

· 研究论文 ·

大豆蛋白编码基因起源与进化

唐康, 杨若林*

西北农林科技大学生命科学院, 杨凌 712100

摘要 物种基因组是一个高度动态的进化过程, 其中相对较近起源的种系和物种特异性基因会持续整合到包含古老基因的原始基因网络中。新基因在塑造基因组结构中发挥重要作用, 能提高物种适应性。基因复制和新基因的从头起源是产生新基因及改变基因家族大小的2种方式。目前, 大豆(*Glycine max*)基因起源时间与进化模式的相互联系很大程度上还未被探索。该研究选择19种具有代表性的被子植物基因组, 分析基因含量动态性与大豆基因起源之间的潜在联系。采用基因出现法, 研究显示约58.7%的大豆基因能追溯到大约1.5亿年前, 同时有21.7%的基因为最近起源的orphan基因。研究结果表明, 与新基因相比, 古老基因受到更强的负选择压并且更加保守。此外, 古老基因的表达水平更高且更可能发生选择性剪切。此外, 具有不同拷贝数的基因在上述特征中也具有明显差异。研究结果有助于认识不同年龄基因的进化模式。

关键词 被子植物, 基因复制, 基因家族, 基因起源, 大豆

唐康, 杨若林 (2019). 大豆蛋白编码基因起源与进化. 植物学报 54, 316–327.

基因组变异是生物表型多样性的主要来源, 而新基因的出现是基因组进化及物种间遗传差异的重要因素之一(Long et al., 2003; Kaessmann, 2010; Chen et al., 2013)。新基因起源是一个动态变化过程, 物种中的每个基因都“诞生”于某个特定的进化节点, 即大部分基因只出现在特定的种系或物种中, 这些新基因最初以很快的速度进化, 直到被整合到原始基因网络中, 表明这些基因可能与物种形成或适应性有关(Lynch and Conery, 2000; Long et al., 2003; Kaessmann, 2010; Chen et al., 2013)。

真核生物中, 基因复制(gene duplication, GD)是新基因的主要来源, 可作为研究新功能进化的原材料, 也是增加基因家族大小的重要因素(Ohno, 1970; Zhang, 2003; 孙红正和葛颂, 2010)。基因复制后, 由于拷贝数的增加会使该基因先经历短暂的选择压放松期, 此时拷贝序列将累积突变, 如果发生了有害突变, 则会导致基因功能减弱, 最终被清除出基因组; 也可能由于遗传漂变的原因, 使得假基因化的某一拷贝得以在群体中保留, 而突变与假基因化是导致基因家族减小的重要原因(Albalat and Cañestro, 2016)。基因某一拷贝上偶尔也会累积有利突变, 在达尔文正

选择作用下驱动该拷贝进化出新的生物学功能, 即所谓的新功能化(neo-functionalization), 而另一拷贝则执行原有的功能(Zhang, 2003)。此外, 对于多功能的基因还有可能出现复制产生的2个拷贝以功能互补的方式各自承担起父本基因的部分角色, 即亚功能化(sub-functionalization) (Zhang, 2003)。基因复制的机制主要包括全基因组复制(whole genome duplication, WGD)、片段重复(segmental duplication)、串联重复(tandem duplication)和转座诱导重复(transposon-induced duplication) (Freeling, 2009; Panchy et al., 2016), 每种机制对重复基因的功能、进化命运及基因组结构产生不同的影响。此外, 基因的从头起源能产生新的基因家族, 其中非编码区的从头演变是新基因出现的重要模式(Tautz and Domazet-Lošo, 2011)。

随着高通量技术的快速发展, 有来自不同分类群的100多种植物相继完成全基因组测序(Michael and Jackson, 2013; Michael and VanBuren, 2015), 海量的基因组序列和功能数据为我们在全基因组水平研究基因进化模式带来了前所未有的契机。基因出现法被用于研究单个物种或特定生物类群的基因基于

收稿日期: 2018-08-14; 接受日期: 2018-12-10

基金项目: 陕西省“百人计划”(No.SXBR8025)

* 通讯作者。E-mail: desert.ruolin@gmail.com

系统发育水平的进化起源(Domazet-Lošo et al., 2007; Cai et al., 2009; Quint et al., 2012; Guo, 2013)。该方法根据被调查物种基因在一组亲缘关系由近及远的代表物种基因组中是否有序列相似的同源基因, 将基因映射到不同的系统发育层级, 称为phylostratigraphy或phylostrata (PS), 代表相关基因大致的起源时间区间。目前, 该方法已广泛应用于追溯基因的进化历程, 及研究基因的起源时间与一些重要生物学过程(如胚胎发育)的关系(Quint et al., 2012)。

豆科是被子植物中物种数目最多的家族之一, 有20 000多个种(Doyle and Luckow, 2003)。大豆(*Glycine max*)作为该科非常重要的经济作物, 其基因组已经完成测序(Schmutz et al., 2010)。除大豆外, 本研究还选择了18种具有代表性的被子植物, 包括1种基部被子植物、5种单子叶植物和12种双子叶植物, 通过将大豆蛋白编码基因与这些基因组进行序列比对, 对大豆基因的起源时间进行推测, 并在此基础上, 对不同年龄的基因家族(和基因)进行比较分析: (1) 基因家族大小分布模式以及不同拷贝数基因家族的相对比例; (2) 基因序列水平的进化速率和选择压大小; (3) 基因在转录及转录后水平的特征, 包括在不同组织不同发育阶段的表达水平、表达模式以及选择性剪切等。综上, 我们通过对19种被子植物进行比较基因组学研究, 以期揭示大豆基因在不同起源时期不同复制状态下的进化动态。

1 材料与方法

1.1 材料

1.1.1 基因组序列、蛋白组序列及注释数据

除大豆(*Glycine max* (Linn.) Merr.)外, 我们还选择了18种具有代表性且已完成测序的被子植物(图1A), 其基因组序列文件、蛋白组序列文件、GTF文件及CDS序列文件下载自Ensembl Plants v.33和Phytozome v.11数据库。这19种被子植物包括1种基部被子植物无油樟(*Amborella trichopoda*)、5种单子叶植物和13种双子叶植物。其中单子叶植物包含1种凤梨科植物(菠萝(*Ananas comosus*))、4种禾本科植物(水稻(*Oryza sativa*)、二穗短柄草(*Brachypodium distachyon*)、高粱(*Sorghum bicolor*)及玉米(*Zea mays*)); 双子叶植物中有2种茄科植物(马铃薯

