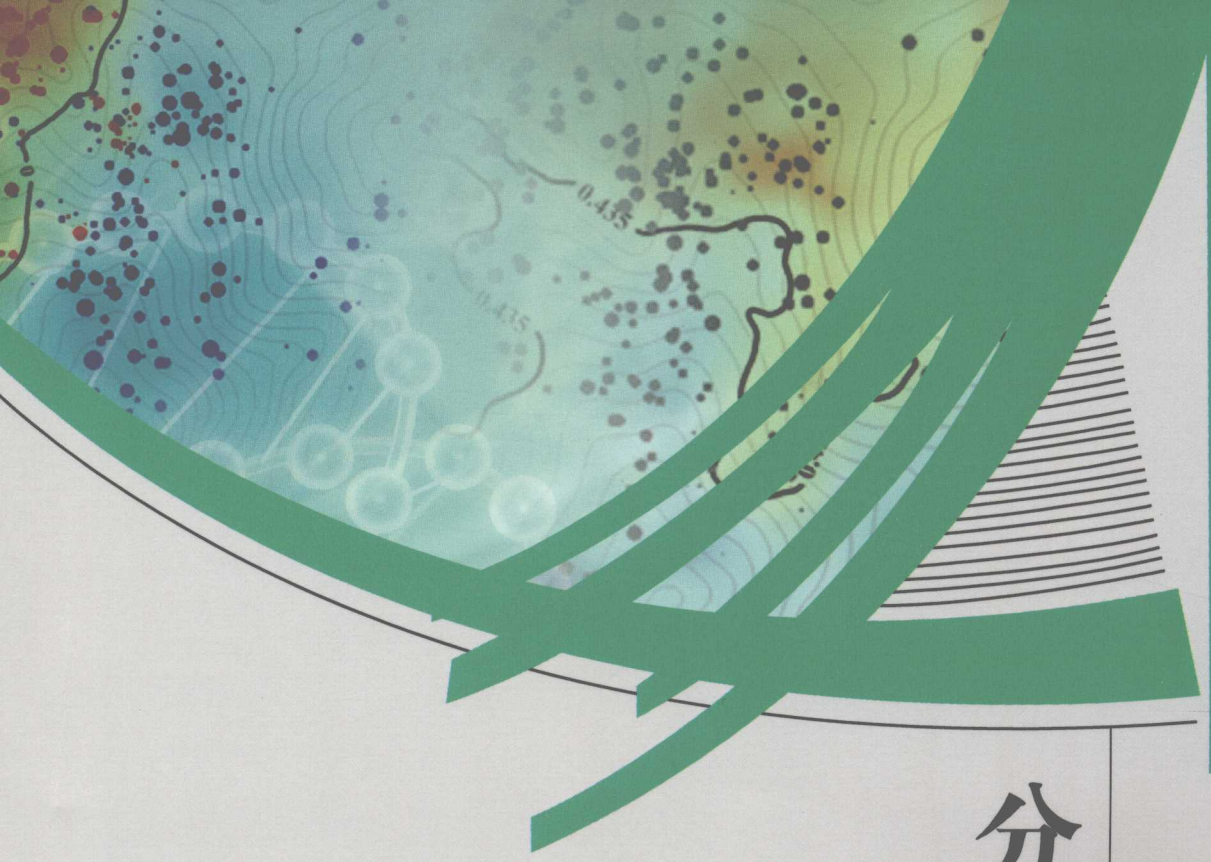

研究生创新教育系列丛书

分子生态学与数据分析基础

王峥峰 著

科学出版社

研究生创新教育系列丛书

分子生态学与数据分析基础

王峥峰 著

科学出版社
北京

内 容 简 介

本书首先从理论上介绍了分子生态学基本研究内容和手段，并总结作者以往的研究工作，较全面地概括了分子生态学理论内涵；然后从实践角度介绍了分子生态学数据获得的方法与具体分析内容和步骤，特别是采用图解法一步一步对分子生态学用到的各种主流分析软件进行过程讲解（包括作者编写的程序），不但有各类分析提示，还提供演示数据。本书具有极强的理论性和实践操作性，对于促进我国分子生态学发展，利用分子遗传标记手段进行物种经营、保护及资源利用和环境规划具有重要的推动作用。

