

F-box基因拷贝数目变异的机制研究： 以12种果蝇为例

李 安^{1,2} 徐桂霞¹ 孔宏智^{1*}

1 (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)

2 (中国科学院研究生院, 北京 100049)

摘要: 拷贝数目变异是一种对表型变异和生物进化具有重要意义的基因组结构变异。以前的研究表明不同物种中F-box基因的拷贝数目差异较大。为了深入探索拷贝数目变异的式样和机制, 我们以12个果蝇近缘种为研究对象, 分析了F-box基因的系统发育关系、进化式样以及它们在染色体上的位置。结果发现, 虽然各个物种中F-box基因的拷贝数目差别不大(42–47个), 但是仍然存在着很多引起拷贝数目变异的基因获得和丢失事件。这说明表面上变化不大的拷贝数目在一定程度上掩盖了频繁发生的基因获得和丢失事件。通过比较这些基因在染色体上的位置, 发现只有在亲缘关系很近的物种之间才能鉴定出有明显微共线性关系的基因组区段。我们还发现, 造成F-box基因拷贝数目增加的主要机制是散在重复和串联重复, 而反转录转座和新基因的非编码区起源也是两种值得注意的机制。此外, 序列变异导致的外显子边界变化以及外显子丢失是引起拷贝数目减少的两种机制。在12种果蝇的最近共同祖先中, F-box基因的拷贝数目与现存物种基本相似, 但是基因的获得和丢失事件使得现存物种中的F-box基因在构成上已经有了明显的差别。对数目变异的式样及其与基因功能的关系的研究表明, 拷贝数目变异是F-box基因家族“生与死”的进化在基因组层面的系统反映, 并有可能为表型变异提供了原始材料。

关键词: 拷贝数目变异, F-box基因, 果蝇, 直系同源基因

Mechanisms underlying copy number variation in F-box genes: evidence from comparison of 12 *Drosophila* species

An Li^{1, 2}, Guixia Xu¹, Hongzhi Kong^{1*}

1 State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093

2 Graduate University of the Chinese Academy of Sciences, Beijing 100049

Abstract: Copy number variation (CNV) is a special type of mutational change that plays important roles in phenotypic variation and organismal evolution. To explore the mechanisms underlying copy number variation and to understand its biological significance, we analyzed the phylogenetic relationships, evolutionary patterns and chromosomal locations of F-box genes in 12 closely-related *Drosophila* species. A total of 541 F-box genes were identified and phylogenetic analyses suggested that they are members of 48 gene clusters (or orthologous groups). Although we observed no drastic changes in the total numbers of F-box genes among the 12 extant *Drosophila* species (42–47), we found many gene gain and loss events that have caused copy number variation. These results demonstrated that the similarity in the total numbers of F-box genes among different species has, to a certain degree, masked the frequent and independent gain and loss events. Comparisons of the chromosomal locations of orthologous genes showed that extensive microsynteny could be detected only between very closely-related sibling species. We also found that the main mechanisms that caused the increase in gene number were dispersed duplication and tandem duplication, while retroposition and *de novo* origination from non-coding sequences were two other noteworthy mechanisms. Mutations that

caused shifts in exon-intron boundaries and/or losses of exons seemed to be the main mechanisms that underlie decreases in copy number. Although the most recent common ancestor (MRCA) of the 12 *Drosophila* species had a similar number of genes as the extant species we studied, gains of new genes and losses of existing ones have caused changes in the makeup of F-box genes in descendent species. Our study suggested that variations in the numbers of gene copies is a reflection of “birth-and-death” evolution at the genomic level and have provided raw materials for phenotypic and physiological diversification.

Key words: copy number variation, F-box gene, *Drosophila*, orthologs

变异是生命的基本特征, 是形成生物多样性的前提, 并为生物的进化提供原始材料。变异可以发生在多个水平上, 但发生在基因组层面的变化最为重要。在基因组层面上, 变异既表现为单个碱基的替换、插入和缺失, 又表现为基因组(或者染色体)结构的变化。基因拷贝数目变异(copy number variation, CNV), 或称拷贝数目多态性(copy number polymorphism, CNP), 是一种特殊的变异类型, 是指基因组在亚显微(submicroscopic)水平上的DNA片段多态性(Redon *et al.*, 2006; Beckmann *et al.*, 2007)。已有研究表明, 在真核生物中, 拷贝数目变异发生的频率远远高于其他类型的染色体结构变异, 在基因组中所覆盖的核苷酸总数也大大超过了单核苷酸多态性(single nucleotide polymorphisms, SNP)所覆盖的总数(Redon *et al.*, 2006; Hinds *et al.*, 2006); 这表明拷贝数目变异可能在遗传多样性、表型多样性以及系统演化中发挥着重要作用。

2006年11月23日, 由多个国家的研究人员组成的研究小组在 *Nature* 杂志上公布了人类基因组第一代拷贝数目的变异图谱(Redon *et al.*, 2006), 获得了1,447个拷贝数目变异区域(copy number variable regions, CNVRs), 覆盖区域约为360 Mb(占整个基因组的12%)。这些拷贝数目变异位点包含了数百个功能基因、多个致病基因座位及功能调控因子。这一成果是对基因组变异认知的一个里程碑。此后, 关于拷贝数目变异的研究在其他物种的种内和种间展开, 取得一系列新的进展, 成为当前国际上的研究热点和学术前沿。

拷贝数目变异的研究在动物(尤其是人类)基因组中的研究最为深入。研究发现, 40%的拷贝数目变异位于基因沙漠区(gene deserts)(Derti *et al.*, 2006)。存在拷贝数目变异的基因常影响人体对外界环境的反应, 在细胞连接、感觉认知、化学刺激和神经生理等过程中发挥重要作用。不存在拷贝数目变异

的基因往往是剂量敏感(dosage sensitive)基因, 参与维持细胞的生长发育, 包括细胞信号传导、增殖、激酶化和磷酸化等保守过程(Redon *et al.*, 2006)。在拷贝数目变异与自然选择的关系方面的研究也已取得丰硕成果。例如, 对唾液淀粉酶基因(salivary amylase gene, 即*AMY1*)拷贝数目多态性的研究显示, *AMY1*基因的数目在以淀粉为主食的人群和低淀粉摄入的人群间存在显著差异, 基因的拷贝数目与其在唾液腺中的表达水平呈显著正相关。由此推断, *AMY1*在进化过程中可能承受着强的选择压力, 拷贝数目的增加提高了基因的表达水平(Perry *et al.*, 2007)。但也有研究者认为, 拷贝数目变异是由突变和遗传漂变决定的, 其在基因组中的分布是随机的, 并不是选择的结果(Sharp *et al.*, 2005; Small *et al.*, 2007)。而且, 另有一些研究表明, 拷贝数目变异在大多数情况下并不引起适合度的不同(Niimura & Nei, 2006; Nozawa & Nei, 2007), 是随机的基因组漂变(genomic drift)的结果(Dharmasiri *et al.*, 2005; Niimura & Nei, 2007)。因此, 尽管人们对拷贝数目变异的式样已经有了较多研究, 但对导致拷贝数目变异的机制和生物学意义还不清楚, 更多更细致的研究尚需开展。

F-box蛋白广泛存在于真核生物中, 是细胞信号传导、转录调节以及细胞周期等许多生理活动的关键调节蛋白。除了N端的F-box结构域之外, F-box蛋白一般还在C端包括一些与蛋白相互作用密切相关的二级结构(如LRR、WD40和Kelch等结构域的重复序列等), 负责对底物的特异性识别(Kipreos & Pagano, 2000)。已有的研究表明, 不同物种间F-box基因的数目差异很大(Bai *et al.*, 1994; Kipreos & Pagano, 2000)。例如, 在出芽酵母(*Saccharomyces cerevisiae*)、线虫(*Caenorhabditis elegans*)、黑腹果蝇(*Drosophila melanogaster*)和人类中, F-box基因的数目分别为14、337、24和38个(Kipreos & Pagano,

2000)。在植物中, F-box基因的数目更多, 可达七、八百个(如拟南芥有692个, 水稻有779个), 而且最近的研究表明许多基因类型实际上是在很短的时间内获得的(Xu *et al.*, 2009)。因此, 对F-box类基因的进化研究, 不仅有助于阐明该类基因本身的进化历程和变化规律, 而且有望在一定程度上揭示导致拷贝数目变异产生的机制。但是, 由于已有研究中所涉及的物种都具有较远的系统发育关系, 它们在进化过程中往往会经历一些复杂的进化事件(如基因组加倍等), 从而使得研究这类问题需要考虑多种因素。为更准确地弄清拷贝数目变异的机制, 我们需要从较小的进化尺度上去了解更多的细节问题。因此, 我们需要选择更近缘的物种进行研究。

果蝇属12个近缘物种的全基因组测序和注释工作的相继完成, 为我们开展这一方面的研究提供了非常难得的机会(*Drosophila 12 Genomes Consortium*, 2007; Stark *et al.*, 2007)。此外, 由于这些物种在亲缘关系、行为方式和地理分布等多方面均存在着不同程度的差异, 我们还可以通过基因组水平上的比较分析来研究近缘种间拷贝数目的差异与上述物种分化的关系以及造成这种差异的可能机制。在本研究中, 通过对12个果蝇近缘种中的F-box基因的系统发育关系、进化式样和在染色体上的位置进行的分析, 我们发现虽然各个物种中F-box基因的总数差别很小, 但实际上却在进化过程中经历了多次的基因获得和丢失事件; 其中导致基因获得的机制有串联重复(tandem duplication)、散在重复(dispersed duplication)、反转录转座(retroposition)和基因的从头起源(*de novo origination*), 而导致基因丢失的机制则有外显子-内含子的边界变化和外显子丢失。果蝇近缘种中F-box基因的拷贝数目变异可能为表型变异提供了原始材料。

1 材料与方法

1.1 F-box蛋白质序列的获取和鉴定

本文用到的12个果蝇近缘种的基因组和蛋白质序列从FlyBase数据库(<http://flybase.org>)下载获得。然后, 利用HMMER 2.3.2软件包(Eddy, 1998)中的Hmmssearch程序, 获得各个物种中含有F-box结构域(PF00646)的蛋白序列, 得到最初的序列库。其中, 搜索过程中所用到的F-box结构域的隐马尔可夫模型(Hidden Markov Model, HMM)文件(Bateman

et al., 2002)从Pfam(<http://pfam.sanger.ac.uk/>)网站上下载得到。在所得序列库中, 有一些基因在DNA水平上的相似度超过了95%, 如果这些基因在染色体上的座位(locus)不同, 则全部保留; 如果相同, 则说明它们是等位基因或是由选择性剪切(alternative splicing)产生的不同转录本, 我们仅保留有表达序列标签(expressed sequence tag, EST)支持或其mRNA序列与直系同源基因(orthologs)更相似的一个序列(附录I)。在剔除了冗余序列后, 利用SMART(<http://smart.embl-heidelberg.de/>)和Pfam网站上的结构域检索工具对每一条序列所含有的结构域进行进一步的确认。对于两种工具均未检测到F-box结构域的蛋白质, 如果其序列与其他F-box基因的一致性不高, 则直接从最初的序列库中删除; 如果其序列与其他F-box基因具有较高的一致性, 则利用Clustal X 1.83 (Thompson *et al.*, 1997)、GeneDoc v.2.6.002 (Nicholas & Nicholas, 1997)、DNAMAN version 6.0和aa2dna (<http://www.bio.psu.edu/People/Faculty/Nei/Lab/software.htm>)等工具, 综合考虑其基因、蛋白质、EST及上下游序列等信息进行重新注释, 进一步进行结构域分析。若其具有F-box结构域, 则保留, 并加入已有的序列库中(附录II)。

