# 一种基于候选基因研究树木生态适应的方法
## ——以川滇高山栎为例

王玉垚[1], 张　悦[1], 方剑火[2], 杜　芳[1]

1. 北京林业大学 林学院, 北京 100083
2. 清华大学 基因合成与测序平台, 北京 100082

**摘　要：**应用二代测序技术, 结合对川滇高山栎的生态适应研究, 介绍一种基于候选基因的研究方法, 其优点: 对受选择基因进行研究, 可以得到最为直接的证据来描述物种的生态适应模式; 利用标签序列, 混合所有个体样本进行二代测序, 极大地节约了建立测序文库的成本, 降低了研究费用, 可使研究者利用多个候选基因对大量样本开展实验和研究, 更好地阐明物种生态适应的机制.

**关键词：**生态适应; 候选基因; 混样测序; 标签序列; 川滇高山栎
**中图分类号：**Q-31　　**文献标识码：**A　　**文章编号：**0455-2059(2018)06-0811-06
**DOI:** 10.13885/j.issn.0455-2059.2018.06.014

# A research method for ecological adaptation of tree species based on candidate genes: with *Quercus aquifolioides* as an example

Wang Yu-yao[1], Zhang Yue[1], Fang Jian-huo[2], Du Fang[1]

1. College of Forestry, Beijing Forestry University, Beijing 100083, China
2. Genomics & Synthetic Biology Center, Tsinghua University, Beijing 100082, China

**Abstract:** The next generation sequencing was applied to introduce a new procedure of developing candidate genes in a non-model tree species *Quercus aquifolioides*. The method described here showed two advantages: being able to obtain the most direct evidence to describe the patterns of ecological adaptation through selected genes and reduce the total costs by using barcodes for preparation of sequencing libraries. These advantages enable researchers to elucidate the mechanism of ecological adaptation using a large number of samples and genes.

**Key words:** ecological adaptation; candidate gene; pool-sequencing; barcode; *Quercus aquifolioides*

快速的气候变化正在影响着物种的分布范围, 甚至威胁到一些物种的生存[1−3]. 这种影响对树木尤为显著, 因为树木的灭绝不仅仅是失去一个物种, 而且对整个生态系统及其多样性都会产生深远的负面影响[4−6].

植物在长期演化过程中有3种应对气候变化的特有模式: 1) 短期内改变生活史或者通过表型可塑性产生形态变化; 2) 通过远距离迁移的方式

**收稿日期：**2017-09-11　　**修回日期：**2017-11-15
**基金项目：**国家自然科学基金项目(41671039); 北京市科技新星项目(Z151100000315056)
**作者简介：**杜　芳(1981−), 女, 甘肃兰州人, 副教授, 博士, e-mail: dufang325@bjfu.edu.cn, 研究方向为种群遗传学、森林生态学, 通信联系人.

扩散至更适宜其生存的地区; 3) 通过生态适应(ecological adaptation)演化获得对环境适合度更高的新基因型, 避免物种局部或大面积灭绝[7]. 目前关于植物应对气候变化模式的研究主要集中在表型可塑性和物种扩散两个方面. 对于生活周期较长、固着生长的木本植物来说, 第1和2种模式的影响非常有限, 更多地通过适应性演化来应对气候变化. 因此, 越来越多的科学家将注意力转移到适应性基因, 即受选择基因的遗传变异上, 尝试解读自然居群中适应性遗传变异的空间结构及其与气候变化的关系, 从而得到直接有力的证据来阐明物种对气候变化的响应模式.

