

· 特邀综述 ·

## 重复基因的进化——回顾与进展

孙红正<sup>1,2</sup>, 葛颂<sup>1\*</sup>

<sup>1</sup>中国科学院植物研究所系统与进化植物学国家重点实验室, 北京 100093

<sup>2</sup>中国农业科学院作物研究所, 作物种质资源与生物技术重点实验室, 北京 100081

**摘要** 基因重复是普遍存在的生物学现象, 是基因组和遗传系统多样化的重要推动力量, 在生物进化过程中发挥着极其重要的作用。基因重复有何利弊, 基因发生重复后, 2个重复子拷贝的保留在基因功能方面是否存在偏好性, 子拷贝在表达和进化速率上如何分化, 以及重复基因为什么会被保留下来一直是进化生物学领域研究的热点问题之一。该文对以上重复基因研究的热点问题进行了介绍, 并对重复基因的进化机制和理论模型及其近年来的一些主要研究进展进行了综述。

**关键词** 进化, 基因重复, 基因组

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.

基因重复(gene duplication)是指在一个基因组内存在2个或者2个以上拷贝的同源基因序列, 是非常普遍的生物学现象。在已完成或即将完成全基因组测序的三大界生物(细菌、古菌和真核生物)基因组中, 业已发现存在大量的重复基因(Zhang, 2003)。基因重复一般有以下几种来源方式: 基因组内的不等交换、反转录插入、大规模的染色体片段重复和全基因组重复(Zhang, 2003)。在植物界, 基因重复现象更为普遍, 仅在被子植物中就有大约70%–80%的物种存在基因重复或多倍化过程, 且发生在不同的进化阶段, 许多物种是多次全基因组重复的产物(图1)(Masterson, 1994; Bowers et al., 2003)。例如, 针对模式植物拟南芥(*Arabidopsis thaliana* (L.) Heynh.)全基因组数据的分析表明, 该种至少经历了3次全基因组加倍(Ku et al., 2000; Ermolaeva et al., 2003)。最近葡萄(*Vitis vinifera* L.)全基因组测序的完成进一步表明, 大部分双子叶物种都起源于六倍体祖先物种, 在随后的进化中, 不同的支系又发生过四倍体化过程(Jaillon et al., 2007)。在单子叶模式植物水稻(*Oryza sativa* L.)的基因组中, 有62%的位点属于重复位点(Moore and Purugganan, 2005), 目前发现水稻基因组内发生过2次大的重复事件: 禾本科分化前(约70 MYA, million years ago)的全基因组重复和发

生在5–8 MYA的涉及第11和12号染色体的部分染色体片段重复(Paterson et al., 2004; Wang et al., 2005; Yu et al., 2005)。另外一个重要的禾本科作物玉米(*Zea mays* L.)也在约11 MYA发生了一次多倍化事件(Gaut et al., 2000)。每发生一次全基因组或部分染色体片段重复, 都有一部分基因消亡, 而另一部分基因则被保留下来成为重复基因。尽管揭示基因重复现象的报道日益增多, 但迄今为止针对基因重复的起源方式和机制以及基因重复后重复拷贝的进化命运仍缺乏足够的研究。为此, 本文总结了近年来重复基因研究的一些进展, 对重复基因的生物意义和进化命运, 尤其是重复基因分子进化的一些特点及其理论模型进行介绍, 以期推动国内该领域的研究。

### 1 基因重复的利和弊

基因重复在生物进化过程中发挥着极其重要的作用, 是基因组和遗传系统分化的重要推动力量(Ohno, 1970)。基因重复发生后, 其中一个拷贝继续发挥其原来的功能, 而另一个拷贝可以不再受遗传选择的约束, 随机地积累突变, 为新基因和新功能的出现提供可能(Hughes, 1994; Wendel, 2000; Zhang, 2003; 高远等, 2008), 从而促进复杂基因表达调控网络和

收稿日期: 2008-12-18; 接受日期: 2009-03-13

基金项目: 国家自然科学基金(No. 30430030)和国家基础项目研究基金(No.2007CB815704)

\* 通讯作者。E-mail: gesong@ibcas.ac.cn

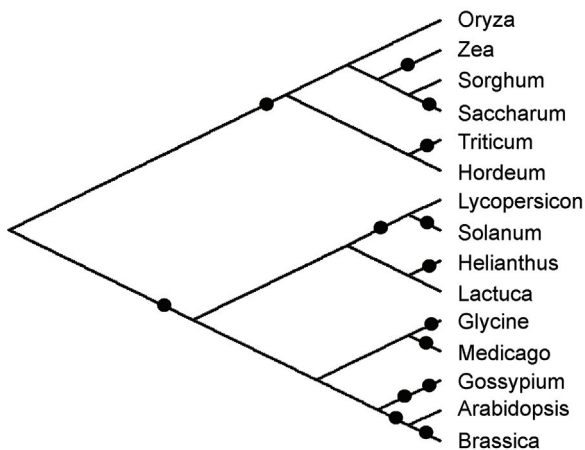


图1 被子植物主要进化谱系中发生的多倍化事件(Adams and Wendel, 2005b)

Figure 1 Polyploidization events in the main lineages of angiosperm evolution (Adams and Wendel, 2005b)