(*Solanum tuberosum*)和番茄(*S. lycopersicum*))、1种葡萄科植物(葡萄(*Vitis vinifera*))、1种杨柳科植物(毛果杨(*Populus trichocarpa*))、1种锦葵科植物(雷蒙德氏棉(*Gossypium raimondii*))、1种番木瓜科植物(番木瓜(*Carica papaya*))、2种十字花科植物(拟南芥(*Arabidopsis thaliana*)和琴叶拟南芥(*A. lyrata*))、1种葫芦科植物(黄瓜(*Cucumis sativus*))、1种蔷薇科植物(碧桃(*Prunus persica*))及3种豆科植物(蒺藜苜蓿(*Medicago truncatula*)、大豆和菜豆(*Phaseolus vulgaris*))。

1.1.2 大豆转录组数据

大豆28个不同发育阶段组织样本的RNA-seq数据从NCBI SRA数据库筛选得到, 其项目编号为SRP038111。28个样本包括: 萌发阶段的子叶、根、茎和叶芽; 三叶期的子叶、茎、叶芽和叶; 发芽分化阶段的叶芽、叶、花和芽分生组织; 开花期收集了3种花型, 花芽、花(2个样本)和开花后5天的花; 种子萌发后2周、3周和4周的荚粒及3周、4周和5周豆荚; 萌发后大约3周、5周、6周、8周和10周; 衰老期的叶。

1.2 方法

1.2.1 直系同源基因家族的鉴定

为了获得高质量的蛋白质序列数据以鉴定基因的同源关系, 我们对上述19种被子植物的蛋白质组数据按以下条件进行过滤: (1) 去除长度小于50个氨基酸残基的蛋白质; (2) 对于由可变剪切产生的多个转录本所翻译的蛋白质, 只保留每个基因最长转录本对应的蛋白质。过滤之后, 所有19个物种的共641 473条蛋白质序列作为输入数据提交至OrthoMCL v2.0.9 (Li et al., 2003)进行蛋白聚类。OrthoMCL运行中的2个关键步骤是: (1) All-against-all BLASTP, 即用blastp v2.6.0将每个蛋白与所有其它蛋白进行比对($E\text{-value} < 1 \times 10^{-6}$), 产生原始的blast输出; (2) 使用马尔科夫聚类算法(Markov Cluster algorithm, MCL)对解析的blast结果构建马尔科夫矩阵, 然后产生最终的基因家族(Enright et al., 2002)。MCL聚类时的重要参数(膨胀系数)设为1.5。本研究中我们共鉴定到34 010个直系同源基因家族。

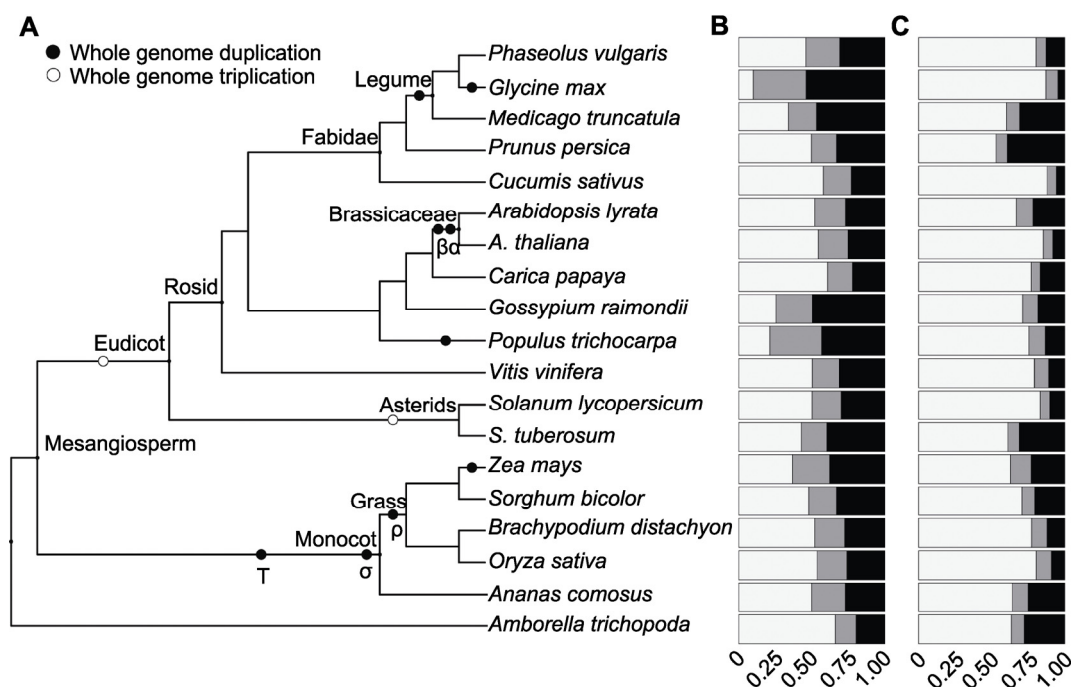


图1 19种被子植物基因家族大小分布

(A) 系统发育树代表19种被子植物的进化关系; (B) 直系同源基因家族大小; (C) orphan基因家族大小。白、灰、黑分别代表单拷贝、两拷贝和多拷贝基因所占比例。

Figure 1 Gene family size distribution of 19 angiosperm species

(A) Phylogenetic tree showing the relationships between the 19 angiosperm species used in this study; (B) Homologous gene family sizes; (C) Gene family sizes of orphan genes. The colors indicate the proportions of genes, white for singletons, grey for two-genes and black for multigenes.