为了在自然居群中鉴定与气候相关的候选基因, 最重要的依据是其遗传变异与气候梯度相关[8]. 近年来对树木居群的研究使用了全基因组扫描(whole-genome scan)的方法来发现与气候梯度显著相关的单核苷酸多态性(single-nucleotide polymorphisms, SNP)位点, 由此来推定受选择的基因区域[9−12]. 然而树木由于生长周期长、多为异交且基因组非常大, 全基因组扫描的方法受到限制, 必须通过候选基因(candidate gene)测序的手段获得与气候梯度显著相关的SNP位点[13]. 一种取代的方法是参与气候适应的基因序列[11, 14−17], 优点在于这些功能基因直接参与控制观察到的适应性表型变异[13, 18], 越来越多的树木全基因组及其注释信息的发表(表1), 使得这种方法变得更加直接有效.

在第2代测序技术(next generation sequencing, NGS)成熟之前, 很多研究选择使用一代测序的手段来对扩增的候选基因SNP位点进行分型[19−21]. 但是由于一代测序通量小、费用高、耗时长的局限性, 不能满足大批量检测候选基因变异信息的需求, 无法得到充足的样本数据来得出结论[22]. NGS成熟以后, 其通量高、价格低的优点[22], 使以上问题得以完美地解决并已在该类研究中得到应用[23−26].

栎类植物是北半球森林生态系统的重要组成部分, 是陆地生物多样性的主要驱动力之一[27]. 川滇高山栎(*Quercus aquifolioides*)为广泛分布于四川、贵州、云南和西藏的硬叶常绿乔灌木, 海拔2 000~4 300 m均有分布, 为当地的优势种和建群种[28]. 关于该物种基于中性分子标记的叶绿体基因片段和核微卫星(nSSR)标记的谱系地理学研究已经完成, 揭示了川滇高山栎的遗传结构、居群动态, 并对适应性分化进行了推测[29]. 由于川滇高山栎生境异质性明显, 是研究生态适应的理想材料. 本研究利用NGS手段对候选基因进行SNP分型, 对川滇高山栎进行生态适应性研究, 并着重介绍可有效降低实验成本的混样测序(pool-sequencing)方法.

## 1 实验设计

### 1.1 获得候选基因

研究的第1步是获取与研究目的相关的基因(例如与抗旱、抗寒密切联系的基因)作为候选基因. 从数据库网站下载候选基因, 是最为直接、便捷、有效的方法. 目前, 基因组测序和基因功能注释迅速发展, 越来越多物种的基因组信息得以公开发表,可以很便捷地下载目的物种或近缘种的候选基因及其基因功能注释信息(表1), 也可从其他途径获取所需要的候选基因, 如从转录组序列中挑选需要的基因[24], 并使用Blastx (https://blast.ncbi.nlm.nih.gov/Blast.cgi)进行功能注释.

所挑选的基因片段长度不能过长, 应尽量短于测序仪读长, 以免造成有效数据丢失. 如果计划使用Illumina Pair End 250进行测序, 若序列长度大于500 bp, 序列中间位置的信息无法获得, 从而造成数据丢失.

### 1.2 引物设计、引物筛选和标签设计

将候选基因序列作为模板进行引物设计.

**表1 树种候选基因数据库举例**
**Table 1 Some candidate gene databases of tree species**

| 类群 | 数据库 | 网址 |
|---|---|---|
| 云杉属 *Picea* | Spruce Genome Project | http://congenie.org/start |
| 松属 *Pinus* | Pine Reference Sequence | http://www.pinegenome.org/pinerefseq |
| 杨属 *Populus* | The Populus Genome Integrative Explorer | http://popgenie.org |
| 桦木属 *Betula* | The Dwarf Birch Genome Project | http://birchgenome.org |
| 桉属 *Eucalyptus* | *Eucalyptus camaldulensis* Genome Database | http://www.kazusa.or.jp/eucaly |
| 栎属 *Quercus* | Oak Genome Sequencing | http://www.oakgenome.fr |
| 橡胶属 *Hevea* | Rubber Tree Genome | http://www4a.biotec.or.th/rubber |
| 桑属 *Morus* | Morus Genome Database | http://morus.swu.edu.cn/morusdb |

由于下载的目的基因可能来源于近缘物种,因此需要选择4~8个研究物种的样本个体对设计的引物进行PCR验证,以确保其可用性.若扩增失败,则需重新设计新的引物.将成功扩增的PCR产物进行Sanger测序.测序数据需经过MEGA 7.0[30]或其他软件比对,保留有多样性的基因.最后,保留Sanger测序序列作为参考序列,用于后续开发SNP等步骤.

