中国科学: 生命科学 2015 年 第 45 卷 第 5 期: 488 ~ 497
SCIENTIA SINICA Vitae www.scichina.com life.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

论 文

辣木(*Moringa oleifera* Lam.)的高质量参考基因组

田洋[①⑩⑬†], 曾严[④†], 张静[⑧†], 杨承光[⑨], 严亮[①⑤], 王宣军[⑬], 史崇颖[②], 谢静[③], 戴天浥[②], 彭磊[②], 曾寰宇[①], 徐安妮[①], 黄业伟[⑬], 张佳进[⑪⑫], 马啸[⑬], 董扬[⑦⑩], 郝淑美[⑥*], 盛军[⑬*]

① 吉林大学生命科学学院, 长春 130012;
② 云南农业大学食品学院, 昆明 650201;
③ 云南农业大学动物学院, 昆明 650201;
④ 中国科学院大学生命科学学院, 北京 100049;
⑤ 普洱茶研究院, 普洱 665000;
⑥ 云南大学农学院, 昆明 650091;
⑦ 昆明理工大学生命科学与技术学院, 昆明 650093;
⑧ 华中科技大学生命科学学院, 武汉 430074;
⑨ 武汉大学生命科学学院, 武汉 430072;
⑩ 云南省高原特色农业研究院云南辣木研究所, 昆明 650201;
⑪ 云南农业大学信息科学与工程学院, 昆明 650201;
⑫ 中国科学院动物研究所, 遗传资源与进化国家重点实验室, 昆明 650223;
⑬ 云南农业大学, 普洱茶学教育部重点实验室, 昆明 650201
† 同等贡献
* 联系人, E-mail: haosm@sina.com; shengjunpuer@163.com

收稿日期: 2015-01-07; 接受日期: 2015-02-27

摘要 辣木因其具有高蛋白含量和对干旱的适应在许多发展中国家作为多年生的作物广泛种植. 本文完成了高质量的辣木基因组草图, 组装出预测基因组 91.78%的大小, 注释出来 19465 个蛋白质编码基因. 此外, 本文对辣木基因组和其他一些物种进行了比较基因组分析, 验证了辣木的系统发生地位, 同时鉴别了辣木的一些物种特异的基因家族和受正选择的基因, 这些基因可能帮助进一步鉴别与辣木的高蛋白、快速生长和抗逆相关的基因. 这个参考基因组将开拓对辣木的研究, 促进应用基因组学手段对辣木的育种和改良.

关键词
辣木
基因组
测序

虽然从 1950 年起, 谷物的产量已经翻倍, 但是世界上仍有 1/7 的人营养不良, 而且这些人主要集中在欠发达国家. 造成食物匮乏的原因很多, 其中一个原因就是人类食物主要是一年生的农作物, 如玉米(*Zea mays*)、小麦(*Triticum aestivuml*)、水稻(*Oryza sativa*)和蔬菜, 这些作物需要每年播种, 非常耗费人力和资源, 而且很多发展中国家的气候和生态条件并不适合种植这些一年生的农作物. 然而, 多年生的植物播种一次可以存活多年, 水分和营养的循环利用效率较高, 可以适应非常广泛的生态条件, 可能替

引用格式: 田洋, 曾严, 张静, 等. 辣木(*Moringa oleifera* Lam.)的高质量参考基因组. 中国科学: 生命科学, 2015, 45: 488–497
英文版见: Tian Y, Zeng Y, Zhang J, et al. High quality reference genome of Drumstick tree (*Moringa oleifera* Lam.), a potential perennial crop. Sci China Life Sci, 2015, 58, in press

代传统的一年生作物. 但是到目前为止, 多年生的作物除了香蕉(*Musa nana*)、可可(*Theobroma cacao*)、木豆(*Cajanus cajan*)外, 基本都没有得到广泛的种植和食用.

近期一种原产于印度、巴基斯坦、尼泊尔的中小型的常青树辣木(*Moringa oleifera* Lam.), 也称鼓槌树、油赖木, 在农业和工业上越来越受到关注[1~5], 辣木的各个部位可以作为食物、入药或者用于工业生产, 也因其较高的蛋白、维他命、矿物质含量而在发展中国家广泛种植[2,4,6,7]. 辣木可以在海拔 0~1800 m、年降雨量 500~1500 mm 的地区种植, 适合干旱和半干旱的地区, 而这样的地区占地球陆地面积的 37.0%, 发展中国家的干旱地区比重更大. 许多地区正在大力推广种植辣木, 但关于辣木的基础研究还有许多空白, 这也限制了对辣木的进一步开发利用. 本课题组首次对辣木进行了全基因组测序, 组装了高质量的基因组并且完成了注释工作, 这些工作将极大程度地促进对多年生植物辣木的利用和研究.

1 材料与方法

1.1 材料

测序使用的 DNA 从云南普洱栽培的一年树龄的辣木叶中提取. 共提取了 50 μg DNA 用于建库.

1.2 测序数据的产生和处理

使用全基因组鸟枪法测序的策略, 用 Illumina Hiseq2500TM(美国)共建了 7 个片段大小不同的库, 分别为 177, 222, 390, 503, 3500, 11500 和 15000 bp. 测序前还需连接适配序和制备 DNA 簇. 经图像分析、碱基识别、序列分析 3 个步骤共得到 202 Gb 数据. 然后过滤掉其中质量值不高的读取、含有“N”的读取、重复的读取以及含有适配序列的读取. 所有的读取都去掉末端 2 个碱基. 然后用 SOAPec2.01[8]工具进行 *K*-mer 分析和纠错来减少由于测序错误导致的低频读取.

