

# 谷子和玉米基因组多倍化进化比较分析

张 琼, 王振怡, 马雪莲, 聂林曼, 汪厚龙, 王金朋\*

(河北联合大学 生命科学学院 基因组学与计算生物学研究中心, 河北 唐山 063009)

**摘要:** 以谷子和玉米为研究对象, 基于比较基因组学, 利用改进的基因同源共线性方法对其基因组结构和基因同源信息进行比对分析, 确定了物种基因组加倍的规模和加倍后的进化规律。结果表明: 谷子中有 3 846 对(10.9%)、玉米中有 6 016 对(18.5%)由多倍化产生的重复基因, 它们之间的直系同源基因对为 13 652 对; 通过同源共线基因间的同义核苷酸置换率分析证实, 玉米与谷子不仅共同经历了一次古老的全基因组加倍, 而且在约 12 个百万年前玉米独立发生了一次较近的全基因组加倍; 加倍后产生的大量重复基因片段常分布在染色体末端; 基因组进化分析发现, 玉米基因组进化速率比谷子快 1/14。

**关键词:** 谷子; 玉米; 多倍化; 重复基因; 基因共线性

**中图分类号:** S513 S515 **文献标志码:** A **文章编号:** 1004-3268(2014)06-0010-06

## Comparative Analysis of Paleopolyploidy Evolution in Genomes of *Setaria italica* and *Zea mays*

ZHANG Qiong, WANG Zhen-yi, MA Xue-lian, NIE Lin-man, WANG Hou-long, WANG Jin-peng\*

(Center for Genomics and Computational Biology, School of Life Science, Hebei United University,  
Tangshan 063009, China)

**Abstract:** Based on the comparative genomics and the improved homologous gene collinearity method, the genomic structure and gene homology information of *C<sub>4</sub>* plant *Setaria italica* and *Zea mays* were comparatively analyzed in order to determine the doubled scale of genomes and the evolution law of polyploidy. The result showed that 3 846 pairs of paralogs(10.9%) in *Setaria italica* genome and 6 016 pairs of paralogs(18.5%) in *Zea mays* genome were generated by whole-genome duplication(WGD), and 13 652 pairs of orthologs were identified between them. The analysis of synonymous nucleotide substitution rate between homologous colinear genes indicated that *Zea mays* and *Setaria italica* shared an ancient WGD, and a maize-specific WGD event occurred 12 million years ago. The duplicated genes often distributed near the ends of chromosomes. The analysis of genome evolution showed that the gene evolution rate was 1/14 faster in *Zea mays* than in *Setaria italica*.

**Key words:** *Setaria italica*; *Zea mays*; paleopolyploidy; duplicated genes; gene collinearity

多倍化(polyploidy)可迅速使物种基因组加倍, 创造大量的重复基因, 引发大规模的基因组变化, 如染色体重排、基因倒位、基因丢失等<sup>[1-2]</sup>。基因倍增

是物种演化最重要的动力源泉<sup>[3-4]</sup>。研究表明, 重复的基因拷贝有不同的进化方式, 分别是新功能化(neofunctionalization)、亚功能化(subfunctionaliza-

收稿日期: 2013-12-28

基金项目: 国家自然科学基金项目(31100913); 河北联合大学科学研究基金项目(z201230, z201235); 河北联合大学大学生创新性实验计划项目(X2013044)

作者简介: 张 琼(1993-), 女, 河北石家庄人, 本科, 研究方向: 生物信息学、比较基因组学。E-mail: zhangqiong201502@163.com

\* 通讯作者: 王金朋(1984-), 男, 河北邯郸人, 讲师, 硕士, 主要从事生物信息学、比较基因组学研究。

E-mail: wangjinpeng1010@gmail.com

tion)、亚新功能化(subneofunctionalization)<sup>[1,5]</sup>,重复基因的这一系列变异为物种的适应性进化提供了很大机会。对禾本科植物全基因组加倍进行研究,有利于对其家族组成、分化以及基因功能的解析。