1.2 序列比对和系统发育分析

利用Clustal X 1.83软件(Thompson *et al.*, 1997)对所得到的F-box蛋白进行全局序列比对, 比对参数选择默认值。此时得到的矩阵是依据F-box蛋白序列全长范围内的一致性而产生的。由于F-box蛋白除了F-box结构域之外其他部分的序列差异较大, 我们还利用HMMER 2.3.2软件包(Eddy, 1998)中的Hmmalign程序仅对F-box结构域进行了比对。

分别用全长的F-box蛋白序列和F-box结构域中的位点构建系统发育树。其中, 前者是利用ClustalX 1.83软件构建的邻接(neighbor joining, NJ)树, 邻接树的支持率以自展(bootstrap)法获得, 重复取样1,000次。在利用F-box结构域的位点进行系统发育分析时, 用MEGA 4.0软件(Tamura *et al.*, 2007)构建了NJ树, 用PHYML 2.4软件(Guindon & Gascuel, 2003)构建了最大似然(maximum likelihood, ML)树。构建邻接树时使用的是p-距离(p-distance)模型, 并选择了成对删除(pairwise deletion)空位(gap)的选项。邻接树的支持率以自展法获得, 重复取样1,000

次。在构建最大似然树时,用Modeltest v.3.8软件(Posada & Crandall, 1998)估测出适合该序列矩阵的模型(即HKY模型和gamma参数),然后构建无根树,最后再以自展法100次重复取样得到了各个分支的支持率。

1.3 拷贝数目变异的分析

在系统发育树上,每个由来自果蝇不同物种的基因聚集在一起形成的进化支客观上对应于一组直系同源基因,将每组直系同源基因依次顺序编号为Clade1, Clade2, …… , Clade48。分别统计各组F-box基因在各物种中的拷贝数目。将所有F-box基因分别按照物种、进化支和C末端的结构域进行分类并比较拷贝数目。对有拷贝数目增加的基因,首先利用其在染色体上的位置信息,结合基因之间的系统发育关系,判断它们是不是通过串联重复产生的。如果不是,则选取基因自上游到下游共100 kb的序列,利用DOTLET(Junier & Pagni, 2000)程序进行Dotplot分析,查看增加的拷贝与原拷贝是否位于重复片段上,也就是判断它们是否是通过片段重复产生的。如若两者都不是,则进一步比较不同拷贝的基因结构是否符合反转录转座的特征。如果不符合,则将其归结为通过散在重复产生的。对拷贝数目减少的基因,则在缺失的物种数据库中重新进行同源序列的检索,如果能找到与该类基因一致性较高的区域,则进一步将该区域翻译为蛋白质序列,看其是否本身就没有F-box结构域,还是因突变而导致了F-box结构域没有正常翻译。另外,在研究中,需要频繁选择直系同源基因作为外类群进行系统发育分析及序列比较,从而判断变异的极性。

进行基因的共线性比较时,先利用基因组信息,将所有F-box基因定位于染色体上,从而直观地比较直系同源基因之间以及旁系同源基因(paralogs)之间的关系。我们以基因组拼装得最好且研究得较成熟的*D. melanogaster*的染色体为基准来定位所有F-box基因(Reese *et al.*, 2000; Celniker *et al.*, 2002; <http://flybase.org>)。首先将*D. melanogaster*中所有F-box基因从各条染色体的左臂到右臂依次标定(Nozawa & Nei, 2007),再将其他物种中与*D. melanogaster*对应的染色体片段依次排列(参考FlyBase中的保守基因区段的共线性,即synteny blocks的资料),然后将各自的F-box基因定位在染色体片段上。由于共线性不佳的区域相当多,为了

增加可读性,我们将部分染色体片段倒置进行比较。被倒置的染色体片段分别为:*D. ananassae*的C(对应*D. melanogaster*的2R)和D(对应*D. melanogaster*的3L); *D. pseudoobscura*的B(对应*D. melanogaster*的2L)、C(对应*D. melanogaster*的2R)和A/D(对应*D. melanogaster*的3L); *D. persimilis*的B(对应*D. melanogaster*的2L)、C(对应*D. melanogaster*的2R)和A/D(对应*D. melanogaster*的3L); *D. grimshawi*的B(对应*D. melanogaster*的2L)和C(对应*D. melanogaster*的2R); *D. mojavensis*的B(对应*D. melanogaster*的2L)和C(对应*D. melanogaster*的2R); *D. virilis*的C(对应*D. melanogaster*的2R)和E(对应*D. melanogaster*的3R)。

2 结果

2.1 果蝇近缘种中F-box基因的拷贝数目

我们从果蝇的12个近缘种基因组中鉴定出了541个F-box基因,并利用其编码蛋白的F-box区域和蛋白质全长分别重建了这些基因之间的系统发育关系。结果显示,用蛋白质全长序列构建的系统树与用F-box结构域构建的系统树在拓扑结构上有一定的不同,但在直系同源基因的分组上十分相似。由于前者的自展值普遍较高,所以本研究中仅选用了用全长序列构建的系统树(附录III)。此外,由于核苷酸水平上的各种变化(如外显子区的插入/缺失以及外显子-内含子结构变化等),我们构建的系统树有时也并非完全合理;在这种情况下,核苷酸序列被用来作进一步的分析。例如,核苷酸序列比对的结果表明, *pseFBpp0275460*基因(为了描述方便,用每个果蝇物种拉丁名中种加词的前3个字母代表各个物种,下同)与第23个进化支(Clade23)中 *pseFBpp0274105*基因的关系较近,因而在随后的分析中将其归入Clade23。同理, Clade25中的 *virFBpp0228425* 和 *wilFBpp0244199* 应分别归入Clade11和Clade17。经分析发现,12个物种中的F-box基因并没有聚集成物种特异的分支,而是分散于48个分支中,每个分支由2-16个基因组成。各个物种中F-box基因的拷贝数目是不同的,变异范围为42-47。在48个分支中,拷贝数目在物种间没有差异的为29个,占60.4%。在其余的19个(占总数的39.6%)分支中,基因的拷贝数目在物种间有差异(图1);这说明拷贝数目在近缘种间有差异的F-box

基因占相当高的比例。

除了N端的F-box结构域, 在我们鉴定出的541个F-box基因所编码的蛋白质中, 其C末端的结构域不仅包括已知的LRR、WD40、FTH、Kelch、DUF590、SPRY、IBR、UBCc或Znf_UBR1等类型, 还有一些未知或者未被定义的类型(图1)。系统发育分析的结果显示, 虽然也有例外, 但绝大多数结构相似的F-box蛋白还是聚在了一起(附录III)。在48个分支中, C端无已知结构域的F-box基因最多, 有24支, 其中拷贝数目发生变化的占12支(50.0%); 说明这类基

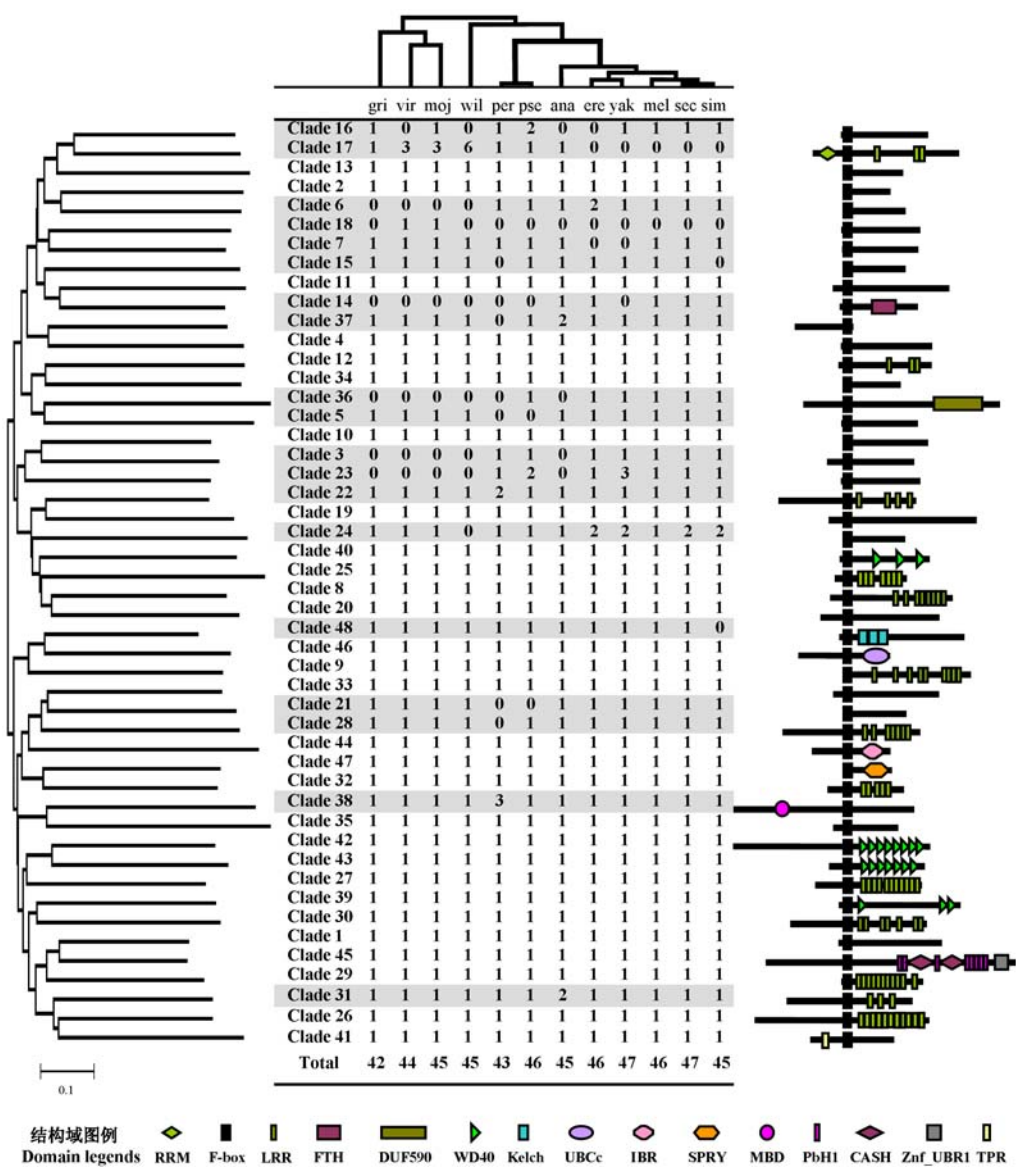


图1 果蝇12个近缘种的F-box基因的拷贝数目和对应的结构域组成。统计表的上方是果蝇12个近缘种的物种树(修改自<http://rana.lbl.gov/drosophila/index.html>), 分支末端对应各自的物种名(用种名的前三个字母代替), 即拷贝数目统计表的第一行。统计表左侧是用48个直系同源基因分支中的代表序列构建的系统树, 各支末端对应统计表中该基因在12个物种中的拷贝数目; 右侧则为该基因编码的蛋白质的结构域组成模式图。灰色标出的行是有拷贝数目变异的基因。

Fig. 1 Copy number and domain structure of F-box proteins from 12 *Drosophila* species. The phylogenetic tree of the 12 *Drosophila* species (modified from <http://rana.lbl.gov/drosophila/index.html>) is above the table. The first three letters of each specific epithet are used as the abbreviation for each species. On the left side of the table is a neighbor-joining (NJ) tree for 48 clades of F-box proteins whose domain structure are shown on the right. Numbers in the table indicate the copy numbers of F-box genes belonging to each clade. Clades with copy number variations are shaded.