1.3 基因组组装

使用 Platanus 1.2.1[9]把读取序列组装成 contig. Platanus 是一个专门针对高杂合度的基因组软件. 参数设置如下: initial *K*-mer size 41, step size 10, maximum difference for branch cutting 0.3, maximum difference for bubble crush 0.15, *K*-mer coverage cutoff 5. 然后使用 SSPACE v2.0[10]来组装 scaffold. 然后使用 SOAPdenovo[8]含有的 Gapcloser v1.12 对 scaffold 进行补洞, 得到最后的组装版本. 组装完之后, 使用 SOAPaligner 2.18 把所有的读取序列都与基因组比对, 用于评估组装的质量.

1.4 重复序列的注释

使用 Tandem Repeats Finder (TRF) 4.04[11]鉴别辣木基因组中的串联重复序列, 使用 Repeatmasker 3.3.0 和 RepeatProteinMask 分别从 DNA 和蛋白质水平把重复序列与 Repbase[12]比对. 这些同源预测的结果, 结合使用 LTR_FINDER 1.05[13]和 RepeatScout[14]的从头预测结果, 最后得到对辣木基因组重复序列的注释.

1.5 蛋白编码基因的注释

结合同源预测和从头预测的方法, 共预测出 19465 个蛋白编码基因. 同源预测方法使用了拟南芥(*Arabidopsis thaliana*)[15]、大豆(*Glycine max*)[16]、水稻(*Oryza sativa*)[17]、白杨树(*Populus trichocarpa*)[18]、高粱(*Sorghum bicolor*)[19]、卷柏(*Selaginella moellendorffii*)[20] 6 个物种的蛋白质序列, 选择这 6 个物种是因为这些物种都具有组装、注释良好的基因组, 并且涵盖了从裸子植物到被子植物的各大分支, 选择这些物种也便于分析辣木的进化过程. 同源预测时, 首先进行 tBLASTN 比对, e-value 设置为 1×10^{-5}. 比对上的序列上下游各延伸 2000 bp, 然后和蛋白序列用 GeneWise[21]比对识别基因结构. 从头预测基因的方法使用了 AUGUSTUS 2.5.5[22], Genscan 和 Glimmer-HMM 3.0.1[23]. 从头预测的基因与拟南芥的蛋白序列比对, 重叠率阈值设置为 0.5. 然后, 同源预测和从头预测的基因序列使用 GLEAN 软件, 根据不同的基因结构信息, 得到一致的基因集.

1.6 基因功能注释

通过把已注释的基因与 TrEMBL[24], KEGG[25]和 InterProscan[26]数据库比对, 注释出基因的潜在功能.

1.7 非编码基因的注释

注释 tRNA 使用 tRNAscan-SE v1.23[27]软件. 把 Rfam[28]数据库下载的序列作为参考序列, 用同源预测的手段注释出 rRNA. INFERNAL v0.81[29]用于鉴别

snRNA 和 miRNA.

1.8 miRNA 靶基因分析

从 miRbase[30]下载成熟的 miRNA 序列, 然后和预测的 miRNA 序列比对, 比对长度大于 16 bp 的序列挑选出来作为潜在的 miRNA 序列, 然后使用在线工具 psRNATarget[31]预测这些 miRNA 序列的靶位点序列.

1.9 基因家族分析

使用葡萄(*Vitis vinifera*)、木豆(*Cajanus cajan*)、番木瓜(*Carica papaya*)、苹果(*Malus pumila*)以及辣木的基因组序列, 先两两之间做 BLASTP 比对, e-value 阈值设置为 1×10^{-5}, 然后用 OrthoMCL 1.4[32]做基因家族分类, 参数缺省.

1.10 分类关系和分歧时间分析

使用 5 种木本植物(辣木、葡萄、木豆、番木瓜、苹果)的基因做基因家族分析, 得到单拷贝的基因, 并且使用单拷贝的基因用 MUSCLE 3.8.31[33] 做多序列比较. 然后分析四重简并位点, 并且把每个物种的四重简并位点的序列连接成一条线性序列, 再用 PhyML 3.0 来做邻接树. 最后使用 http://www.timetree.org/上已知的分歧时间数据和 MCMCTREE[34]软件来估计邻接树上各个物种之间的分歧时间.

1.11 基因家族扩张收缩

本文使用 CAFE2.1[35]来研究基因家族的收缩扩张历史.

1.12 正选择分析

使用番木瓜的基因作为参考来研究辣木基因的正选择情况. 首先用 BLAST 把辣木的基因和番木瓜的基因相互比对, 找到最佳的比对, 再找出直系同源基因对, 一共 5601 对. 再用 LASTZ 把这些直系同源基因对比对, 输出的结果使用 KaKs_Calculator 1.2[36]分析, 得到每个基因对的 *Ka*/*Ks* 比率. 画图使用 ClustalX[37](图 3, 网络版附图 1).

2 结果

2.1 基因组组装

共使用了 457×覆盖度的 DNA 序列数据. 测序数据汇总见网络版附表 1, 数据的 *K*-mer 分析见网络版附图 2. 根据 17-mer 的频率分布, 预测辣木的基因组大小为 315 M(网络版附表 2), 使用流式细胞术预测辣木的核基因组大小(C 值)略小于水稻. 最终组装出的 contig 和 scaffold 的 N50 值分别为 123 kb 和 1.14 Mb (表 1), 基因组总大小为 289 Mb, 总长度 80%以上(231 Mb)的序列集中在 262 个 scaffold 上. 辣木的基因组的 N50 和近期发表的一些高质量的植物基因组相近[20]. 而且 95.67%的测序序列能够重新比对到辣木的基因组上, 这进一步验证了辣木基因组的质量(网络版附表 3).