谷子(*Setaria italica*)和玉米(*Zea mays*)作为禾本科植物家族的成员,不仅是重要的粮食和饲料作物,而且可以作为 C<sub>4</sub> 生物燃料。目前,它们的全基因组测序工作相继完成<sup>[6-8]</sup>,为比较基因组学研究提供了良好的数据材料。有研究基于基因共线性建立了算法,用于推断物种多倍化后产生的重复基因片段<sup>[9-11]</sup>。然而,由于物种在进化过程中可能发生过不止一次的多倍化现象,导致物种基因组结构复杂化,因此,准确地确定物种全基因组加倍的程度,特别是确定物种间基因的同源关系,是当前比较基因组学的一个主要难题。多重序列比对工具 McScan 可以比对多个基因组,搜索同源基因片段<sup>[12]</sup>,而基于共线性方法的共线性片段显著性分析工具 ColinearScan,可以用来评估获得共线性片段的显著性<sup>[13]</sup>,但是如何将获得的同源染色体区进行分组,明确基因组内和基因组间的同源关系,直接影响重复基因进化的研究结果,如基因置换(gene conversion)<sup>[14-15]</sup>、物种基因组重要基因家族进化、物种之间基因进化及功能差异等<sup>[16]</sup>。基于此,本研究建立了一种新的方法用以确定种内和种间复杂的同源共线信息,首先,利用多重序列比对工具 McScan 和 ColinearScan 获得可信的同源共线染色体片段,然后结合谷子、玉米全基因组同源结构分析,确定其全基因组加倍事件和种间同源共线关系,分析多倍化发生的规模和规律以及物种基因组进化的时间,为阐明禾本科植物基因组多倍化进化提供理论依据。

## 1 材料和方法

### 1.1 物种基因组数据

玉米(*Zea mays*)和谷子(*Setaria italica*)最新的全基因组序列从植物基因组数据库 Phytozome (<http://www.phytozome.net/>)下载获得。

### 1.2 推断基因共线性

首先,对种内和种间基因组进行相似性比对分析,得到基因同源性;然后,利用多重序列比对工具 McScan 寻找基因组内和基因组间同源共线 DNA 片段。同源共线区域中每对匹配基因的打分为  $\min(-\log_{10}^{E\text{-value}}, 50)$ ,失配得分为 -1,打分路径中匹配

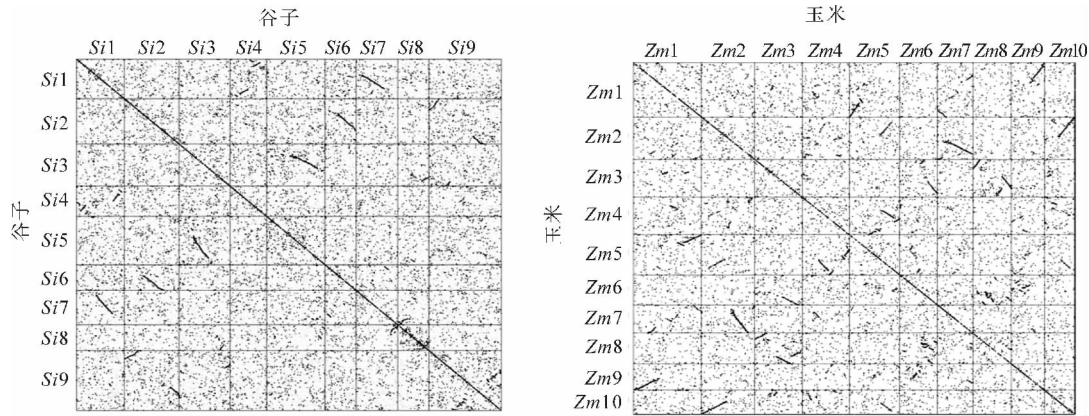
基因对与邻近匹配基因对之间的距离超过 10 000 bp 时得分为 -1,要求共线性区域的打分大于 300 时保留,并利用加入统计估计的共线性分析工具 ColinearScan 对获得的基因组共线区域进行统计,将显著性大于  $1 \times 10^{-10}$  的区域去掉,确定基因组内重复基因和基因组间的同源共线信息。最后,生成基因组同源点阵图(dotplot map),画出基因组内和基因组间的同源基因点阵图,对同源区分组,分析不同物种间基因组点阵图,结合同源基因相似性,理清不同进化事件产生的同源片段,并把不同的同源组分组,利用 Perl 语言编程实现这一过程;选定谷子基因组为参考,因为谷子基因组没有再次的基因加倍,相对于玉米更好地保持了祖先基因组的结构。综合上述过程生成两基因组间的联合比对结果。