因的拷贝数目最不稳定。另外一类相对不保守的F-box基因在C端含有LRR结构域,共有13支,拷贝数目发生变化的分支有4个(30.8%)。其他11个分支的F-box基因在进化上高度保守,拷贝数目在12个果蝇近缘种中没有变化。这些结果表明,F-box基因的保守程度与其编码蛋白的C端结构域类型具有很明显的相关性。

2.2 F-box基因拷贝数目变异的式样

拷贝数目的变化表明不同的果蝇物种在进化中获得了新的基因或者丢失了已有基因。为了进一步了解拷贝数目变化的程度和过程,我们重建了上述48个分支中F-box基因的进化历史,推断了基因获得或者丢失事件发生的分支及次数。我们发现,有的基因只在某一个物种中获得或丢失了一个拷贝,而有的基因则经历了较为复杂的进化历史。例如,Clade31的基因在*D. ananassae*中有2个拷贝,但在其余11个物种中只有1个拷贝(图2);系统发育分析的结果表明该基因在进化中经历了物种特异的基因重复事件。又如,Clade7的基因在10个物种中存在,但在*D. yakuba*和*D. erecta*两个物种中丢失(图2);这说明基因的丢失发生在这两个物种分化之前。当然,还有更为复杂的例子。如Clade24的基因在果蝇这12个物种中的分布很不均匀,拷贝数目从0到2不等(图2),经分析表明,该基因很可能在这些物种的最近共同祖先(the most recent common ancestor, MRCA)中存在,但在*D. willistoni*中丢失,并在*D. sechellia*、*D. simulans*、*D. melanogaster*、*D. yakuba*和*D. erecta*等5个物种的最近共同祖先中发生了一次基因重复,从而获得了一个新拷贝,但随后*D. melanogaster*又丢失了一个基因拷贝(图2)。

在48个F-box基因分支中,我们发现6个分支(Clade3、Clade6、Clade14、Clade23、Clade18和Clade36)基因的起源时间存在不确定性(附录III)。其中,Clade6、Clade36、Clade3和Clade23基因存在于物种树上部8个物种(即*melanogaster*组和*obscura*组)构成的单系分支中(附录III),Clade14基因在上部6个物种构成的单系分支中除*D. yakuba*以外的其余物种中存在(图2),Clade18基因仅存在于*D. mojavensis*和*D. virilis*构成的单系分支中(图2)。如果这6个分支的基因在祖先物种中是存在的,则在多个物种中发生了独立的基因丢失事件;如果它们在祖

先物种中不存在,则是在后面的物种分化过程中新产生的。例如,假定Clade14基因在祖先物种中存在,则在基部3个物种(即*Drosophila*亚属)的最近共同祖先、*D. pseudoobscura*和*D. persimilis*的最近共同祖先、*D. willistoni*以及*D. yakuba*中分别丢失,需要假设的基因增减事件为4次;如果其在祖先物种中不存在,则是在物种树上部6个物种中获得的,随后又在*D. yakuba*中丢失,需要假设的基因增减事件为2次(图2)。由此可见,后者更符合简约原则。由于其余5个分支的基因都是存在于由2个、6个或8个物种分别构成的单系类群中,因而,按照后期起源的假说进行解释都是最简约的。因此,我们倾向于认为这6个分支的基因在祖先物种中都不存在,也就是说祖先物种具有42个F-box基因。

经过上述分析,我们对F-box基因在12个果蝇近缘种中的进化历史作如下推断:Clade6、Clade36、Clade3和Clade23基因是在物种树上部8个物种即*melanogaster*组和*obscura*组(Singh *et al.*, 2009)的最近共同祖先中获得的,因此此处拷贝数目的增减事件为+4/-0,此节点的拷贝数目为46(图3)。Clade14基因是在物种树上部6个物种的最近共同祖先中获得,因此此处增减事件为+1/-0,此节点的拷贝数目为47。*D. ananassae*分别获得Clade37基因、Clade31基因的一个新拷贝,而丢失了Clade16基因、Clade23基因、Clade3基因和Clade36基因,因此增减事件为+2/-4(图3)。由于物种树上部5个物种的最近共同祖先获得了Clade24的新拷贝,而Clade17基因在其最近共同祖先中丢失,因此此处增减事件为+1/-1,此节点的拷贝数目为47(图3)。同理,可以推断出其他节点和现存物种中的拷贝数目以及各进化阶段的增减事件。总体来看,与*Drosophila*亚属的3个物种相比,*Sophophora*亚属9个物种中F-box基因的拷贝数目较多,发生导致拷贝数目变异的事件也明显较多。在所有12个物种中,*D. persimilis*所经历的基因获得和丢失事件最多,分别为8次和7次,使得祖先物种的42个基因中有约1/6已不复存在,虽然F-box基因的总数仅增加了1个。

2.3 F-box基因拷贝数目变异的机制

近缘物种中同源基因的共线性研究能够揭示基因组在进化过程中伴随的各种序列重排现象。为了深入研究拷贝数目变异的机制,我们对12个果蝇近缘种中F-box基因的共线性关系进行了分析。总体

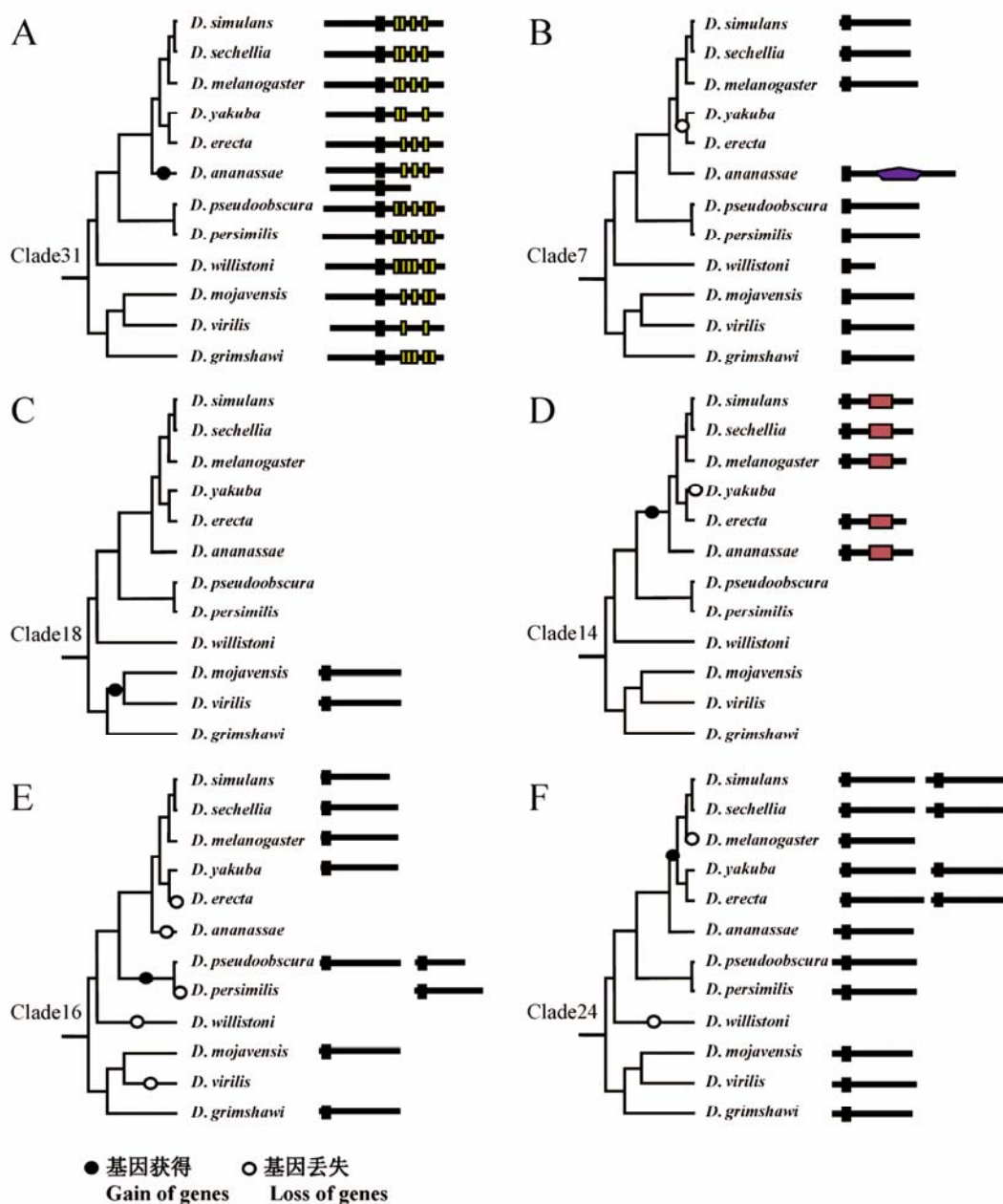


图2 F-box基因拷贝数目变异的典型情况。结构域的图例与图1的一致。A: 该基因在12个物种中均存在, 且在*D. ananassae*中又获得了一个拷贝; B: 该基因存在于10个物种中, *D. yakuba*和*D. erecta*的最近共同祖先丢失了这个基因; C: 该基因仅在*D. mojavensis*和*D. virilis*中存在, 说明此基因是在这两个物种的最近共同祖先中获得的; D: 物种树上部6个物种的最近共同祖先获得了该基因, 又在*D. yakuba*中发生丢失。E: 该基因在*D. erecta*、*D. ananassae*、*D. willistoni*和*D. virilis*这4个物种中分别丢失, 而在*D. pseudoobscura*和*D. persimilis*的最近共同祖先中发生了基因重复, 产生了第二个拷贝, 后来*D. persimilis*中的第一个拷贝丢失; F: 该基因在11个物种中存在, 在*D. willistoni*中丢失。物种树上部5个物种的最近共同祖先中增加了1个拷贝, 但*D. melanogaster*又丢掉了1个拷贝。

Fig. 2 Examples of copy number variation of F-box genes. Domain legends are the same as those in Fig.1. A, *D. ananassae* gained a new copy of F-box gene; B, The most recent common ancestor (MRCA) of *D. yakuba* and *D. erecta* lost the ortholog of Clade7; C, MRCA of *D. mojavensis* and *D. virilis* gained a new gene; D, MRCA of the top six species gained a new copy of F-box gene and *D. yakuba* lost this gene after their divergence; E, Four gene loss events occurred independently in *D. erecta*, *D. ananassae*, *D. willistoni*, and *D. virilis*. A gene duplication event happened in the MRCA of *D. pseudoobscura* and *D. persimilis*, generating two copies, but then one copy lost in *D. persimilis*; F, *D. willistoni* lost Clade24 gene. The MRCA of the top five species gained a new copy of this gene, but then *D. melanogaster* lost the new copy.