在经过验证且具有多样性的基因引物一端或两端加上不同的标签序列,分别用于扩增不同的样本个体,可以使不同个体的序列在混样及二代测序后通过标签序列进行区分(图1).此步是整个流程的最关键所在,可以使几十甚至上百个样本个体的PCR产物混合到一个文库中进行测序,极大地提高效率,降低建库成本.根据对实验费用的预估,设计适当数量的标签序列,用于此实验.标签序列的设计原则与引物设计原则类似:1)标签长度一般为6 bp,每个标签之间至少有3个碱基的差异,以进行有效地区分;2)嘌呤和嘧啶的组成大致相当;3)避免连续3个碱基的重复;4)避免形成引物二聚体;5)避免与引物一起形成发卡结构等.将引物与标签进行连接,组成"引物+标签"组合,交由生物公司合成,用于后续实验.

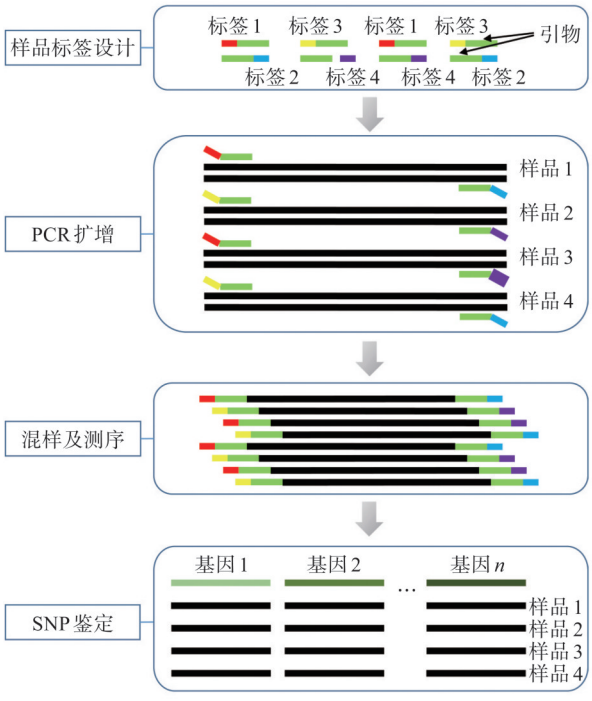


图1 混样测序方法的流程

Fig. 1 Flow of pool-sequencing method

### 1.3 批量PCR扩增、混合PCR产物以及二代测序

在PCR时,将连接着不同标签的引物分别与每个个体对应,这样每个样本的PCR产物前后均被"贴上了"代表样本个体的"标签"(图1).从所有PCR产物中分别抽取1~2 μL,混匀后,进行建库测序.

### 1.4 SNP鉴定

测序前混样建库使得下机数据混合了所有样本所有候选基因的序列信息.由于事先在引物前加上了用于区分样本个体的标签序列,所以可利用标签序列信息将每个样本个体的序列挑选出来,生成包含每个样本序列信息的文件.可以使用perl语言编写合适的脚本,利用标签序列分别将个体逐个进行分离.

使用Trimmomatic-0.36[31]或其他软件去除低质量的序列及上机测序前加上的测序平台专用接头.使用GATK pipeline[32]对序列进行比对,去重复和获取变异位点信息后,根据需要对插入/缺失(insertion-deletion)和多核苷酸多态性(multi-nucleotide polymorphism)等特殊变异结构进行处理(直接过滤掉或者将其作为一个或多个SNP).最终得到的包含高质量变异位点信息的.vcf文件,用于后续居群遗传变异、遗传分化等研究.此步骤同样可以使用SAMtools[33]、FreeBayes (https://www.geneious.com/plugins/freebayes/)等软件进行处理,可根据实际情况进行选择.