1.2.2 大豆基因的起源分析

我们参考Domazet-Lošo等(2007)的方法对大豆蛋白编码基因进行起源时间预测。具体而言, 根据OrthoMCL聚类结果, 对于每个大豆蛋白基因, 我们按照其出现在上述18种植物基因组中的分布模式, 将其起源时间范围指定到以下7个系统发育层级之一。这7个基因分类为: (1) Soybean (PS7), 含有大豆特有的基因; (2) Phaseoleae (PS6), 起源于大豆与菜豆的最近共同祖先(most recent common ancestor, MRCA); (3) Legume (PS5), 起源于豆科植物大豆、菜豆与苜蓿的共同祖先; (4) Rosid (PS4), 在豆科植物与葡萄的共同祖先起源; (5) Eudicot (PS3), 在蔷薇目植物与茄科植物的最近共同祖先中起源; (6) Mesangiosperm (PS2), 在双子叶植物与单子叶植物的MRCA中起源; (7) Angiosperm (PS1), 基因出现在核心被子植物与无油樟的MRCA中。

1.2.3 选择压分析

为了估算出大豆中基因的进化速率, 我们从OrthoMCL的聚类结果中提取出大豆与菜豆的直系同源基因对; 用ClustalW v2.1软件(Larkin et al., 2007)进行直系同源蛋白比对; 再利用PAL2NAL v14.0 (Suyama et al., 2006)以比对好的蛋白序列为指导, 将相应的核苷酸编码序列进行比对; 最后将上述序列文件提交至PAML v4.9b (Yang, 2007)程序以计算每对直系同源基因间的非同义替换速率(dN)、同义替换速率(dS)及选择压值(dN/dS)。

1.2.4 基因表达分析

RNA-seq产生的原始fastq格式文件从NCBI SRA数据库中筛选获得, 然后采用Trimmomatic v0.36软件(Bolger et al., 2014)去除原始读段(raw reads)两端低质量序列, 并保证读段最小长度为50 bp。我们进一步

使用Hisat v2.1.0软件(Kim et al., 2015)将clean reads比对到大豆(Williams 82)参考基因组(v1.0.33)上, 最大内含子长度设为参考基因模型(gene model)的最大内含子长(由R包GenomicFeatures计算得出), 其它参数均使用软件默认值。本研究只保留能比对到参考基因组的最好结果。最后使用StringTie v1.3.3b软件(Pertea et al., 2015)对每个样本进行转录组构建, 并计算出标准化后的基因及转录水平。

本研究中, 基因在28个被调查的不同发育阶段组织样本的至少1个样本中满足FPKM值大于1, 即认为该基因已表达, 并使用上述已表达基因在28个样本中的FPKM值的中值来估计基因总体的表达水平。

1.2.5 基因表达的组织和发育阶段特异性指数(r)

指数 r 定义为:

$$r = \frac{\sum_{i=1}^N \left(1 - \frac{\log(x_i + 1)}{\log(x_{\max} + 1)} \right)}{N - 1}$$

其中, N 代表参与计算的不同发育阶段的所有组织样本的数量, x_i 代表基因在每个样本中的表达量, x_{\max} 代表基因在所有样本中的最大表达量。 r 值在0–1范围内, 一般将 $r \geq 0.85$ 的基因作为组织和发育阶段特异性表达基因, 而 $r < 0.15$ 则视为持家基因(Yanai et al., 2005)。

1.2.6 基因选择性剪切事件的鉴定

利用Cufflinks (Trapnell et al., 2010)程序中的Cuffcompare将组装好的转录本与参考基因模型(gene model)做对比, 并进行命名。其中class codes为“c”、“j”、“=”、“e”或“o”的转录本被提取出来(Merkin et al., 2012)。利用ASTALAVISTA v4.0软件(<http://genome.crg.es/astalavista/>, Foissac and Sammeth, 2007)对组装出来的转录本进行选择剪切事件鉴定。

1.2.7 基因GO注释和功能分析

使用在线工具GOSlim (www.geneontology.org)对大豆不同起源基因进行功能富集分析。

2 结果与分析

2.1 基因家族大小分布

利用OrthoMCL软件(Li et al., 2003), 我们共获得19个物种的34 010个直系同源基因家族, 物种间基因家族总数变异较大, 从11 407个(无油樟)至16 957个(拟南芥)不等。根据基因家族大小, 我们将基因归为单拷贝基因(singleton)、两拷贝基因(two-genes)或多拷贝基因(multigenes, 家族大小 ≥ 3), 发现各物种基因组中不同拷贝数的基因家族之间的相对比例各不相同。如表1和图1B所示, 与其它物种相比, 大豆中两拷贝(36.2%)和多拷贝(53.9%)基因家族占很大比例, 这可能与大豆基因组发生过2次最近的全基因组复制事件(分别为大约5 900万和1 300万年前)有关(Schmutz et al., 2010)。

我们初步统计了上述19种植物中orphan基因占总蛋白编码基因数目的相对比例, 结果显示各物种特有的基因在总的蛋白编码基因中的比例差异显著, 菜豆中仅为9.2%, 而无油樟中高达45.5%, 其中大豆基因组中orphan基因所占比例为21.1%。单拷贝orphan基因占总orphan基因的比例在物种间变异不大(表2; 图1C), 大豆中该比例为87.1%, 表明orphan基因主要以单拷贝形式存在。

2.2 大豆基因起源

在54 174个大豆基因中, 97.3% (52 733个)被定位到7个主要关注的系统发育层级(表3; 图2A)。本研究显示, 超过58.7%的基因定位到PS1, 表明一半以上的大豆蛋白编码基因至少起源于无油樟与核心被子植物分歧之前, 同时有超过21.7%的基因为大豆特异性基因(PS7)(表3; 图2B)。有趣的是, 起源晚的基因更多以单拷贝形式存在, 单拷贝基因占PS1到PS7中基因总数的比例逐渐从6.4%上升至87.1% (图2C)。

对于定位到7个不同系统发育层级的52 733个基因, 73.4% (38 730个)的基因具有GO注释(图2D)。有趣的是, 基因起源时间与包含GO注释信息的基因比例之间存在很强的负相关关系, 起源于PS1–PS7的基因中, 分别有81.4%–56.9%不等的基因被注释。