## 2 结果与分析

### 2.1 谷子和玉米基因组内同源结构

为了获得基因组内同源结构,分别对谷子和玉米蛋白序列进行了 Blastp 比对分析,根据基因间的同源信息以及基因在染色体上的物理位置,画出了基因组内的点阵图(图 1)。在图 1 中可以看到,每个物种基因组中有大量的共线性同源基因片段存在。相对于谷子基因组,玉米中有更多的共线性片段,但同源共线性片段有较多的短片段。谷子基因组中较长的共线性基因片段分别存在于 Si1-Si4、Si1-Si7、Si2-Si6、Si2-Si9、Si3-Si5、Si9-Si9;玉米基因组中较长的共线性基因片段分别存在于 Zm1-Zm5、Zm1-Zm9、Zm2-Zm7、Zm2-Zm10、Zm3-Zm8、Zm4-Zm5、Zm6-Zm8、Zm6-Zm9。

上述共线区域分别存在两物种基因组中的重复基因,这些重复基因由全基因组加倍产生,但在物种进化过程中保留了较好的共线性片段,如谷子的 1、4、7 号染色体是由其共同祖先物种的同一染色体加倍而来,其中 1、4 号染色体是基因组加倍之后的祖先染色体发生断裂而形成的 2 条染色体序列;相对于谷子基因组,玉米基因组有更多的同源染色体片段,且染色体片段较短,在点图中呈现灰色的线,与黑色的线上基因对间相似性相比差一些,这些重复基因是玉米和谷子祖先共同加倍产生<sup>[17-18]</sup>,但在后来的进化过程中基因序列变异导致相似性降低,并且伴随大量的基因丢失,较长的由黑色点构成的染色体片段由较近的基因加倍产生<sup>[19]</sup>。