变异位点信息是后续群体遗传学分析的基础,可以以.vcf文件为基础,使用Arlequin 3.5[34]、STRUCTURE 2.3[35]、FDIST2[36]、BayEnv[9]等软件进行遗传多样性、遗传结构检测、变异位点异常值检测以及遗传因子与环境因子的相关性分析等各式分析.

## 2 实验费用预测

标签可以接到一端引物上,也可以加到两端引物上配合使用.当使用单端标签时,需要使用的标签数量必须等于样本数量(假设建立一个测序文库);使用双端标签时,用$m+n$条标签,便可用于区分$m\times n$个样本.使用单端标签时,仅需要再合成单端的"引物+标签"序列,另一端可使用上一步合成的引物.使用双端标签时,则需要再合成两端的"引物+标签"序列.虽然单端"引物+标签"的方法只需要合成一端,但是一个"引物+标签"只

能区分一个样本个体, 因此当样本数量比较多时, 需要合成的“引物+标签”也会很多, 而双端“引物+标签”配合使用, 则不需要合成太多.

预测了假设的3种测序方法的总费用: 一代测序、单端加标签的NGS和双端加标签(图2a), 总花费包括引物合成、“引物+标签”合成、PCR、建库和测序费用. 以100个基因为例, 需要设计100对引物, 基因序列长度设定为500 bp; 引物长度预计为20 bp, 标签长度预计为6 bp, 合成费用为0.5元/bp; PCR费用预计5元/反应; 建库费用以1 000元为例; 测序深度为50倍, 测序费用为500元/G; 一代测序的费用不包括“引物+标签”合成和建库费用, 测序费用为15元/反应. 从图2a可以看出, 随着样本数量的增多, 一代测序的费用始终最高, 而双端加标签的NGS费用始终最低.

以上预测是建立在只有一个测序文库的前提之下的, 也可以建立多个文库, 以减少标签的使用数量, 从而减少“引物+标签”的巨额合成费用. 假设使用500个样本, 建库的数量会根据标签使用数量的改变而改变. 例如使用10条单端标签, 这时需要建立50个测序文库, 当使用20条单端标签时, 只需要建立25个测序文库. 双端标签也是类似的情况, 例如当使用5对双端标签时, 需要建立20个测序文库, 当使用10对双端标签时, 只需要建立5个测序文库. 此时再对总费用进行估测(图2b). 由于样本数量固定, PCR费用不变, 因此在进行本项预测时排除PCR费用. 从图2b可以看出, 一代测序仍然是费用最高的方法; 两种NGS方法花费差别不大, 但双端加标签测序方法的最低费用依然比单端加标签方法的最低费用低2万元左右(单端标签数量为25时, 费用最低, 为59 250元; 双端标签数量为15时, 费用最低, 为36 750元). 这两种方法相比, 标签使用数目一样的情况下, “引物+标签”的合成费用一致, 引物合成、测序费用以及没有被统计的PCR费用均一致, 唯一的区别为建库费用. 因此, 这种可以区分更多样本个体而建立更少测序文库的双端加标签的测序方法更加节约成本. 然而, 标签使用数量并不是越多越好. 在设计实验时, 需要根据实际情况进行费用预估, 使用适当数量的标签进行后续实验.