表1 19种被子植物中直系同源基因家族(及基因)数目

Table 1 Number of homologous gene families (and genes) in 19 angiosperm species

Species	Singletons	Two-gene families	Multigene families	Total gene families	Maximum gene family size
<i>Amborella trichopoda</i>	9823	1061(2122)	523(2935)	11407	207
<i>Ananas comosus</i>	9059	2087(4174)	1007(4916)	12153	124
<i>Oryza sativa</i>	11966	2269(4538)	1167(5805)	15402	64
<i>Brachypodium distachyon</i>	11455	2264(4528)	1209(6066)	14928	50
<i>Sorghum bicolor</i>	12663	2529(5058)	1399(8749)	16591	416
<i>Zea mays</i>	10277	3568(7136)	1964(10539)	15809	297
<i>Solanum tuberosum</i>	11592	2390(4780)	1399(10741)	15381	1051
<i>S. lycopersicum</i>	12210	2448(4896)	1371(7277)	16029	72
<i>Vitis vinifera</i>	10408	1931(3862)	1104(6483)	13443	100
<i>Populus trichocarpa</i>	6550	5476(10952)	2337(13368)	14363	108
<i>Gossypium raimondii</i>	7582	3700(7400)	2960(14806)	14242	90
<i>Carica papaya</i>	10776	1505(3010)	667(3948)	12948	194
<i>Arabidopsis thaliana</i>	13278	2485(4970)	1194(6144)	16957	125
<i>A. lyrata</i>	12767	2605(5210)	1327(6596)	16699	67
<i>Cucumis sativus</i>	10152	1691(3382)	795(4038)	12638	38
<i>Prunus persica</i>	10822	1876(3752)	1106(7192)	13804	217
<i>Medicago truncatula</i>	9936	2812(5624)	1948(13673)	14696	308
<i>Glycine max</i>	4241	7735(15470)	4206(23027)	16182	153
<i>Phaseolus vulgaris</i>	11324	2873(5746)	1430(7626)	15569	132

表2 19种被子植物中的orphan基因家族(及基因)数目

Table 2 Number of orphan gene families (and genes) in 19 angiosperm species

Species	Singletons	Two-gene families	Multigene families	Species-specific genes	Maximum gene family size
<i>Amborella trichopoda</i>	7892	547(1094)	502(3447)	12433	105
<i>Ananas comosus</i>	5685	483(966)	297(2224)	8875	94
<i>Oryza sativa</i>	10774	686(1372)	292(1224)	13370	29
<i>Brachypodium distachyon</i>	3485	235(470)	125(548)	4503	15
<i>Sorghum bicolor</i>	5682	350(700)	254(1644)	8026	103
<i>Zea mays</i>	7253	813(1626)	552(2643)	11522	65
<i>Solanum tuberosum</i>	7278	471(942)	376(3688)	11908	163
<i>S. lycopersicum</i>	7836	308(616)	177(950)	9402	51
<i>Vitis vinifera</i>	7238	445(890)	229(1006)	9134	44
<i>Populus trichocarpa</i>	7923	593(1186)	281(1398)	10507	31
<i>Gossypium raimondii</i>	5495	408(816)	293(1406)	7717	26
<i>Carica papaya</i>	7680	307(614)	224(1653)	9947	88
<i>Arabidopsis thaliana</i>	2751	105(210)	57(261)	3222	21
<i>A. lyrata</i>	5413	461(922)	366(1759)	8094	83
<i>Cucumis sativus</i>	3458	125(250)	54(223)	3931	13
<i>Prunus persica</i>	3347	242(484)	195(2483)	6314	838
<i>Medicago truncatula</i>	12763	962(1924)	820(6524)	21211	145
<i>Glycine max</i>	9961	476(952)	118(523)	11436	23
<i>Phaseolus vulgaris</i>	2013	85(170)	58(318)	2501	19

表3 定位到每个系统发育层级的大豆基因家族(和基因)数目

Table 3 Number of soybean gene families (and genes) assigned to each phylostratum

Phylostratum internode	Genes (%)	Singletons	Two-genes	Multigenes
Angiosperm (PS1)	30932(58.7%)	1982	5150(10300)	3400(18650)
Mesangiosperm (PS2)	4057(7.7%)	508	708(1416)	359(2133)
Eudicot (PS3)	2356(4.5%)	303	521(1042)	206(1011)
Rosid (PS4)	582(1.1%)	109	181(362)	31(111)
Legume (PS5)	1780(3.4%)	460	452(904)	87(416)
Phaseoleae (PS6)	1590(3.0%)	568	400(800)	49(222)
Soybean (PS7)	11436(21.7%)	9961	476(952)	118(523)

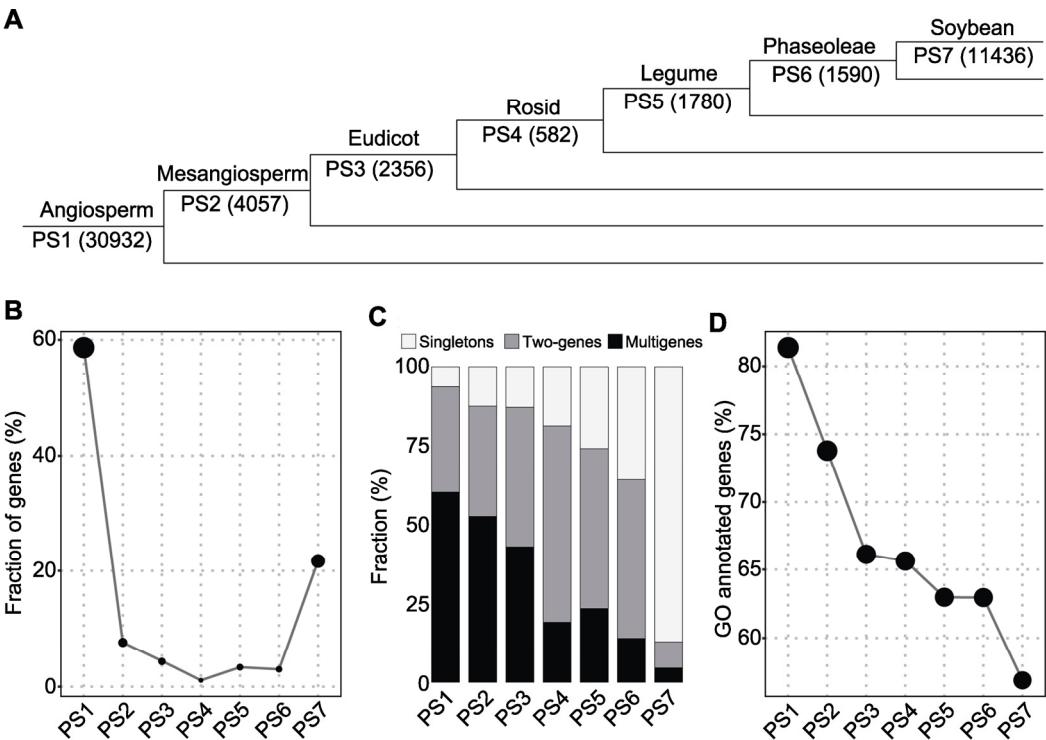


图2 大豆基因起源
(A) 不同起源节点(PS1–PS7)基因数目; (B) 基因比例; (C) 基因拷贝数状态; (D) 基因GO注释

Figure 2 Origination of soybean genes

(A) Numbers in parenthesis denote the number of genes per phylostratum (PS1–PS7); (B) Gene fraction; (C) Gene copy status; (D) Gene Ontology annotation