木本植物的基因组大小分布很广, 从梅花(*Prunus mume*)[38]的 280 Mb 到火炬松(*Pinus taeda*)[39]的 221.8 Gb. 辣木的基因组大小在木本植物中是很小的, 与梅花相近, 比水稻的略小. 小基因组对辣木来说, 不仅使它具有了快速生长、种子产量高、适应干旱以及半干旱地区等特性, 还让辣木成为一个潜在的研究木本植物基因组特性的突破口.

2.2 辣木的基因组注释

使用同源预测和从头预测的方法, 在辣木基因组中共注释出 19465 个基因, 每个基因的平均长度为 3354.22 bp, 包含 5.42 个外显子(网络版附表 4, 网络版附图 3 和 4). 对蛋白质编码基因进行基于基因结构的分析, 结果显示, 93.74%的辣木基因在 TrEMBL 蛋白质数据库中有同源序列, 72.67%的序列可以在 Swiss-Prot[24]中分门别类. 共有 94.01%的基因有已知的同源基因或者可以通过 InterPro, GO, KEGG, Swiss-Prot, TrEMBL[40]数据库进行功能注释分类(网络版附表 5).

依据结构以及同源序列的分析, 共鉴定了 148820058 bp 的重复原件, 涵盖了大部分的植物转座子种类. 大部分的重复序列都是从头预测出来的, 只有 10.1%的重复序列是通过同源查找的方式找到, 侧面验证了辣木与其他已经发表的植物在进化上的关系很远. 转座子的总长度达到 136 Mb, 占组装出来的基因组大小的 47.10%, 再加上其他的很多重复序列, 这些重复原件共占基因组的 51.45%(网络版附表 6 和 7). 网络版附图 5 和 6 展示了转座子的分歧率分布分析结果. 在网络版附表 8 中总结了非编码基因的注释情况, 其中注释了 87 个成熟的 miRNA 和 369 个潜在的 miRNA 靶基因(网络版附表 9). 使用

表 1　辣木基因组组装总览

统计量	contig		scaffold	
	长度(bp)	数量	长度(bp)	数量
N90	4165	4362	5792	1382
N80	30989	1914	150929	262
N70	60562	1261	396940	147
N60	91660	880	736902	93
N50	123008	611	1140476	61
最长	1070888	–	6788971	–
平均长度	6911	–	8677	–
总数量(>1000 bp)	–	13512	–	10494
综合	287419725	41586	289241074	33332