本书可供高等学校、研究机构从事分子生态学和相关领域研究的师生，以及农、林、牧、副、渔、医各行业利用分子遗传标记开展研究的工作人员阅读参考。

图书在版编目（CIP）数据

分子生态学与数据分析基础 / 王峥峰著. —北京：科学出版社, 2016.1

（研究生创新教育系列丛书）

ISBN 978-7-03-046478-1

Ⅰ.①分… Ⅱ.①王… Ⅲ. ①分子生物学–生态学–数据处理 Ⅳ.①Q145

中国版本图书馆 CIP 数据核字(2015)第 282673 号

责任编辑：王海光 高璐佳 / 责任校对：彭珍珍

责任印制：徐晓晨 / 封面设计：陈 敬

科 学 出 版 社 出版

北京东黄城根北街 16 号

邮政编码: 100717

http://www.sciencep.com

北京京华虎彩印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2016 年 1 月第 一 版 开本：787×1092 1/6

2016 年 1 月第一次印刷 印张：14 1/4

字数：326 000

定价：88.00 元

(如有印装质量问题，我社负责调换)

前　言

分子生态学是通过分子生物学（分子标记、分子遗传学、生物信息学等）手段来研究生物和环境关系的科学。目前，分子生态学的研究主要还是集中在群体遗传学方面，即应用分子标记手段开展物种种群遗传多样性、基因流等方面的研究，是群体遗传学研究的延展。但随着高通量测序技术、分子遗传学和生物信息学研究的发展，不仅大量功能基因被挖掘，而且物种整个基因组功能被完整解析，将使得通过基因、基因组（整体）表达调控以了解环境对生物的作用及个体、种群环境响应适应成为开展生态学机制研究的最主要方面，也将使得分子生态学成为了解生态系统、生物进化和进行生物资源合理利用及保护的重要利器。

近来，利用分子手段开展生态学研究的一个重要分支学科——保护遗传学（conservation genetics）也方兴未艾。但从研究内容看，分子生态学涵盖了保护遗传学研究内容，而且分子生态学研究落脚点其实也是生物资源的保护和合理利用。例如，对于“是否把物种保护在自然保护区和植物园就万事大吉了呢？”这样的问题，是保护遗传学的研究内容，也是分子生态学关心的焦点之一，包括小种群问题、基因流的问题、群落谱系生态学（phylogenetic ecology）问题等。因此分子生态学研究既具有很强的理论性，又具有丰富的实践内涵，可以说没有分子生态学参与的生态学是不完全和粗糙的。对于分子生态学和保护生态学这两方面的研究，我都曾进行过相关总结，读者可参考王峥峰等（2001，2002），王峥峰和彭少麟（2003a，2003b），王峥峰和葛学军（2009）的研究。但这些总结主要发表在学术期刊上，篇幅较短小，很多收集的资料和想法不能很好展现。而且随着分子生态学发展逐渐成熟，我觉得应该在国内出版一本介绍分子生态学的书籍。同时我在做研究期间，不断学习各种新的方法和技术，总想如能把我自己摸索学到的关于数据分析方面的心得介绍给对分子生态学感兴趣的新人，对于他们的成长和将来学科的发展将有非常大的作用。因此，我就把多年来读的文献进行总结，并面向初学者就分子生态学数据的获得和分析进行详细的介绍，形成此书。因此，本书分成两部分，第一部分介绍了分子生态学研究内容及其最基本的原理，第二部分是如何进行数据分析。

由于我主要以植物为研究对象开展分子生态的研究，因此书中介绍的主要研究案例是植物；但分子生态学研究的材料是遗传物质 DNA 或者 RNA，针对的是遗传变异、表达调控等，因此对于动物也好，植物也罢，两者并没有很大的差异，理论和方法是通用的。同时本书介绍的数据分析以微卫星体展开，对于那些用分子序列或者单核苷酸多态性（single nucleotide polymorphism，SNP）做研究的工作者来说可能觉得不适用。但不论用什么分子标记方法，最终的目的都是通过标记发现个体或种群间的遗传差异，然后在遗传差异的基础上进行各种关联性分析，因此这些标记在数据的后期处理过程中并无不同。另外，我在书中还简单介绍了 ArcGIS 软件、R 语言等内容，这些内容是对地理数

据或者模型分析方法的介绍，不论采用何种分子标记都可以用。

本书的出版，要感谢科学出版社相关工作人员付出的辛勤劳动。感谢安树青老师、王伯荪老师、张宏达老师、余世效老师、彭少麟老师和叶万辉老师。感谢同事练琚愉和陈红锋。感谢其他亲人、同事对我的帮助和支持。本书得到国家自然科学基金（31170352，41371078，31100312）资助。本书是作者以往研究成果的总结，这些课题包括：International Foundation For Science（瑞典，AD/13076）、国家自然科学基金（30300055）、广东省自然科学基金（031264）、中国科学院知识创新工程重要方向项目（KSCX2-YW-Z-023）、973 项目（2007CB411606）、中国科学院生命科学领域基础前沿研究专项（KSCX2-EW-J-28），在此一并感谢！