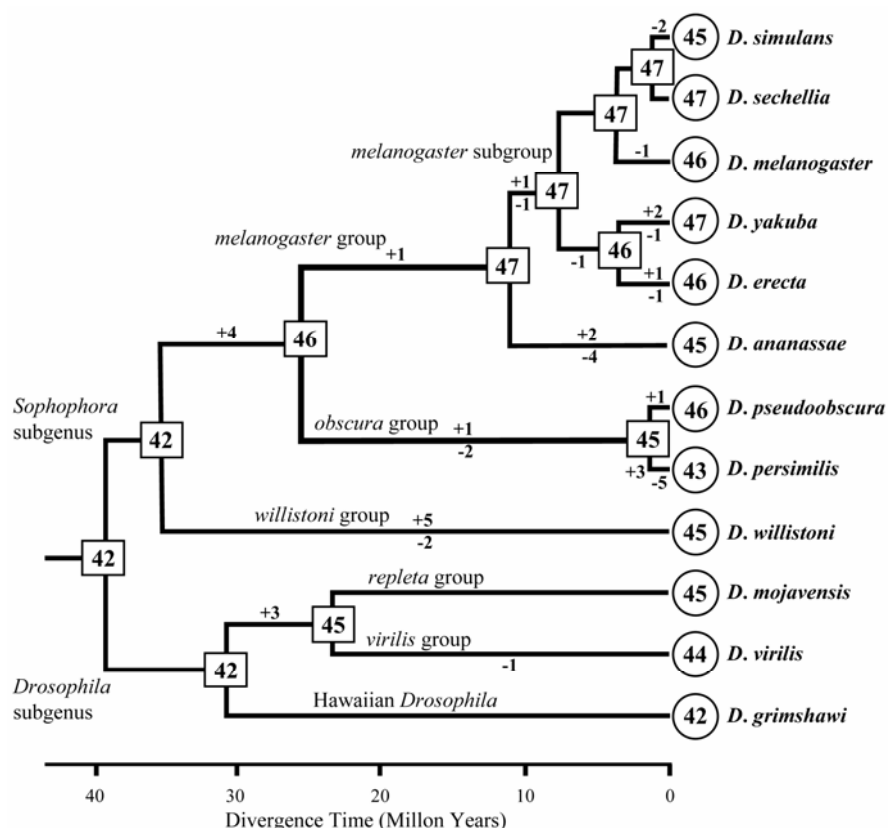


图3 F-box基因的拷贝数目在果蝇进化历史中的变化。圆圈和方框中的数字分别表示12个现存物种及其祖先中的拷贝数目。系统树分支上的+/-及数字表示基因拷贝数目的增减事件及增减的拷贝数目。系统发育树修改自Assembly/Alignment/Annotation of 12 related *Drosophila* species [http://rana.lbl.gov/drosophila/index.html]。

Fig. 3 Evolutionary change of the copy number of F-box genes in *Drosophila* species. The numbers in circles and rectangles represent the copy numbers of genes in extant and ancestral species, respectively. Numbers above and below each branch indicate the numbers of gains (+) and losses (-) of genes, respectively. The phylogenetic tree is modified from Assembly/Alignment/Annotation of 12 related *Drosophila* species [http://rana.lbl.gov/drosophila/index.html]。

来看, F-box基因在12个果蝇近缘种染色体上的排布并未呈现良好的共线性(图4)。可能是由于倒位、易位等各种序列重排和染色体片段的合并导致了果蝇的不同种间的核型(karyotype)有较大差异所致(*Drosophila* 12 Genomes Consortium, 2007; Schaeffer *et al.*, 2008)。在12个果蝇近缘种的基因组中, F-box基因广泛分布在除了点状染色体(*D. melanogaster*中为4号染色体F)之外的各条染色体上。例如, 与*D. melanogaster*的X染色体对应的各条染色体上, 除*D. persimilis*之外均有2个拷贝的F-box基因, *D. persimilis*的5个拷贝是由于两个F-box基因分别经历了一次和两次基因重复事件。与*D. melanogaster*的2号染色体左臂对应的各条染色体上, 各自分布着8-11个拷贝的F-box基因。虽然近缘种中F-box基因

的总拷贝数差别不大, 但同源拷贝之间共线性关系较差, 仅在亲缘关系很近的物种(如*obscura*组的两个物种)之间维持了良好的共线性关系。因而很难判断片段重复的贡献。在对拷贝数目变异的初步探索性研究中, 我们不研究基因组重排的影响, 只对基因组重排后仍能判断的拷贝数目变异机制进行研究。我们总结出了除片段重复之外能引起拷贝数目变异的机制。导致拷贝数目增加的机制有串联重复、散在重复、反转录转座以及新基因的非编码区起源等(附录IV), 导致拷贝数目减少的机制有点突变、插入/缺失等导致的外显子边界变化和在外显子丢失。我们把同属一个进化支且位于同一染色体区域(彼此之间的相隔少于20个基因)(Xu *et al.*, 2009)的一对序列最相似的基因推断为串联重复基因。在本

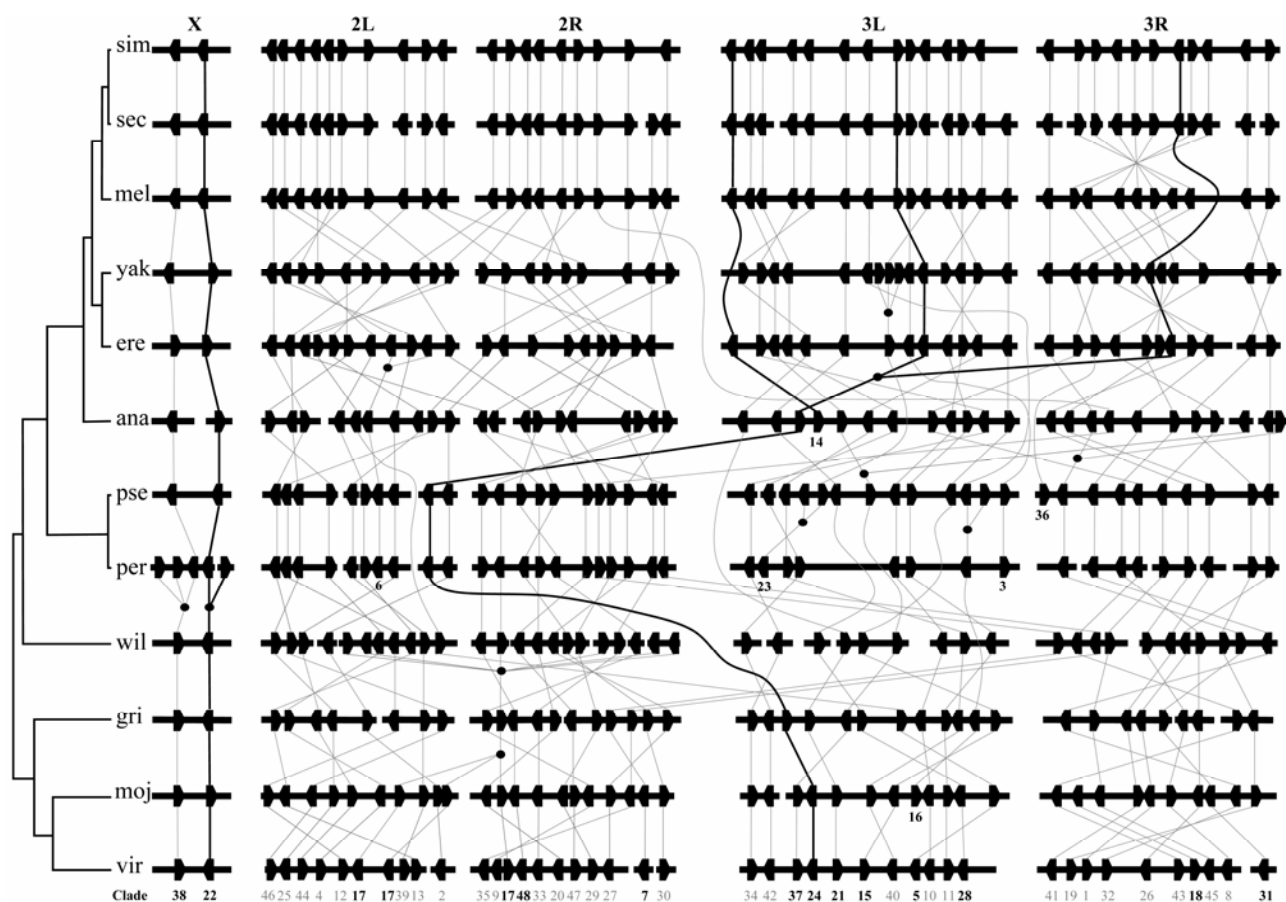


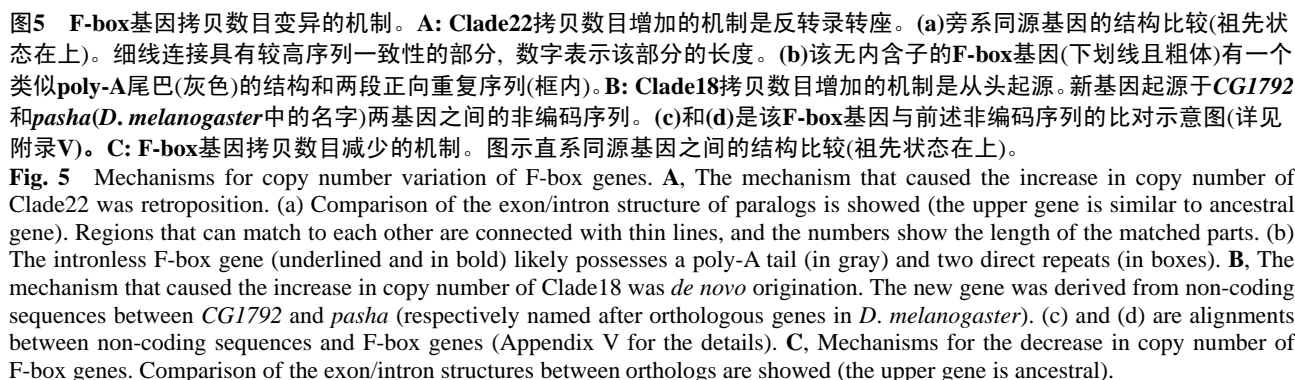
图4 12个果蝇近缘种基因组中F-box基因的染色体定位及相互之间的直系同源与旁系同源关系。横向的长条表示12个果蝇的染色体臂/支架/片段。L和R分别指示染色体的左、右臂。五边形的黑色方块表示F-box基因。直系同源基因之间用直线相连, 实心圆圈表示基因重复事件。曲线表示其所穿过的染色体上已经丢失相应的直系同源基因。基因下面的数字代表其所在进化支的编号(粗体表示有拷贝数目变异的进化支)。由于这12个种的核型不同, 我们在*D. melanogaster*中将基因按从X染色体到3号染色体右臂的顺序展示, 并将其他物种的染色体按与*D. melanogaster*相对应的顺序排列。加粗的线指示了拷贝数目变异的三个例子。分别包括了一次获得(Clade22)、一次获得和一次丢失(Clade14)以及一次获得和两次丢失(Clade24)事件。

Fig. 4 Chromosomal locations of F-box genes and their orthologous and paralogous relationships in the 12 *Drosophila* genomes. Horizontal bars represent the chromosomal arms/scaffolds/segments of the 12 *Drosophila* genomes. L and R indicate the left and right arms of chromosomes, respectively. Hexagonal lumps designate F-box genes. Orthologs are connected by lines. Filled circles indicate gene duplication events. Orthologs connected by curve do not exist in the species between them. Numbers below genes are the No. of clades (those in bold indicate that the copy numbers are different among the 12 species). Since the 12 species are different in karyotype, we arranged the genes from chromosomes X to 3R in *D. melanogaster*, and the chromosomes in other species are arranged corresponding to *D. melanogaster*. Lines in bold indicate three examples of copy number variation, which show one gain (Clade22), one gain and one loss (Clade14), as well as one gain and two losses (Clade24), respectively.