相互作用网络的进化,使各种调控更加精确且特异(Wagner, 1998; He and Zhang, 2005; 吕山花和孟征, 2007),也有可能致基因出现种属特异性而使群体之间形成生殖隔离,加速种间分化(Lynch and Conery, 2000)。在基因突变和自然选择的作用下,重复基因的进化可使生物更好地适应环境变化,如高粱(*Sorghum vulgare* Pers)基因组测序的研究结果发现,高粱对干旱的耐受性可能与基因和microRNA的重复有关(Paterson et al., 2009)。另外,基因重复对保持遗传系统的稳定性及缓冲外界环境对遗传系统的影响也有重要的作用,即使一个拷贝因突变而失活,另外一个拷贝仍然可以执行原功能,弥补另一个拷贝失活所造成的伤害(Gu, 2003; Chapman et al., 2006)。但是,基因重复并不是完全没有弊端,基因重复以后,重复基因成为基因组的额外负担,特别是全基因组复制后,细胞的遗传结构和有丝分裂变得更加复杂,染色体不能正常分配到子代细胞中,造成不育或者后代生活力低下(Comai, 2005)。由此可见,基因重复对生物个体和物种的影响是一个利与弊共存的现象,其利和弊与生物本身的特性及其在进化中的地位以及与环境相互作用有关,对其进行全面揭示有待于来自更多类群的研究。

## 2 重复基因保留的偏好性

基因重复以后,并不是所有的重复拷贝都能被保留下来,有很大一部分基因因为积累了有害突变导致功能丧失而成为假基因,并在长期的进化过程中消亡。有研究表明,一个基因如果在一次基因重复后被保留下来,那么它在新一轮基因重复中也很有可能被保留下来(Adams and Wendel, 2005b),而有些基因则在几次独立的全基因组重复事件后又回到原来的单拷贝状态(Paterson et al., 2006),说明重复基因的保留与丢失并不是一个完全随机的过程,但究竟哪一类重复基因在重复发生以后比较容易保留下来呢?

Seoighe和Wolfe(1999)在研究酵母基因组中的重复基因时发现,按功能分类重复基因在各个类别中的分布并不随机,其更倾向存在于细胞周期相关酶和信号转导相关酶类中,如蛋白质激酶、磷酸化酶、GTP酶、GTP酶激活蛋白和鸟嘌呤核苷酸交换因子,而转录因子类基因则比较少。Kondrashov等(2002)的研究结果表明,细菌的重复基因中有一部分编码膜蛋白或者分泌蛋白;在酵母中,重复基因以跨膜蛋白和逆境胁迫相关基因较多,而在多细胞真核生物中,则以受体蛋白和信号转导相关蛋白为主。Blanc和Wolfe(2004)在拟南芥基因组中发现与信号转导和转录相关的基因被优先保留下来,而DNA修复相关基因则被优先淘汰。Thomas等(2006)也在拟南芥基因组内发现,对剂量敏感的基因(包括蛋白质修饰基因、信号转导基因、编码核糖体蛋白和转录因子的基因)在重复基因内所占比例较大。Wu等(2008)在水稻基因组中也发现,调控表达基因、信号转导基因、糖代谢相关基因和合成酶类基因被保留下来的频率比较高。以上研究都认为转录因子类基因是比较容易在基因重复后保留下来的,但Barker等(2008)对菊科多倍化事件的研究表明,在重复拷贝中,结构蛋白类基因较为丰富而调控类基因则相对较少。除了在功能类别上发现重复基因的保留存在偏好性以外, Jordan等(2004)对几个真核和原核生物基因组中重复基因的研究发现,发生重复的基因,其进化速率显著低于单基因的进化速率,也就是说进化速率越慢的基因越容易在基因重复中保留下来,而Davis和Petrov(2004)对线虫(*Caenorhabditis elegans*)和酵母基因组中重复基因的研究也表明,被保留下来的重复基因在进化上都是比较保守的,与不存在重复的单拷贝基因相

比, 这些重复基因的进化速率显著小于单拷贝基因。另外, Chapman等(2006)通过对水稻和拟南芥基因组中重复基因和单基因的比较发现, 编码较长和较为复杂的蛋白质的基因被优先保留下来; 而Aury等(2006)在原生动物纤毛虫(*Paramecium tetraurelia*)基因组中发现表达量高的基因在重复基因中占有较高的比例。这两种现象被认为与保持遗传系统的稳定和剂量效应有关(Gu, 2003; Aury et al., 2006; Semon and Wolfe, 2007)。

### 3 基因重复后在表达上的分化

基因重复后2个拷贝在表达上的分化被认为是重复基因功能分化产生新基因的重要一步(Ohno, 1970; Ferris and Whitt, 1979; Li et al., 2005), 2个重复拷贝在功能上的分化可能需要很长一段时间才会出现, 但是在表达上的分化可以在重复后很短时间内就开始发生(Wendel, 2000; Gu et al., 2002; Adams and Wendel, 2005a; Chain et al., 2008)。在拟南芥(Comai et al., 2000)和棉花(*Gossypium* spp.) (Adams et al., 2003)的人工合成同源多倍体中, 1个基因拷贝的沉默和2个拷贝在不同组织中表达上的偏好性在多倍体合成后的第1和第2代就开始出现。