图中每个点是 1 对同源基因,根据基因对在染色体上的位置打点;同一基因和最相似的基因打点为黑色,相似性次之的基因打点为灰色,其他为浅灰色

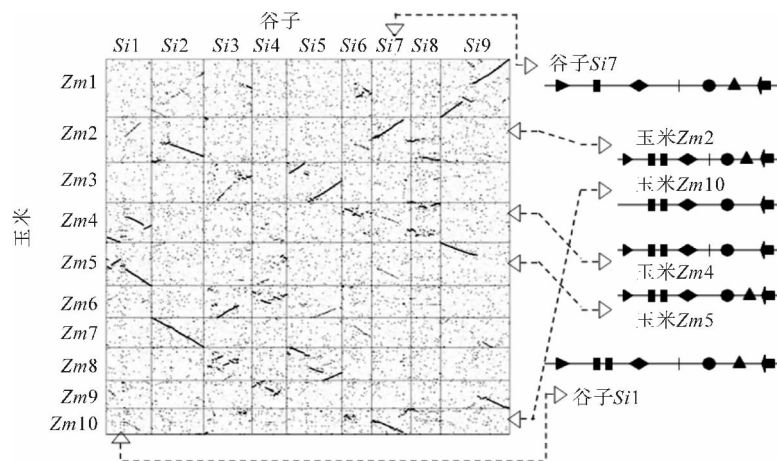
图 1 谷子、玉米同源基因结构点阵图

## 2.2 谷子和玉米基因组间同源结构

由图 2 可以发现,每条谷子染色体均对应 2 条完整的染色体片段,如谷子的 7 号染色体对应最相似的玉米 2 条染色体分别是 2 和 10 号染色体,它们之间的同源染色体片段在图中形成了 2 条黑色的线,谷子的 7 号和 1 号染色体是 1 对旁系同源基因对,并且这一规律在其他染色体上也都存在,这一结果表明,玉米在其进化过程中与谷子共同发生过一次古老的全基因组加倍,之后独立发生过再次的全基因组加倍<sup>[17]</sup>。同源基因对在图中形成的黑色线代表两物种之间真正的直系同源基因片段,如谷子 7 号和玉米的 2 号染色体上有一直系同源染色体片

段;但同一染色体区域上的灰色线是由种外的旁系同源基因构成,如谷子的 7 号和玉米的 5 号染色体之间的同源染色体片段。

玉米独立发生的基因加倍导致现在有 2 套重复的基因组存在。比较发现,2 套基因组中基因保留并不一致,其中的 1 套基因组偏向于丢失较多的基因,而另外 1 套基因组中的重复基因偏向于保留下来,支持物种的适应性进化,这可能是由于物种全基因组加倍之后产生的重复基因具有不同的功能所导致<sup>[19]</sup>。有的重复基因保留可能在生物体中具有较高的表达,而有些基因可能参与的生物过程比较少,或者表达不明显,在进化过程中逐渐丢失。



黑色的线是 Blast 比对两物种基因组序列获得的最相似的基因对,表示物种间直系同源染色体片段;灰色是由次好相似基因对构成的线,表示种间旁系同源染色体片段;图中共线性上相同形状的基因为同源基因,谷子的 7 号和 1 号染色体上 1 对旁系同源染色体片段与它们各自在玉米中 2、10、4 以及 5 号染色体上的 2 个直系同源染色体片段共同构成了同源染色体六联子片段

图 2 谷子、玉米间同源基因结构点阵图

## 2.3 谷子和玉米基因组同源共线性

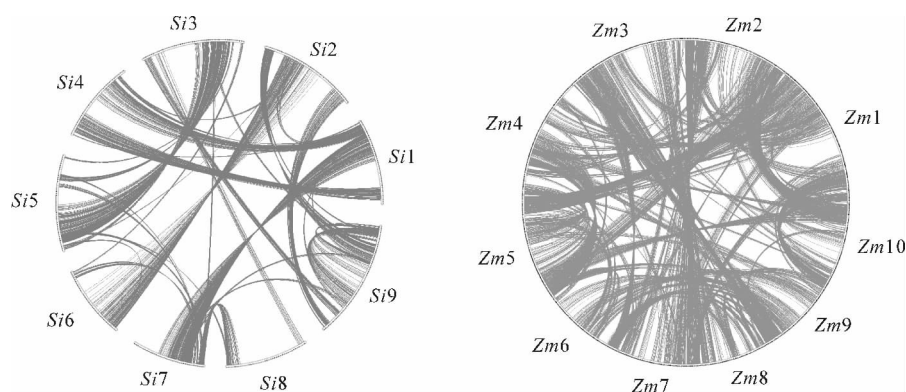
在谷子基因组内共有 402 个共线性区域,包含 3 846 对旁系同源基因对,占其全基因组的 10.9%,

其中最长的区域存在于 3 号与 5 号染色体之间,有 279 对旁系同源基因对,长度大于 10 对旁系同源基因对的区域有 60 个,大于 50 对的区域有 8 个;玉米

