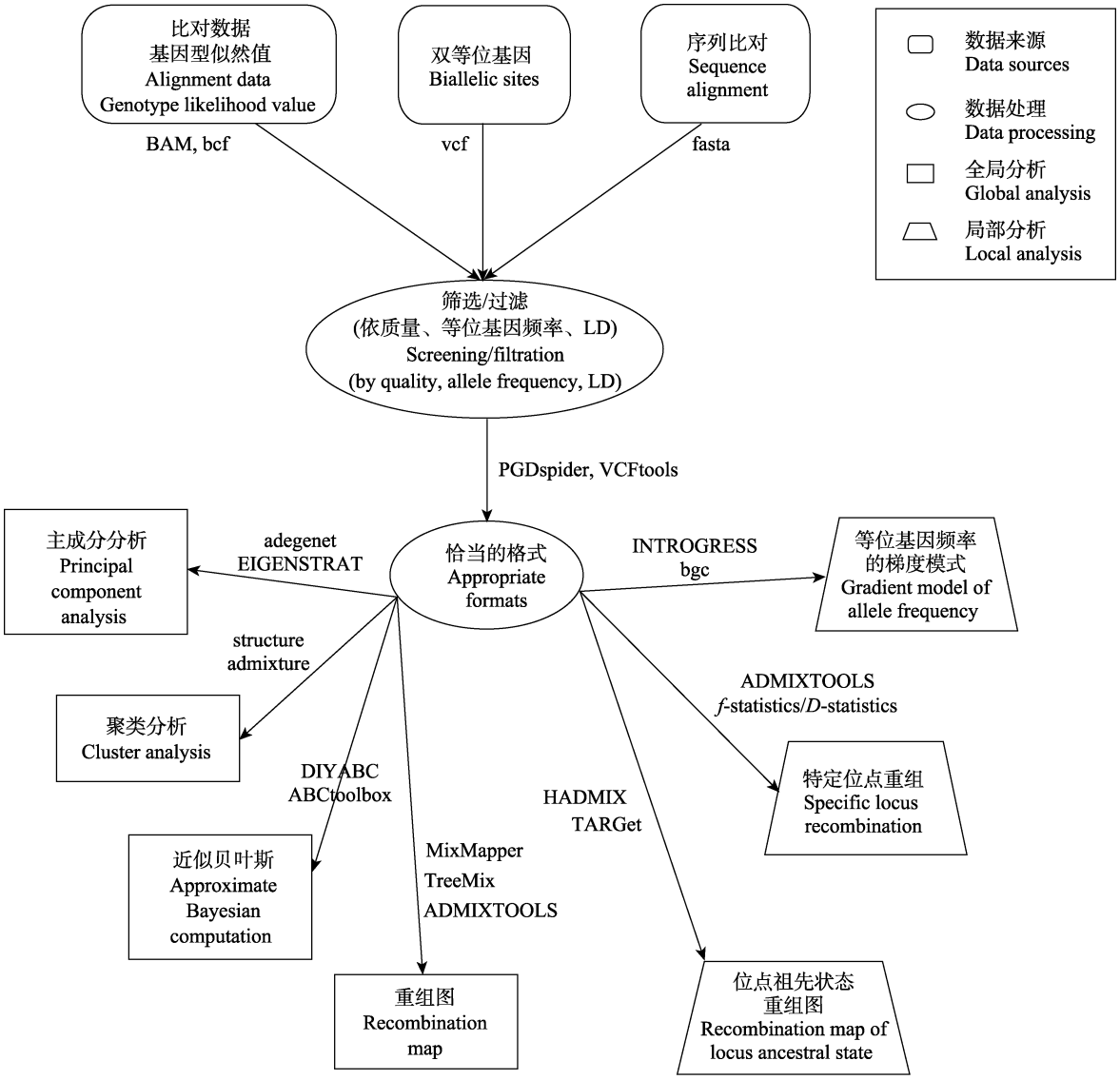

图3 检验杂交群体遗传分析的基本流程。图中给出了利用当前流行的群体基因组学技术，通过数据收集、数据处理和分析等环节，对杂交的存在和模式进行全局和局部检验的基本流程，并列举出了数据格式类型和部分可利用的软件。

Fig. 3 The current population genomic work-flow used to test hybridization. The basic process of global and local testing for the existence and mode of hybridization through data collection, data processing and analysis are presented. The data format and some of the available softwares are listed.

这类聚类方法只是假定祖先群体的个数，没有对各种可能的群体模型进行评估，因此当存在复杂的杂交重组历史时，可能会给出错误的结果(Patterson et al, 2012)。此外，这里全局性的祖先重建方法也会受到抽样的影响，比如当某些遗传分化类群没有被抽样，或不同类群间抽样不均衡，都会影响对祖先群体遗传组分(ancestry components)的估算(Mcvean, 2009)。

3.3.3 等位基因频率的梯度变化模式

等位基因频率在连续分布的地理群体上的梯度变异模式很早就被用来识别杂交带和判断杂交带的存在形式(Barton, 1983; Szymura & Barton, 1986)，它也被用于寻找适应性变异和生殖隔离的遗传基础(Gompert & Buerkle, 2009, 2011; Fitzpatrick, 2013)。INTROGRESS (Gompert & Buerkle, 2010)、bgc (Gompert & Buerkle, 2012)和Hlest (Fitzpatrick, 2013)等软件可以实现这类计算。但这类检测方法来源于早期的杂交带研究，它们多关注分化程度不大的类群或群体间的杂交，对更为古老的杂交事件，它们可能无法识别。

3.3.4　几种统计参数和检验

上面提到的主成分分析、聚类分析(比如基于structure和admixture的分析)可以用来澄清群体遗传结构, 并对杂交的存在提供全局性的线索(那些看似杂交的迹象可能由杂交以外的其他因素导致), 但它们都没有提供杂交是否存在的检验。比如, 与距离关联的分化(isolation by distance)可以产生在PCA上的梯度变异。基于structure和admixture的结果也很难做出对群体历史的推算, 因为它们没有对特定的群体历史模型进行检验, 而是简单地假定抽样群体都是从某特定群体快速辐射分化而来。Patterson等(2012)综合已有工作(Reich et al, 2009; Green et al, 2010; Durand et al, 2011; Moorjani et al, 2011), 归纳出了几个针对群体间杂交历史的检验和相应的参数, 并提供了实现有关计算的软件包(ADMIXTOOLS)。这些参数和检验包括1个三群体检验(*f*3-statistic, the three-population test)和2个四群体检验(*D*-statistics或者ABBA-BABA test, 以及*f*4-statistics或*F*4-ratio)。这里的*f*3-statistic和*f*4-statistics也被统称为*F*-statistics。

三群体检验基于对群体间等位基因频率关联性进行评估, 可以对群体间即便是非常近期发生的杂交事件给出清晰的验证。研究人员成功地利用*f*3-statistic估算出尼安德特人(*Homo neanderthalensis*)贡献了非洲现代人群1.5–2.1%的遗传变异(Prufer et al, 2014)。通过选用外类群, *f*3-statistic也被用来揭示多个不同类群群体的多重起源历史(Haak et al, 2015; Haber et al, 2016; Mörseburg et al, 2016)。