那么重复拷贝的沉默和偏好是随机的还是可重复的呢? Wang等(2004)在拟南芥的4个多倍体人工合成谱系中发现, 有一些基因在所有的合成系中都表达, 而有些基因则在1或2个合成系中沉默或者表达量降低。另外, 同一合成系同一世代的不同姊妹株之间也存在基因表达上的分化, 这些证据支持随机沉默模式。但Adams等(2004)比较了3个具有相同母本不同父本的棉花多倍体人工合成系, 在所研究的3个基因中, 1个单加氧酶(monoxygenase)基因在3种基因型中都表现出极为相似的器官特异性表达和基因沉默模式, 而另外2个在3个合成系中的表现却有差异。Soltis等(2004)在研究自然多倍体物种*Tragopogon*时, 在父母本正交和反交形成的多倍体中发现不同拷贝在表达上存在差异, 可能是因为父母本效应的影响, 也可能是父母本的多态性或者随机因素造成的。所以, 根据目前的研究结果, 重复拷贝在后代中的表达模式既有随机因素也有固定因素。近来有研究认为, 这种重复拷贝短期内表达模式的快速分化是由DNA甲基化等表观遗传(epigenetic)因素决定的(Lee and

Chen, 2001; Richards and Elgin, 2002; Liu and Wendel, 2003; Osborn et al., 2003; Adams and Wendel, 2005b)。

短期内的表达分化是重复拷贝在基因组内保留完整、没有突变和消亡的情况下, 在基因组的调控和修饰下实现的。对于在长期进化过程中保留下来的基因, 随着进化时间的不断推移, 2个拷贝之间所共享的调控元件呈逐步减少的趋势(Papp et al., 2003), 而且发生在不同物种之间; 有些重复基因的表达谱不如单基因的表达谱保守, 重复基因的表达谱变窄且特异性增加(Huminiecki and Wolfe, 2004)。通过对酵母、线虫、拟南芥和人类基因组内的重复基因进行研究发现, 重复基因在表达上的分化程度与发生重复的时间( $d_s$ )之间存在正相关关系, 即重复发生的时间越久远, 其在表达上的分化越明显(Gu et al., 2002; Makova and Li, 2003; Blanc and Wolfe, 2004; Zhang et al., 2004; Li et al., 2005)。

### 4 基因重复后进化速率的分化

一般认为, 基因重复后一个拷贝受到的选择限制减轻, 进化速率会发生改变, 但是从现有的研究结果来看, 并不是所有重复拷贝都会发生进化速率的改变。四倍体蛙*Xenopus laevis*起源于约30 MYA的多倍化事件, Hughes和Hughes(1993)用哺乳动物、人和小鼠的同源基因作为外类群比较了四倍体蛙中17对重复基因的进化速率, 结果发现这些重复基因的进化速率都没有显著差异。在其它一些已完成全部或部分全基因组测序的物种中, 用全基因组搜索的方法找出物种重复基因对, 通过和外类群同源基因的比较, 发现重复基因对之间进化速率显著不同的基因所占比例从0%–100%不等(表1)。Gayral等(2007)研究了老鼠基因组中10对反转录重复基因, 用近缘鼠类物种的同源基因作为外类群, 比较发现所有10对重复基因都表现出进化速率的差异, 源拷贝的进化速率没有受到基因重复的影响, 而反转座拷贝的进化速率则发生了显著的加速。进化速率的显著不同可能暗示着基因功能发生了变化, 进化加快的拷贝有可能会出现新的功能(Chain and Evans, 2006; Semon and Wolfe, 2007), 但是He和Zhang (2005)认为这种进化速率的不同也可能是由于2个子代拷贝分别继承了祖先功能不同的结构域, 而不同功能域所受的选择压力不一样

**表1** 用相对速率检测(relative rate test)研究重复基因进化速率差异的例子**Table 1** Examples of evolutionary rate differences between duplicates by using relative rate test

物种	基因对	速率差异的基因对	百分比(%)	参考文献
蛙	17	0	0	Hughes and Hughes, 1993
鱼, 哺乳动物	19	4	21	Robinson-Rechavi and Laudet, 2001
斑马鱼	26	13	50	Van de Peer et al., 2001
细菌, 古菌和拟南芥	101	7	7	Kondrashov et al., 2002
人类	250	145	60	Zhang et al., 2003
芽殖酵母	22	6	27	Conant and Wagner, 2003
裂殖酵母	14	3	21	Conant and Wagner, 2003
果蝇	44	13	30	Conant and Wagner, 2003
线虫	164	46	28	Conant and Wagner, 2003
酵母	457	76	17	Kellis et al., 2004
鱼	2 466	610	25	Steinke et al., 2006
老鼠	10	10	100	Gayral et al., 2007

同样会导致产生不同的进化速率。

以上研究都采用重复基因与单拷贝外类群同源基因相比较的方法, 用相对速率检验(relative rate test)来检测2个拷贝之间在进化速率上是否显著不同。外类群比较方法有2个弊端: 一是所选外类群亲缘关系的远近影响检测的灵敏度; 二是外类群能检测到已经发生较大分化的2个重复基因拷贝, 但对于分化较小或者正在发生的分化可能不太灵敏。通过分析重复基因在近缘物种内的进化可以更灵敏地反映其在进化速率上的差异, 而群体遗传学方法则能反映重复基因短期进化时间内在物种水平上发生的事件, 如假基因化、正选择和选择压力的松弛等, 所以通过物种间进化比较和物种内群体遗传学相结合的方法可以更灵敏地反映重复基因的进化历史, 是未来研究的方向。