基因组内共有 468 个共线性区域,包含 6 016 对旁系同源基因对,占其全基因组的 18.5%,其中最长的区域存在于 2 号与 7 号染色体之间,共有 256 对旁系同源基因,长度大于 10 对旁系同源基因的区域有 132 个,大于 50 对的区域有 15 个。谷子和玉米基因组之间直系同源区域有 216 个,包含 13 652 对直系同源基因对,其中最长的区域存在于谷子 1 号与玉米 5 号染色体之间,共有 763 对直系同源基因对,种外旁系同源区域有 81 个,包含 2 454 对中外旁系同源基因对,其中最长的区域存在于谷子 3 号与玉米 3 号染色体之间,共有 212 对。

比较发现,谷子基因组中的共线区域少于玉米,这是由玉米的 2 次基因组加倍所导致,如果没有重复基因丢失,它们之间的重复基因数量理论上相差

1 倍。数据显示,玉米中由全基因组加倍产生的重复基因数量不足谷子的 2 倍,原因在于物种基因组加倍之后的基因丢失不一致,这可能是导致物种分化的一个重要动力<sup>[20]</sup>。重复基因不均匀地分布在每条染色体上,且常常是分布在靠近染色体的末端位置(图 3),原因在于远离着丝粒位置上存在许多保守或者冗余的序列,并且在这一区域上的基因密度比较高。远离着丝粒的重复基因,在进化过程中可能更容易发生变异<sup>[21]</sup>,使得基因序列变异为新的基因,进而出现新的功能,支持物种生命的生存<sup>[22]</sup>,但也可能在进化过程中使得该基因序列丢失,或者是死亡,因为对于一个生物体来说往往不需要 2 个相同的拷贝基因存在。植物基因组中大量重复基因的存在决定了植物基因组比动物进化快<sup>[18]</sup>。



圆周按逆时针依次排列的是谷子的 9 条染色体和玉米的 10 条染色体,每条灰色曲线连接的是基因组加倍产生的 1 对重复基因,这里去掉了共线性区域少于 10 对旁系同源对的基因

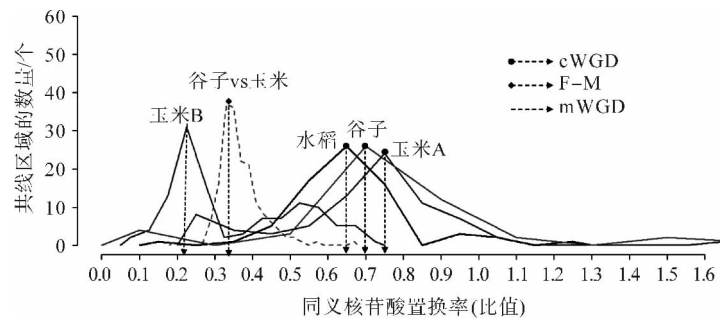
图 3 谷子、玉米中的重复基因

## 2.4 基因组进化事件时间推断

为了推断谷子和玉米基因组的进化,以水稻基因组为参考揭示相对的进化过程。水稻是禾本科植物进行比较基因组学研究的典型模式生物,因为它的进化速度较慢,更好地保持了祖先基因组序列的特征;利用获得的基因组内和基因组间共线性基因推断进化时间,将玉米 2 次加倍产生的重复基因区域进行区分,用以推断出更为准确的加倍时间。物种多倍化的发生可以用基因组内旁系同源共线区域推断加倍的相对时间,利用基因组间直系同源共线区域推断物种分化时间。核苷酸同义置换率( $K_s$ )是利用物种进化分析软件 PAML<sup>[23]</sup>中包含的 Nei-Gojobori 方法计算<sup>[24]</sup>,由于禾本科基因同义核苷酸置换率为每年  $6.1 \times 10^{-8}$ <sup>[25]</sup>,因此对共线性区域旁系或直系同源基因对间核苷酸同义置换率的均值进行统计,用来估计进化事件的相对时间。