研究中发现的24次引起拷贝数目增加的事件中, 有7次属于串联重复, 15次属于散在重复(附录IV)。显然, 串联重复和散在重复是导致新基因产生的主要方式。

此外, 反转录转座和新基因的非编码区起源也是两种值得注意的机制。如Clade22基因在12个物种中均存在, 且在*D. persimilis*中有两个拷贝(perFBpp0191069和perFBpp0177108)(附录III)。虽然

这两个拷贝在基因组上的准确位置还未知, 但是将分别包括这两个拷贝在内的上、下游共10 kb的片段进行比对, 我们发现: 有约1.5 kb(包括本基因在内)的序列相似性超过96%, 其余部分则相似性不高。比较这两个拷贝的基因结构, 发现perFBpp0191069仅有1个外显子, 并且与perFBpp0177108的后5个外显子具有很高的 consistency(图5)。这一现象符合由反转录转座产生的基因与其祖先基因之间的序列关系。



由于反转录转座的机制是由祖先基因产生的mRNA反转录形成cDNA后随机插入到基因组中,因此,与祖先基因相比,新产生的基因除了没有内含子之外,还应在3'端具有poly-A尾巴,并在序列两端有短的正向重复序列。我们的确发现了另外两个特征:在perFBpp0191069的3'端下游138–157位点中有17个腺嘌呤(A),可能是poly-A结构的遗迹;紧邻该poly-A结构下游(即基因3'端下游158–166位点)的序列为“CCTGTGGGA”,与该基因5'端上游从–467到–459位点的序列“CCTGTGCCA”疑似正向重复。因此,perFBpp0191069很可能是通过perFBpp0177108的反转录转座产生的,说明这种机制对F-box基因拷贝数目的增加也起了一定的作用。

在本研究中,我们还发现了一个从头起源的新F-box基因。前面已经提到,Clade18基因仅存在于*D. mojavensis*和*D. virilis*两个物种中,并根据最简约原则推断该基因起源于这两个物种分化之前。那么这个基因是如何产生的?我们发现,分别位于此F-box基因上、下游的基因CG1792(以*D. melanogaster*中该基因的名字为代表)和pasha在其他10个物种中都有同源基因,但在这两基因之间却没有任何已注释的基因,只有从0.14 kb到2.76 kb长度不一的非编码序列(图5)。其中,*D. melanogaster*中的此片段最短(0.14 kb),并且在其他9个物种中相应的片段均含有与*D. melanogaster*中的pasha基因第一外显子前46 bp同源的DNA序列。我们还发现,在较短的此间隔区片段中出现的一些序列在相应的较长的片段中往往有多个重复。例如,*D. simulans*的该基因间隔区中含有的一段长170 bp的序列在*D. erecta*的相应片段中重复了三次,还有一些短的重复序列(如aataaattagg)。更有意思的是,序列比对结果显示,*D. virilis*的F-box基因竟然与*D. grimshawi*相应间隔区中的一段序列有着较高的同源性(图5,附录V);这说明此F-box基因应该不是由整段序列的插入产生的,而可能是由非编码区序列转变而来,即该基因为从头起源。那么,是否可能是*D. mojavensis*、*D. virilis*和*D. grimshawi*三者的最近共同祖先中插入了一段序列而形成F-box基因,但在三者分化之后,*D. grimshawi*丢掉了该基因呢?我们发现,*D. mojavensis*中该新基因不但与*D. grimshawi*的非编码序列有同源性,其与*D. willistoni*的基因间非编码序列也有较高相似度(附录V)。这说明此段序列不可能

起源于这3个物种的最近共同祖先中,而应该是本身就存在于12个果蝇近缘种的最近共同祖先中,只是上部8个物种的最近共同祖先中丢失了此段序列,而在*D. mojavensis*和*D. virilis*两者的最近共同祖先中通过从头起源机制演化为新的F-box基因。

对引起F-box基因拷贝数目减少的机制,我们发现了序列变异导致的外显子–内含子边界变化以及外显子丢失两种机制。例如,Clade37基因在除了*D. persimilis*的其他11个种中都存在。序列比较表明,*D. persimilis*的同源基因中本该编码F-box结构域的第二个外显子因为序列变异出现终止密码子“TGA”而导致翻译提前终止,不能形成F-box结构域。也可以说,第二个外显子的边界缩短导致了拷贝数目的减少(图5)。又如Clade7基因在10个物种中存在,在*D. yakuba*和*D. erecta*两个物种中丢失了。基因结构的比较显示,*D. melanogaster*的第一个外显子和第二个外显子(含编码F-box结构域的序列)在*D. yakuba*和*D. erecta*中发生了碱基缺失,从而使*D. yakuba*和*D. erecta*的原F-box基因变成了假基因(图5)。另外,前文提到的Clade14基因在*D. yakuba*中的丢失是由外显子丢失导致的。

3 讨论

在本研究中,我们发现果蝇中F-box基因的拷贝数目变化包括拷贝数目减少和增加两类事件。F-box基因拷贝数目的减少可以通过外显子的丢失、外显子的边界变化等丧失基因的功能区段来实现。其拷贝数目的增加则有串联重复、散在重复、反转录转座、非编码区起源等多种途径。这些途径可以按照来源分成两类:从已有基因中起源和从头起源。我们比较了各种拷贝数目增加机制的贡献(附录IV),发现63%的拷贝数目增加事件是散在重复引起的,29%是通过串联重复实现的。虽然我们研究的样本(24次事件)较小,但仍然可以说明散在重复和串联重复是果蝇中F-box基因的拷贝数目增加的主要机制。而且,由于序列重排在果蝇基因组中广泛存在,祖先物种中发生的串联重复可能转化为现存物种中散在形式的重复,所以本研究可能低估了串联重复的贡献。

基因的从头起源机制曾一度被认为很难发生而被人们所忽视,人们认为几乎所有的基因都是以现有的基因为原材料,通过重复、分裂、融合等多

种机制产生的(Long *et al.*, 2003)。Levine等(2006)通过对*D. melanogaster*的基因组序列及相关物种的序列进行的比较研究,首次报道了5个从非编码序列起源的基因。这一发现使得人们对通过从头起源产生新基因的机制有了新的思考和认识,并引发了更多的相关研究。例如,Cai等(2008)在酵母中也发现了一个从头起源的基因,并通过群体遗传学、表达谱、蛋白质组以及联合致死等实验证明该基因编码了有功能的蛋白质。随后,Zhou等(2008)在对12个果蝇近缘种的全基因组序列的研究中,发现有11.9%左右的新基因是由非编码序列变异而来的。近来,人们又基于进化、遗传、细胞和生化等方面的证据说明了*MDF1*基因的从头起源机制(Li *et al.*, 2010)。而本研究发现的在*D. mojavensis*和*D. virilis*最近共同祖先中从头起源的新F-box基因也证明了从头起源是新基因起源不可忽略的重要机制。另外,由于在其他10个近缘种中,该新基因对应的非编码区段含有部分序列的多次重复和相似性较高的区段,尤其是*D. mojavensis*的该新基因与*D. willistoni*的基因间非编码序列以及*D. virilis*的该新基因与*D. grimshawi*的基因间非编码序列都有较高的一致性,说明新基因从非编码区起源的过程有可能是渐变和具有一定的偶然性的。事实上,Cai等(2008)对从头起源的酵母新基因*BSC4*的研究中发现,近缘种中与新基因对应的序列都有RNA水平的表达,他们推测蛋白编码基因的从头进化需要先经历由非编码区到拥有转录活性区的进化,然后由转录活性区的序列再进一步进化成蛋白编码基因。

前述结果已表明,果蝇中F-box基因的拷贝数目在现存物种和祖先物种中差别不大。但是F-box基因在果蝇进化过程中频繁发生基因的获得和丢失事件,说明果蝇中F-box基因的进化历史也是生与死(birth-and-death evolution)的进化过程。另外,获得和丢失事件虽未导致F-box基因在各物种中的拷贝数目具有显著差异,但却导致了各物种的F-box基因家族的成员有了较大不同。*D. grimshawi*的F-box基因拷贝数目相对于12个近缘种的祖先来说没有变化,并且各个基因分支中也未发现导致F-box基因拷贝数目变异的事件。也就是说,*D. grimshawi*中的F-box基因家族的构成与12个近缘种的最近共同祖先相同。与*D. melanogaster*等广泛分布于世界各地且食物来源多样的物种不同,*D.*

*grimshawi*仅分布于太平洋的夏威夷群岛上,而且只靠腐烂的夏威夷植物的树皮生存(Markow & O'Grady, 2005)。由于生物的进化与环境以及食物来源有一定程度的关联,那么,或许环境的稳定和食物来源的单一导致了*D. grimshawi*中F-box基因的高度保守。但是,只分布在印度洋的塞舌尔岛且食物来源仅为橘叶巴戟(*Morinda citrifolia*)的*D. sechellia*在整个进化历史中发生了7次F-box基因的丢失/获得事件。然而,值得注意的是,这7次事件均发生在该物种分化之前,在*D. sechellia*与*D. simulans*分开之后,其F-box基因家族的成员并无变化。此外,分布相对广泛一些但食物来源也很单一的*D. mojavensis*在与*D. virilis*分开之后也没有发生F-box基因家族组成的变化。

为了帮助了解F-box基因的拷贝数目变异与功能之间的关系,我们还分析了有功能方面实验证据的F-box基因。在*D. melanogaster*中有9个已研究过功能的基因,其中2个(Clade21和Clade22)在直系同源基因中有拷贝数目的变异,比例为22.2%,低于39.6%的平均值。7个无拷贝数目变异的基因多与有丝分裂等基本的生理过程相关(附录VI)。Clade21基因在10个物种中存在而在*D. pseudoobscura*和*D. persimilis*的最近共同祖先中发生丢失,资料显示Clade21基因参与了凋亡细胞的吞噬过程(Silva *et al.*, 2007)。该过程对维持正常的表型应该是必要的,我们推测*D. pseudoobscura*和*D. persimilis*中可能有其他基因参与类似的生理过程,或者凋亡细胞的吞噬过程在*D. pseudoobscura*和*D. persimilis*中发生了变化。一般而言,进化式样的不同部分是由环境和行为的差异所导致的(Nei, 2007)。*D. pseudoobscura*和*D. persimilis*的食物来源是12个近缘种中最多样化的,有多种类型的宿主,而不像其他物种只能集中在某种或某些腐烂的水果或腐烂的树上。这种取食上的差异可能对其体内的生理活动有一定的影响。Clade22基因在12个物种中均存在,且在*D. persimilis*中有两个拷贝。该基因通过反馈调节钙离子激活GPCRs的活性来避免外界刺激终止后,过量钙离子涌入对细胞造成的毒害作用(Han *et al.*, 2006)。显然,这一过程直接受外界环境的影响,*D. persimilis*中该F-box基因拷贝数目的增加可能与其所处的环境有关。除了保证基本的生理需求之外,拷贝数目变异可能增加了果蝇F-box基因与多种底

物作用的潜力, 从而为表型进化提供更多机会。

本研究在基因组层面上比较近缘种间基因的拷贝数目, 并总结了造成拷贝数目变异的机制, 探讨了拷贝数目变异的生物学意义, 这有助于我们全面理解遗传、变异和进化的本质, 为阐明物种形成和物种多样性产生的机制奠定基础。