## 5 基因重复后的命运以及进化模型

基因重复以后, 发生重复的个体有可能在群体中被保留下来并扩散, 也有可能消失在群体中。假如基因重复是中性的, 重复基因在二倍体基因组中被固定下来的概率只有 $1/2Ne$ ( $Ne$ 为有效群体大小)。这样, 很多基因重复个体在很短的时间内就被丢失, 而那些固定下来的基因重复个体在群体中的固定也需要很长的过程, 因为对于一个中性基因来说, 被固定下来平均需要 $4Ne$ 个世代(Kimura, 1983)。在生物界各个物种中, 重复基因的普遍性是无法用随机因素解释的。重复基因如何被固定下来以及重复基因长期的进化命

运是生物进化研究的热点问题, 形成了各种不同的观点, 并提出了多个模型来解释重复基因的进化命运(图2)。

早在1970年Ohno就提出一个假设, 认为基因重复以后2个基因会有不同的进化命运, 其中一个基因拷贝保持原来的功能不变, 而另一个拷贝则有2种不同的命运(Ohno, 1970)。第1种命运是这个拷贝作为一个冗余拷贝, 不受选择压力约束而发生随机突变, 以至于不能转录或者翻译, 再或者翻译后无功能, 这样导致在长期进化过程中成为完全无功能的DNA序列而被丢失或因突变而变得面目全非难以确认。这个过程被称为假基因化过程(non-functionalization), 一般在基因重复后的几百万年内发生(Lynch and Force, 2000)。第2种命运是重复拷贝在突变过程中产生有利突变, 赋予这个重复拷贝以新的功能, 使生物能更好地适应环境, 即正选择(positive selection), 这个过程称为新基因化过程(neofunctionalization, NF), 新基因化的经典例证来自于灵长类动物的EDN和ECP基因(Zhang et al., 1998)。研究表明, EDN和ECP基因起源于旧世界猴中的一次基因重复事件, 而在新世界猴和类人猿中只有1个EDN拷贝。EDN和ECP基因都属于RNase A基因家族, 所不同的是ECP基因出现了抗菌性的新功能, 而且这种新功能不依赖于RNase活性, 人类和类人猿的EDN基因则不具备ECP基因的抗菌活性(Rosenberg and Dyer, 1995)。进化分析表明, ECP基因的新功能是因为基因重复后编码蛋白发生了大量精氨酸替代, 这些带正电

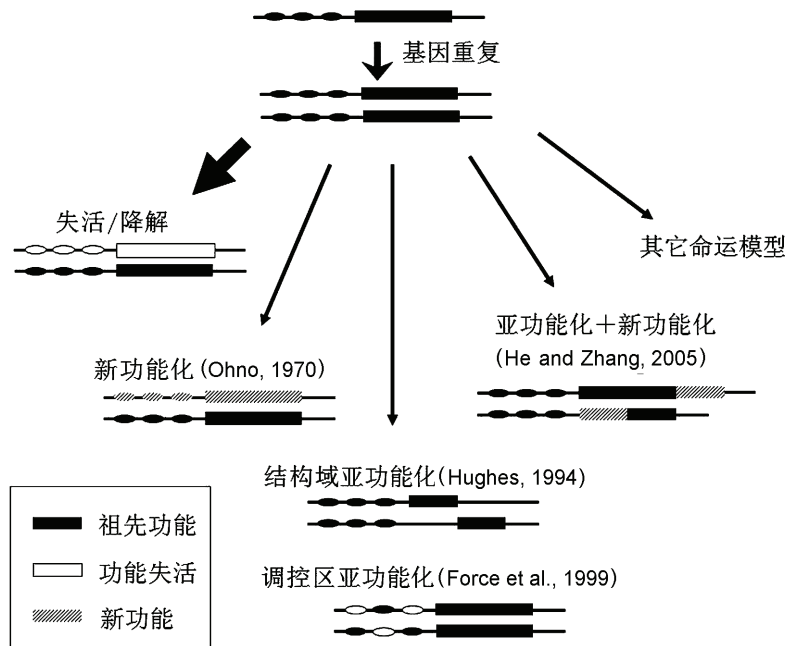


图2 基因重复后的进化模型

Figure 2 Schematic illustration of gene duplicates evolution models

荷的精氨酸更有利于其结合带负电荷的细菌细胞膜,并进一步破坏细菌细胞,从而达到抗菌效果(Zhang et al., 1998)。

鉴于Ohno提出其经典模型时分子数据和功能上的证据都很少,因此他只强调了失去功能/消亡和新功能化这2种命运。随着分子数据和基因组数据的不断增多,虽然也有支持Ohno经典模型的例子,但是用经典模型不能解释的数据也逐渐增多。Hughes(1994)首先从现有的数据中发现了不符合这种新功能化模型的现象:四倍体蛙类*Xenopus laevis*中,在全基因组重复发生约30 MY之后,重复基因仍受到强烈的负选择,而不是有一个重复基因拷贝可以自由突变。根据同一个基因编码具有2个不同功能域的蛋白质这一现象,他提出了基因复制后重复拷贝的另一种命运:亚功能化(subfunctionalization, SF),即一个基因在基因重复发生之前具有2个功能区域,发生重复以后2个拷贝分别保留其中一个功能区域,并且在自然选择的作用下,2个拷贝都积累一定的非同义突变,使它们更适合于特定的功能。这个模型在