2.3 大豆基因的进化速率及所受选择压大小

为了分析大豆基因在进化过程中所受选择约束和进化速率,我们对大豆与菜豆直系同源基因对的dN、dS和dN/dS值进行计算。如图3A所示,起源时期不同的基因尽管同义位点的核苷酸进化速率(dS在0.345 8 (PS1)–0.398 4 (PS6)之间)差异不大;但从dN/dS来看,不同组的基因所受到的选择压存在显著差异,其dN/dS值的中值分布在0.194 5 (PS1)–0.358 0 (PS6)

之间。因此,本研究显示,古老基因受到的选择压较大,而新起源基因受到的选择压较小,这可能是由于古老基因中富集了更多的被子植物生存所必需的基因。与这一推测相吻合的是,在PS1–PS4中,单拷贝基因比多拷贝基因受到的选择压要小,这表明重要基因在基因组中进化出更多拷贝数以提高基因组鲁棒性(robustness)的趋势。

非同义替换速率(dN值)也是衡量基因进化速率的重要因素。如图3C所示, dN值的中值范围在

0.0767 5 (PS1)–0.138 3 (PS6)之间。与此同时, 基因在不同拷贝状态下的进化速率不同, 其中多拷贝基因dS和dN值的中值最大, 单拷贝基因次之, 两拷贝基因最小(图3B, C)。

2.4 大豆基因表达

对于起源时间指定到7个不同系统发育层级的52 733个大豆基因, 其中有42 591个(80.1%)在至少1个组

织样本中表达(Shen et al., 2014)。起源于PS1–PS7的基因中所表达的基因比例差别较大, 其中在PS1中表达的比例(93.2%)最高, 而PS7中表达的比例(49.5%)最低(图4A)。与此现象一致的是, 从基因总体表达水平来看, PS1中最高, 而PS7中最低(图4B)。

我们进一步比较了不同年龄基因表达的组织和发育阶段特异性。结果显示, 起源较早的基因趋于广谱表达, 而起源较近的基因则具有较高组织和发育阶

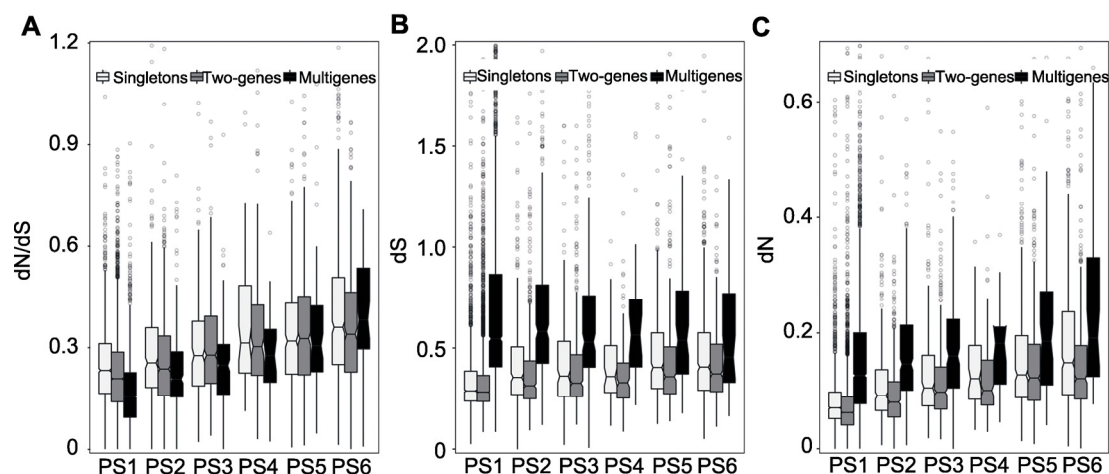


图3 大豆基因分歧程度

通过大豆与菜豆同源基因对来评估选择压(dN/dS) (A)、同义替换率(dS) (B)和非同义替换率(dN) (C)。

Figure 3 Divergence degrees of soybean genes

Estimated between soybean and common bean selection pressure (dN/dS) (A), synonymous substitution rate (dS) (B) and nonsynonymous substitution rate (dN) (C).