*F*4-ratio检验可以用来推断杂交造成的遗传重组的比例, 即使在不知道祖先群体的情况下, 也可以依据对系统发育关系的假定进行推断。四群体检验不但可以对杂交是否存在给出证据, 还可以提供基因流的方向。这些四群体检验也同样被广泛用于检测现代人类基因组中的古人类成分和相关的历史过程(Green et al, 2010; Fu et al, 2014, 2016; Prufer et al, 2014; Meyer et al, 2016)。最近对非洲丽鱼的研究将该方法进一步拓展为五群体检验, 或称为*f*5-statistic (Meier et al, 2017), 该研究表明古老的杂交事件带来遗传变异, 成为后期适应性分化和物种形成的重要推动力。

与*f*3-statistic和*f*4-statistics不同, *D*-statistics不需要对群体进行细致的取样以得到等位基因频率的准确信息, 它可以应用于一个类群只有一条序列的情况。对这些参数和检验的详细描述请参考Patterson等(2012)的文章。

3.3.5　重组图

重组图(admixture graph)是一个允许类群间基因交流的系统发育树的拓展。基于上面提到的*F*-statistics就可以构建这样的图, 目前有几种工具可以利用。MixMapper (Lipson et al, 2014)是一个半自动的工具。它首先基于由一对明显没有杂交的群体间等位基因频率(*F*2)求算的遗传聚类生成邻接树(neighbor-joining tree), 然后在允许基因流的情况下把剩下的群体添加到图上来(Lipson et al, 2013, 2014)。TreeMix (Pickrell & Pritchard, 2012)实现了与MixMapper在理论上类似的算法。在假定重组或者基因流个数的情况下, 它可以全自动地计算出重组图。不同的重组次数的假定会对结果带来极大影响, 选取时需谨慎。上面提到的ADMIXTOOLS软件(Patterson et al, 2012)中也提供了类似的工具。但这里的重组图工具更加稳健, 因为它提供了假设中的重组模式是否和数据相符的检验。重组图的手段已经在人类遗传领域广泛应用, 曾被用来揭示北美原著民的遗传组成和来源(Raghavan et al, 2014b)、丹尼索瓦人(Denisovans)对现代人群的遗传贡献(Meyer et al, 2012)、新世界寒带地区的人类定居历史(Raghavan et al, 2014a)。

3.3.6　位点的祖先状态

与杂交关联的是不同特征的基因组间的重组, 这种重组会带来一个嵌合的基因组, 在嵌合的基因组上不同的区域体现着不同的进化历程。识别这些位点不但对澄清种间或群体间的基因交流有意义(Green et al, 2010; Reich et al, 2010; Prufer et al, 2014), 同时也可以提供重组模式和适应性分化的信息(Kim & Rothschild, 2014)。这些位点的识别对于揭示生殖隔离的遗传基础和找寻物种形成基因(speciation genes)有重要意义(Wu & Ting, 2004; Nosil & Schluter, 2011; Burri et al, 2015), 同时对建立针对濒危物种的保护策略有应用价值(der Sarkissian et al, 2015)。基于基因组水平的滑动窗(sliding window)分析可以追溯不同染色体区段的祖先状态, 上面提到过的*D*-statistics就是一个很好的工具(Kronforst et al, 2013; Smith & Kronforst, 2013)。

已经有一些估算位点祖先状态的计算工具,

Padhukasahasram (2014)就进行了较为详尽的综述。一般来说, 这些计算工具需要借助隐马尔科夫模型(hidden Markov models, HMMs)来确定不同亲本来源的基因组区段的界限。HAPMIX软件可以利用未定相(unphased)的基因型数据确定杂种个体特定位点的祖先来源(Price et al, 2009)。CRFs应用一个广义隐马尔科夫模型(generalized hidden Markov model), 借助训练集数据实现了对祖先来源的估算(Sankararaman et al, 2012)。实现这类分析的软件还包括: SABER (Tang et al, 2006)、HAPA (Sundquist et al, 2008)、LAMP (Sankararaman et al, 2008)、LAMP-LD/LAMP-HAP (Baran et al, 2012)、WINPOP (Pasaniuc et al, 2009)、SupportMix (Omberg et al, 2012)、ASPCA (Moreno-Estrada et al, 2013)、ALLOY (Rodriguez et al, 2013)、RFMix (Maples et al, 2013)、Lanc-CSV (Brown & Pasaniuc, 2014)和EILA (Yang et al, 2013)。EILA没有对输入数据位点间连锁平衡的前提假设, 且比LAMP和HAPMIX更准确, 运行速度也不慢。Lanc-CSV实现了经典算法, 但对超大样本数据有速度上的优势。

在位点水平祖先状态和重组历史的估算上, 祖先重组图(ancestral recombination graph, ARG)也是一个重要策略。祖先重组图试图将所有样本的所有位点经历的溯祖过程(coalescence)和重组历史(recombination events)呈现出来, 是详细解析杂交历史的重要工具(Siepel, 2009)。这类方法在计算上的挑战不小, 目前仅有少量工具可用, 比如ARGWeaver (Rasmussen et al, 2014)、Beagle (Song & Hein, 2005)和TARGet (Cámara et al, 2016)。TARGet是一个最新也是唯一声称可以分析数百基因组特定位点数据的工具。

3.3.7 ABC算法

基于最大似然算法的策略往往很复杂, 需要大量的计算。近似贝叶斯算法(approximate Bayesian computation, ABC)不需要计算似然值, 它提供了一个检验各种复杂进化过程是否存在的平台(Beaumont, 2010)。近似贝叶斯计算中, 首先对待检测过程依据一定的期望建立多重模拟, 然后将真实数据和模拟产生的数据进行比较, 再通过选取那些与真实数据最类似的模型来实现对待检验假设的推断。近似贝叶斯计算需要完成数据模拟、统计计算、模型比较等步骤, 目前有多种针对性的工具可以利用, 可参考一些重要综述文章(Csilléry et al, 2010; Sunnåker et al, 2013; Lintusaari et al, 2016)。DIYABC (Cornuet et al, 2014)、ABCtoolbox (Wegmann et al, 2010)和abc (Csilléry et al, 2011)等是通常用来实现ABC计算的平台。利用基因组水平未定向数据和ABC算法揭示杂交事件存在的一个研究实例来自针对胡蜂(*Biorhiza pallida*)的工作(Robinson et al, 2014)。

4 总结

作为重要的进化过程, 杂交对于生物多样性产生和维持的意义受到越来越多的关注。针对如何检测杂交存在这个问题, 本文对来自多个学科的理论和技术发展作了综述, 以基因组时代的基因型获取和数据分析技术为重点, 归纳了各新兴技术的特点, 作了多方面的比较, 在一些重要应用领域提出了可能存在的问题和解决对策。虽然本文所综述的检测杂交是否存在的研究方法多来自研究较为透彻的高等动植物类群, 但鉴于这些方法的通用性, 亦可拓展到其他生物类群。

致谢: 感谢专辑组织者、两位审稿人和编委的建设性意见, “生物进化与系统学论坛” QQ群和2016年在中国科学院上海辰山植物科学研究中心举办的“自然杂交与生物多样性研讨会”给了作者很多重要的启发, 在此一并表示感谢。

参考文献

Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJE, Bierne N, Boughman J, Brelsford A, Buerkle CA, Buggs R, Butlin RK, Dieckmann U, Eroukhmanoff F, Grill A, Cahan SH, Hermansen JS, Hewitt G, Hudson AG, Jiggins C, Jones J, Keller B, Marczewski T, Mallet J, Martinez-Rodriguez P, Möst M, Mullen S, Nichols R, Nolte AW, Parisod C, Pfennig K, Rice AM, Ritchie MG, Seifert B, Smadja CM, Stelkens R, Szymura JM, Väinölä R, Wolf JBW, Zinner D (2013) Hybridization and speciation. Journal of Evolutionary Biology, 26, 229–246.