古细菌的tRNA内切酶基因中得到了支持(Tocchini-Valentini et al., 2005)。

Force等(1999)进一步对以前的经典模型提出了疑问。第1,在很多研究中,重复基因被保存下来的比例通常要比用经典模型所预测的高。例如,在四倍体鱼类中,30%–75%的重复基因在50–100 MY的进化过程中被保留下来而没有被假基因化或丢失;玉米中72%的重复基因被保留了11 MY;在*Xenopus laevis*中几乎所有重复基因都被保留了30 MY。而在经典模型中,重复基因被保留下来的原因是出现了新功能化,但基因出现新功能化的几率很小,不可能出现如此大比例的重复基因被保留。第2,重复基因一般都受到负选择(同Hughes)。第3,在群体中没有发现无效等位基因,这种现象与经典模型中认为冗余拷贝可以随机发生突变这一假说相背离。基于这些原因,Force等(1999)在经典模型的基础上对原来的模型进行了修订,提出DDC(Duplication-Degeneration-Complementation)模型,也称为亚功能化(SF)模型。Force的模型认为基因重复发生以后,2个拷贝都会产

生突变,所不同的是这2个拷贝的突变可发生在2个重复基因的不同调节区,从而都被随机的遗传漂变固定下来,造成2个重复基因必须互补合作才能完成原来祖先基因所能完成的功能。这种互补可以是质的互补,也可以是量的互补。质的互补是指2个重复基因不同的调节功能域都有不同的完全功能丧失,一个拷贝互补另一个拷贝所丧失的功能。而量的互补是2个拷贝的调节区域都有部分功能丧失,导致2个拷贝的表达水平都有所降低,只有2个拷贝都表达才能维持基因重复前的表达水平。Force等(1999)还通过前人的研究发现了几个支持这种模型例子,如斑马鱼中的*eng1*、*eng1b*基因以及玉米的*zag1*、*zmm2*基因。Hughes和Force的模型都认为基因的亚功能化是重复基因保留的重要机制,但是Hughes的亚功能化侧重于蛋白质结构的亚功能化,而Force的亚功能化则侧重于表达模式的亚功能化。因为在生物的生长发育过程中,起调控作用的基因的功能既包括不同基因编码蛋白质本身的功能,又包括基因表达在时间和空间上的调控功能,所以Hughes和Force的模型都可以归入基因的亚功能化模型(He and Zhang, 2005)。

Hughes和Force等提出的亚功能化模型极大地完善了Ohno的经典模型,能够很好地解释更多的重复基因进化现象,也都找到了支持各自学说的例子,但是He和Zhang(2005)用酵母和人类基因组数据来验证NF和SF模型时,发现单独用NF或SF模型都不能很好地解释基因组数据,于是就提出了一个综合模型SNF (subneofunctionalization)模型。Ohno的NF模型认为重复基因以获得新功能为最终结局,而He和Zhang认为这种模型过于狭隘,并在他们的SNF模型里对这种NF进行了扩展,认为获得新功能的重复基因可以分为3种类型:新基因保留了原来基因的全部功能(NF-I),新基因完全失去了原基因功能(NF-II)和只保留了部分原基因功能(NF-III)。SNF模型的另一个观点是,基因重复后发生SF是一个相当快速的过程,在长期进化过程中这些发生亚功能化的基因有可能各自进化出新的基因功能(NF)。

除了上述模型外,还有一些其它模型来解释重复基因的进化。Rastogi和Liberles(2005)用计算机模拟进化的方法研究了基因重复后基因的进化,他们也认为SF在重复基因的进化过程中发挥了重要作用,但是SF只是作为NF的一个中间转换状态,而不是最终状态;模拟结果显示基因重复后2个拷贝的命运会出