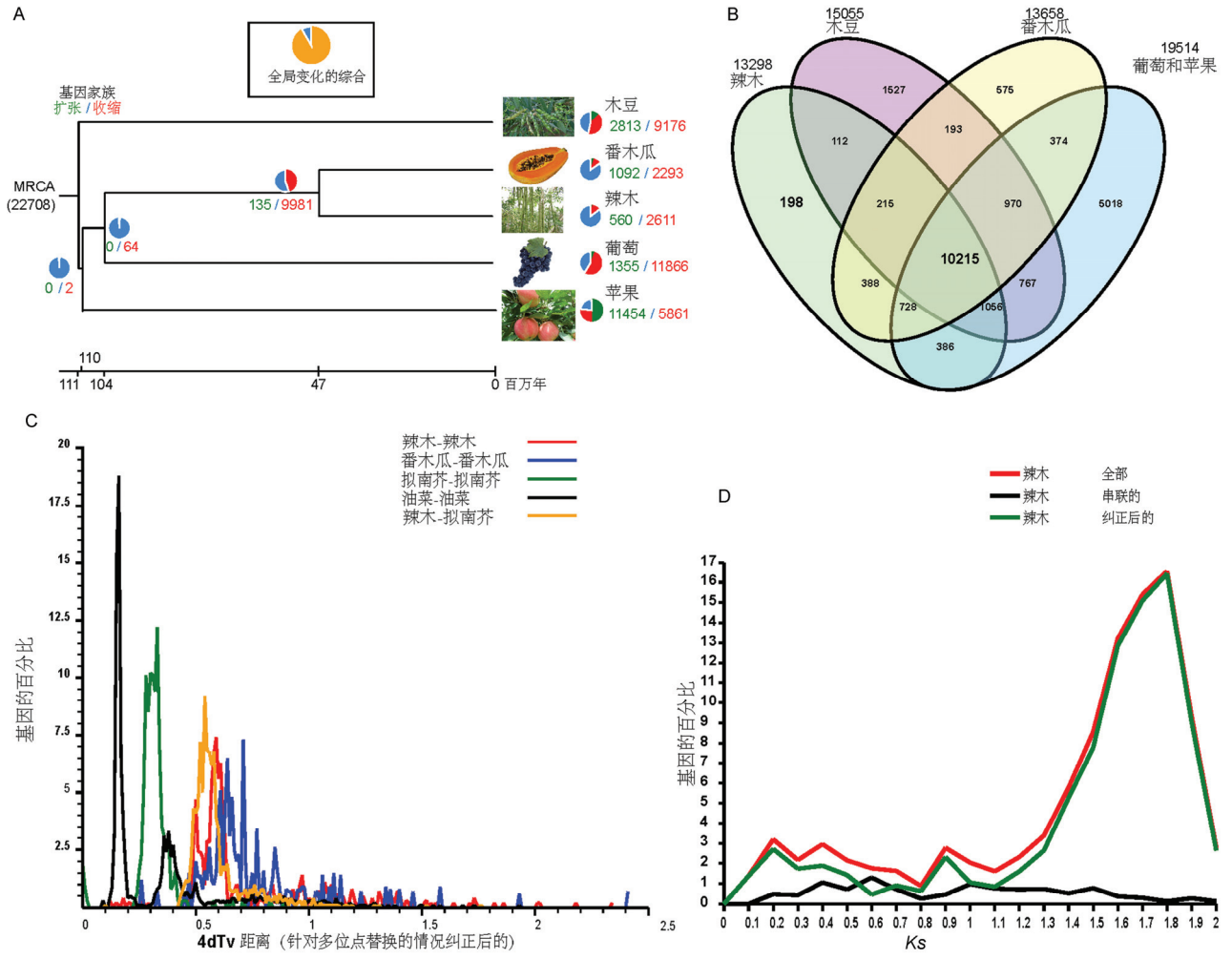

图 1　辣木的比较基因组学分析

A: 辣木、葡萄、木豆、番木瓜、苹果这 5 种木本植物的系统发生树，预测的物种分歧时间在每个节点标出，饼状图展示了基因家族的扩张收缩情况; B: 辣木、葡萄、木豆、番木瓜、苹果的基因家族分析; C: 辣木的 4dTv 距离分析，分别计算了各个物种内（辣木、拟南芥、油菜、番木瓜）以及辣木与拟南芥之间的 4dTv; D: 旁系同源基因之间的 *Ks* 分布

Ontologizer[41]对这 369 个基因做基因本体(gene ontology, GO)[42]富集分析, 结果 26 个富集的条目中有 25 个集中在细胞生物过程调节方面(网络版附图 7).

前人的研究显示, 在细胞内的 tRNA 水平可能与 tRNA 基因的拷贝数相关[43], 本文在辣木的基因组中注释出 1777 个 tRNA 基因, 之前发表的番木瓜基因组中有 388 个 tRNA 基因, 葡萄中有 600 个 tRNA 基因. tRNA 基因的数目异常或许与辣木非常强的蛋白合成能力相关.

2.3 系统发生以及全基因组复制分析

辣木在中国植物志中记载为罂粟目植物, 然而很多的分子生物学证据显示辣木是十字花目[44], 本文用 4 种双子叶植物葡萄[45]、木豆[46]、番木瓜[47]、苹果[48], 再加上辣木, 做系统发生分析. 图 1A 展示了系统发生树以及估计的各个分支的分歧时间, 其中与辣木最近的物种是番木瓜, 番木瓜属于十字花目, 这也支持了辣木属于十字花目这一说法. 本文又选用了十字花目的几个代表性物种, 拟南芥、油菜、番木瓜, 与辣木一起做系统发生分析(网络版附图 8). 再对这些物种做全基因组复制分析, 结果显示十字花目中全基因组复制事件发生了多次. 这些全基因组复制事件可以帮助理清许多问题, 之前已知番木瓜没有经历过 At-β全基因组复制事件[47], 现在数据显示辣木和番木瓜都没有经历过近期的全基因组复制事件(图 1C). 这 2 个物种最近的一次全基因组复制事件是 At-γ, 发生在这 2 个物种的祖先与拟南芥分歧之前. 对于辣木的旁系同源基因的 *Ks* 计算分析也验证了这一点, *Ks* 分布图只在 $Ks \approx 1.8$ 的地方有 1 个明显的峰, 说明最近的一次全基因组复制事件已经非常久远(图 1D).

2.4 辣木特有的基因家族和基因

基因家族通常是一些有相似功能基因的集合. 物种特异的基因家族对于物种的特性有重要的意义[49,50]. 本课题组对蛋白质编码基因进行了基因家族的聚类分析. 结果显示, 辣木、葡萄、木豆、番木瓜和苹果的基因家族数目相似, 共有 10215 个共享的基因(图 1B). 然而, 辣木的单拷贝基因家族和未聚类的基因数目明显少于其他物种, 网络版附表 10 和附图 9 展示了基因家族聚类的结果. 辣木共有 12298 个基因家族, 其中 198 个基因家族是辣木特异的, 包括 812 个基因, 网络版附图 10 展示了对这些基因的 GO 富集分析结果. 系统发生树上各个分支的基因家族的扩张收缩情况显示, 辣木的 560 个基因家族扩张了, 是这 5 个物种中数目最少的; 2611 个基因家族收缩了, 让辣木的基因组更加精小.

另外 4 个 *SKP1* 基因和 18 个含有 F-box 域的基因被鉴定为辣木特有的基因家族. SKP1 蛋白对细胞的周期控制非常重要[51], 它能协调周期特异蛋白的泛素化和降解从而维持正常的细胞周期, 主要是因为 F-box 的结构能够维持这些蛋白之间的联系[52]. 而且同时在辣木基因组中还有其他 7 个 *SKP1* 和 104 个含有 F-box 域的基因未被归为辣木特异的基因家族. 理论上可能的原因至少有 2 个: (ⅰ) 这些被认为是物种特异的基因家族的基因是新衍生出来的; (ⅱ) 这些基因对于辣木已经不再重要, 积累了很多突变. 本文也发现 *Betv1* 基因被归为单拷贝基因家族. Betv1 首次发现于杨树的花粉中[53], 后来陆续发现了一些新功能, 如具有甾类载体的功能[54]. *Betv1* 基因可能与辣木的快速生长特性有关, 因为 Betv1 可以与很多种的配体结合, 包括 ABA, 脂质和甾类化合物. 这些辣木特有的基因的功能可能与辣木的特性存在某种联系, 值得进一步研究.