Abby SS, Tannier E, Gouy M, Daubin V (2012) Lateral gene transfer as a support for the tree of life. Proceedings of the National Academy of Sciences, USA, 109, 4962–4967.

Ackermann RR, Mackay A, Arnold ML (2016) The hybrid origin of “Modern” humans. Evolutionary Biology, 43, 1–11.

Adams RP (1982) A comparison of multivariate methods for

the detection of hybridization. Taxon, 31, 646–661.
Albert AYK, Schluter D (2005) Selection and the origin of species. Current Biology, 15, 283–288.
Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ (2007) Direct selection of human genomic loci by microarray hybridization. Nature Methods, 4, 903–905.
Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 19, 1655–1664.
Anderson E, Hubricht L (1938) Hybridization in *Tradescantia*. III. The evidence for introgressive hybridization. American Journal of Botany, 25, 396–402.
Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. Nature Reviews Genetics, 17, 81–92.
Arnold ML (2016) Divergence with Genetic Exchange. Oxford University Press, Oxford.
Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE, 3, e3376.
Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, Rodriguezcintron W, Chapela R, Ford JG, Avila PC (2012) Fast and accurate inference of local ancestry in Latino populations. Bioinformatics, 28, 1359–1367.
Barton NH (1983) Multilocus clines. Evolution, 37, 454–471.
Bauer DC (2011) Variant calling comparison CASAVA1.8 and GATK. Nature Precedings, doi: 10.1038/npre. 2011.6107.1/.
Beaumont MA (2010) Approximate Bayesian computation in evolution and ecology. Annual Review of Ecology, Evolution, and Systematics, 41, 379–406.
Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D (2004) Ultraconserved elements in the human genome. Science, 304, 1321–1325.
Boussau B, Szöllosi GJ, Duret L, Gouy M, Tannier E, Daubin V (2013) Genome-scale coestimation of species and gene trees. Genome Research, 23, 323–330.
Branstetter MG, Longino JT, Ward PS, Faircloth BC (2017) Enriching the ant tree of life: enhanced UCE bait set for genome-scale phylogenetics of ants and other *Hymenoptera*. Methods in Ecology and Evolution, 8, 768–776.
Brown R, Pasaniuc B (2014) Enhanced methods for local ancestry assignment in sequenced admixed individuals. PLoS Computational Biology, 10, e1003555.
Bujarski J (2013) Genetic recombination in plant-infecting messenger-sense RNA viruses: overview and research perspectives. Frontiers in Plant Science, 4, doi: 10.3389/fpls. 2013.00068.
Burri R, Nater A, Kawakami T, Mugal CF, Olason PI, Smeds L, Suh A, Dutoit L, Bureš S, Garamszegi LZ (2015) Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula flycatchers*. Genome Research, 25, 1656–1665.
Cámara PG, Levine AJ, Rabadán R (2016) Inference of ancestral recombination graphs through topological data analysis. PLoS Computational Biology, 12, e1005071.
Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution, 17, 540–552.
Castric V, Bechsgaard J, Schierup MH, Vekemans X (2008) Repeated adaptive introgression at a gene under multiallelic balancing selection. PLoS Genetics, 4, e1000168.
Chan YC, Roos C, Inouemurayama M, Inoue E, Shih CC, Pei JC, Vigilant L (2010) Mitochondrial genome sequences effectively reveal the phylogeny of *Hylobates* gibbons. PLoS ONE, 5, e14419.
Chester M, Leitch AR, Soltis PS, Soltis DE (2010) Review of the application of modern cytogenetic methods (FISH/GISH) to the study of reticulation (polyploidy/hybridisation). Genes, 1, 166–192.
Choler P, Erschbamer B, Tribsch A, Gielly L, Taberlet P (2004) Genetic introgression as a potential to widen a species' niche: insights from alpine *Carex curvula*. Proceedings of the National Academy of Sciences, USA, 101, 171–176.
Cohan FM (2002) Sexual isolation and speciation in bacteria. Genetica, 116, 359–370.
Cohan FM, Kane M (2001) Bacterial species and speciation. Systematic Biology, 50, 513–524.
Cornuet JM, Pudlo P, Veyssier J, Dehnegarcia A, Gautier M, Leblois R, Marin JM, Estoup A (2014) DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. Bioinformatics, 30, 1187.
Crawford NG, Faircloth BC, Mccormack JE, Brumfield RT, Winker K, Glenn TC (2012) More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biology Letters, 8, 783–786.
Csilléry K, Blum MGB, Gaggiotti OE, François O (2010) Approximate Bayesian computation (ABC) in practice. Trends in Ecology and Evolution, 25, 410–418.
Csilléry K, François O, Blum MGB (2011) abc: an R package for approximate Bayesian computation (ABC). Methods in Ecology and Evolution, 3, 475–479.
Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics, 12, 499–510.
Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. PLoS Genetics, 2, e68.
Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic and the multispecies coalescent. Trends in

Ecology and Evolution, 24, 332–340.

Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biology, 3, e314.

Delport W, Poon AFY, Frost SDW, Pond SLK (2010) Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics, 26, 2455–2457.

Depotter JR, Seidl MF, Wood TA, Thomma BP (2016) Interspecific hybridization impacts host range and pathogenicity of filamentous microbes. Current Opinion in Microbiology, 32, 7–13.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly M (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics, 43, 491–498.

der Sarkissian C, Ermini L, Schubert M, Yang MA, Librado P, Fumagalli M, Jónsson H, Bargal GK, Albrechtsen A, Vieira FG (2015) Evolutionary genomics and conservation of the endangered Przewalski's horse. Current Biology, 25, 2577–2583.

Duncan KE, Istock CA, Graham JB, Ferguson N (1989) Genetic exchange between *Bacillus subtilis* and *Bacillus licheniformis*: variable hybrid stability and the nature of bacterial species. Evolution, 43, 1585–1609.

Durand EY, Patterson N, Reich D, Slatkin M (2011) Testing for ancient admixture between closely related populations. Molecular Biology and Evolution, 28, 2239–2252.

Earl AM, Losick R, Kolter R (2008) Ecology and genomics of *Bacillus subtilis*. Trends in Microbiology, 16, 269–275.

Ebersberger I, Strauss S, von Haeseler A (2009) HaMStR: profile hidden Markov model based search for orthologs in ESTs. BMC Evolutionary Biology, 9, doi:10.1186/1471-2148-9-157.

Edwards S (2016) Species trees. In: Encyclopedia of Evolutionary Biology (ed. Kliman R), pp. 236–244. Academic Press, Oxford.

Edwards SV, Liu L, Pearl DK (2007) High-resolution species trees without concatenation. Proceedings of the National Academy of Sciences, USA, 104, 5936–5941.

Ellstand NC, Elam D (1993) Population genetic consequences of small population size: implications for plant conservation. Annual Review of Ecology, Evolution, and Systematics, 24, 217–242.

Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA (2011) Local *de novo* assembly of RAD paired-end contigs using short sequencing reads. PLoS ONE, 6, e18561.