最后，由于我开展的研究主要围绕种群遗传变异，因此理论介绍部分不涉及在分子生态学中对基因的研究。所有的理论、数据分析方法是针对我的研究内容的总结，涉及的面还是有限。抛砖引玉，期望读者更正、补充、完善。针对此书如有任何问题、批评和建议，请发电子邮件 wzf1973@21cn.com 给我，并访问我的个人网站 www.molecular-ecologist.com 查看问题解答或错误更正。非常感谢！

王峥峰

2015 年 7 月

目　录

第一部分

分子生态学理论

第一章　分子生态与遗传变异

第一节　遗传变异的产生

点突变是产生遗传变异的主要原因。另外染色体的倒置（inversion）、易位（translocation），有性生殖的重组，序列的插入缺失等都会导致遗传变异。引起突变的内因有基因组的不稳定、基因的相互作用等；外因有环境因素如光、温、化学物质、辐射等。一般说来，自然界中每个细胞循环中每一个碱基突变率的数量级为 10^{-10}～10^{-9}，考虑每一位点（loci，研究的某段特定序列或基因都可以称为位点）有 100～1000 个碱基，那么每一位点突变率的数量级为 10^{-7}～10^{-6}（Baur & Schmid，1996）。

由于植物体大部分细胞是体细胞，因此大部分突变为体细胞突变，而只有生殖细胞中的突变才会遗传到后代。生殖细胞和体细胞的分化在个体发育早期就开始，与体细胞不同的是，为了避免更大的遗传负荷，生殖细胞的分裂次数远小于体细胞，仅几十到几百次。然而即使如此，对于庞大的基因组来说，如拥有 10^9bp 碱基的基因组，在减数分裂阶段，生殖细胞也会有 100bp 的碱基可能发生了突变（$10^{-9}\times10^9\times10^2$，假设生殖细胞分裂了 10^2 次）（Vida，1994）。

突变本身并不驱动种群进化。基因可能是多拷贝的，因此如果多拷贝基因中的一个或少数几个发生了突变，并不影响这一基因整体的功能。

当新的遗传变异产生，它可能对个体性状没有影响，即中性变异；也可能是有有害影响，即有害变异，导致个体死亡或适应性降低；也可能是有有益影响，促进适应。判断遗传变异是中性还是有害或有益并非易事，它受环境、种群大小状况、种群遗传历史等影响（Hedrick，2004）。例如，在某个种群中，有些遗传变异可能是有利于种群抵抗病原菌的，但对生长在缺乏这种病原菌环境的另一个种群，这一遗传变异可能是有害的，因为基因是具有多效性的。

如果遗传变异是发生在同一位点上，就会有多个等位基因（allele）形成。例如，某段序列 AATACCTCCCTACAACTCATG 中第三个位置发生了点突变，由“T”变为“A”，那么这段序列就有了两个等位基因，一个是前面的，一个是 AAAACCTCCCTACAACTCATG。这两个等位基因，可以分别用“*A*”和“*B*”表示，也可以用“*1*”和“*2*”表示，也可以用“*wa*”或者“*wb*”表示，只要能区分两者就可以。如果原始序列在第 10 个位置再次发生突变，由“C”变为“G”，那么这段序列就有了三个等位基因，一个是原始的，一个是第三个位置由“T”变为“A”的，一个是 AATACCTCCGTACAACTCATG。值得注意的是，上面的例子是以这个完整序列作为等位基因判断的标准。但假如只考虑突变点，如这个序列的第三个位置，那么这个位置只有两个等位基因，一个等位基因的碱基形式是“A”，另一个等位基因的碱基形式是“T”。而对于这个序列第 10 个位置，它也是只有两个等位基因，一个等位基因的碱基形式是“C”，另一个等位基因的碱基形

式是“G”。在实际的研究中我们需要弄清。

对于微卫星体序列，其序列的特点是包含了一段重复序列。如 AAATGGGAGTGCGGGAGATTGCCAGTGAGGGTATAGAGG**GAGAGAGAGAGAGAGAGAGA**GAAACAGCGAGCAAAGGCAGCAAAGAGGGACGGAGAG 这个序列就包含了重复序列单元“GA”，重复了 10 次。不同于点突变，微卫星体的遗传变异主要是由它所包含的重复序列的增减而产生，例如，如果上述序列中“GA”这个重复序列单元由重复 10 次变为重复了 12 次，就得到两个等位基因，一个是$(GA)_{10}$，一个是$(GA)_{12}$。