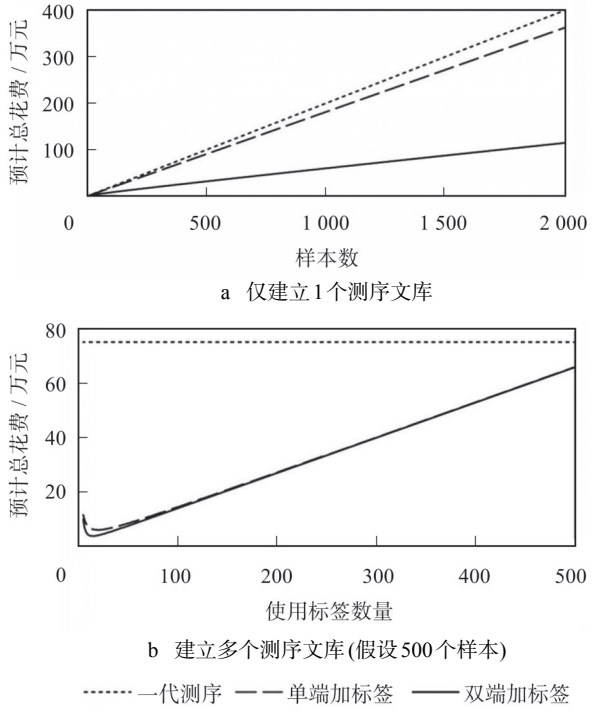


图2 预测一代测序、单端加标签的NGS和双端加标签的NGS的总费用

Fig. 2 Estimation of total costs using sanger sequencing, NGS with single-end barcode, NGS with pair-end barcodes

## 3 对川滇高山栎的生态适应性研究

从川滇高山栎分布区域内采集到60个居群587个样本, 提取每个样本的全基因组DNA, 用于生态适应研究. 从已公开发表的欧洲栎树基因组(http://www.oakgenome.fr/)和注释的功能基因组网站EvolTree (http://www.evoltree.eu/)选取180个抗干旱、寒冷的基因, 设计引物以及使用4个川滇高山栎样品进行筛选后, 保留可用的有多样性的引物共73对, 用于后续实验.

标签的使用数目不同, 导致“引物+标签”合成费用不同、建库费用不同, 从而总花费不同. 因此提前对实验的花费进行预估, 包括引物合成、“引物+标签”合成、PCR、建库和测序费用. 从图3可知, 使用20条标签用于双端加标签的NGS方法费用最少.

将每对引物均与10对标签连接, 组成10对“引物+标签”组合, 交由生物公司合成. 将每一个基因的10对“引物+标签”两两组合, 共100种组合方式, PCR扩增时, 分别与100个样本对应. 批量扩增以后, 从100个样本个体的所有候选基因PCR产物中分别抽取1.5 μL, 均匀混合后, 交由生物公司进行建库测序. 对川滇高山栎的研究一共用到587个样本个体, 因此共建立了6个文库, 实际每个文库分别混合了100、100、99、96、96、96个样本个体的PCR产物.

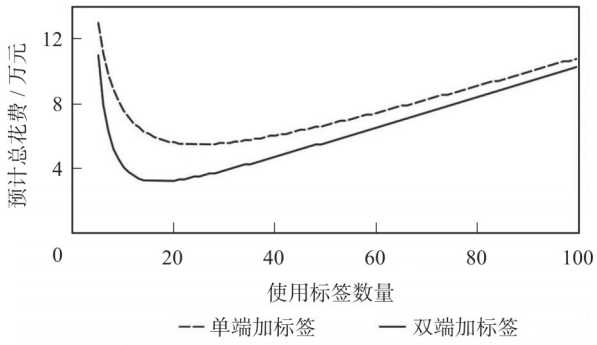


图3 预测使用不同数目的标签所需要的总费用
Fig. 3 Estimation of total costs using different number of barcodes

测序下机数据为6对代表每个测序文库的双端.fastq文件，每对序列文件中都混合了100个样本所有候选基因的序列信息.利用双端标签序列信息，使用perl语言编写的脚本，将个体的序列从6个库的文件中逐个挑选出来，共分离出587对代表样本个体的.fastq序列文件. 去除低质量的序列以及测序平台专用引物后，使用GATK pipeline鉴定SNP位点信息，得到包含700个高质量SNP位点信息的.vcf文件. 根据最小等位基因频率(minor allele frequency)>2.5%的限制条件对SNP进行过滤，排除假阳性的可能性，最终得到378个SNP位点信息，分布于64个基因中. 之后所有的遗传变异信息均以此378个SNP位点信息为基础，通过各种分析方法得到.