2.5 辣木基因组中受正选择的基因

受到正选择的基因通常在这个物种的适应性方面都有重要贡献. 为了找出那些可能与辣木的特性相关的基因, 本课题组对辣木基因组进行了正选择分析. 使用 BLAST 和 KaKs_Calculator[55]把辣木与番木瓜、葡萄、苹果逐一比较分析, 分别发现了 566, 399, 112 个 $Ka/Ks>1$ 的基因($P<0.05$, 网络版附表 11~13). 进一步发现, 有 4 个基因在这 3 组分析中都出现了(图 2). 本实验还发现 2 个基因(lamu_GLEAN_10016878, lamu_GLEAN_10011614)受正选择区域的长度超过基因总长的 1/2 以上, 说明这 2 个基因在辣木中受到了强烈的选择.

lamu_GLEAN_10016878 基因的功能注释为 “Myb/SANT-like DNA binding domains”. SANT 域在染色质调节蛋白中很常见, 参与组蛋白的乙酰化、去乙酰化以及依赖 ATP 的染色质重塑过程[56]. 更重要的是, 许多含有 Myb/SANT 域的蛋白质具有与 DNA 结合的能力, 与基因表达调控相关. 这些对不同功能

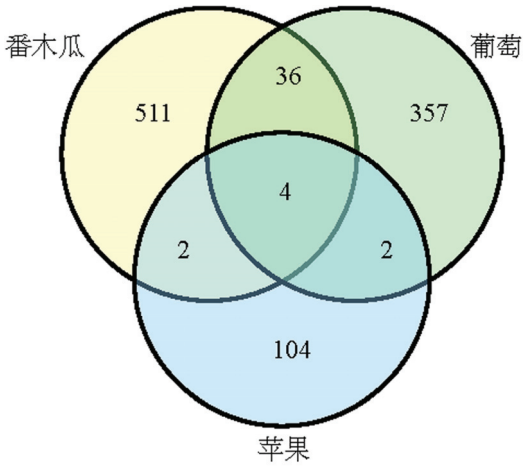

图 2　与番木瓜、葡萄、苹果相比，辣木受正选择的基因

的基因起调控作用的原件通常差异很大. 然而基因的两端和中间区域的一些片段比较保守，只有中间的一些区域不尽相同[57,58](网络版附图 1).

lamu_GLEAN_10011614 基因编码核糖体蛋白 S6e，在脊椎动物和真菌中非常保守[59]. 真核生物中核糖体蛋白在细胞质中合成后转运到核仁中，然后与新转录的 pre-rRNA 结合、相互作用，形成一个 90S 复合物，随后这个复合物被加工成一个 60S 和一个 40S 的核糖体亚结构，然后被排到细胞质中[60]. 核糖体蛋白辅助了 pre-18SRNA 和核糖体的成熟与行使功能[61]. 根据 Kundu-Michalik 等人[59]的研究，S6e 的氨基酸序列有 2 个 Nobis(nucleolar binding sequence)和多个 NLS(nuclear localization signal). 通过识别(G)RVRL 氨基酸序列鉴定了 Nobis 1 的 N 端，通过其长度再确定 C 端. 基于 Kundu-Michalik 的研究，本文还大致推测了 Nobis 2 的框架. 其他的元件如 NLS 和磷酸化位点也猜测性地在图 3 中给出了. 磷酸化状态的 S6 经常通过磷酸化级联反应来调控细胞中的过程[62]. 这个基因受到的强烈正选择，可能会导致辣木的蛋白质合成机制发生演化、重构，从而增强蛋白质的合成.

2.6　转录因子家族的分析

转录因子调控基因的表达，所以从微生物到高等动植物中转录因子都很重要，而且种类很多[63,64]. 分析之前已经鉴定出来的转录因子，可以得出大量的基因转录调控的信息. 下载 TAIR 数据库的转录因子家族数据(http://arabidopsis.org/browse/genefamily/index.jsp)[65]作为参考，使用 BLASTP(P-value< 1×10^{-20})共鉴定出 939 个转录因子(网络版附表 14). 之前的正选择分析中，辣木相对葡萄、番木瓜、苹果，共有 43 个转录因子受正选择. 编码 939 个转录因子的基因分属不同的家族，包括 *ABI3VP1*, *AP2-EREBP*, *Alfin-like*, *C2C2-Dof*, *C2C2-Gata*, *C2H2*, *C3H*, *CPP*, *E2F-DP*, *G2-like*, *GRAS*, *Homeobox*, *MADS*, *MYB*, *NAC*, *PHD*, *Trihelix*, *WRKY*, *bHLH*. 在这些种类中，WRKY 转录因子在抗逆方面具有非常重要的作用，如抗寒、抗旱、耐高温、耐盐、耐营养匮乏、适应多变的光照条件. 本文找到了 5 个受正选择的 *WRKY* 基因. C2H2 转录因子是一个超级家族，在防御机制和许多其他的生理过程中有重要作用，根据