致谢: 感谢山红艳、高大海和国春策对本文提出的宝贵意见和建议。

参考文献

- Bai C, Richman R, Elledge SJ (1994) Human cyclin F. *The EMBO Journal*, **13**, 6087–6098.
- Bateman A, Birney E, Cerruti L, Durbin R, Ewinger L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL (2002) The Pfam protein families database. *Nucleic Acids Research*, **30**, 276–280.
- Beckmann JS, Estivill X, Antonarakis SE (2007) Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nature Reviews Genetics*, **8**, 639–646.
- Cai J, Zhao R, Jiang H, Wang W (2008) *De novo* origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics*, **179**, 487–496.
- Celniker SE, Wheeler DA, Kronmiller B, Carlson JW, Halpern A, Patel S, Adams M, Champe M, Dugan SP, Frise E, Hodgson A, George RA, Hoskins RA, Lavery T, Muzny DM, Nelson CR, Pacle JM, Park S, Pfeiffer BD, Richards S, Sodergren EJ, Svirskas R, Tabor PE, Wan K, Stapleton M, Sutton GG, Venter C, Weinstock G, Scherer SE, Myers EW, Gibbs RA, Rubin GM (2002) Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biology*, **3**, research0079.1-0079.14.
- Derti A, Roth FP, Church GM, Wu CT (2006) Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nature Genetics*, **38**, 1216–1220.
- Dharmasiri N, Dharmasiri S, Weijers D, Lechner E, Yamada M, Hobbie L, Ehrismann JS, Jürgens G, Estelle M (2005) Plant development is regulated by a family of auxin receptor F box proteins. *Developmental Cell*, **9**, 109–119.
- Drosophila* 12 Genomes Consortium (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, **450**, 203–218.
- Eddy SR (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
- Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**, 696–704.
- Han J, Gon P, Reddig K, Mitra M, Guo P, Li HS (2006) The fly CAMTA transcription factor potentiates deactivation of rhodopsin, a G protein-coupled light receptor. *Cell*, **127**, 847–858.
- Hinds DA, Klok AP, Jen M, Chen XY, Frazer KA (2006) Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nature Genetics*, **38**, 82–85.
- Junier T, Pagni M (2000) Dotlet: diagonal plots in a web browser. *Bioinformatics*, **16**, 178–179.
- Kipreos ET, Pagano M (2000) The F-box protein family. *Genome Biology*, **1**, REVIEWS3002.
- Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ (2006) Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proceedings of the National Academy of Sciences, USA*, **103**, 9935–9939.
- Li D, Dong Y, Jiang Y, Jiang H, Cai J, Wang W (2010) A *de novo* originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell Research*, **20**, 408–420.
- Long M, Betran E, Thornton K, Wang W (2003) The origin of new genes: glimpses from the young and old. *Nature Reviews Genetics*, **4**, 865–875.
- Markow TA, O'Grady PM (2005) *Drosophila: A Guide to Species Identification and Use*, P. 250. Elsevier Academic, London.
- Nei M (2007) The new mutation theory of phenotypic evolution. *Proceedings of the National Academy of Sciences, USA*, **104**, 12235–12242.
- Nicholas K, Nicholas HB Jr (1997) GeneDoc: a tool for editing and annotating multiple sequence alignments. Distributed by the author.
- Niimura Y, Nei M (2006) Evolutionary dynamics of olfactory and other chemosensory receptor genes invertebrates. *Journal of Human Genetics*, **51**, 505–517.
- Niimura Y, Nei M (2007) Extensive gains and losses of olfactory receptor genes in mammalian evolution. *PLoS One*, **2**, e708.
- Nozawa M, Nei M (2007) Evolutionary dynamics of olfactory receptor genes in *Drosophila* species. *Proceedings of the National Academy of Sciences, USA*, **104**, 7122–7127.
- Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, Carter NP, Lee C, Stone AC (2007) Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*, **39**, 1256–1260.
- Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.
- Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shaper MH, Carson AR, Chen WW, Cho EK, Dallaire S, Freeman JL, González JR, Gratacòs M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armenogol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME (2006) Global variation in copy number in the human ge-

- nome. *Nature*, **444**, 444–454.
- Reese MG, Hartzel G, Harris NL, Ohler U, Abril JF, Lewis SE (2000) Genome annotation assessment in *Drosophila melanogaster*. *Genome Research*, **10**, 483–501.
- Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, Rohde C, Valente VL, Aguadé M, Anderson WW, Edwards K, Garcia AC, Goodman J, Hartigan J, Kataoka E, Lapoint RT, Lozovsky ER, Machado CA, Noor MA, Papacit M, Reed LK, Richards S, Rieger TT, Russo SM, Sato H, Segarra C, Smith DR, Smith TF, Strelets V, Tobari YN, Tomimura Y, Wasserman M, Watts T, Wilson R, Yoshida K, Markow TA, Gelbart WM, Kaufman TC (2008) Polytene chromosomal maps of 11 *Drosophila* species: the order of genomic scaffolds inferred from genetic and physical maps. *Genetics*, **179**, 1601–1655.
- Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Valente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE (2005) Segmental duplications and copy-number variation in the human genome. *The American Journal of Human Genetics*, **77**, 78–88.
- Silva E, Au-Yeung HW, Van Goethem E, Burden J, Franc NC (2007) Requirement for a *Drosophila* E3-ubiquitin ligase in phagocytosis of apoptotic cells. *Immunity*, **27**, 585–596.
- Singh ND, Larracuente AM, Sackton TB, Clark AG (2009) Comparative genomics on the *Drosophila* phylogenetic tree. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 459–480.
- Small KS, Brudno M, Hill MM, Sidow A (2007) Extreme genomic variation in a natural population. *Proceedings of the National Academy of Sciences, USA*, **104**, 5698–5703.
- Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, Ruby JG, Brennecke J, Harvard FlyBase curators, Berkeley Drosophila Genome Project, Hodges E, Hinrichs AS, Caspi A, Paten B, Park S, Han MV, Maeder ML, Polansky BJ, Robson BE, Aerts S, Helden J, Hassan B, Gilbert DG, Eastman DA, Rice M, Weir M, Hahn MW, Park Y, Dewey CN, Pachter L, Kent WJ, Haussler D, Lai EC, Bartel DP, Hannon GJ, Kaufman TC, Eisen MB, Clark AG, Smith D, Celniker SE, Gelbart WM, Kellis M (2007) Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature*, **450**, 219–232.
- Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, **24**, 1596–1599.
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*, **25**, 4876–4882.
- Xu G, Ma H, Nei M, Kong H (2009) Evolution of F-box genes in plants: different modes of sequence divergence and their relationships with functional diversification. *Proceedings of the National Academy of Sciences, USA*, **106**, 835–840.
- Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W (2008) On the origin of new genes in *Drosophila*. *Genome Research*, **18**, 1446–1455.

(责任编辑: 王文 责任编辑: 闫文杰)

附录I 冗余序列的筛选表

Appendix I Redundant sequences and the one finally selected.

(<http://www.biodiversity-science.net/CN/article/downloadArticleFile.do?attachType=PDF&id=9490>)

附录II 重新注释的基因

Appendix II Re-annotated genes.

(<http://www.biodiversity-science.net/CN/article/downloadArticleFile.do?attachType=PDF&id=9490>)

附录III 果蝇的12个近缘种中F-box蛋白质的系统发育关系。树中基因的名字用对应的物种的种加词前3个字母、蛋白质序列名和该蛋白质的C端结构域组成(其中N即none)。

Appendix III Phylogenetic tree of F-box proteins from 12 *Drosophila* species. The name of each protein is composed of the first three letters of the specific epithet, followed by the name of the sequence and the C-terminal domain (N means none).

(<http://www.biodiversity-science.net/CN/article/downloadArticleFile.do?attachType=PDF&id=9490>)

附录IV 引起果蝇F-box基因拷贝数目增加的事件

Appendix IV Events caused increase in copy number of F-box genes in *Drosophila*.

(<http://www.biodiversity-science.net/CN/article/downloadArticleFile.do?attachType=PDF&id=9490>)

附录V *D. mojavensis*(a)和*D. virilis*(b)中的Clade18基因分别与*D. willistoni*和*D. grimshawi*中的非编码区序列的比对

Appendix V Sequence alignments of Clade18 genes from *D. mojavensis* and *D. virilis* with non-coding sequences from *D. willistoni* and *D. grimshawi*, respectively.

(<http://www.biodiversity-science.net/CN/article/downloadArticleFile.do?attachType=PDF&id=9490>)

附录VI F-box蛋白质的功能与拷贝数目变异的关系

Appendix VI Relationships between functions and copy number variation of F-box proteins.

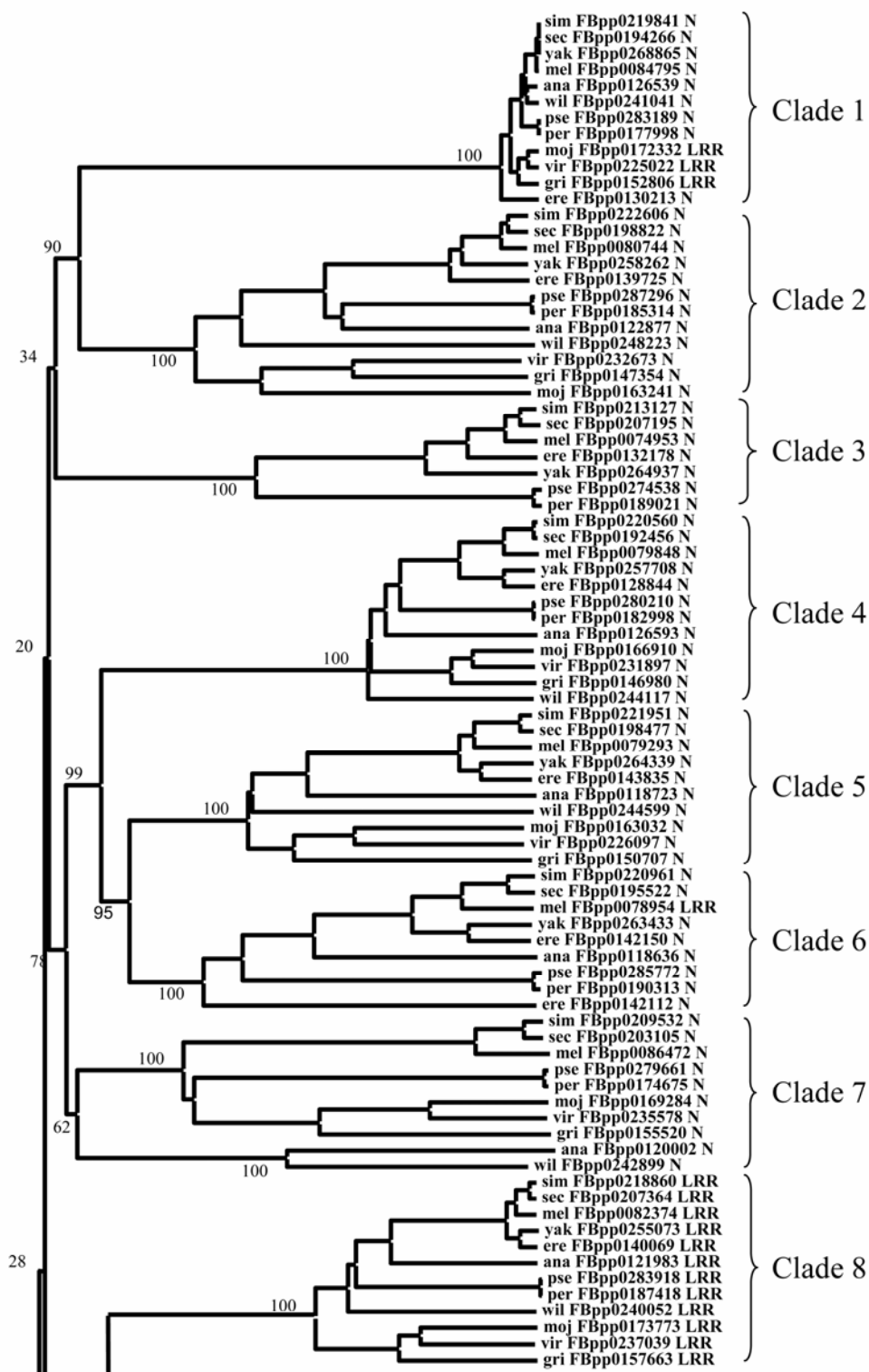
(<http://www.biodiversity-science.net/CN/article/downloadArticleFile.do?attachType=PDF&id=9490>)

附录I 冗余序列的筛选列表
Appendix I Redundant sequences and the one finally selected

冗余序列 Redundant sequences	所选的序列 Selected sequences
FBpp0073103, FBpp0073101, FBpp0073102	FBpp0073102
FBpp0078592, FBpp0110535, FBpp0078594, FBpp0078593	FBpp0078593
FBpp0084796, FBpp0084795	FBpp0084795
FBpp0110196, FBpp0078693	FBpp0078693
FBpp0087190, FBpp0111982, FBpp0111980, FBpp0111981	FBpp0111981
FBpp0086875, FBpp0086876	FBpp0086876
FBpp0087274, FBpp0087275, FBpp0087276	FBpp0087276
FBpp0083216, FBpp0083215	FBpp0083215
FBpp0099839, FBpp0099383	FBpp0099383
FBpp0268227, FBpp0264637	FBpp0264637
FBpp0259602, FBpp0267478	FBpp0267478
FBpp0267102, FBpp0267103	FBpp0267103
FBpp0153606, FBpp0157663	FBpp0157663
FBpp0157330, FBpp0144755	FBpp0144755
FBpp0157205, FBpp0157374	FBpp0157374