Falush D, Stephens M, Pritchard JK (2003) Inference of population sructure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics, 164, 1567–1587.

Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Systematic Zoology, 27, 401–410.

Felsenstein J (2004) Inferring Phylogenies. Sinauer Associates, Sunderland.

Fitzpatrick BM (2013) Alternative forms for genomic clines. Ecology and Evolution, 3, 1951–1966.

Freeling M, Subramaniam S, Yano M, Tuberosa R (2009) Conserved noncoding sequences (CNSs) in higher plants. Current Opinion in Plant Biology, 12, 126–132.

Frichot E, Mathieu F, Trouillon T, Bouchard G, Francois O (2014) Fast and efficient estimation of individual ancestry coefficients. Genetics, 196, 973–983.

Fu QM, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PLF, Petri AA, Prüfer K, Filippo CD (2014) The genome sequence of a 45,000-year-old modern human from western Siberia. Nature, 514, 445–449.

Fu QM, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, Furtwängler A, Haak W, Meyer M, Mittnik A (2016) The genetic history of Ice Age Europe. Nature, 534, 200–205.

Gambette P, Berry V, Paul C (2012) Quartets and unrooted phylogenetic networks. Journal of Bioinformatics & Compututational Biology, 10, 1250004–1250023.

Gao H, Williamson S, Bustamante CD (2007) A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics, 176, 1635–1651.

Gao J, Wang B, Mao JF, Ingvarsson P, Zeng QY, Wang XR (2012) Demography and speciation history of the homoploid hybrid pine *Pinus densata* on the Tibetan Plateau. Molecular Ecology, 21, 4811–4827.

Gnirke A, Melnikov A, Maguire J, Rogov P, Leproust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotechnology, 27, 182–189.

Godden G, Jordon-Thaden I, Chamala S, Crowl AA, García N, Germain-Aubrey C, Heaney JM, Latvis M, Qi XS, Gitzendanner MA (2012) Making next-generation sequencing work for you: approaches and practical considerations for marker development and phylogenetics. Plant Ecology & Diversity, 5, 427–450.

Gompert Z, Buerkle CA (2012) bgc: software for Bayesian estimation of genomic clines. Molecular Ecology Resources, 12, 1168–1176.

Gompert Z, Buerkle CA (2009) A powerful regression-based method for admixture mapping of isolation across the genome of hybrids. Molecular Ecology, 18, 1207–1224.

Gompert Z, Buerkle CA (2010) INTROGRESS: a software package for mapping components of isolation in hybrids. Molecular Ecology Resources, 10, 378–384.

Gompert Z, Buerkle CA (2011) Bayesian estimation of genomic clines. Molecular Ecology, 20, 2111–2127.

Grant PR, Grant BR (2014) Evolutionary biology: speciation undone. Nature, 507, 178–179.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH (2010) A draft sequence of the Neandertal genome. Science, 328, 710–722.

Griffin PC, Robin C, Hoffmann AA (2011) A next-generation sequencing method for overcoming the multiple gene copy problem in polyploid phylogenetics, applied to *Poa* grasses. BMC Biology, 9, 1–18.

Gross BL, Rieseberg LH (2005) The ecological genetics of homoploid hybrid speciation. Journal of Heredity, 96, 241–252.

Grunewald S, Spillner A, Bastkowski S, Bogershausen A (2013) SuperQ: computing supernetworks from quartets. Computational Biology & Bioinformatics IEEE/ACM Transactions, 10, 151–160.

Guillot G, Mortier F, Estoup A (2004) Geneland: a program for landscape genetics. Molecular Ecology Notes, 5, 712–715.

Guillot G, Estoup A, Mortier F, Cosson JF (2005) A spatial satistical model for landscape genetics. Genetics, 170, 1261–1280.

Gunnarsdóttir ED, Li M, Bauchet M, Finstermeier K, Stoneking M (2011) High-throughput sequencing of complete human mtDNA genomes from the Philippines. Genome Research, 21, 1–11.

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, Fu QM, Mittnik A, Banffy E, Economou C, Francken M, Friederich S, Pena RG, Hallgren F, Khartanovich V, Khokhlov A, Kunst M, Kuznetsov P, Meller H, Mochalov O, Moiseyev V, Nicklisch N, Pichler SL, Risch R, Rojo GMA, Roth C, Szecsenyi-Nagy A, Wahl J, Meyer M, Krause J, Brown D, Anthony D, Cooper A, Alt KW, Reich D (2015) Massive migration from the steppe was a source for Indo-European languages in Europe. Nature, 522, 207–211.

Haber M, Mezzavilla M, Xue YL, Comas D, Gasparini P, Zalloua P, Tyler-Smith C (2016) Genetic evidence for an origin of the Armenians from Bronze Age mixing of multiple populations. European Journal of Human Genetics, 24, 931–936.

Hapke A, Thiele D (2016) GIbPSs: a toolkit for fast and accurate analyses of genotyping-by-sequencing data without a reference genome. Molecular Ecology Resources, 16, 979–990.

Hegarty MJ, Hiscock SJ (2008) Genomic clues to the evolutionary success of polyploid plants. Current Biology, 18, R435–R444.

Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. Molecular Biology and Evolution, 27, 570–580.

Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, Myers S (2014) A genetic atlas of human admixture history. Science, 343, 747–751.

Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ (2007) Genome-wide *in situ* exon capture for selective resequencing. Nature Genetics, 39, 1522–1527.

Huang CH, Sun RR, Hu Y, Zeng LP, Zhang N, Cai LM, Zhang Q, Koch MA, Ihsan AS, Edger PP (2015) Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. Molecular Biology and Evolution, 33, 394–412.

Huson DH, Scornavacca C (2011) A survey of combinatorial methods for phylogenetic networks. Genome Biology and Evolution, 3, 23–35.

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics, 24, 1403–1405.

Jones G, Sagitov S, Oxelman B (2013) Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. Systematic Biology, 62, 467–478.

Kenny EM (2011) Multiplex target enrichment using DNA indexing for ultra-high throughput SNP Detection. DNA Research, 18, 31–38.

Kim C, Guo H, Kong WQ, Chandnani R, Shuang LS, Paterson AH (2016) Application of genotyping by sequencing technology to a variety of crop breeding programs. Plant Science, 242, 14–22.

Kim ES, Rothschild MF (2014) Genomic adaptation of admixed dairy cattle in East Africa. Frontiers in Genetics, 5, doi: 10.3389/fgene.2014.00443.

Kim M, Cui ML, Cubas P, Gillies A, Lee K, Chapman MA, Abbott RJ, Coen E (2008) Regulatory genes control a key morphological and ecological trait transferred between species. Science, 322, 1116–1119.

Kosakovsky PSL, Posada D, Gravenor MB, Woelk CH, Frost SDW (2006) GARD: a genetic algorithm for recombination detection. Bioinformatics, 22, 3096–3098.

Kronforst MR, Hansen MEB, Crawford NG, Gallant JR, Zhang W, Kulathinal RJ, Kapan DD, Mullen SP (2013) Hybridization reveals the evolving genomic architecture of speciation. Cell Reports, 5, 666–677.

Kubatko LS, Carstens BC, Knowles LL (2009) STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics, 25, 971–973.

Kubatko LS, Degnan JH (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. Systematic Biology, 56, 17–24.

Kumar S (2012) Statistics and truth in phylogenomics. Molecular Biology and Evolution, 29, 457–472.