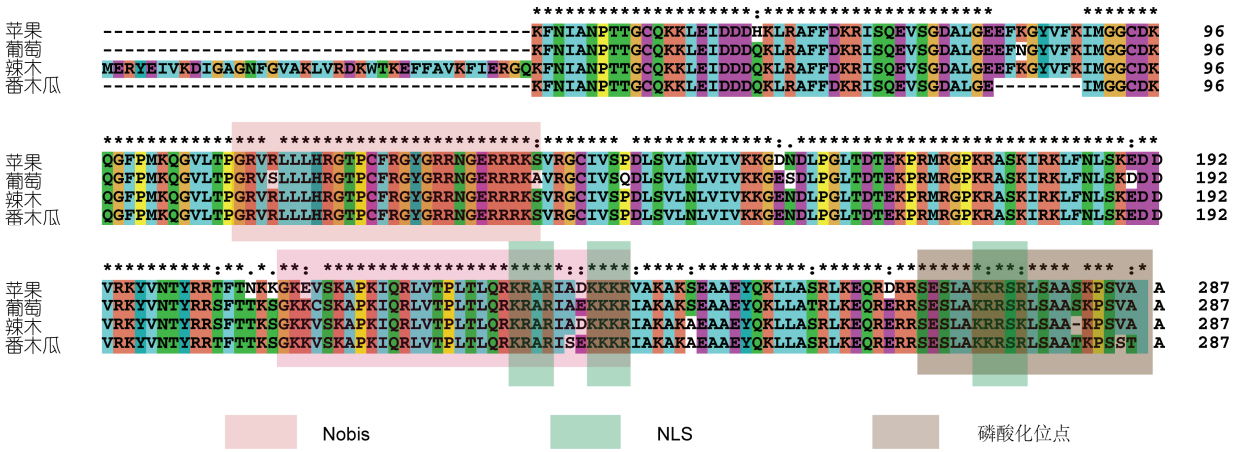

图 3　辣木基因 lamu_GLEAN_10011614 与其在葡萄、木豆、番木瓜、苹果中的直系同源基因比对，推测了基因上的 Nobis, NLS, 磷酸化位点等功能区域

正选择分析, 共有 4 个 *C2H2* 基因受到正选择. AP2-EREBP 转录因子, 在严寒、干旱等非生物的外界压力下与激素、糖、还原信号等密切相关, 正选择分析显示有 2 个拷贝的基因受正选择. 部分的 C3H 转录因子与抗旱特性相关, 本课题组发现了 2 个拷贝的基因受正选择(网络版附表 15). 这些受到正选择的与抗逆相关的基因, 或许与辣木的抗旱、耐高温的特性相关.

2.7 *HSP* 基因

高温环境对作物的产量是一个巨大的威胁, 然而随着全球变暖, 高温会越来越严重. 因此, 辣木的耐高温特性非常有价值. 热激蛋白 (heat shock protein, HSP)或者是环境压力诱导的蛋白参与了许多重要的抗逆反应, 包括抗旱、抗寒、耐盐、耐热、耐化学剂[66~68]. 使用从 HSPIR(http://pdslab.biochem.iisc.ernet.in/hspir/chaperone.php)[69]下载的拟南芥的 HSPs 序列作为参考, 一共鉴定了 133 个 HSP. 根据它们的功能特性和分子大小, HSPs 被大致划分为 6 个家族, HSP70 (在辣木基因组中有 25 个拷贝), HSP40 (J-proteins, 辣木基因组中有 52 个拷贝), HSP60 (chaperonins, 辣木基因组中有 17 个拷贝), HSP90 (在辣木基因组中有 3 个拷贝), HSP100(Clp proteins, 在辣木中有 9 个拷贝)以及小 HSP(辣木基因组中有 27 个拷贝)(网络版附表 16). 进一步检查辣木和番木瓜的 HSP 基因的 *Ka*/*Ks* 比率, 发现 HSP 的 *Ka*/*Ks* 值比背景基因高(网络版附表 17). 与番木瓜、葡萄、苹果相比, 辣木中受正选择的 *HSP* 基因在网络版附表 18 中列出, 这些基因可能与辣木的耐高温特性相关.

2.8 油菜素内酯信号转导途径

油菜素内酯(brassinosteroid, BR)是一种能够调节细胞增长和细胞分裂的植物激素, BR 也与植物的抗寒、抗旱、耐热相关. 本文研究了辣木的 BR 的信号转导途径, 发现 *BAK1*(BRI1 associated receptor kinase 1)基因在辣木中扩张至 29 个拷贝, 而拟南芥只有 5 个拷贝(网络版附图 11). 其中有一个拷贝的 BAK1 受正选择(葡萄为参考). BAK1 蛋白在 BR 的信号转导过程中起主导作用, *BAK1* 基因的敲除会导致植株的瘦弱矮小[70].

2.9 γ-氨基丁酸和谷甾醇的合成途径

γ-氨基丁酸(γ-aminobutyric acid, GABA)和谷甾醇的合成通路是重要的植物激素合成通路. 本实验研究了辣木的通路, 并且注释了通路中的所有基因. GABA 是一种非氨基酸的四碳化合物, 广泛存在于高等动植物以及细菌、真菌中, 在植物中, GABA 的浓度受缺氧、温度骤变、物理损伤以及植物激素等的激发[71,72]. 发现, GABA 是由谷氨酸脱羧酶(GAD, lamu_GLEAN_10006873, lamu_GLEAN_10006874, lamu_GLEAN_10004957, lamu_GLEAN_10007711, lamu_GLEAN_10007712, lamu_GLEAN_10007713)催化 L-glutamate 发生不可逆的 脱羧反应, 然后再加工产生[73,74]. 脱羧反应之后, 被 GABA 转氨酶(GAGA-田 T, lamu_GLEAN_10002543)和琥珀酸半醛脱氢酶(SSADH, lamu_GLEAN_10008793, lamu_GLEAN_10008794)转化为琥珀酸, 一种重要的三羧酸循环的产物[74]. 已知谷氨酸的代谢中被钙离子调节的酶只有谷氨酸脱氢酶(GDH, lamu_GLEAN_10005665), 是一种线粒体酶(网络版附图 12).

为了了解辣木的固醇合成相关基因, 本文展示了在大部分高等植物中都存在的生物合成途径[75]. 谷甾醇和菜油甾醇对细胞膜的生成很重要[76]. 发现了 2 个 *STM2* 基因, *STM2* 通过调节菜油甾醇和谷甾醇的比率来平衡生长和细胞膜的完整性(网络版附图 13).