现8种情况。Yang等(2006)在研究杨树(*Populus trichocarpa*)、拟南芥和水稻*Dof*基因家族的重复后进化时也发现了基因进化模式的多样化。他们把基因重复后重复拷贝的命运分为3种:消亡(D)、保留(R)和新功能化(N)。每个拷贝有3种命运,则2个拷贝组合起来就有6种模式:RR、RD、RN、DD、NN和ND,他们用寻找结构域的方法在*Dof*基因家族内找到了分别支持这6种模式的基因。Gu等(2003)通过对酵母基因组内重复基因的逐个删除,认为大约有1/4的酵母重复基因丢掉1个拷贝后没有出现表型上的差异,被删除基因的功能被另外一个基因拷贝互补,因此重复基因的存在可能更利于遗传系统保持稳定,并由此提出缓冲模型(buffering model)。这种模型认为重复基因的存在可能是为了应对潜在的有害突变,当一个拷贝上发生有害突变时,因为另外一个拷贝可以执行同样的功能,所以有害突变对生物体的影响不会是致命的(Gu, 2003; Chapman et al., 2006; Semon and Wolfe, 2007)。从群体水平上来说,发生有害突变的个体保留下来以后,所产生配子中有害等位基因在群体中会不断地被其它个体所产生的配子稀释,有害等位基因频率逐渐降低以致最后消亡,从而保证了物种水平上遗传系统的稳定。在全基因组重复后的短时期内,属于同一个调控网络或代谢途径的相互作用蛋白之间要保持相对剂量而被保留下来,这种模型被称为剂量效应模型(stoichiometry effect model),但这种模型可能只适用于基因组重复发生的较短时间内,对于时间久远的重复来说则并不适用(Semon and Wolfe, 2007)。另外,根据转录因子和信号转导基因在大规模基因组重复中保留的比例较高这一现象提出的平衡模型(balancing model)与剂量效应模型有类似之处,它们都认为重复基因的存在是基因调控网络之间相互平衡的结果(Maere et al., 2005; Blomme et al., 2006; Semon and Wolfe, 2007),但平衡模型强调的是调控因子(如抑制因子和激活因子)之间的相互制约。这种模型也得到了对一些研究结果的支持(Li et al., 2006; Qian and Zhang, 2008)。

上述各个重复基因进化模型都是在针对特定基因或者特定基因组的研究中所提出的,生物体内基因相互作用和进化模式各异,再加上不同物种与不同环境互作的特殊性,因此每一种模型都不可能很好地解释所有重复基因的进化模式,而且随着基因组和基因功能的深入研究,也有可能出现现有模型不能解释



的数据。因此, 重复基因的进化模式还会得到进一步的补充和完善。随着大量生物全基因组测序的完成, 可以在同一物种内和不同物种间2个层面上发现更多的重复基因。根据分子钟和物种分化时间估算, 这些重复基因发生的时间可以从近期的重复事件延伸到比较古老的重复事件, 从而更有利于我们揭示重复基因在短期、中期以及较长进化时间内的进化动态和规律, 而通过microarray表达分析、knock-out/RNAi功能缺失分析以及酵母双杂交蛋白互作等技术手段对这些重复基因进行更深入的研究, 则可以对重复基因在新基因形成、物种分化以及遗传系统稳定性的维持等方面的作用予以更好地阐释。

## 参考文献

- 高远, 田李, 秦松 (2008). 植物进化中的正选择作用. *植物学通报* **25**, 401–406.
- 吕山花, 孟征 (2007). MADS-box基因家族基因重复及其功能的多样性. *植物学通报* **24**, 60–70.
- Adams KL, Cronn R, Percifield R, Wendel JF (2003). Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc Natl Acad Sci USA* **100**, 4649–4654.
- Adams KL, Percifield R, Wendel JF (2004). Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. *Genetics* **168**, 2217–2226.
- Adams KL, Wendel JF (2005a). Novel patterns of gene expression in polyploid plants. *Trends Genet* **21**, 539–543.
- Adams KL, Wendel JF (2005b). Polyploidy and genome evolution in plants. *Curr Opin Plant Biol* **8**, 135–141.
- Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V, Aiach N, Arnaiz O, Billaut A, Beisson J, Blanc I, Bouhouche K, Camara F, Duharcourt S, Guigo R, Gogendeau D, Katinka M, Keller AM, Kissmehl R, Klotz C, Koll F, Le Mouel A, Lepere G, Malinsky S, Nowacki M, Nowak JK, Plattner H, Poulain J, Ruiz F, Serrano V, Zagulski M, Dessen P, Betermier M, Weissenbach J, Scarpelli C, Schachter V, Sperling L, Meyer E, Cohen J, Wincker P (2006). Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* **444**, 171–178.
- Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, Rieseberg LH (2008). Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol Biol Evol* **25**, 2445–2455.
- Blanc G, Wolfe KH (2004). Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* **16**, 1679–1691.
- Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y (2006). The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol* **7**, R43.
- Bowers JE, Chapman BA, Rong J, Paterson AH (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**, 433–438.
- Chain FJ, Ilieva D, Evans BJ (2008). Duplicate gene evolution and expression in the wake of vertebrate allopolyploidization. *BMC Evol Biol* **8**, 43–43.
- Chain FJJ, Evans BJ (2006). Multiple mechanisms promote the retained expression of gene duplicates in the tetraploid frog *Xenopus laevis*. *PLoS Genet* **2**, 478–490.
- Chapman BA, Bowers JE, Feltus FA, Paterson AH (2006). Buffering of crucial functions by paleologous duplicated genes may contribute cyclicity to angiosperm genome duplication. *Proc Natl Acad Sci USA* **103**, 2730–2735.
- Comai L (2005). The advantages and disadvantages of being polyploid. *Nat Rev Genet* **6**, 836–846.
- Comai L, Tyagi AP, Winter K, Holmes-Davis R, Reynolds SH, Stevens Y, Byers B (2000). Phenotypic instability and rapid gene silencing in newly formed Arabidopsis allotetraploids. *Plant Cell* **12**, 1551–1568.
- Conant GC, Wagner A (2003). Asymmetric sequence divergence of duplicate genes. *Genome Res* **13**, 2052–2058.
- Davis JC, Petrov DA (2004). Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol* **2**, e55.
- Ermolaeva MD, Wu M, Eisen JA, Salzberg SL (2003). The age of the *Arabidopsis thaliana* genome duplication. *Plant Mol Biol* **51**, 859–866.
- Ferris SD, Whitt GS (1979). Evolution of the differential regulation of duplicate genes after polyploidization. *J Mol Evol* **12**, 267–317.
- Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545.
- Gaut BS, Le Thierry d'Ennequin M, Peek AS, Sawkins MC (2000). Maize as a model for the evolution of plant nuclear genomes. *Proc Natl Acad Sci USA* **97**, 7008–