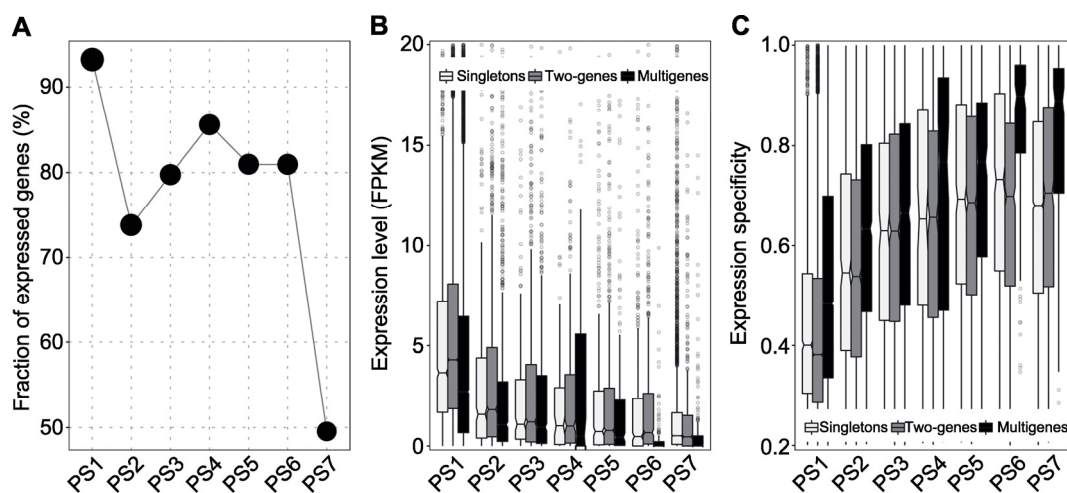


图4 大豆基因表达

(A) 已表达基因; (B) 表达水平; (C) 表达特异性

Figure 4 Expression of soybean genes

(A) Expressed genes; (B) Expression level; (C) Expression specificity

段特异性表达(图4C)。有意思的是,从整体水平来看,两拷贝基因具有较高的基因表达水平及组织和发育阶段特异性(图4B, C)。

2.5 大豆选择性剪切

为探明大豆基因复制事件与选择性剪切事件之间的进化关系,我们对28个样本的转录组数据进行分析,发现有61.3% (26 114个)的已表达基因发生选择性剪切,这一比例与前人研究结果一致(Shen et al., 2014),同时发现内含子保留(intron retention, IR)是

最主要的剪切类型,占35.7% (图5A)。

总体水平上,基因的起源时间与发生选择性剪切基因的比例之间存在负相关,发生选择性剪切的基因比例从36.4%–70.8%不等,其中PS6的比例最低,而PS1的比例最高(图5B)。有趣的是,与单拷贝基因和多拷贝基因相比,两拷贝基因更趋向于发生选择性剪切,在不同起源阶段,发生选择性剪切的两拷贝基因占总体两拷贝基因的比例介于39.4% (PS6)–74.6% (PS1)之间;起源较早的基因中,以多拷贝形式存在的基因发生选择性剪切的比例要高于单拷贝基因,起

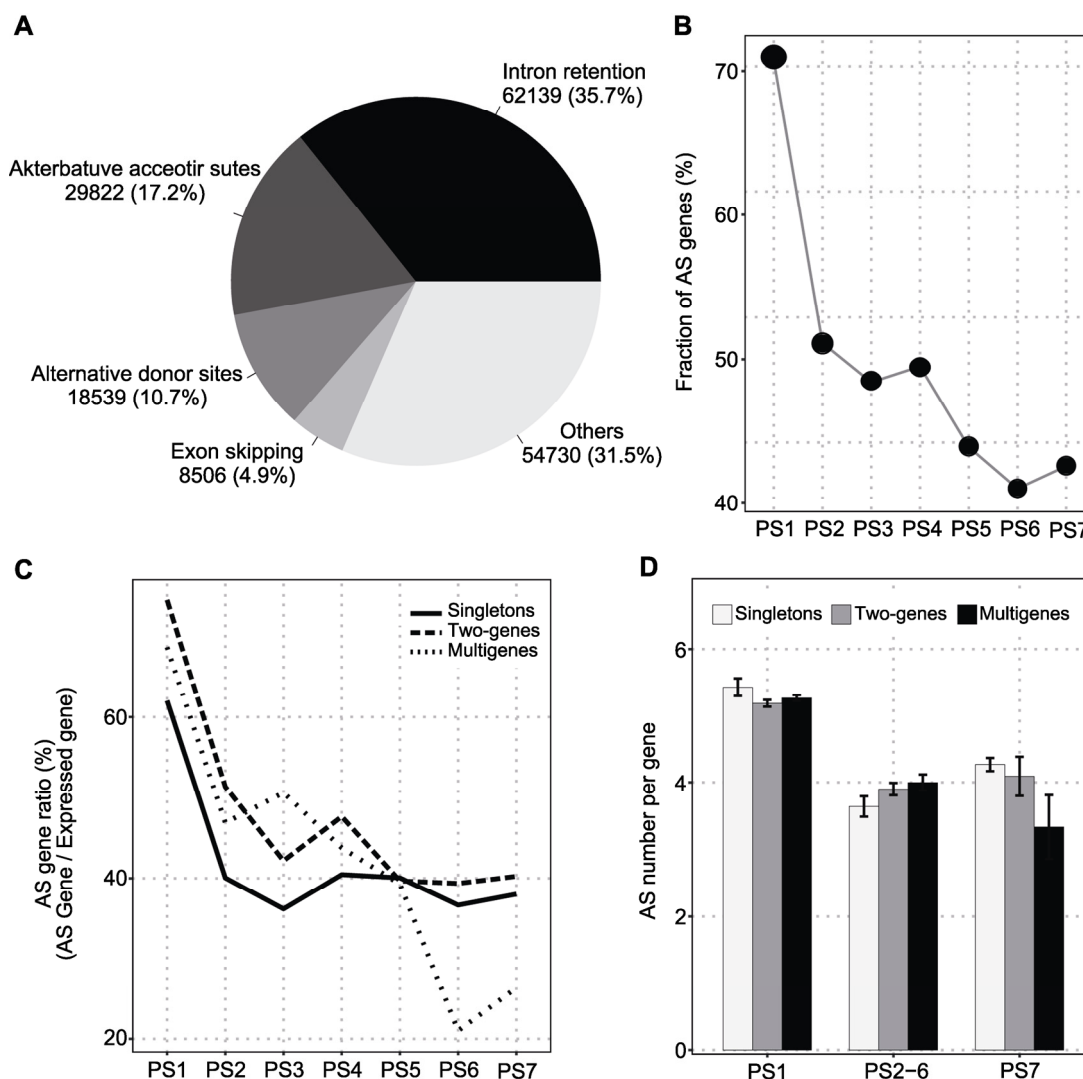


图5 大豆基因的选择性剪切(AS)

(A) 选择性剪切事件; (B) 发生选择性剪切的基因比例; (C) 不同拷贝数状态下发生选择性剪切的基因; (D) 每个基因发生选择性剪切事件的数目

Figure 5 Alternative splicing (AS) of soybean genes

(A) AS event; (B) AS genes ratio; (C) AS genes for different copy status; (D) AS number per gene