## 4 总结与展望

使用候选基因对物种进行研究，能得到直接的生态适应证据，为研究物种的适应机理提供重要的依据. 基于高通量测序技术的混样测序方法可以极大地降低研究成本，使基于候选基因的生态适应研究不再承担不起高昂费用，使研究者们可大胆使用大量样本、多候选基因开展实验，极大地促进该类研究的开展.

尽管存在巨大的优势，但该方法依然存在着不可忽视的问题. 例如，人工混样导致的测序深度不均匀；高通量测序技术依然存在较高的错误率，尤其是标签序列，导致从下机数据分离样本个体的序列时，易造成数据的丢失，导致最后生成的VCF文件存在大量的数据缺失，这两个问题可通过加大测序量和测序深度来尽量避免. 另外，从下机数据中分离样本个体时，方法较为繁琐耗时，开发SNP的方法比较复杂，亟待开发简单方便的方法.

## 参考文献

[1] Hughes I. Biological consequences of global warming: is the signal already apparent?[J]. Trends in Ecology & Evolution, 2000, 15(2): 56-61.

[2] Parmesan C. Ecological and evolutionary responses to recent climate change[J]. Annual Review of Ecology Evolution & Systematics, 2006, 37(1): 637-669.

[3] Parry M L, Canziani O F, Palutikof J P, et al. Climate change 2007: impacts, adaptation and vulnerability. Contribution of working group Ⅱ to the fourth assessment report of the intergovernmental panel on climate change[M]. Cambridge: Cambridge University Press, 2007.

[4] Hughes A R, Inouye B D, Johnson M T, et al. Ecological consequences of genetic diversity[J]. Ecology Letters, 2008, 11(6): 609-623.

[5] Kremer A, Ronce O, Robledoarnuncio J J, et al. Long-distance gene flow and adaptation of forest trees to rapid climate change[J]. Ecology Letters, 2012, 15(4): 378-392.

[6] Sork V L. Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate[J]. Tree Genetics & Genomes, 2013, 9(4): 901-911.

[7] Riordan E C, Gugger P F, Ortego J, et al. Association of genetic and phenotypic variability with geography and climate in three southern California oaks[J]. American Journal of Botany, 2016, 103(1): 73-85.

[8] Endler J A. Natural selection in the wild[M]. Princeton: Princeton University Press, 1986.

[9] Coop G, Witonsky D, Rienzo A D, et al. Using environmental correlations to identify loci underlying local adaptation[J]. Genetics, 2010, 185(4): 1411-1423.

[10] Eckert A J, Bower A D, González-Martínez S C, et al. Back to nature: ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae)[J]. Molecular Ecology, 2010, 19(17): 3789-3805.

[11] Eckert A J, Heerwaarden J V, Wegrzyn J L, et al. Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L. Pinaceae)[J]. Genetics, 2010, 185(3): 969-982.

[12] Bashalkhanov S, Eckert A J, Rajora O P. Genetic signatures of natural selection in response to air pollution in red spruce (*Picea rubens*, Pinaceae)[J]. Molecular Ecology, 2013, 22(23): 5877-5889.

[13] Neale D B, Savolainen O. Association genetics of complex traits in conifers[J]. Trends in Plant Science, 2004, 9(7): 325-330.

[14] González-Martínez S C, Ersoz E, Brown G R, et al. DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for

drought-stress response in *Pinus taeda* L[J]. Genetics, 2006, 172(3): 1915-1926.

[15] Holliday J A, Ritland K, Aitken S N. Widespread, ecologically relevant genetic markers developed from association mapping of climate-related traits in Sitka spruce (*Picea sitchensis*) [J]. New Phytologist, 2010, 188(2): 501-514.