- 7015.
- Gayral P, Caminade P, Boursot P, Galtier N** (2007). The evolutionary fate of recently duplicated retrogenes in mice. *J Evol Biol* **20**, 617–626.
- Gu X** (2003). Evolution of duplicate genes versus genetic robustness against null mutations. *Trends Genet* **19**, 354–356.
- Gu Z, Nicolae D, Lu HH, Li WH** (2002). Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet* **18**, 609–613.
- Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH** (2003). Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**, 63–66.
- He X, Zhang J** (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**, 1157–1164.
- Hughes AL** (1994). The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* **256**, 119–124.
- Hughes MK, Hughes AL** (1993). Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Mol Biol Evol* **10**, 1360–1369.
- Huminiecki L, Wolfe KH** (2004). Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res* **14**, 1870–1879.
- Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Huguency P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyere C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lechardy A, Scarpelli C, Artiguenave F, Pe ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quetier F, Wincker P** (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467.
- Jordan IK, Wolf YI, Koonin EV** (2004). Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol Biol* **4**, 22–22.
- Kellis M, Birren BW, Lander ES** (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**, 617–624.
- Kimura M** (1983). The Neutral Theory of Molecular Evolution. Cambridge: Cambridge University Press.
- Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV** (2002). Selection in the evolution of gene duplications. *Genome Biol* **3**, research0008.1–0008.9.
- Ku HM, Vision T, Liu J, Tanksley SD** (2000). Comparing sequenced segments of the tomato and Arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proc Natl Acad Sci USA* **97**, 9121–9126.
- Lee HS, Chen ZJ** (2001). Protein-coding genes are epigenetically regulated in Arabidopsis polyploids. *Proc Natl Acad Sci USA* **98**, 6753–6758.
- Li L, Huang Y, Xia X, Sun Z** (2006). Preferential duplication in the sparse part of yeast protein interaction network. *Mol Biol Evol* **23**, 2467–2473.
- Li WH, Yang J, Gu X** (2005). Expression divergence between duplicate genes. *Trends Genet* **21**, 602–607.
- Liu B, Wendel JF** (2003). Epigenetic phenomena and the evolution of plant allopolyploids. *Mol Phylogenet Evol* **29**, 365–379.
- Lynch M, Conery JS** (2000). The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155.
- Lynch M, Force A** (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**, 459–473.
- Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y** (2005). Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci USA* **102**, 5454–5459.
- Makova KD, Li WH** (2003). Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res* **13**, 1638–1645.
- Masterson J** (1994). Stomatal size in fossil plants—evidence for polyploidy in majority of angiosperms. *Science* **264**, 421–424.
- Moore RC, Purugganan MD** (2005). The evolutionary dynamics of plant duplicate genes. *Curr Opin Plant Biol* **8**, 122–128.
- Ohno S** (1970). Evolution by Gene Duplication. New York: Springer-Verlag.
- Osborn TC, Pires JC, Birchler JA, Auger DL, Chen ZJ, Lee HS, Comai L, Madlung A, Doerge RW, Colot V, Martienssen RA** (2003). Understanding mechanisms of novel gene expression in polyploids. *Trends Genet* **19**, 141–147.
- Papp B, Pal C, Hurst LD** (2003). Evolution of cis-regulatory elements in duplicated genes of yeast. *Trends Genet* **19**,