源较晚的基因趋势与之相反(图5C)。

本研究中,我们发现不同起源时期基因发生选择性剪切的数目具有差异。如图5D所示,起源较早基因(PS1)其选择性剪切数目较多,而起源较晚的基因发生选择性剪切的数目较少。与此同时,在不同起源时期,选择性剪切数目与基因家族大小之间的关系明显不同,具体表现为:在起源早期(PS1),平均每个基因发生选择性剪切事件的数目较多且与基因家族大小没有显著关联;在起源中期(PS2–6),选择性剪切数目与基因家族大小呈正相关;而在起源晚期(PS7),选择性剪切数目与基因家族大小呈负相关。这可能与基因复制后不同的选择性剪切进化模式有关。

3 讨论

相比前人的研究(Guo, 2013; Jiao and Paterson, 2014),我们选择的19个植物物种中包含了无油樟,其被普遍认为是被子植物中最早分化出来的一个进化分支(图1A)。无油樟基因组已经完成测序(Amborella Genome Project, 2013),这为祖先被子植物基因家族大小的重构提供了重要参考,也有利于推断早期起源的基因家族(基因)更加精细的起源时间。借助这一关键物种基因组数据,本研究阐明高达58.7%的大豆蛋白编码基因起源于被子植物多样化之前。

本研究通过OrthoMCL(Li et al., 2003)聚类共获得34 010个直系同源基因家族,其中所有这些物种共有的基因家族为6 122个,这些家族可能代表了“核心”被子植物功能基因。我们挑选拟南芥基因组中每个基因家族中的一个代表性基因,以此作为参考进行核心基因的功能富集分析,结果显示核心基因的功能主要与氧化还原、跨膜转运和植物组织发育等重要生物学过程相关(图6)。此外,我们发现大豆比其它所有代表种的基因组具有更高比例的两拷贝(30.3%)和多拷贝(43.5%)基因,这可能与大豆基因组发生过2次最近的WGD事件有关,即发生在大约5 900万年前的蝶形花亚科起源处的复制事件和大约1 300万年前大豆属特有的复制事件(Schmutz et al., 2010)。

本研究中,大豆基因组中较古老基因受到更强的选择压且进化速率较慢(图3),这与拟南芥中的研究结果一致(Guo, 2013)。有意思的是,在起源早期基因

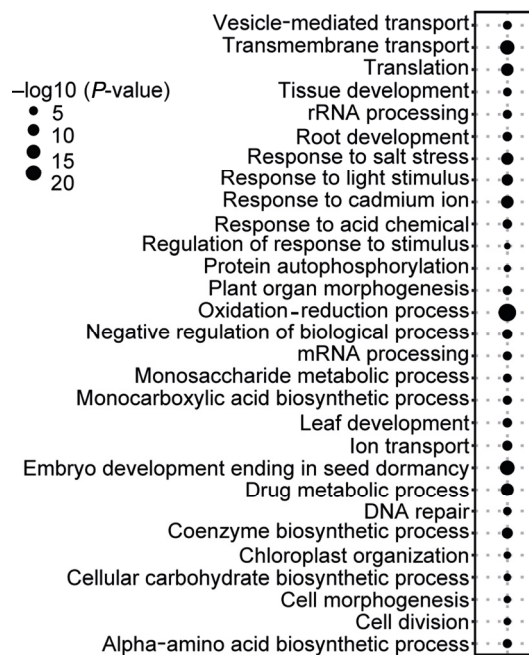


图6 核心被子植物基因的功能富集分析

Figure 6 Functional enrichment analyses of the core angiosperm genes

(PS1–PS4)中,多拷贝基因所占比例较高(在19.1%(PS4)–60.3%(PS1)之间)且受到较大的选择压,这表明古老基因中的复制基因可能在序列上受到更强的功能约束,从而更容易保留下来。不同复制机制(Panchy et al., 2016)产生的基因拷贝保留下来的程度具有差异。一般来说,WGD产生的基因拷贝更倾向于保留下来。例如,WGD同时使所有基因的数量增加1倍,而出于维持剂量平衡(dosage balance)的需要导致参与编码大分子复合体或处在同一生物学网络的基因优先保留(Tasdighian et al., 2017),这些基因一般处在强的负选择压下。而非WGD产生的复制基因一般会导致剂量不平衡,从而发生假基因化或丢失。相较于WGD能大量增加基因数目,非WGD可能会将新拷贝与祖先调控特征分开,并将它们置于新的基因组环境中,这可能更容易产生新的表达模式和新的功能(Jiao and Paterson, 2014)。据此我们推断古老基因中的拷贝更多以WGD形式产生并保留下来。

与基因复制相同,选择性剪切也被认为是提高物种转录组和蛋白组多样性的主要机制(Keren et al., 2010)。随着二代测序技术的发展,研究显示植物中

也存在大量的选择性剪切事件(Reddy et al., 2013)。然而, 选择性剪切和基因复制在增加蛋白组多样性的方式及进化模式上具有明显差异。基因复制之后选择性剪切的模式是如何受到影响的? 前人已经提出3种模型来解释基因组复制和选择性剪切之间的关系(Reddy et al., 2013): (1) 独立模型(independent model), 即基因家族大小与选择性剪切数量之间相互独立; (2) 功能共享模型(function-sharing model), 其预测基因家族大小与选择性剪切数量之间呈反向相关性; (3) 促进选择性剪切模型(accelerated AS model), 其推测每个基因选择性剪切事件数量的增加是由于每个旁系同源基因的放松选择压力引起。与人(*Homo sapiens*)和小鼠(*Mus musculus*)基因组中的研究结果一致(Chen et al., 2011), 不同年龄大豆基因在基因复制后的选择性剪切进化模式不同。本研究中, 我们发现起源早期(PS1)基因的选择性剪切数量与基因家族大小没有显著关联, 起源中期(PS2–6)基因的选择性剪切数量与基因家族大小呈正相关; 而在起源晚期(PS7), 选择性剪切数量与基因家族大小呈负相关(图5D)。这表明起源早期、中期和晚期基因可能分别符合独立模型、促进选择性剪切模型和功能共享模型。

综上, 大豆基因组中不同起源时间的基因具有不同的特征。大多数基因(58.7%)起源于被子植物多样化之前, 较古老基因通常以多拷贝状态存在, 功能注释较完整, 受到更大的选择压且进化速率更小, 更趋向于广谱表达且表达量较高, 有更多比例的基因发生选择性剪切且具有更多的剪切事件。此外, 不同复制状态下基因的特征也具有显著差异, 其中以两拷贝状态存在的基因进化速率较慢、表达水平较高, 越不趋向于特异性表达且更容易发生选择性剪切。本文首次分析了大豆基因在不同起源阶段的进化特征, 对理解大豆基因组的宏进化过程具有重要意义。