那么微卫星体这种重复序列的增减遵循怎样的变异模式呢？目前大致有 4 种解释（Oliveira et al.，2006；Putman & Carbone，2014）。

第一种是无限等位基因模型（infinite alleles model，IAM），即重复序列单元的增减不受限制，如对于上述“GA”单元，一次可以增加 10 个，也可能一次增加 5 个，也可能一次减少 8 个或者 6 个，没有规律。在计算中，重复单元数的多少和亲缘关系远近没有关系。例如，对于$(GA)_{10}$、$(GA)_{14}$和$(GA)_{16}$这三个重复序列，$(GA)_{14}$和$(GA)_{16}$之间的亲缘关系与$(GA)_{10}$和$(GA)_{16}$之间的亲缘关系是一样的，并不因为前两个序列“GA”重复单元只相差两个（16–14=2）碱基就比后面相差 6 个碱基的两个序列亲缘关系更大。

第二种是逐步突变模型（stepwise mutation model，SMM）。这个模型推测微卫星体重复单元是逐渐增加和减少的。这里增加和减少的重复单元可以是多个，也可以是一个。例如，对于两个等位基因，一个重复单元是$(GA)_{10}$，一个是$(GA)_{16}$，那么从 10 个“GA”重复单元变化到 16 个重复单元，可能经历了变为 11 个“GA”重复单元，即$(GA)_{11}$，增加了一个重复单元；再经历 13 个“GA”重复单元，即$(GA)_{13}$，这次增加了 2 个重复单元，最后再增加三个重复单元变为$(GA)_{16}$。当然，这只简单描述了这个模型变异过程，实际计算过程中，这一模型会考虑重复单元不断增减的过程，即在从重复单元 10 到 16 的过程中，并非一直是增加重复单元，变化过程中又会减少重复单元，之后再增加重复单元，增增减减，逐步达到 16。依据这个模型，对于两个微卫星体来说，重复单元的数目越接近，其亲缘关系也越大。如上面的例子中的三个重复序列$(GA)_{10}$、$(GA)_{14}$和$(GA)_{16}$，依据 SMM 模型，它们两两之间的亲缘关系是$(GA)_{14}$和$(GA)_{16}$ >$(GA)_{10}$和$(GA)_{14}$ >$(GA)_{10}$和$(GA)_{16}$。由于这种增增减减的计算过程非常复杂，在数据量稍大时运算会非常缓慢。为此在计算过程中，研究人员一般使用布朗运动模型（Brownian-motion model）取代计算（Blum et al.，2004）。

第三种是两相模型（two phase model）。这个模型和 SMM 相似，但假设重复单元的增减每次只能是一个，一次不能进行多个重复单元的增减。例如，两个重复序列，一个是$(GA)_{10}$，另一个是$(GA)_{16}$，那么从 10 个“GA”重复单元变化到 16 个重复单元，要先经历 11 个“GA”重复单元，即$(GA)_{11}$，再经历$(GA)_{12}$，再经历$(GA)_{13}$，之后依次变为 14、15 个重复单元后才能变为$(GA)_{16}$。

第四种是 *K*-等位基因模型（*K*-alleles model，KAM）。这一模型类似于 IAM，但假设所研究的微卫星体序列只能有 *K* 个等位基因，而 IAM 模型是可以有无限可能的等位基因。

在这 4 个变异模型中，IAM 和 SMM 模型是最常用的。对于“完美的”（perfect）微卫星体，即重复序列很单一的微卫星体，如上面的例子 **GAGAGAGAGAGAGAGAGAGA**，

选择 SMM 模型可能好些。对于不完整（imperfect）和复杂的（compound）微卫星体，其重复序列不单一，如 **GAGAGAGATTGAGAGAGAGAGA**（在 GA 重复序列中间增加了 TT 序列）、**GAGAGAGAGAGAGAGAGAGACCACCACCACCACCACCACCA**（GA 和 CCA 重复序列混合），这时选择 IAM 模型较合适。

第二节 分子标记的种类

分子标记是找到个体间遗传差异的钥匙，是开展分子生态学研究的关键。在每个物种复杂的基因组面前，我们可能并不需要所有的遗传变异来进行分子生态学的研究，选取其中一部分应该可以解决绝大多数问题。在最早期的研究中，同工酶是最广泛使用的，国外在这个方面开展的工作很多，为分子生态学的发展奠定了基础。但随着新技术的发展，这一方法逐渐退出历史舞台，目前主要是基于 DNA 多态性检测的方法进行分子生态学的研究。