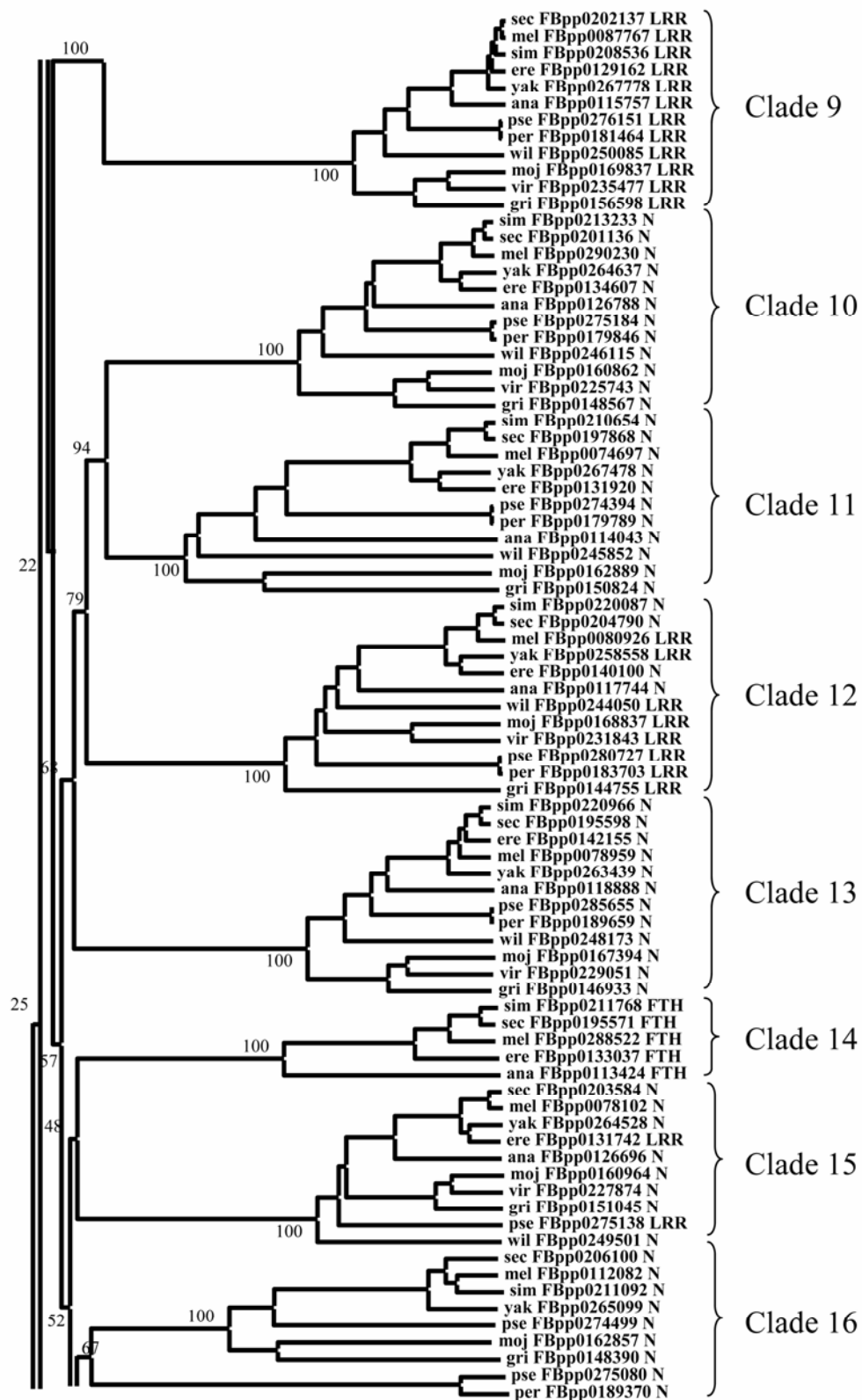
附录II 重新注释的基因
Appendix II Re-annotated genes

重新注释的基因 Re-annotated genes	对应的蛋白质 Corresponding proteins
Dsim\GD12690	FBpp0211092
Dmel\CG14102	FBpp0074697
Dyak\GE20089	FBpp0265099
Dpse\GA21694	FBpp0280727



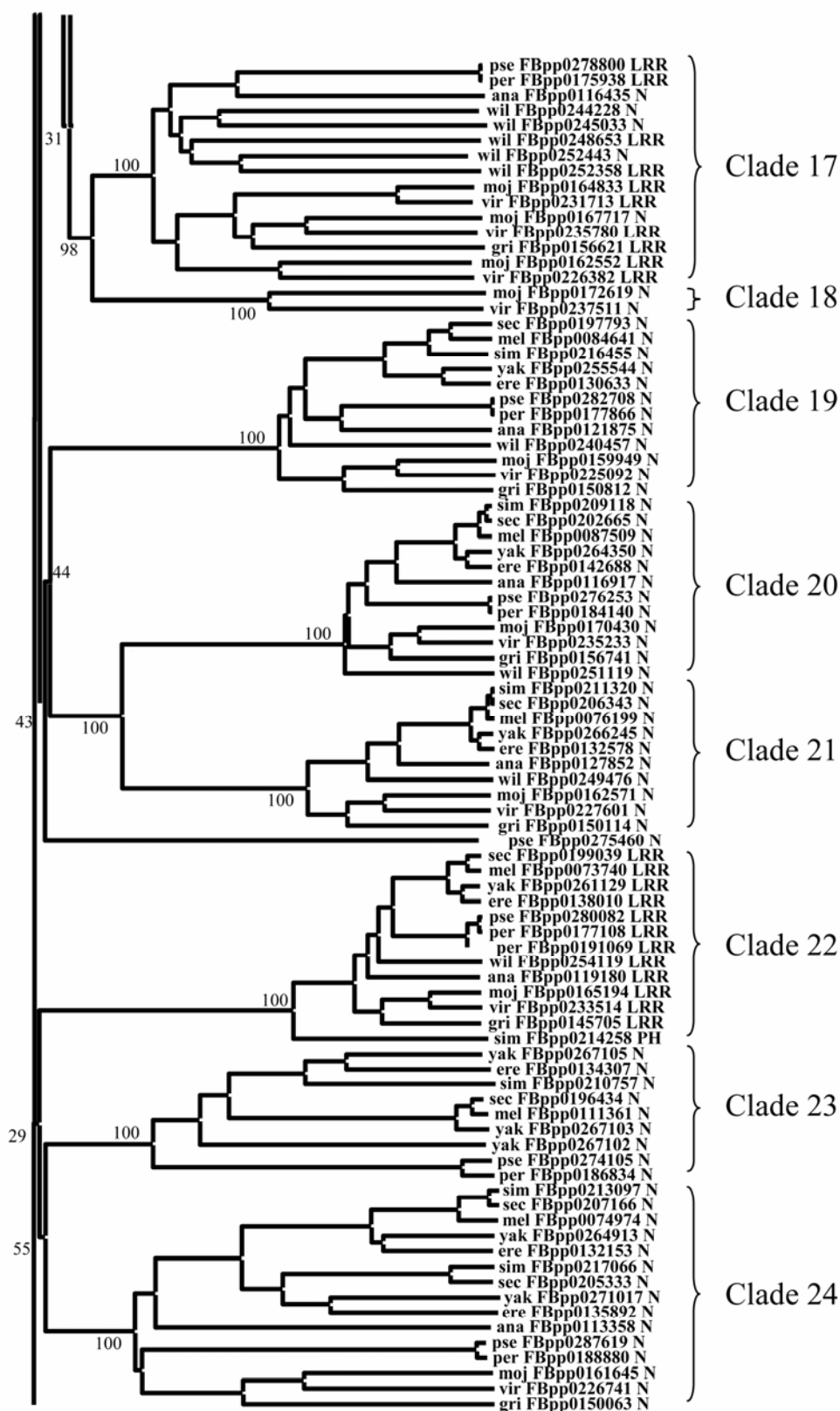
附录III 果蝇的12个近缘种中F-box蛋白质的系统发育关系。树中基因的名字用对应的物种的种加词前3个字母、蛋白质序列名和该蛋白质的C端结构域组成(其中N即none)。

Appendix III Phylogenetic tree of F-box proteins from 12 *Drosophila* species. The name of each protein is composed of the first three letters of the specific epithet, followed by the name of the sequence and the C-terminal domain (N means none).



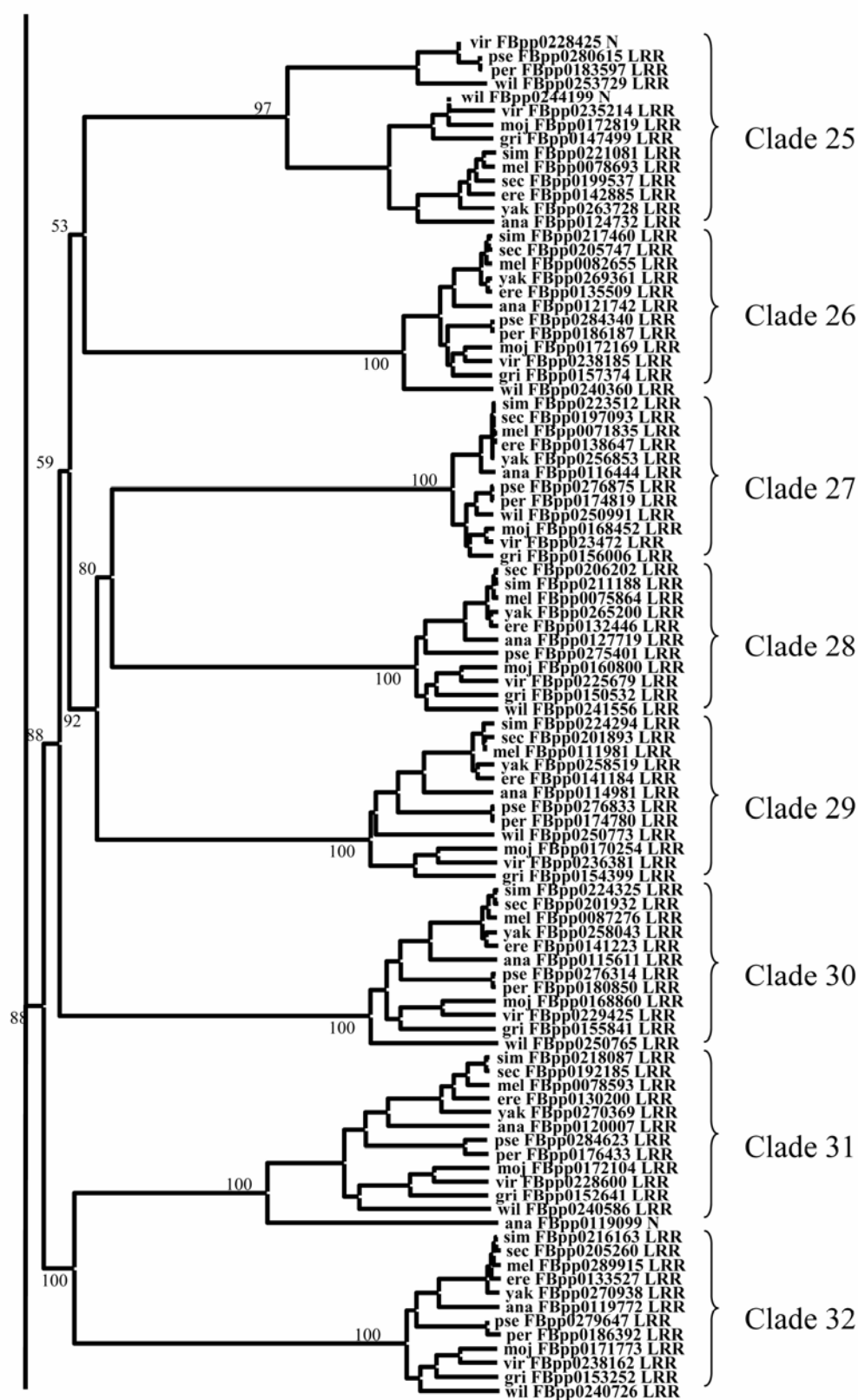
附录III(续) 果蝇的12个近缘种中F-box蛋白质的系统发育关系

Appendix III (continued) Phylogenetic tree of F-box proteins from 12 *Drosophila* species.