3 讨论

至今为止, 还没有辣木科的植物基因组相关文章的发表, 使得辣木的基因组数据不仅对辣木的进一步研究很重要, 也对辣木科其他植物的研究意义非凡. 由于相关研究的匮乏, 本文的研究在许多方面还不能得出确定性结论, 只是对于未来研究辣木有价值的特性建议性地指出方向. 基因家族分析显示, 辣木拥有非常少的单拷贝基因家族和辣木特异基因家族, 以及注释出来的基因数目比一般的高等植物基因数目少很多, 许多基因家族收缩, 还有辣木的基因组大小非常小, 这暗示了辣木拥有非常精小的基因组, 也可能是辣木生长速度比较快的原因. 本文专注于辣木基因组的特征以及辣木的特征形状, 找出了一些可能与高蛋白、耐热、耐旱、快速生长有关的基因. 本文提供的基因列表不仅对将来辣木的功能研究, 也对未来辣木的育种、改良非常重要, 这或许能帮助辣木尽快被世界上粮食短缺以及干旱的地方作为多年生作物种植, 让辣木物尽其用.

参考文献

1 Desaintsauveur A. *Moringa*, a multipurpose tree for the Sahel. Physiologie des arbres et arbustes en zones... Proceedings. John Libbey Eurotext, 1993. 441–446
2 Olson M E, Fahey J W. *Moringa oleifera*: a multipurpose tree for the dry tropics. Revista Mexicana De Biodiversidad, 2011, 82: 1071–1082
3 Horwath M, Benin V. Theoretical investigation of a reported antibiotic from the “Miracle Tree” *Moringa oleifera*. Comput Theor Chem, 2011, 965: 196–201
4 Makkar H P S, Becker K. Nutrients and antiquality factors in different morphological parts of the *Moringa oleifera* tree. J Agr Sci, 1997, 128: 311–322
5 Palada M C. *Moringa* (*Moringa oleifera* Lam): a versatile tree crop with horticultural potential in the subtropical United States. Hortscience, 1996, 31: 794–797
6 Oliveira J T A, Silveira S B, Vasconcelos I M, et al. Compositional and nutritional attributes of seeds from the multiple purpose tree *Moringa oleifera* Lam. J Sci Food Agr, 1999, 79: 815–820
7 Amaglo N K, Bennett R N, Lo Curto R B, et al. Profiling selected phytochemicals and nutrients in different tissues of the multipurpose tree *Moringa oleifera* Lam, grown in Ghana. Food Chem, 2010, 122: 1047–1054
8 Luo R, Liu B, Xie Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. Gigascience, 2012, 1: 18
9 Kajitani R, Toshimoto K, Noguchi H, et al. Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res, 2014, 24: 1384–1395
10 Boetzer M, Henkel C V, Jansen H J, et al. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics, 2011, 27: 578–579
11 Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res, 1999, 27: 573–580
12 Jurka J, Kapitonov V V, Pavlicek A, et al. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res, 2005, 110: 462–467
13 Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res, 2007, 35: W265–W268
14 Price A L, Jones N C, Pevzner P A. *De novo* identification of repeat families in large genomes. Bioinformatics, 2005, 21: i351–i358
15 *Arabidopsis* Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature, 2000, 408: 796–815
16 Schmutz J, Cannon S B, Schlueter J, et al. Genome sequence of the palaeopolyploid soybean. Nature, 2010, 463: 178–183
17 Goff S A, Ricke D, Lan T H, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science, 2002, 296: 92–100
18 Tuskan G A, Difazio S, Jansson S, et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science, 2006, 313: 1596–1604
19 Paterson A H, Bowers J E, Bruggmann R, et al. The *Sorghum bicolor* genome and the diversification of grasses. Nature, 2009, 457: 551–556
20 Banks J A, Nishiyama T, Hasebe M, et al. The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. Science, 2011, 332: 960–963
21 Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res, 2004, 14: 988–995
22 Stanke M, Steinkamp R, Waack S, et al. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res, 2004, 32: W309–W312
23 Majoros W H, Pertea M, Salzberg S L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics, 2004, 20: 2878–2879
24 Boeckmann B, Bairoch A, Apweiler R, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res, 2003, 31: 365–370
25 Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 2000, 28: 27–30
26 Quevillon E, Silventoinen V, Pillai S, et al. InterProScan: protein domains identifier. Nucleic Acids Res, 2005, 33: W116–W120
27 Lowe T M, Eddy S R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res, 1997, 25: 955–964
28 Burge S W, Daub J, Eberhardt R, et al. Rfam 11.0: 10 years of RNA families. Nucleic Acids Res, 2013, 41: D226–D232
29 Nawrocki E P, Eddy S R. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics, 2013, 29: 2933–2935
30 Griffiths-Jones S, Saini H K, van Dongen S, et al. miRBase: tools for microRNA genomics. Nucleic Acids Res, 2008, 36: D154–D158
31 Dai X, Zhao P X. psRNATarget: a plant small RNA target analysis server. Nucleic Acids Res, 2011, 39: W155–W159