[16] Alberto F J, Derory J, Boury C, et al. Imprints of natural selection along environmental gradients in phenology-related genes of *Quercus petraea*[J]. Genetics, 2013, 195(2): 495-512.

[17] 张腾国, 毛玉珊, 常燕, 等. 油菜*ICE*1基因表达载体构建及转化烟草[J]. 兰州大学学报: 自然科学版, 2016, 52(2): 277-282.

[18] Gonzálezmartínez S C, Krutovsky K V, Neale D B. Forest-tree population genomics and adaptive evolution[J]. New Phytologist, 2006, 170(2): 227-238.

[19] Eckert A J, Wegrzyn J L, Pande B, et al. Multilocus patterns of nucleotide diversity and divergence reveal positive selection at candidate genes related to cold hardiness in coastal Douglas fir (*Pseudotsuga menziesii* var. *menziesii*) [J]. Genetics, 2009, 183(1): 289-298.

[20] Keller S R, Levsen N, Olson M S, et al. Local adaptation in the flowering-time gene network of balsam poplar, *Populus balsamifera* L[J]. Molecular Biology and Evolution, 2012, 29(10): 3143-3152.

[21] Zhou Yong-feng, Zhang Li-rui, Liu Jian-quan, et al. Climatic adaptation and ecological divergence between two closely related pine species in Southeast China[J]. Molecular Ecology, 2014, 23(14): 3504-3522.

[22] 王兴春, 杨致荣, 王敏, 等. 高通量测序技术及其应用[J]. 中国生物工程杂志, 2012, 32(1): 109-114.

[23] Sork V L, Squire K, Gugger P F, et al. Landscape genomic analysis of candidate genes for climate adaptation in a California endemic oak, *Quercus lobata*[J]. American Journal of Botany, 2016, 103(1): 33-46.

[24] Roschanski A M, Csilléry K, Liepelt S, et al. Evidence of divergent selection for drought and cold tolerance at landscape and local scales in *Abies alba* Mill. in the French Mediterranean Alps[J]. Molecular Ecology, 2016, 25(3): 776-794.

[25] Rellstab C, Zoller S, Walthert L, et al. Signatures of local adaptation in candidate genes of oaks(*Quercus* spp.) with respect to present and future climatic conditions[J]. Molecular Ecology, 2016, 25(23): 5907-5924.

[26] 令利军, 何楠, 白雪, 等. 基于高通量测序的玉米秸秆自然发酵过程中细菌菌群结构特征[J]. 兰州大学学报: 自然科学版, 2017, 53(4): 526-533.

[27] Kremer A, Abbott A G, Carlson J E, et al. Genomics of Fagaceae[J]. Tree Genetics & Genomes, 2012, 8(3): 583-610.

[28] 周浙昆. 中国栎属的起源演化及其扩散[J]. 植物分类与资源学报, 1992, 14(3): 227-236.

[29] Du Fang, Hou Meng, Wang Wen-ting, et al. Phylogeography of *Quercus aquifolioides* provides novel insights into the Neogene history of a major global hotspot of plant diversity in south-west China[J]. Journal of Biogeography, 2017, 44(2): 294-307.

[30] Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets[J]. Molecular Biology & Evolution, 2016, 33(7): 1870-1874.

[31] Bolger A M, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data[J]. Bioinformatics, 2014, 30(15): 2114-2120.

[32] McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data[J]. Genome Research, 2010, 20(9): 1297-1303.

[33] Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools[J]. Bioinformatics, 2009, 25(16): 2078-2079.

[34] Excoffier L, Lischer H E. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under linux and windows[J]. Molecular Ecology Resources, 2010, 10(3): 564-567.

[35] Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software structure: a simulation study[J]. Molecular Ecology, 2005, 14(8): 2611-2620.

[36] Beaumont M A, Balding D J. Identifying adaptive genetic divergence among populations from genome scans[J]. Molecular Ecology, 2004, 13(4): 969-980.

(责任编辑: 蔡红霞)