针对分子标记检测方法，可以大致分为以下 4 类。

1. 随机引物为基础的分子标记

随机引物为基础的分子标记包括以下几类。

DNA 扩增指纹印迹（DNA amplification fingerprinting，DAF），使用 5～8 个碱基的单个随机引物进行 DNA 多态性扩增。见 Caetano-Anolles 等（1991）的文章。

简单重复序列间区（inter-simple sequence repeat，ISSR），使用的随机引物为微卫星体内部重复序列，如$(GA)_n$，扩增的是重复序列之间的 DNA 片段。见 Zietkiewicz 等（1994）的文章。

随机扩增多态性 DNA（random amplified polymorphic DNA，RAPD），使用 8～10 个碱基的单个随机引物进行 DNA 多态性扩增。见 Williams 等（1990）的文章。

2. 非随机引物为基础的分子标记

非随机引物为基础的分子标记包括以下几类。

扩增片段长度多态性（amplified fragment length polymorphism，AFLP），使用限制性内切核酸酶酶切基因组 DNA，如位点的碱基序列发生变化（如突变），就导致不同个体基因组 DNA 酶切片段长度上的差异，即多态性。这一方法还需通过人工合成特定 DNA 片段连接到 DNA 酶切片段上，再用特异引物 PCR 扩增检测这种多态性。见 Vos 等（1995）的文章。

酶切扩增多态性序列（cleaved amplified polymorphic sequences，CAPS），对 PCR 扩增产物进行酶切，观测酶切片段长度多态性。见 Akopyanz 等（1992）的文章。

序列特征化扩增区域（sequence characterized amplified region，SCAR），对 RAPD 扩增产物进行克隆和测序，设计特定引物，再进行特异性扩增，比较多态性。见 Paran 和 Michelmore（1993）的文章。

简单序列重复（simple sequence repeat，SSR），即微卫星体，其序列中含如$(AC)_n$、$(AG)_n$、$(AT)_n$样的重复序列，重复序列长度为 1～5 个碱基，其多态性来源于这些重复序列的增加和减少。见 Beckmann 和 Soller（1990）的文章，以及 Akkaya 等（1992）的文章。

可变数串联重复序列（variable number tandem repeat，VNTR），多指小卫星体（minisatellite），其所含的重复序列长度长于微卫星体，为 11～60 个碱基。见 Jeffreys 等（1985）的文章。

单核苷酸多态性（single nucleotide polymorphism，SNP），是指在染色体基因组水平上单个核苷酸的变异引起的 DNA 序列多态性。

3. 杂交为基础的分子标记

限制性片段长度多态性（restriction fragment length polymorphism，RFLP），使用限制性内切核酸酶酶切基因组 DNA，如位点的碱基序列发生变化（如突变），就导致不同个体基因组 DNA 酶切片段长度上的差异，即多态性。通过用凝胶电泳分离这些片段，再用特异标记的探针和这些片段杂交检测这种多态性。见 Botstein 等（1980）的文章。随着技术的发展，这一方法也已很少被使用了。

4. 测序为基础的序列分子标记

测序为基础的序列分子标记即一段序列。如叶绿体的 *trnL-trnF* 片段，线粒体的 *COI* 片段等。

这些标记中，近年来，应用较多的是微卫星体、SNP 和序列分子标记。这三种标记间并无绝对优劣之分，可按照所需解决的问题进行有针对性的选择。如果是进行物种、种群的进化分析，可用序列分子标记；如果是进行个体鉴定（亲本分析）、亲缘关系、基因流等分析，可以采用微卫星体和 SNP。微卫星体由于其多态性更高，进行个体鉴定更好些，使用较少的标记就可以进行大量个体的区分；而采用 SNP 可以辅助寻找相关目的基因，如找到受选择作用的基因。

当然，由于物种的基因组较大，基因组不同区域进化状况不同，因此不同区域得到的分子标记在遗传变异计算结果上也会不同。Defaveri 等（2013）以三棘刺鱼（*Gasterosteus aculeatus*）为研究对象，对比了基因内部的微卫星体和 SNP、非基因内部的微卫星体和 SNP 在衡量遗传多样性（H_E）、种群间遗传分化（F_{ST}）上的差异（图 1-1）。结果表明基因内部和非基因内部的微卫星体所计算的 H_E（或 F_{ST}）结果差异较大，而基因内部和非基因内部的 SNP 所计算的 H_E（或 F_{ST}）结果差异不大。微卫星体在小尺度（fine-scale）上对种群遗传结构的分析效果优于 SNP 标记（如果 SNP 标记相对较少的时候）。同时，在开展种群适应性遗传进化研究时，SNP 标记虽优于微卫星体标记，但当大规模 SNP 标记成本太高无法进行时，利用基因内部的微卫星体标记也可得到令人满意的结果。最后，他们的研究还发现，除了衡量种群间遗传分化指标，微卫星体与 SNP 所得的分析结果不太相关。因此我们在开展研究前，首先确定研究目的，有针对性地选择相关标记非常重要。