Lanfear R, Calcott B, Ho SY, Guindon S (2012) Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Molecular Biology and Evolution, 29, 1695–1701.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, Berger B, Economou C, Bollongino R, Fu QM, Bos KI, Nordenfelt S, Li H, de Filippo C, Prufer K, Sawyer S, Posth C, Haak W, Hallgren F, Fornander E, Rohland N, Delsate D,

Francken M, Guinet JM, Wahl J, Ayodo G, Babiker HA, Bailliet G, Balanovska E, Balanovsky O, Barrantes R, Bedoya G, Ben-Ami H, Bene J, Berrada F, Bravi CM, Brisighelli F, Busby GBJ, Cali F, Churnosov M, Cole DEC, Corach D, Damba L, van Driem G, Dryomov S, Dugoujon JM, Fedorova SA, Gallego RI, Gubina M, Hammer M, Henn BM, Hervig T, Hodoglugil U, Jha AR, Karachanak-Yankova S, Khusainova R, Khusnutdinova E, Kittles R, Kivisild T, Klitz W, Kucinskas V, Kushniarevich A, Laredj L, Litvinov S, Loukidis T, Mahley RW, Melegh B, Metspalu E, Molina J, Mountain J, Nakkalajarvi K, Nesheva D, Nyambo T, Osipova L, Parik J, Platonov F, Posukh O, Romano V, Rothhammer F, Rudan I, Ruizbakiev R, Sahakyan H, Sajantila A, Salas A, Starikovskaya EB, Tarekegn A, Toncheva D, Turdikulova S, Uktveryte I, Utevska O, Vasquez R, Villena M, Voevoda M, Winkler CA, Yepiskoposyan L, Zalloua P, Zemunik T, Cooper A, Capelli C, Thomas MG, Ruiz-Linares A, Tishkoff SA, Singh L, Thangaraj K, Villems R, Comas D, Sukernik R, Metspalu M, Meyer M, Eichler EE, Burger J, Slatkin M, Paabo S, Kelso J, Reich D, Krause J (2014) Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature, 513, 409–413.

Leaché AD, Rannala B (2010) The accuracy of species tree estimation under simulation: a comparison of methods. Systematic Biology, 60, 126–137.

Lefeuvre P, Moriones E (2015) Recombination as a motor of host switches and virus emergence: geminiviruses as case studies. Current Opinion in Virology, 10, 14–19.

Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM (2009) The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Systematic Biology, 58, 130–145.

Lemmon AR, Moriarty EC (2004) The importance of proper model assumption in Bayesian phylogenetics. Systematic Biology, 53, 265–277.

Lemmon EM, Lemmon AR (2013) High-throughput genomic data in systematics and phylogenetics. Annual Review of Ecology, Evolution, and Systematics, 44, 99–121.

Liang D, Mao JF, Zhao W, Zhou XQ, Yuan HW, Wang LM, Xing FQ, Wang XR, Li Y (2013) Seedling performance of *Pinus densata* and its parental population in the habitat of *P. tabuliformis*. Chinese Journal of Plant Ecology, 37, 150–163. (in Chinese with English abstract) [梁冬, 毛建丰, 赵伟, 周先清, 袁虎威, 王黎明, 邢芳倩, 王晓茹, 李悦 (2013) 高山松及其亲本种群在油松生境下的苗期性状. 植物生态学报, 37, 150–163.]

Lintusaari J, Gutmann MU, Dutta R, Kaski S, Corander J (2016) Fundamentals and recent developments in approximate Bayesian computation. Systematic Biology, 66, 66–82.

Lipson M, Loh PR, Levin A, Reich D, Patterson N, Berger B (2013) Efficient moment-based inference of admixture parameters and sources of gene flow. Molecular Biology and Evolution, 30, 1788–1802.

Lipson M, Loh PR, Patterson N, Moorjani P, Ko YC, Stoneking M, Berger B, Reich D (2014) Reconstructing Austronesian population history in Island Southeast Asia. Nature Communications, 5, doi: 10.1038/ncomms5689.

Liu L (2008) BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics, 24, 2542–2543.

Liu LQ, Gu ZJ (2011) Genomic *in situ* hybridization identifies genome donors of *Camellia reticulata* (Theaceae). Plant Science, 180, 554–559.

Liu L, Yu LL (2011) Estimating species trees from unrooted gene trees. Systematic Biology, 60, 661–667.

Liu L, Yu LL, Edwards SV (2010) A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evolutionary Biology, 10, doi: 10.1186/1471-2148-10-302.

Livak KJ (2003) SNP genotyping by the 5'-nuclease reaction. Methods in Molecular Biology, 212, 129–147.

Mörseburg A, Pagani L, Ricaut FX, Yngvadottir B, Harney E, Castillo C, Hoogervorst T, Antao T, Kusuma P, Brucato N (2016) Multi-layered population structure in Island Southeast Asians. European Journal of Human Genetics, 24, 1605–1611.

Ma XF, Szmidt AE, Wang XR (2006) Genetic structure and evolutionary history of a diploid hybrid pine *Pinus densata* inferred from the nucleotide variation at seven gene loci. Molecular Biology and Evolution, 23, 807–816.

Mallet J (2005) Hybridization as an invasion of the genome. Trends in Ecology and Evolution, 20, 229–237.

Mallet J, Besansky N, Hahn MW (2016) How reticulated are species? BioEssays, 38, 140–149.

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ (2010) Target-enrichment strategies for next-generation sequencing. Nature Methods, 7, 111–118.

Mao JF, Li Y, Wang XR (2009) Empirical assessment of the reproductive fitness components of the hybrid pine *Pinus densata* on the Tibetan Plateau. Evolutionary Ecology, 23, 447–462.

Maples BK, Gravel S, Kenny EE, Bustamante CD (2013) RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. American Journal of Human Genetics, 93, 278–288.

Marczewski T, Ma YP, Zhang XM, Sun WB, Marczewski AJ (2016) Why is population information crucial for taxonomy? A case study involving a hybrid swarm and related varieties. AoB Plants, 8, doi:10.1093/aobpla/plw070.

Maricic T, Whitten M, Pääbo S (2010) Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PLoS ONE, 5, e14004.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Re-

search, 18, 1509–1517.

Martin DP, Murrell B, Golden M, Khoosal A, Muhire B (2015) RDP4: detection and analysis of recombination patterns in virus genomes. Chemical Research in Toxicology, 1, doi: 10.1093/ve/vev003.

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nature Reviews Genetics, 12, 671–682.

Martin NH, Bouck AC, Arnold ML (2006) Detecting adaptive trait introgression between *Iris fulva* and *I. brevicaulis* in highly selective field conditions. Genetics, 172, 2481–2489.

Martinsen GD, Whitham TG, Turek RJ, Keim P (2001) Hybrid populations selectively filter gene introgression between species. Evolution, 55, 1325–1335.

Mayer C, Sann M, Donath A, Meixner M, Podsiadlowski L, Peters RS, Petersen M, Meusemann K, Liere K, Wägele JW, Misof B, Bleidorn C, Ohl M, Niehuis O (2016) BaitFisher: a software package for multispecies target DNA enrichment probe design. Molecular Biology and Evolution, 33, 1875–1886.

Mccormack JE, Al E (2011) Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences. Molecular Phylogenetics & Evolution, 62, 397–406.