禾本科植物基因组共同经历的一次古老全基

因组加倍可能是在白垩纪—第三纪(Cretaceous-Tertiary)约 65 个百万年前<sup>[26]</sup>的大灭绝时代,基因组加倍之后的物种适应了环境变迁得以生存。比较水稻、谷子、玉米物种基因组内同源染色体片段间  $K_s$  分布(图 4)发现,水稻  $K_s$  分布在 0.65,而谷子和玉米  $K_s$  分别分布在 0.70 和 0.75,这表明谷子和玉米基因进化速率分别高于水稻 1/13 和 2/13。值得注意的是,玉米经历了 2 次全基因组加倍后,进化速率不仅高于水稻,而且高于谷子 1/14,这表明物种全基因组加倍可能加速了物种进化。谷子和玉米之间的  $K_s$  分布对应应在 0.32 位置,对其进行校正后应该是 0.245,若物种全基因组加倍是在 65 个百万年前,那么谷子和玉米分化的时间约在 24.5 个百万年前;玉米第 2 次加倍之后产生的重复基因间的  $K_s$  分布在 0.22,对其进行校正后应该是 0.12,那么玉米的最近加倍是在约 12 个百万年前发生。



cWGD 是禾本科物种共同祖先经历的全基因组加倍, mWGD 是玉米基因组在物种分化之后独立发生的全基因组加倍, F-M 是谷子和玉米物种分化

图 4 物种基因同义核苷酸置换率分布

### 3 结论

物种多次多倍化为其遗传创新提供了重要材料,同时也增加了基因组同源结构的复杂性。本研究通过基因共线性的多重序列比对,获得同源染色体片段,根据直系同源片段高于旁系同源染色体片段间的相似性,对种内和种间基因组同源结构进行分析,建立了一种新的方法用以确定种内和种间复杂的同源共线信息。该方法应用于  $C_4$  植物谷子和玉米基因组的比较分析,有效获得了谷子和玉米基因组内和基因组间的同源共线信息,在谷子基因组中的 10.9%(3 846 对)、玉米基因组中的 18.5%(6 016 对)重复基因由多倍化产生,并且鉴定了它们之间直系同源基因对为 13 652 对,为进一步研究重复基因进化提供了良好的数据材料。统计重复基因对间的同义核苷酸置换率发现,玉米重复基因对间核苷酸置换率分布具有 2 个峰值,揭示玉米与谷子不仅同时经历了一次古老的全基因组加倍,而且在物种分化之后玉米在约 12 个百万年前独立发生了一次较近的全基因组加倍,此次加倍增加了玉米基因组中的重复基因,同时也加快了物种进化,研究表明,玉米基因组进化速率比谷子快 1/14。

基因组加倍产生的重复基因不均匀地保留在每条染色体上,靠近染色体末端的重复基因偏向于保留在物种基因组中,虽然全基因组加倍可以导致染色体重排,出现非同源染色体重构,但这些重复基因在一定程度上维持了基因组重排后部分同源染色体间的序列相似性,为基因重组提供了机会。虽然重复基因维持了序列相似性,但比较谷子和玉米重复基因保留情况发现,不同物种基因丢失存在极大差异,这种差异性可能成为物种分歧的驱动力。关于物种基因组加倍后产生的大