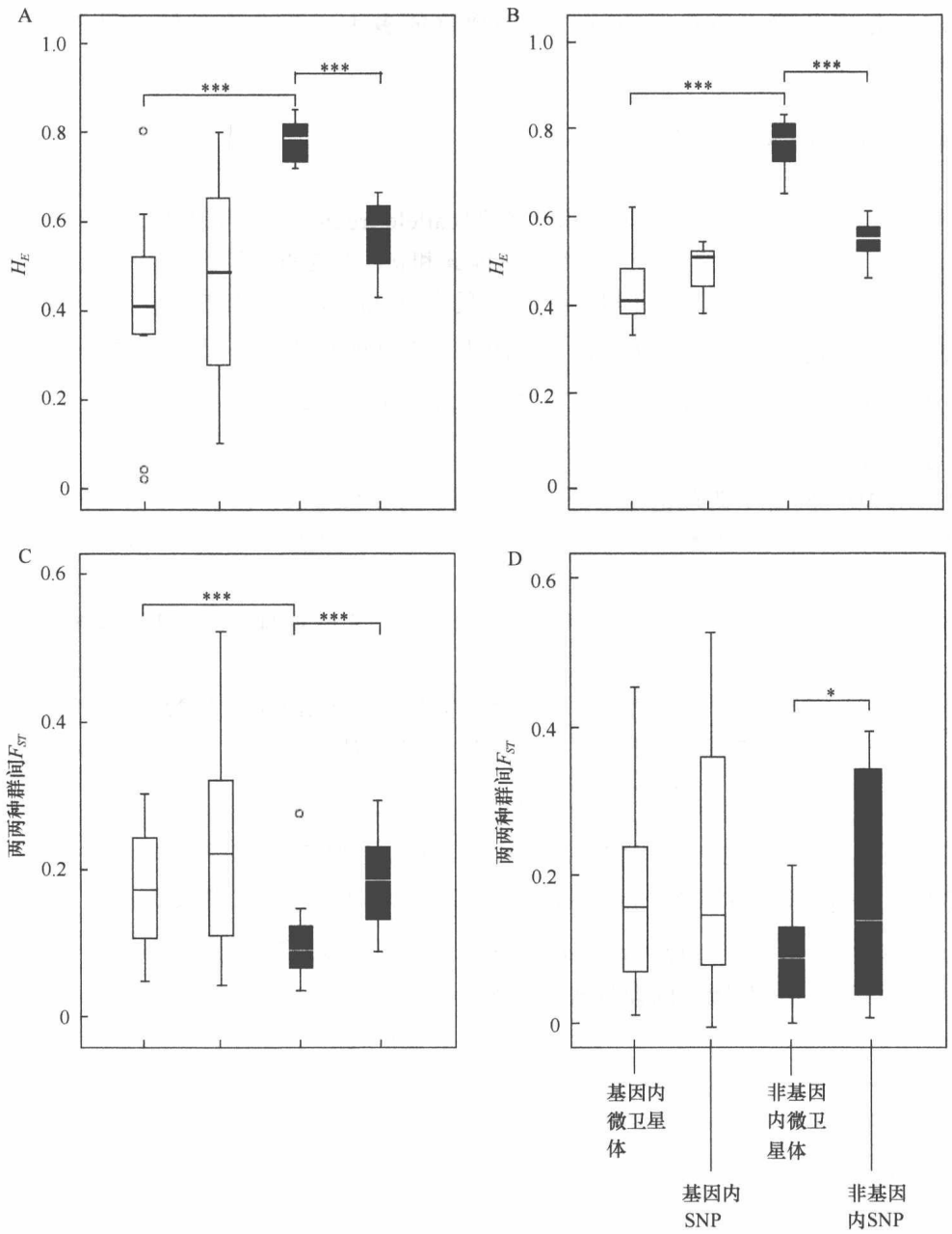

图 1-1　对比微卫星体和 SNP 遗传标记计算的遗传多样性（H_E）和种群间遗传分化（F_{ST}）结果（引自 Defaveri et al.，2013）

A、C. 对位点计算的结果；B、D. 对种群计算的结果；A、B、C 中的 4 个盒体标记见 D；*P<0.05，*** P <0.001

第三节　遗传变异的衡量

检测到个体遗传变异后，就要对这些变异在群体水平上的状况进行分析。这包括两个最基本的衡量指标：等位基因多样性和杂合度（heterozygosity）。等位基因多样性较容