Mcguire G, Wright F, Prentice MJ (1997) A graphical method for detecting recombination in phylogenetic data sets. Molecular Biology and Evolution, 14, 1125–1131.

McKinney GJ, Waples RK, Seeb LW, Seeb JE (2016) Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. Molecular Ecology Resources, 17, 656–669.

Mcvean G (2009) A genealogical interpretation of principal components analysis. PLoS Genetics, 5, e1000686.

Meier JI, Marques DA, Mwaiko S, Wagner CE, Excoffier L, Seehausen O (2017) Ancient hybridization fuels rapid cichlid fish adaptive radiations. Nature Communications, 8, doi: 10.1038/ncomms14363.

Menelaou A (2013) Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. Bioinformatics, 29, 84–91.

Meng C, Kubatko LS (2009) Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. Theoretical Population Biology, 75, 35–45.

Menozzi P, Piazza A, Cavallisforza L (1978) Synthetic maps of human gene frequencies in Europeans. Science, 201, 786–792.

Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C (2012) A high-coverage genome sequence from an archaic *Denisovan* individual. Science, 338, 222–226.

Meyer M, Arsuaga JL, Filippo CD, Nagel S, Aximupetri A, Nickel B, Martínez I, Gracia A, Castro JMBD, Carbonell E (2016) Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. Nature, 531, 504–507.

Meyer M, Stenzel U, Hofreiter M (2008) Parallel tagged sequencing on the 454 platform. Nature Protocol, 3, 267–278.

Misof B, Misof K (2009) A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. Systematic Biology, 58, 21–34.

Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, Burns E, Ostrer H, Price AL, Reich D (2011) The history of African gene flow into southern Europeans, Levantines, and Jews. PLoS Genetics, 7, e1001373.

Moreno-Estrada A, Gravel S, Zakharia F, Mccauley JL, Byrnes JK, Gignoux CR, Ortiz-Tello PA, Martínez RJ, Hedges DJ, Morris RW (2013) Reconstructing the population genetic history of the Caribbean. PLoS Genetics, 9, 569–573.

Morgan JAT, Harry AV, Welch DJ, Street R, White J, Geraghty PT, Macbeth WG, Tobin A, Simpfendorfer CA, Ovenden JR (2012) Detection of interspecies hybridisation in Chondrichthyes: hybrids and hybrid offspring between Australian (*Carcharhinus* tilstoni) and common (*C. limbatus*) blacktip shark found in an Australian fishery. Conservation Genetics, 13, 455–463.

Morin PA, Archer FI, Foote AD, Vilstrup J, Allen EE, Wade P, Durban J, Parsons K, Pitman R, Li L (2010) Complete mitochondrial genome phylogeographic analysis of killer whales (*Orcinus orca*) indicates multiple species. Genome Research, 20, 908–916.

Morrison DA (2011) Estimating species trees: practical and theoretical aspects. Systematic Biology, 60, 562–564.

Nabholz B, Künstner A, Wang R, Jarvis ED, Ellegren H (2011) Dynamic evolution of base composition: causes and consequences in avian phylogenomics. Molecular Biology and Evolution, 28, 2197.

Nakhleh L (2013) Computational approaches to species phylogeny inference and gene tree reconciliation. Trends in Ecology and Evolution, 28, 719–728.

Nelson RR (1963) Interspecific hybridization in the fungi. Annual Reviews in Microbiology, 17, 31–48.

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics, 12, 443–451.

Nosil P, Schluter D (2011) The genes underlying the process of speciation. Trends in Ecology and Evolution, 26, 160–167.

Novikova PY, Hohmann N, Nizhynska V, Tsuchimatsu T, Ali J, Muir G, Guggisberg A, Paape T, Schmid K, Fedorenko OM, Holm S, Sall T, Schlotterer C, Marhold K, Widmer A, Sese J, Shimizu KK, Weigel D, Kramer U, Koch MA, Nordborg M (2016) Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. Nature Genetics, 48, 1077–1082.

Omberg L, Salit J, Hackett N, Fuller J, Matthew R, Chouchane L, Rodriguez-Flores JL, Bustamante C, Crystal RG, Mezey

JG (2012) Inferring genome-wide patterns of admixture in Qataris using fifty-five ancestral populations. BMC Genetics, 13, 49–59.

O'Neill EM, Schwartz R, Bullock CT, Williams JS, Shaffer HB, Aguilar-Miguel X, Parra-Olea G, Weisrock DW (2013) Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex. Molecular Ecology, 22, 111–129.

Ozsolak F (2011) RNA sequencing: advances, challenges and opportunities. Nature Reviews Genetics, 12, 87–98.

Padhukasahasram B (2014) Inferring ancestry from population genomic data and its applications. Frontiers in Genetics, 5, doi: 10.3389/fgene.2014.00204.

Pagel M, Meade A (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Systematic Biology, 53, 571–581.

Pagel M, Meade A (2008) Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. Philosophical Transactions of the Royal Society of London, 363, 3955–3964.

Parks M (2009) Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biology, 7, doi: 10.1186/1741-7007-7-84.

Pasaniuc B, Sankararaman S, Kimmel G, Halperin E (2009) Inference of locus-specific ancestry in closely related populations. Bioinformatics, 25, 213–221.

Patterson N, Moorjani P, Luo YT, Mallick S, Rohland N, Zhan YP, Genschoreck T, Webster T, Reich D (2012) Ancient admixture in human history. Genetics, 192, 1065–1093.

Pavy N, Gagnon F, Deschênes A, Boyle B, Beaulieu J, Bousquet J (2016) Development of highly reliable in silico SNP resource and genotyping assay from exome capture and sequencing: an example from black spruce (*Picea mariana*). Molecular Ecology Resources, 16, 588–598.

Payseur BA, Rieseberg LH (2016) A genomic perspective on hybridization and speciation. Molecular Ecology, 25, 2337–2360.

Pease JB, Haak DC, Hahn MW, Moyle LC (2016) Phylogenomics reveals three sources of adaptive variation during a rapid radiation. PLoS Biology, 14, e1002379.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. PLoS ONE, 7, e37135.

Pérez-Losada M, Arenas M, Galán JC, Palero F, González-Candelas F (2015) Recombination in viruses: mechanisms, methods of study, and evolutionary consequences. Infection, Genetics and Evolution, 30, 296–307.

Philippe H, Delsuc F, Brinkmann H, Lartillot N (2005) Phylogenomics. Annual Review of Ecology, Evolution, and Systematics, 36, 541–562.

Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genetics, 8, e1002967.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics, 38, 904–909.

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genetics, 5, e1000519.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics, 155, 945–959.

Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, Li H, Mallick S, Dannemann M, Fu QM, Kircher M, Kuhlwilm M, Lachmann M, Meyer M, Ongyerth M, Siebauer M, Theunert C, Tandon A, Moorjani P, Pickrell J, Mullikin JC, Vohr Samuel H, Green RE, Hellmann I, Johnson PLF, Blanche H, Cann H, Kitzman JO, Shendure J, Eichler EE, Lein ES, Bakken TE, Golovanova LV, Doronichev VB, Shunkov MV, Derevianko AP, Viola B, Slatkin M, Reich D, Kelso J, Paabo S (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. Nature, 505, 43–49.

Pyron RA, Hsieh FW, Lemmon AR, Lemmon EM, Hendry CR (2016) Integrating phylogenomic and morphological data to assess candidate species-delimitation models in brown and red-bellied snakes (Storeria). Zoological Journal of the Linnean Society, 177, 937–949.