32 Li L, Stoeckert C J Jr, Roos D S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res, 2003, 13: 2178–2189
33 Edgar R C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res, 2004, 32: 1792–1797
34 Yang Z H. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol, 2007, 24: 1586–1591
35 De Bie T, Cristianini N, Demuth J P, et al. CAFE: a computational tool for the study of gene family evolution. Bioinformatics, 2006, 22: 1269–1271
36 Zhang Z, Li J, Zhao X Q, et al. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics, 2006, 4: 259–263
37 Thompson J D, Gibson T J, Plewniak F, et al. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res, 1997, 25: 4876–4882
38 Zhang Q, Chen W, Sun L, et al. The genome of *Prunus mume*. Nat Commun, 2012, 3: 1318
39 Kovach A, Wegrzyn J L, Parra G, et al. The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences. BMC Genomics, 2010, 11: 420
40 Camon E, Barrell D, Brooksbank C, et al. The Gene Ontology Annotation (GOA) project—application of GO in SWISS-PROT, TrEMBL and InterPro. Comp Funct Genomics, 2003, 4: 71–74
41 Bauer S, Grossmann S, Vingron M, et al. Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics, 2008, 24: 1650–1651
42 Ashburner M, Ball C A, Blake J A, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000, 25: 25–29
43 Percudani R, Pavesi A, Ottonello S. Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. J Mol Biol, 1997, 268: 322–330
44 Beilstein M A, Nagalingum N S, Clements M D, et al. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. Proc Natl Acad Sci USA, 2010, 107: 18724–18728
45 Jaillon O, Aury J M, Noel B, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature, 2007, 449: 463–467
46 Varshney R K, Chen W, Li Y, et al. Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nat Biotechnol, 2012, 30: 83–89
47 Ming R, Hou S, Feng Y, et al. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature, 2008, 452: 991–996
48 Velasco R, Zharkikh A, Affourtit J, et al. The genome of the domesticated apple (*Malus×domestica* Borkh.). Nat Genet, 2010, 42: 833–839
49 Christophides G K, Zdobnov E, Barillas-Mury C, et al. Immunity-related genes and gene families in *Anopheles gambiae*. Science, 2002, 298: 159–165
50 Shuai B, Reynaga-Pena C G, Springer P S. The lateral organ boundaries gene defines a novel, plant-specific gene family. Plant Physiol, 2002, 129: 747–761
51 Connelly C, Hieter P. Budding yeast *SKP1* encodes an evolutionarily conserved kinetochore protein required for cell cycle progression. Cell, 1996, 86: 275–285
52 Bai C, Sen P, Hofmann K, et al. *SKP1* connects cell cycle regulators to the ubiquitin proteolysis machinery through a novel motif, the F-box. Cell, 1996, 86: 263–274
53 Breiteneder H, Pettenburger K, Bito A, et al. The gene coding for the major birch pollen allergen Betvl, is highly homologous to a pea disease resistance response gene. Embo Journal, 1989, 8: 1935–1938
54 Markovic-Housley Z, Degano M, Lamba D, et al. Crystal structure of a hypoallergenic isoform of the major birch pollen allergen Betv1 and its likely biological function as a plant steroid carrier. J Mol Biol, 2003, 325: 123–133
55 Wang D, Zhang Y, Zhang Z, et al. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics, 2010, 8: 77–80
56 Boyer L A, Latek R R, Peterson C L. The SANT domain: a unique histone-tail-binding module? Nat Rev Mol Cell Biol, 2004, 5: 158–163
57 Barg R, Sobolev I, Eilon T, et al. The tomato early fruit specific gene *Lefsm1* defines a novel class of plant-specific SANT/MYB domain proteins. Planta, 2005, 221: 197–211
58 Mohrmann L, Kal A J, Verrijzer C P. Characterization of the extended Myb-like DNA-binding domain of trithorax group protein Zeste. J Biol Chem, 2002, 277: 47385–47392
59 Kundu-Michalik S, Bisotti M A, Lipsius E, et al. Nucleolar binding sequences of the ribosomal protein S6e family reside in evolutionary

highly conserved peptide clusters. Mol Biol Evol, 2008, 25: 580–590
60 Fromont-Racine M, Senger B, Saveanu C, et al. Ribosome assembly in eukaryotes. Gene, 2003, 313: 17–42
61 Ferreira-Cerca S, Poll G, Gleizes P E, et al. Roles of eukaryotic ribosomal proteins in maturation and transport of pre-18S rRNA and ribosome function. Mol Cell, 2005, 20: 263–275
62 Ruvinsky I, Meyuhas O. Ribosomal protein S6 phosphorylation: from protein synthesis to cell size. Trends Biochem Sci, 2006, 31: 342–348
63 Matys V, Fricke E, Geffers R, et al. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res, 2003, 31: 374–378
64 Riechmann J L, Heard J, Martin G, et al. *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. Science, 2000, 290: 2105–2110
65 Poole R L. The TAIR database. Methods Mol Biol, 2007, 406: 179–212
66 Morimoto R I. Cells in stress: transcriptional activation of heat shock genes. Science, 1993, 259: 1409–1410
67 Lindquist S, Craig E A. The heat-shock proteins. Annu Rev Genet, 1988, 22: 631–677
68 Lindquist S. The heat-shock response. Annu Rev Biochem, 1986, 55: 1151–1191
69 Breiteneder H, Pettenburger K, Bito A, et al. HSPIR: a manually annotated heat shock protein information resource. Bioinformatics, 2012, 28: 2853–2855
70 Nam K H, Li J. BRI1/BAK1, a receptor kinase pair mediating brassinosteroid signaling. Cell, 2002, 110: 203–212
71 Bown A W, Shelp B J. The metabolism and functions of γ-aminobutyric acid. Plant Physiol, 1997, 115: 1–5
72 Narayan V S, Nair P M. Metabolism, enzymology and possible roles of 4-aminobutyrate in higher plants. Phytochemistry, 1990, 29: 367–375
73 Chung I, Bown A W, Shelp B J. The production and efflux of 4-aminobutyrate in isolated mesophyll cells. Plant Physiol, 1992, 99: 659–664
74 Tuin LG, Shelp B J. *In situ* [^{14}C] glutamate metabolism by developing soybean cotyledons 1. metabolic routes. J Plant Physiol, 1994, 143: 1–7
75 Benveniste P. Biosynthesis and accumulation of sterols. Annu Rev Plant Biol, 2004, 55: 429–457
76 Schaeffer A, Bronner R, Benveniste P, et al. The ratio of campesterol to sitosterol that modulates growth in *Arabidopsis* is controlled by *STEROL METHYLTRANSFERASE 2;1*. Plant J, 2001, 25: 605–615