易理解，是指各位点所包括的等位基因数目的多少，而杂合度的度量需要先知道Hardy-Weinberg平衡。

1. Hardy-Weinberg平衡

Hardy-Weinberg平衡涉及等位基因频率（allele frequency）和基因型频率（genotype frequency），即假设某一位点有两个等位基因A和a，由这两个等位基因可组成三个基因型AA、$Aa(aA)$和aa。其中AA和aa是纯合的，而$Aa(aA)$是杂合的。假设等位基因A和a的频率分别是p和q（这里$p+q=1$，因为只有这两个等位基因），那么基因型AA的期望频率将为p^2，$Aa(aA)$的期望频率为$2pq$，aa的期望频率为q^2，并且$p^2+2pq+q^2=1$。

Hardy-Weinberg平衡理论是一种理想状态下的平衡理论，它假设：①物种为二倍体；②有性繁殖；③世代不重叠；④种群大小无限（无遗传漂变）；⑤没有种群个体的迁入和迁出；⑥种群个体间随机交配；⑦无突变；⑧无自然选择。

在此状况下，如果种群各基因型的期望频率等于实际观测频率，即Hardy-Weinberg平衡。

举例如下：在对某种群调查后得知，种群中基因型为AA的个体数频率为0.25，基因型为$Aa(aA)$的个体数频率为0.5，而基因型为aa的个体数频率为0.25。假设等位基因A的频率为p，等位基因a的频率为q，则

$$p=p(p+q)=p^2+2pq/2=0.25+0.5/2=0.5$$

[注：$p+q=1$，所以$p=p(p+q)$就是$p\times1$；p^2是AA基因型频率。]

$$q=q(p+q)=q^2+2pq/2=0.25+0.5/2=0.5$$

由此反过来计算基因型的期望频率，即

$$AA=p^2=0.5^2=0.25$$

$$Aa=2pq=2\times0.5\times0.5=0.5$$

$$aa=q^2=0.5^2=0.25$$

可以看到期望基因型频率和观测基因型频率是相同的，因此种群处于Hardy-Weinberg平衡。

在实际中，上述理想的种群是不存在的，因此任何不符合上述假设的因素都有可能导致期望基因型频率和观测基因型频率结果的不同，使种群偏离Hardy-Weinberg平衡，其偏离程度可通过χ^2检测或其他方法检测，在后面的数据分析部分会介绍。

2. 基于Hardy-Weinberg平衡理论的遗传多样性度量

由上可知，对于拥有两个等位基因（A和a）的某一位点来说，其共有三种基因型AA、$Aa(aA)$和aa，其基因型频率分别为p^2、$2pq$和q^2。

显然，三种基因型当中，只有基因型 $Aa(aA)$ 是杂合型的，由 $p^2+2pq+q^2=1$，可得 $2pq=1-p^2-q^2$，即 $2pq=1-\sum_{i=1}^{2}x_i^2$（即用 x_i 代表 p^2 和 q^2）。进一步，如果这一位点有 k 个等位基因，上述杂合型基因型的频率将为

$$1-\sum_{i=1}^{k}x_i^2$$

式中，x_i 是第 i 个等位基因的频率。这就是通常用来衡量种群遗传多样性的主要指标，称为期望杂合度（expected heterozygosity，H_E），即基因多样性（gene diversity）。

对于 m 个位点来说，种群的期望杂合度为

$$1-\frac{1}{m}\sum_{l=1}^{m}\sum_{i=1}^{k}x_i^2$$

在本书中，为避免混淆，专用遗传多样性指代杂合度，用遗传变异代表物种（种群）整体遗传上的变异，包括等位基因多样性和杂合度等。

3. Wahlund 效应

Wahlund 效应（Wahlund effect）是指种群内由于存在非随机交配而产生小的亚种群（就像分成了一个个小组一样。组内是随机交配的，但组间没有随机交配）导致种群平均遗传多样性（杂合度）降低的现象。具体说，一个物种种群存在隔离的亚种群，如果我们分亚种群分别计算遗传多样性，然后相加取平均值，这种“平均遗传多样性”将不能代表种群实际拥有的遗传多样性。举个极端的例子来说明：一个种群有 30 个个体，每 10 个组成一个亚种群，对于某个检测到的位点（包括两个等位基因 A 和 a），三个亚种群分别检测到的基因型见表 1-1。