Raghavan M, DeGiorgio M, Albrechtsen A, Moltke I, Skoglund P, Korneliussen TS, Gronnow B, Appelt M, Gullov HC, Friesen TM, Fitzhugh W, Malmstrom H, Rasmussen S, Olsen J, Melchior L, Fuller BT, Fahrni SM, Stafford TJ, Grimes V, Renouf MA, Cybulski J, Lynnerup N, Lahr MM, Britton K, Knecht R, Arneborg J, Metspalu M, Cornejo OE, Malaspinas AS, Wang Y, Rasmussen M, Raghavan V, Hansen TV, Khusnutdinova E, Pierre T, Dneprovsky K, Andreasen C, Lange H, Hayes MG, Coltrain J, Spitsyn VA, Gotherstrom A, Orlando L, Kivisild T, Villems R, Crawford MH, Nielsen FC, Dissing J, Heinemeier J, Meldgaard M, Bustamante C, O'Rourke DH, Jakobsson M, Gilbert MT, Nielsen R, Willerslev E (2014a) The genetic prehistory of the new world Arctic. Science, 345, doi: 10.1126/science.1255832.

Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford TW Jr, Orlando L, Metspalu E (2014) Upper Palaeolithic Siberian genome reveals dual ancestry of native Americans. Nature, 505, 87–91.

Rannala B, Yang Z (2008) Phylogenetic inference using whole genomes. Annual Review of Genomics & Human Genetics,

9, 217–231.

Rasmussen MD, Hubisz MJ, Gronau I, Siepel A (2014) Genome-wide inference of ancestral recombination graphs. PLoS Genetics, 10, e1004342.

Rasmussen MD, Kellis M (2012) Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Research, 22, 755–765.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. Nature, 461, 489–494.

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature, 468, 1053–1060.

Reneker J, Lyons E, Conant GC, Pires JC, Freeling M, Shyu CR, Korkin D (2012) Long identical multispecies elements in plant and animal genomes. Proceedings of the National Academy of Sciences, USA, 109, 1183–1191.

Rieseberg LH, Archer MA, Wayne RK (1999) Transgressive segregation, adaptation and speciation. Heredity, 83, 363–372.

Rieseberg LH, Widmer A, Arntz AM, Burke B (2003) The genetic architecture necessary for transgressive segregation is common in both natural and domesticated populations. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences, 358, 1141–1147.

Robinson JD, Bunnefeld L, Hearn J, Stone GN, Hickerson MJ (2014) ABC inference of multi-population divergence with admixture from unphased population genomic data. Molecular Ecology, 23, 4458–4471.

Rodríguezezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H (2007) Detecting and overcoming systematic errors in genome-scale phylogenies. Systematic Biology, 56, 389–399.

Rodriguez JM, Bercovici S, Elmore M, Batzoglou S (2013) Ancestry inference in complex admixtures via variable-length Markov chain linkage models. Journal of Computational Biology, 20, 199–211.

Rusk N (2009) Focus on next-generation sequencing data analysis. Nature Methods, 6, doi: 10.1038/nmeth.f.271.

Sankararaman S, Patterson N, Li H, Pääbo S, Reich D (2012) The date of interbreeding between Neandertals and modern humans. PLoS Genetics, 8, e1002947.

Sankararaman S, Mallick S, Dannemann M, Prufer K, Kelso J, Paabo S, Patterson N, Reich D (2014) The genomic landscape of Neanderthal ancestry in present-day humans. Nature, 507, 354–357.

Sankararaman S, Sridhar S, Kimmel G, Halperin E (2008) Estimating local ancestry in admixed populations. American Journal of Human Genetics, 82, 290–303.

Schmickl R, Liston A, Zeisek V, Oberlander K, Weitemier K, Straub SCK, Cronn RC, Dreyer LL, Suda J (2016) Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African *Oxalis* (Oxalidaceae). Molecular Ecology Resources, 16, 1124–1135.

Schwenk K, Brede N, Streit B (2008) Introduction: extent, processes and evolutionary impact of interspecific hybridization in animals. Philosophical Transactions of the Royal Society B: Biological Sciences, 363, 2805–2811.

Shafer ABA, Peart CR, Tusso S, Maayan I, Brelsford A, Wheat CW, Wolf JBW (2016) Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. Methods in Ecology and Evolution, doi: 10.1111/2041-210X.12700.

Shen XX, Liang D, Feng YJ, Chen MY, Zhang P (2013) A versatile and highly efficient toolkit including 102 nuclear markers for vertebrate phylogenomics, tested by resolving the higher level relationships of the caudata. Molecular Biology and Evolution, 30, 2235–2248.

Siepel A (2009) Phylogenomics of primates and their ancestral populations. Genome Research, 19, 1929–1941.

Sims GE, Jun SR, Wu GA, Kim SH (2009) Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proceedings of the National Academy of Sciences, USA, 106, 2677–2682.

Skotte L, Korneliussen TS, Albrechtsen A (2013) Estimating individual admixture proportions from next generation sequencing data. Genetics, 195, 693–702.

Smith J, Kronforst MR (2013) Do *Heliconius* butterfly species exchange mimicry alleles? Biology Letters, 9, 20130503.

Sneath PHA (1975) Cladistic representation of reticulate evolution. Systemtic Zoology, 24, 360–368.

Solis-Lemus C, Ane C (2016) Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS Genetics, 12, e1005896.

Soltis PS, Soltis DE (2009) The role of hybridization in plant speciation. Annual Review of Plant Biology, 60, 561–588.

Song BH, Wang XQ, Wang XR, Ding KY, Hong DY (2003) Cytoplasmic composition in *Pinus densata* and population establishment of the diploid hybrid pine. Molecular Ecology, 12, 2995–3001.

Song BH, Wang XQ, Wang XR, Sun LJ, Hong DY, Peng PH (2002) Maternal lineages of *Pinus densata*, a diploid, hybrid. Molecular Ecology, 11, 1057–1063.

Song YS, Hein J (2005) Constructing minimal ancestral recombination graphs. Journal of Computational Biology, 12, 147–169.

Stenz NWM, Larget B, Baum DA, Ané C (2015) Exploring tree-like and non-tree-like patterns using genome sequences: an example using the inbreeding plant species *Arabidopsis thaliana* (L.) Heynh. Systematic Biology, 64, 809–823.

Strimmer K, Moulton V (2000) Likelihood analysis of phylogenetic networks using directed graphical models. Molecular Biology and Evolution, 17, 875–881.

Su S, Wong G, Shi WF, Liu J, Lai ACK, Zhou JY, Liu WJ, Bi YH, Gao GF (2016) Epidemiology, genetic recombination,

and pathogenesis of coronaviruses. Trends in Microbiology, 24, 490–502.

Sullivan J, Joyce P (2005) Model selection in phylogenetics. Annual Review of Ecology, Evolution, and Systematics, 36, 445–466.

Sundquist A, Fratkin E, Do CB, Batzoglou S (2008) Effect of genetic divergence in identifying ancestral origin using HAPAA. Genome Research, 18, 676–682.

Sunnåker M, Busetto AG, Numminen E, Corander J, Foll M, Dessimoz C (2013) Approximate Bayesian computation. PLoS Computational Biology, 9, e1002803.