参考文献

- 孙红正, 葛颂 (2010). 重复基因的进化——回顾与进展. *植物学报* **45**, 13–22.
- Albalat R, Cañestro C (2016). Evolution by gene loss. *Nat Rev Genet* **17**, 379–391.
- Amborella Genome Project (2013). The *Amborella* genome and the evolution of flowering plants. *Science* **342**, 124–1089.
- Bolger AM, Lohse M, Usadel B (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120.
- Cai JJ, Borenstein E, Chen R, Petrov DA (2009). Similarly strong purifying selection acts on human disease genes of all evolutionary ages. *Genome Biol Evol* **1**, 131–144.
- Chen SD, Krinsky BH, Long MY (2013). New genes as drivers of phenotypic evolution. *Nat Rev Genet* **14**, 645–660.
- Chen TW, Wu TH, Ng WV, Lin WC (2011). Interrogation of alternative splicing events in duplicated genes during evolution. *BMC Genomics* **12**(Suppl3), S16.
- Domazet-Lošo T, Brajković J, Tautz D (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet* **23**, 533–539.
- Doyle JJ, Luckow MA (2003). The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. *Plant Physiol* **131**, 900–910.
- Enright AJ, Van Dongen S, Ouzounis CA (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575–1584.
- Foissac S, Sammeth M (2007). ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Res* **35**, W297–W299.
- Freeling M (2009). Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu Rev Plant Biol* **60**, 433–453.
- Guo YL (2013). Gene family evolution in green plants with emphasis on the origination and evolution of *Arabidopsis thaliana* genes. *Plant J* **73**, 941–951.
- Jiao YN, Paterson AH (2014). Polyploidy-associated genome modifications during land plant evolution. *Philos Trans R Soc Lond B Biol Sci* **369**, 20130355.
- Kaessmann H (2010). Origins, evolution, and phenotypic impact of new genes. *Genome Res* **20**, 1313–1326.
- Keren H, Lev-Maor G, Ast G (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet* **11**, 345–355.
- Kim D, Langmead B, Salzberg SL (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357–360.
- Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM,

- Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948.
- Li L, Stoeckert CJ Jr, Roos DS (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178–2189.
- Long M, Betrán E, Thornton K, Wang W (2003). The origin of new genes: glimpses from the young and old. *Nat Rev Genet* **4**, 865–875.
- Lynch M, Conery JS (2000). The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155.
- Merkin J, Russell C, Chen P, Burge CB (2012). Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* **338**, 1593–1599.
- Michael TP, Jackson S (2013). The first 50 plant genomes. *Plant Gen* **6**, 2.
- Michael TP, VanBuren R (2015). Progress, challenges and the future of crop genomes. *Curr Opin Plant Biol* **24**, 71–81.
- Ohno S (1970). Evolution by Gene Duplication. Berlin, Heidelberg: Springer. pp. 1–160.
- Panchy N, Lehti-Shiu M, Shiu SH (2016). Evolution of gene duplication in plants. *Plant Physiol* **171**, 2294–2316.
- Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295.
- Quint M, Drost HG, Gabel A, Ullrich KK, Bönn M, Grosse I (2012). A transcriptomic hourglass in plant embryogenesis. *Nature* **490**, 98–101.
- Reddy ASN, Marquez Y, Kalyna M, Barta A (2013). Complexity of the alternative splicing landscape in plants. *Plant Cell* **25**, 3657–3683.
- Schmutz J, Cannon SB, Schlueter J, Ma JX, Mitros T, Nelson W, Hyten DL, Song QJ, Thelen JJ, Cheng JL, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu SQ, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du JC, Tian ZX, Zhu LC, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA (2010). Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183.
- Shen YT, Zhou ZK, Wang Z, Li WY, Fang C, Wu M, Ma YM, Liu TF, Kong LA, Peng DL, Tian ZX (2014). Global dissection of alternative splicing in paleopolyploid soybean. *Plant Cell* **26**, 996–1008.
- Suyama M, Torrents D, Bork P (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**, W609–W612.
- Tasdighian S, Van Bel M, Li Z, Van de Peer Y, Carretero-Paulet L, Maere S (2017). Reciprocally retained genes in the angiosperm lineage show the hallmarks of dosage balance sensitivity. *Plant Cell* **29**, 2766–2785.
- Tautz D, Domazet-Lošo T (2011). The evolutionary origin of orphan genes. *Nat Rev Genet* **12**, 692–702.
- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511–515.
- Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, Lancet D, Shmueli O (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659.
- Yang ZH (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586–1591.
- Zhang JZ (2003). Evolution by gene duplication: an update. *Trends Ecol Evol* **18**, 292–298.

Origin and Evolution of Soybean Protein-coding Genes

Kang Tang, Ruolin Yang*

College of Life Sciences, Northwest A&F University, Yangling 712100, China

Abstract The evolution of gene composition of a species is a highly dynamic process, wherein lineage- and species-specific genes originated relatively recently are continuously integrated into the original gene network of older genes. These young genes play important roles in shaping the genome architecture, thereby leading to improved adaptation for organisms. Gene duplication and *de novo* origination of new genes are two ways to create new genes, causing different gene families with various copy numbers. To what extent and how the evolutionary pattern of genes depends on the timing of gene origination are still largely unexplored in soybean. In this study, we selected 19 representative angiosperms and analyzed the potential relations of the gene content dynamics with the origination of soybean (*Glycine max*) genes. Using the gene emergence approach, we found that 58.7% of soybean genes could be dated to ~150 million years ago and 21.7% orphan genes had recently originated. As expected, in comparison with young genes, older genes tend to be subjected to stronger purifying selection and were more conserved. In addition, older genes featured higher expression levels and were more likely to undergo alternative splicing. Furthermore, genes with different copy numbers showed a difference in these aspects. These findings may help understand the evolutionary models of genes with different ages.

Key words angiosperms, gene duplication, gene family, gene origin, soybean

Tang K, Yang RL (2019). Origin and evolution of soybean protein-coding genes. *Chin Bull Bot* **54**, 316–327.

* Author for correspondence. E-mail: desert.ruolin@gmail.com

(责任编辑: 朱亚娜)