表 1-1　假设的三个亚种群中某个基因型频率

	AA	*Aa*	*aa*
亚种群 1	10	0	0
亚种群 2	0	10	0
亚种群 3	0	0	10

用上面介绍的等位基因频率和遗传多样性的计算方法，可以算出，亚种群 1 的遗传多样性是 0（等位基因 A 的频率是 1，a 的是 0），亚种群 2 的遗传多样性是 0.5（等位基因 A 的频率是 0.5，a 的也是 0.5），亚种群 3 的遗传多样性是 0（等位基因 A 的频率是 0，a 的是 1）。三个值相加后的平均遗传多样性值是（0+0.5+0）/3=0.5/3。把三个亚种群合在一起，不分亚种群，重新计算种群遗传多样性是 0.5（等位基因 A 的频率是 0.5，a 的也是 0.5）。可以看出三个亚种群的平均值小于整体计算的值。当然上面这个例子中各亚种群不符合随机交配状态，只为说明 Wahlund 效应导致的结果是怎样的。

Wahlund 效应在实际种群中普遍存在，因为即使是很小范围的种群，个体间完全随机交配也很难保证。Wahlund 效应导致估算的种群纯合基因型频率增加，类似近交产生

的纯合子增多的现象（近交系数变大，后面的数据分析中会讲到），但并非种群近交而产生的，个体在亚种群内还是随机交配的。

4. 连锁不平衡

假设有两个位点，各有两个等位基因，即 A、a 和 B、b，等位基因频率分别为 p_1、p_2、q_1、q_2。它们组成 9 个基因型：*AABB*、*AaBB*、*aaBB*、*AABb*、*AaBb*、*aaBb*、*AAbb*、*Aabb* 和 *aabb*。当两个位点的等位基因出现非随机的组合（即连锁），就会导致某些基因型的频率明显大于（或小于）其他基因型频率，这时就称位点间出现了连锁不平衡（linkage disequilibrium，LD）（有时也称配子不平衡）。

连锁不平衡的度量可由下式得出：

$$D = P_{AB}P_{ab} - P_{Ab}P_{aB}$$

式中，P 为配子型频率。

如果 $D>0$，提示等位基因 A 和 B、a 和 b 有连锁的可能；如果 $D<0$，提示等位基因 A 和 b、B 和 a 有连锁的可能；如果 $D=0$，提示等位基因 A、a 和 B、b 之间随机组合，无连锁。

由于重组，连锁不平衡随世代而变化，如图 1-2 所示。

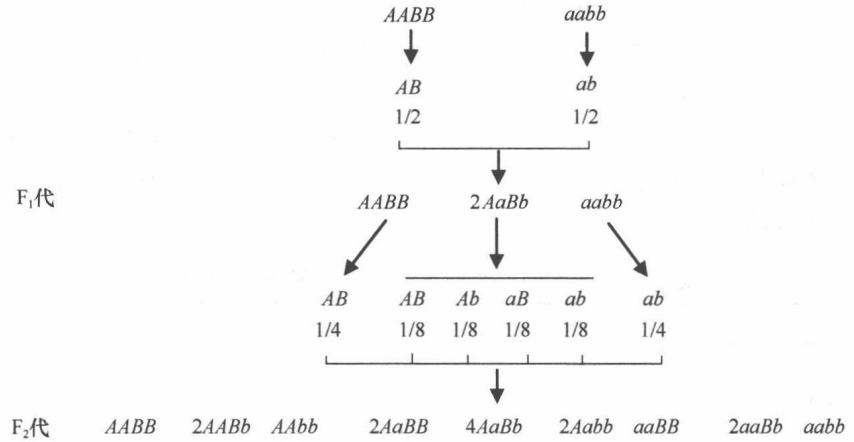

图 1-2　基因型在不同世代重组

即假设父母本的基因型分别是 *AABB* 和 *aabb*，产生 *AB* 和 *ab* 型配子的频率分别为 1/2 和 1/2，产生 *Ab* 和 *aB* 型配子的频率都为 0，由上式得 $D=1/4$；在 F_1 代，产生的 *AB*、*ab*、*Ab* 和 *aB* 型配子的频率将分别为 3/8、3/8、1/8 和 1/8，由上式得 $D=1/8$。

由于 D 可正可负，彼此不好比较，因此我们用一个归一化比值来表示等位基因连锁不平衡状况，即 $D'=D/D_{\max}$。当 $D>0$ 时，$D_{\max}=\min(p_1q_2, p_2q_1)$；当 $D<0$，$D_{\max}=\max(-p_1q_1, p_2q_2)$。

另外一种较多被使用来衡量连锁不平衡的方法是

$$r^2 = \frac{(P_{AB}P_{ab} - P_{Ab}P_{aB})^2}{p_1p_2q_1q_2}$$