- 417–422.
- Paterson AH, Bowers JE, Chapman BA** (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci USA* **101**, 9903–9908.
- Paterson AH, Chapman BA, Kissinger JC, Bowers JE, Feltus FA, Estill JC** (2006). Many gene and domain families have convergent fates following independent whole-genome duplication events in *Arabidopsis*, *Oryza*, *Saccharomyces* and Tetraodon. *Trends Genet* **22**, 597–602.
- Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberler G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Ollillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboobur R, Ware D, Westhoff P, Mayer KF, Messing J, Rokhsar DS** (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556.
- Qian W, Zhang J** (2008). Gene dosage and gene duplicability. *Genetics* **179**, 2319–2324.
- Rastogi S, Liberles DA** (2005). Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol Biol* **5**, 28–28.
- Richards EJ, Elgin SC** (2002). Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects. *Cell* **108**, 489–500.
- Robinson-Rechavi M, Laudet V** (2001). Evolutionary rates of duplicate genes in fish and mammals. *Mol Biol Evol* **18**, 681–683.
- Rosenberg HF, Dyer KD** (1995). Eosinophil cationic protein and eosinophil-derived neurotoxin. Evolution of novel function in a primate ribonuclease gene family. *J Biol Chem* **270**, 21539–21544.
- Semon M, Wolfe KH** (2007). Consequences of genome duplication. *Curr Opin Genet Dev* **17**, 505–512.
- Seoighe C, Wolfe KH** (1999). Yeast genome evolution in the post-genome era. *Curr Opin Microbiol* **2**, 548–554.
- Soltis DE, Soltis PS, Pires JC, Kovarik A, Tate JA, Mavrodiev E** (2004). Recent and recurrent polyploidy in Tragopogon (Asteraceae): cytogenetic, genomic and genetic comparisons. *Biol J Linn Soc Lond* **82**, 485–501.
- Steinke D, Salzburger W, Braasch I, Meyer A** (2006). Many genes in fish have species-specific asymmetric rates of molecular evolution. *BMC Genomics* **7**, 20–20.
- Thomas BC, Pedersen B, Freeling M** (2006). Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res* **16**, 934–946.
- Tocchini-Valentini GD, Fruscoloni P, Tocchini-Valentini GP** (2005). Structure, function, and evolution of the tRNA endonucleases of Archaea: an example of subfunctionalization. *Proc Natl Acad Sci USA* **102**, 8933–8938.
- Van de Peer Y, Taylor JS, Braasch I, Meyer A** (2001). The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. *J Mol Evol* **53**, 436–446.
- Wagner A** (1998). The fate of duplicated genes: loss or new function? *Bioessays* **20**, 785–788.
- Wang J, Tian L, Madlung A, Lee HS, Chen M, Lee JJ, Watson B, Kagochi T, Comai L, Chen ZJ** (2004). Stochastic and epigenetic changes of gene expression in Arabidopsis polyploids. *Genetics* **167**, 1961–1973.
- Wang X, Shi X, Hao B, Ge S, Luo J** (2005). Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytol* **165**, 937–946.
- Wendel JF** (2000). Genome evolution in polyploids. *Plant Mol Biol* **42**, 225–249.
- Wu Y, Zhu Z, Ma L, Chen M** (2008). The preferential retention of starch synthesis genes reveals the impact of whole genome duplication on grass evolution. *Mol Biol Evol* **25**, 1003–1006.
- Yang X, Tuskan GA, Cheng MZ** (2006). Divergence of the *dof* gene families in poplar, Arabidopsis, and rice suggests multiple modes of gene evolution after duplication. *Plant Physiol* **142**, 820–830.
- Yu J, Wang JW, Lin S, Li H, Li J, Zhou P, Ni W, Dong S, Hu C, Zeng J, Zhang Y, Zhang R, Li Z, Xu S, Li X, Li H, Zheng L, Cong L, Lin J, Yin J, Geng G, Li J, Shi J, Liu H, Lv J, Li J, Wang Y, Deng L, Ran X, Shi X, Wang Q, Wu C, Li X, Ren J, Wang X, Wang D, Li D, Liu X, Zhang Z, Ji W, Zhao Y, Sun Z, Zhang J, Bao Y, Han L, Dong J, Ji P, Chen S, Wu J, Liu Y, Xiao D, Bu J, Tan L, Yang C, Ye J, Zhang J, Xu Y, Zhou Y, Yu B, Zhang S, Zhuang H, Wei B, Liu M, Lei H, Yu Y, Li H, Xu S, Wei X, He L, Fang Z, Zhang Y, Zhang X, Huang Z, Su W, Tong J, Li Z, Tong S, Li J, Ye L, Wang L, Fang T, Lei C, Chen H, Chen Z, Xu H, Li H, Huang F, Zhang H, Xu N, Li C, Zhao S, Li L, Dong Y, Huang L, Li Y, Xi Q, Qi W, Li B, Zhang W, Hu Y, Zhang X, Tian Y, Jiao X, Liang J, Jin L, Gao W, Zheng B, Hao S, Liu W, Wang L, Yuan M, Cao**

- J, McDermott R, Samudrala J, Wang G, Wong K, Yang H** (2005). The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol* **3**, e38.
- Zhang J** (2003). Evolution by gene duplication: an update. *Trends Ecol Evol* **18**, 292–298.
- Zhang J, Rosenberg HF, Nei M** (1998). Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci USA* **95**, 3708–3713.
- Zhang P, Gu Z, Li WH** (2003). Different evolutionary patterns between young duplicate genes in the human genome. *Genome Biol* **4**, R56.
- Zhang Z, Gu J, Gu X** (2004). How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution? *Trends Genet* **20**, 403–407.

## Review of the Evolution of Duplicated Genes

Hongzheng Sun<sup>1,2</sup>, Song Ge<sup>1\*</sup>

<sup>1</sup>State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China; <sup>2</sup>Key Laboratory of Crop Germplasm and Biotechnology, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China

**Abstract** Gene duplication is universal across all organisms and plays a key role in driving the evolution of genomes and genetic systems. The fate and mechanism of gene duplication is hotly debated and involves the advantages and disadvantages of gene duplicates, retention bias of daughter copies, divergence of evolution rates and expression patterns, and factors that determine the preservation of duplicated genes. Here, we briefly introduce advances in research into gene duplication and discuss the main issues.

**Key words** evolution, gene duplication, genome

**Sun HZ, Ge S** (2010). Review of the evolution of duplicated genes. *Chin Bull Bot* **45**, 13–22.

---

\* Author for correspondence. E-mail: gesong@ibcas.ac.cn