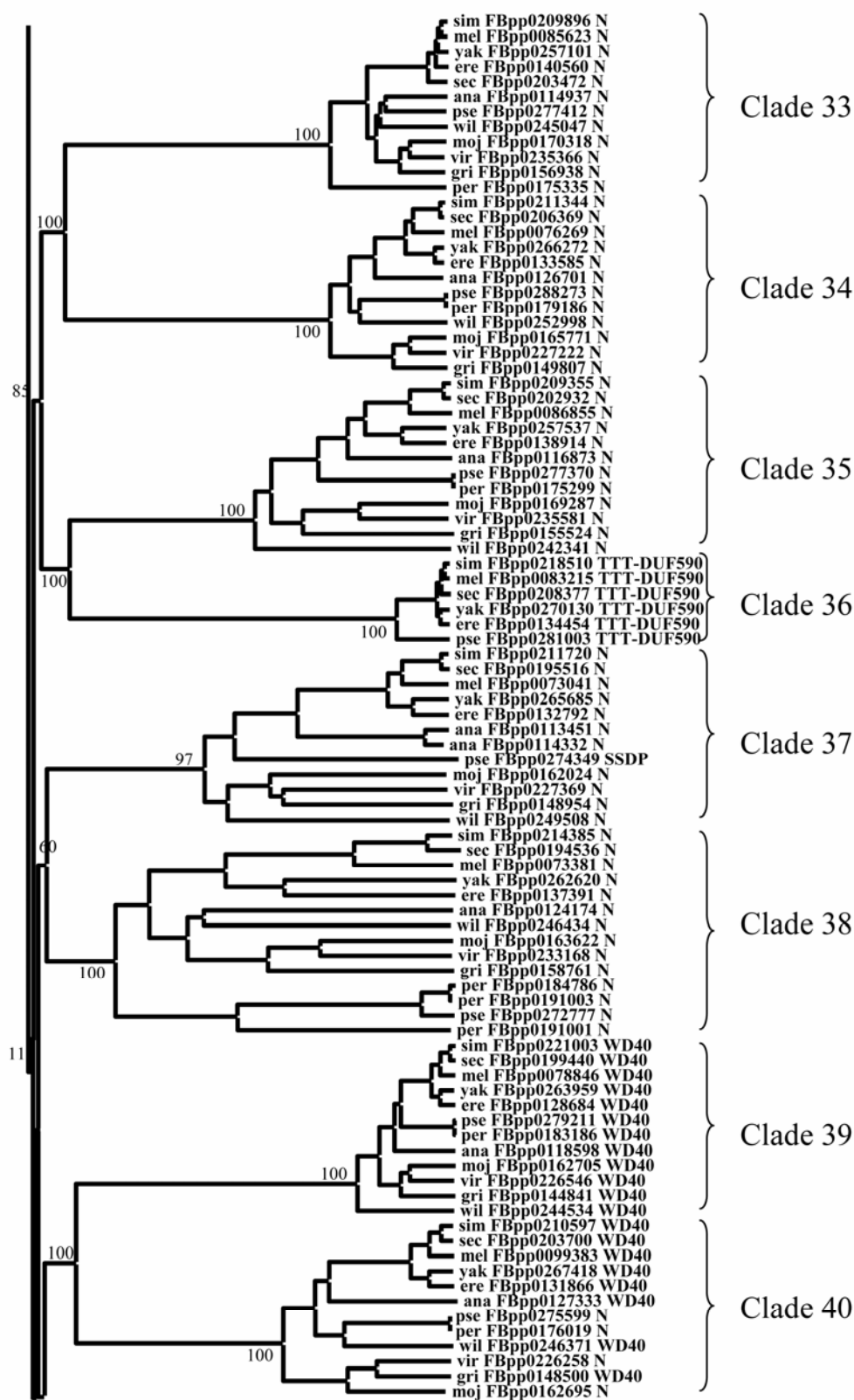
附录III(续) 果蝇的12个近缘种中F-box蛋白质的系统发育关系

Appendix III (continued) Phylogenetic tree of F-box proteins from 12 *Drosophila* species.



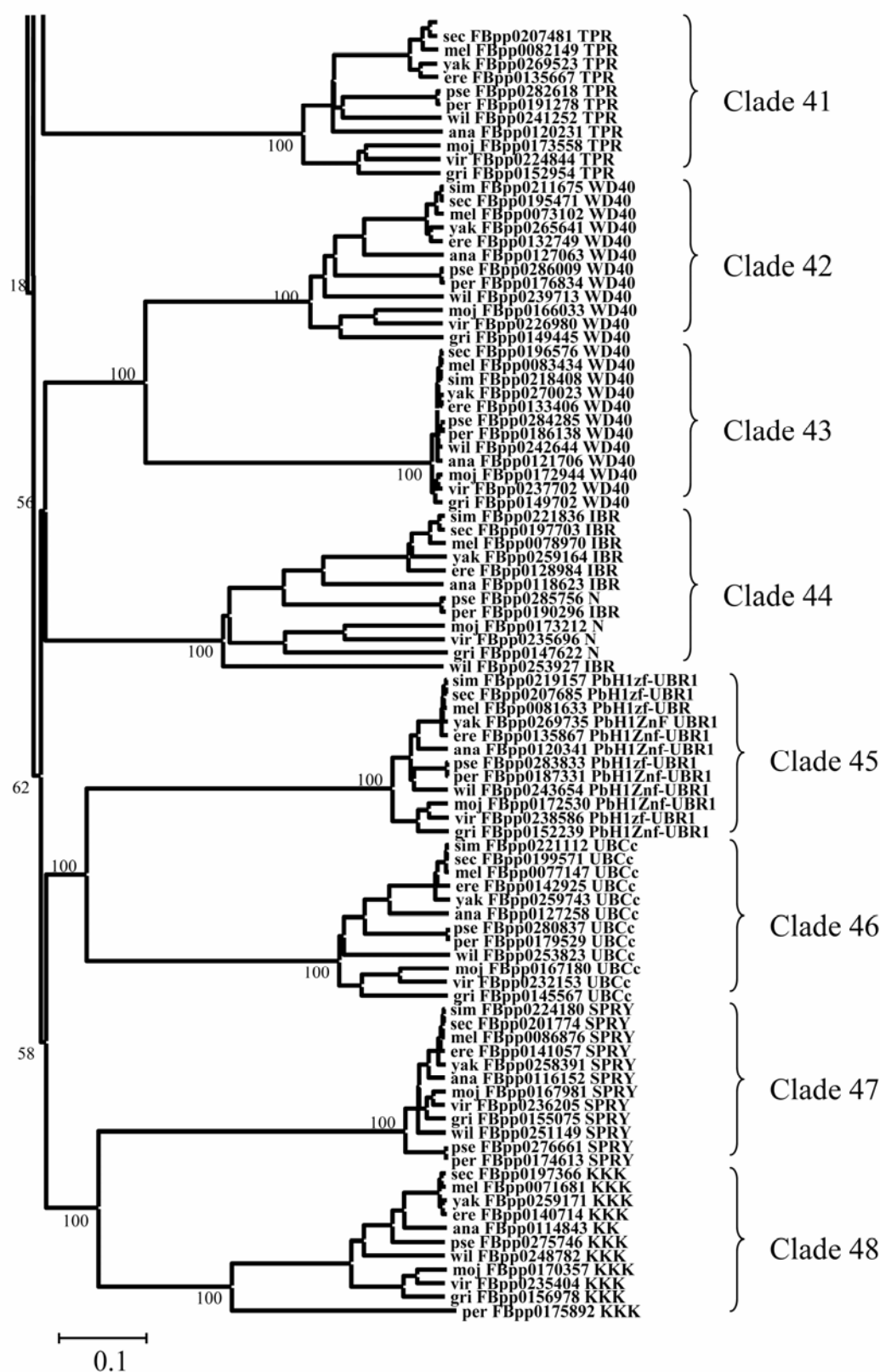
附录III(续) 果蝇的12个近缘种中F-box蛋白质的系统发育关系

Appendix III (continued) Phylogenetic tree of F-box proteins from 12 *Drosophila* species.



附录III(续) 果蝇的12个近缘种中F-box蛋白质的系统发育关系

Appendix III (continued) Phylogenetic tree of F-box proteins from 12 *Drosophila* species.



附录III(续) 果蝇的12个近缘种中F-box蛋白质的系统发育关系

Appendix III (continued) Phylogenetic tree of F-box proteins from 12 *Drosophila* species.

附录IV 引起果蝇F-box基因拷贝数目增加的事件
Appendix IV Events caused increase in copy number of F-box genes in *Drosophila*

	进化枝编号 Clade no.													合计 Total
	3	6	14	16	17	18	22	23	24	31	36	37	38	
串联重复 Tandem duplication				1	1			3					2	7 (29%)
散在重复 Dispersed duplication	1	2	1		6			1	1	1	1	1		15 (63%)
反转录转座 Retroposition							1							1 (4%)
从头起源 De novo origination						1								1 (4%)
合计 Total	1	2	1	1	7	1	1	4	1	1	1	1	2	24 (100%)

[illegible]

附录V *D. mojavenensis*(a)和*D. virilis*(b)中的Clade18基因分别与*D. willistoni*和*D. grimshawi*中的非编码区序列的比对
Appendix V Sequence alignments of Clade18 genes from *D. mojavenensis* and *D. virilis* with non-coding sequences from *D. willistoni* and *D. grimshawi*, respectively.

附录VI F-box蛋白质的功能与拷贝数目变异的关系
Appendix VI Relationships between functions and copy number variation of F-box proteins

蛋白质名称 Protein name	C端结构域 C-terminal domain	拷贝数变异 CNV	参与的生物学过程 Biological process involved in
FBpp0073102	WD40	No	negative regulation of growth; regulation of mitosis; DNA endoreduplication.
FBpp0078846	WD40	No	protein ubiquitination during ubiquitin-dependent protein catabolic process; WD40 protein FBW5 promotes ubiquitination of tumor suppressor TSC2 by DDB1-CUL4-ROC1 ligase.
FBpp0083434	WD40	No	anatomical structure development; ovarian follicle cell development; gamete generation; regulation of biological process; circadian rhythm; learning or memory; cell motion; regulation of cellular component organization; olfactory learning; positive regulation of protein import into nucleus; locomotory behavior; catabolic process; negative regulation of nurse cell apoptosis; regulation of signal transduction; regulation of Wnt receptor signaling pathway.
FBpp0086876	SPRY	No	negative regulation of synaptic growth at neuromuscular junction; neuromuscular synaptic transmission.
FBpp0077147	UBCc	No	apoptosis; induction of compound eye retinal cell programmed cell death.
FBpp0078693	LRR	No	circadian behavior; locomotor rhythm; entrainment of circadian clock by photoperiod.
FBpp0078970	IBR	No	compound eye morphogenesis; negative regulation of protein catabolic process; G2 phase of mitotic cell cycle; G1/S transition of mitotic cell cycle; eye-antennal disc morphogenesis; regulation of mitosis.
FBpp0076199	N	Yes	engulfment of apoptotic cell.
FBpp0073740	LRR	Yes	deactivation of rhodopsin mediated signaling.