Susko E, Spencer M, Roger AJ (2005) Biases in phylogenetic estimation can be caused by random sequence segments. Journal of Molecular Evolution, 61, 351–359.

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic inference. In: Molecular Systematics (eds Hiillis DM, Moritz D, Mable BK), pp. 407–514. Sinauer Associates, Sunderland, Massachusetts.

Szymura JM, Barton NH (1986) Genetic analysis of a hybrid zone between the fire-bellied toads, *Bombina bombina* and *B. variegata*, near Cracow in southern Poland. Evolution, 40, 1141–1159.

Tang H, Choudhry S, Mei R, Morgan M, Rodriguez-Cintron W, Burchard EG, Risch NJ (2007) Recent genetic selection in the ancestral admixture of Puerto Ricans. American Journal of Human Genetics, 81, 626–633.

Tang H, Coram M, Wang P, Zhu X, Risch N (2006) Reconstructing genetic ancestry blocks in admixed individuals. American Journal of Human Genetics, 79, 1–12.

Than C, Ruths D, Nakhleh L (2008) PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics, 9, doi: 10.1186/1471-2105-9-322.

The Genomes Project Consortium, Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Collins FS, de La Vega FM (2010) A map of human genome variation from population-scale sequencing. Nature, 467, 1061–1073.

Torkamaneh D, Laroche J, Bastien M, Abed A, Belzile F (2017) Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. BMC Bioinformatics, 18, doi: 10.1186/s12859-016-1431-9.

Torkamaneh D, Laroche J, Belzile F (2016) Genome-wide SNP calling from genotyping by sequencing (GBS) data: a comparison of seven pipelines and two sequencing technologies. PLoS ONE, 11, e0161333.

Townsend JP (2007) Profiling phylogenetic informativeness. Systematic Biology, 56, 222–231.

Vallejo-Marin M, Hiscock SJ (2016) Hybridization and hybrid speciation under global change. New Phytologist, 211, 1170–1187.

van Tassell CP, Smith TP, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, Haudenschild CD, Moore SS, Warren WC, Sonstegard TS (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods, 5, 247–252.

Wang BS, Mao JF, Gao J, Zhao W, Wang XR (2011) Colonization of the Tibetan Plateau by the homoploid hybrid pine *Pinus densata*. Molecular Ecology, 20, 3796–3811.

Wang S, Meyer E, Mckay JK, Matz MV (2012) 2b-RAD: a simple and flexible method for genome-wide genotyping. Nature Methods, 9, 808–810.

Wang XR, Szmidt AE (1994) Hybridization and chloroplast DNA variation in a *Pinus* species complex from Asia. Evolution, 48, 1020–1031.

Wang XR, Szmidt AE, Savolainen O (2001) Genetic composition and diploid hybrid speciation of a high mountain pine, *Pinus densata*, native to the Tibetan Plateau. Genetics, 159, 337–346.

Wang XR, Szmidt AE, Lewandowski A, Wang ZR (1990) Evolutionary analysis of *Pinus densata* Masters, a putative tertiary hybrid 1, allozyme variation. Theoretical and Applied Genetics, 80, 635–640.

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10, 57–63.

Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCtoolbox: a versatile toolkit for approximate Bayesian computations. BMC Bioinformatics, 11, doi:10.1186/1471-2105-11-116.

Weigel D, Mott R (2009) The 1001 Genomes Project for *Arabidopsis thaliana*. Genome Biology, 10, 107.

Worobey M, Holmes EC (1999) Evolutionary aspects of recombination in RNA viruses. Journal of General Virology, 80, 2535–2543.

Wu CL (1956) The taxonomic revision and phytogeographical study of Chinese pines. Acta Phytotaxonomica Sinica, 5, 131–163. (in Chinese with English abstract) [吴中伦 (1956) 中国松属的分类与分布. 植物分类学报, 5, 131–163.]

Wu CI, Ting CT (2004) Genes and speciation. Nature Reviews Genetics, 5, 114–122.

Wu Y (2012) Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. Evolution, 66, 763–775.

Wu ZY, Raven PH, Hong DY (2014) Flora of China, Vols. 1–25. Science Press, Beijing & Missouri Botanical Garden Press, St. Louis.

Xiang YZ, Huang CH, Hu Y, Wen J, Li SS, Yi TS, Chen HY, Xiang J, Ma H (2017) Well-resolved rosaceae nuclear phylogeny facilitates geological time and genome duplication analyses and ancestral fruit character reconstruction. Molecular Biology and Evolution, 34, 262–281.

Xing FQ, Mao JF, Meng JX, Dai JF, Zhao W, Liu H, Xing Z, Zhang H, Wang XR, Li Y (2014) Needle morphological evidence of the homoploid hybrid origin of *Pinus densata*

based on analysis of artificial hybrids and the putative parents, *Pinus tabuliformis* and *Pinus yunnanensis*. Ecology & Evolution, 4, 1890–1902.

Yang JJ, Li J, Buu A, Williams LK (2013) Efficient inference of local ancestry. Bioinformatics, 29, 2750–2756.

Yang WY, Novembre J, Eskin E, Halperin E (2012) A model-based approach for analysis of spatial structure in genetic data. Nature Genetics, 44, 725–731.

Yang ZH (1996) Maximum-likelihood models for combined analyses of multiple sequence data. Journal of Molecular Evolution, 42, 587–596.

Yu Y, Degnan JH, Nakhleh L (2012) The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genetics, 8, e1002660.

Yu Y, Dong J, Liu KJ, Nakhleh L (2014) Maximum likelihood inference of reticulate evolutionary histories. Proceedings of the National Academy of Sciences, USA, 111, 16448–16453.

Yu Y, Ristic N, Nakhleh L (2013) Fast algorithms and heuristics for phylogenomics under ILS and hybridization. BMC Bioinformatics, 14, 1–10.

Yu Y, Than C, Degnan JH, Nakhleh L (2011) Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. Systematic Biology, 60, 138–149.

Yu Y, Nakhleh L (2015) A maximum pseudo-likelihood approach for phylogenetic networks. BMC Genomics, 16, 1–10.

Zhang LS, Dai JF, Gao Q, Liu H, Zhang H, Zhao W, Mao JF, Li Y (2012) Seedling adaptation of hybrid pine *Pinus densata* and its parental species in the high elevation habitat. Journal of Beijing Forestry University, 34(5), 15–24. (in Chinese with English abstract) [张立沙, 代剑峰, 高琼, 刘灏, 张华, 赵伟, 毛建丰, 李悦 (2012) 高山松与亲本种多种群在高海拔生境下的苗期适应性研究. 北京林业大学学报, 34(5), 15–24.]

Zeng LP, Zhang Q, Sun RR, Kong HZ, Zhang N, Ma H (2014) Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. Nature Communications, 5, doi: 10.1038/ncomms5956.

Zhao W, Meng JX, Wang BS, Zhang LS, Xu YL, Zeng QY, Li Y, Mao JF, Wang XR (2014) Weak crossability barrier but strong juvenile selection supports ecological speciation of the hybrid pine *Pinus densata* on the Tibetan Plateau. Evolution, 68, 3120–3133.

Zheng XH, Lu F, Wang ZY, Zhong F, Hoover J, Mural R (2005) Using shared genomic synteny and shared protein functions to enhance the identification of orthologous gene pairs. Bioinformatics, 21, 703–710.

(责任编委: 卢宝荣　责任编辑: 黄祥忠)