量重复基因间的重组模式,以及其如何影响物种基因组进化,有待于进一步深入研究。

#### 参考文献:

- [1] Gao L Z, Innan H. Very low gene duplication rate in the yeast genome[J]. Science, 2004, 306: 1367-1370.
- [2] Byrne K P, Wolfe K H. The yeast gene order browser: Combining curated homology and syntenic context reveals gene fate in polyploid species[J]. Genome Res, 2005, 15: 1456-1461.
- [3] Chapman B A, Bowers J E, Feltus F A, et al. Buffering of crucial functions by paleologous duplicated genes may contribute cyclicity to angiosperm genome duplication[J]. Proc Natl Acad Sci USA, 2006, 103: 2730-2735.
- [4] Bodt D S, Maere S. Genome duplication and the origin of angiosperms[J]. Trends Ecol Evol, 2005, 20 (11): 591-597.
- [5] He X, Zhang J. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution[J]. Genetics, 2005, 169 (2): 1157-1164.
- [6] Zhang G Y, Liu X, Wang J. Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential[J]. Nature Biotechnology, 2012, 30(6): 549-556.
- [7] Bennetzen J L, Schmutz J, Devos K M, et al. Reference genome sequence of the model plant *Setaria* [J]. Nature Biotechnology, 2012, 30(6): 555-561.
- [8] Schnable P S. The B73 maize genome: Complexity, diversity and dynamics[J]. Science, 2009, 326: 1112-1115.
- [9] Vandepoele K, Saeys Y, Simillion C, et al. The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice[J]. Genome Res, 2002, 12(11): 1792-1801.

- [10] Salse J, Abrouk M, Murat F, *et al.* Improved criteria and comparative genomics tool provide new insights into grass paleogenomics[J]. *Briefings in Bioinformatics*, 2009, 10: 619-630.
- [11] Salse J, Abrouk M, Bolot S, *et al.* Reconstruction of monocotyledonous proto-chromosomes reveals faster evolution in plants than in animals[J]. *Proc Natl Acad Sci USA*, 2009b, 106: 14908-14913.
- [12] Tang H, Wang X, Bowers J E, *et al.* Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps[J]. *Genome Res*, 2008, 18: 1944-1954.
- [13] Wang X, Shi X L, Li Z, *et al.* Statistical inference of chromosomal homology based on gene colinearity and applications to *Arabidopsis* and rice[J]. *BMC Bioinformatics*, 2006, 7: 447-452.
- [14] Wang X, Tang H. Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization[J]. *Genome Res*, 2009, 19(6): 1026-1032.
- [15] Wang X, Tang H A. Gene conversion in angiosperm genomes with an emphasis on genes duplicated by polyploidization[J]. *Genes*, 2011, 10(2): 1-20.
- [16] Ratnaparkhe M B, Wang X, Li J, *et al.* Comparative analysis of peanut NBS-LRR gene clusters suggests evolutionary innovation among duplicated domains and erosion of gene microsynteny[J]. *New Phytol*, 2012, 192: 164-178.
- [17] Paterson A H, Bowers J E, Chapman B A. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics[J]. *Proc Natl Acad Sci USA*, 2004, 101: 9903-9908.
- [18] Wang X, Shi X, Hao B, *et al.* Duplication and DNA segmental loss in the rice genome: Implications for diploidization[J]. *New Phytol*, 2005, 165: 937-946.
- [19] Schnable J C, Springer N M, Freeling M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss[J]. *PNAS*, 2011, 108(10): 4069-4074.
- [20] Gonzalez-Escalona N, Romero J, Espejo R T. Polymorphism and gene conversion of the 16S rRNA genes in the multiple rRNA operons of *Vibrio parahaemolyticus*[J]. *FEMS Microbiol Lett*, 2005, 246: 213-219.
- [21] Galtier N. Gene conversion drives GC content evolution in mammalian histones[J]. *Trends Genet*, 2003, 19: 65-68.
- [22] Wang X. The sequence of rice chromosomes 11 and 12, rich in disease resistance genes and recent gene duplications[J]. *BMC Biol*, 2005, 3: 20.
- [23] Yang Z. PAML4: Phylogenetic analysis by max likelihood[J]. *Mol Biol Evol*, 2007, 24: 1586-1591.
- [24] Nei M, Gojobori T. Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions[J]. *Mol Biol Evol*, 1986, 3: 418-426.
- [25] Gaut B S. Molecular clocks and nucleotide substitution rates in higher plants[J]. *Evol Biol*, 1998, 30: 93-120.
- [26] Fawcett J A, Maere S, Van de Peer Y. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event[J]. *Proceedings of the National Academy of Sciences*, 2009, 14: 6-12.