

# 谷子、玉米重复基因间基因置换的系统发育分析

王金朋, 聂林曼, 王振怡, 马雪莲, 张 琼

(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, 利用系统发育分析和比较基因组学推断了重复基因间可能的基因置换, 旨在对其基因组中由全基因组加倍产生的重复基因的相互作用机制进行研究。结果表明, 谷子中有 192 对(18.9%)、玉米中有 121 对(11.9%)的重复基因在其进化过程中发生了基因置换, 且重复基因对之间的置换常常是以部分基因置换为主, 谷子发生置换的基因对中有 88.4%、玉米中有 76.5%是部分基因置换; 基于动态规划和系统发育相结合的方法, 确定了谷子、玉米多个重复基因对间发生过不止 1 个片段的基因置换; 对基因置换与重复基因在染色体上物理位置的相关性分析发现, 基因置换与基因在染色体上的位置显著相关, 对不同染色体区域重复基因发生置换的频率进行统计分析表明, 靠近染色体末端的重复基因对更易发生置换。

**关键词:** 谷子; 玉米; 重复基因; 基因置换; 系统发育分析

**中图分类号:** S513 S515 **文献标志码:** A **文章编号:** 1004-3268(2014)12-0034-06

## Phylogenetic Analysis of Gene Conversion between Duplicated Genes in *Setaria italica* and *Zea mays*

WANG Jin-peng, NIE Lin-man, WANG Zhen-yi, MA Xue-lian, ZHANG Qiong

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

**Abstract:** *Setaria italica* and *Zea mays* were taken as the materials to study the interaction mechanism of duplicate genes produced by whole genome doubling, and to infer the possible gene conversion between duplicate genes by means of the phylogenetic analysis and comparative genomics. The result showed that there were 192 pairs of duplicate genes in *Setaria italica* and 121 pairs in *Zea mays* that came into gene conversion in the evolution, most of which were partial conversion that accounted for 88.4% in *Setaria italica* and 76.5% in *Zea mays*. Based on the method combining dynamic programming and phylogenetic development, we confirmed that one or more fragments of gene conversion occurred between the multiple duplicate gene pairs in *Setaria italica* and *Zea mays*. The correlation analysis of gene conversion and duplicated genes in the physical location of chromosome showed that there was significant correlation between the gene conversion and the gene location on chromosome. The frequency of gene conversion in different chromosomal regions was analyzed statistically, which showed that the duplicate genes close to the terminal of chromosome were more prone to convert.

**Key words:** *Setaria italica*; *Zea mays*; duplicated genes; gene conversion; phylogenetic analysis

多倍化(polyploidy)作为物种进化过程中的一个重要动力, 可迅速使全基因组加倍(whole ge-

nome duplication), 产生大量的重复基因, 为遗传创新提供了材料来源<sup>[1-2]</sup>。全基因组加倍创造的重复

收稿日期: 2014-06-18

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: 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

染色体或染色体片段常常造成基因组不稳定,导致大量 DNA 片段丢失、基因倒位和整个基因组的 DNA 重排<sup>[3-4]</sup>,基因组重排的结果会产生新的同源染色体对(neo-homologous chromosomes)<sup>[5]</sup>,这一新同源染色体对相对于基因组重排之前并非同源染色体对,而是全基因组加倍之后,物种非同源染色体经历重排保留了由加倍产生的重复基因片段。对谷子、玉米由全基因组加倍产生的重复基因间相互作用的规模、规律进行研究,有利于揭示物种多样化的驱动力,为物种其他相关研究提供重要意义,尤其对于研究有重要经济价值基因的功能进化有重要意义,具有潜在经济价值。

遗传重组(genetic recombination)作为生物进化的一个重要动力,不仅可以对遗传过程中 DNA 序列损伤进行修复,而且可以进行同源序列间信息传递<sup>[6-7]</sup>,即同源重组。相对于同源重组部分同源染色体对之间的重组被称为非正常遗传重组<sup>[5]</sup>。物种多倍化之后产生的新同源染色体对之间重复片段间的重组实质是一种非正常遗传重组,重组可以是相互或者单向传递遗传信息,单向从一个 DNA 片段向其同源 DNA 片段上传递遗传信息的过程为基因置换(gene conversion)<sup>[8]</sup>。有学者基于比较基因组学分析了大鼠、小鼠基因组中由全基因组加倍产生的重复基因,发现其中有 18% 的重复基因对在分化之前发生了置换<sup>[9]</sup>;利用 GENECONV 方法<sup>[10]</sup>搜索重要模式植物拟南芥旁系同源基因对(paralogs)间可能的基因置换,没有发现重复基因对间发生基因置换的证据<sup>[11]</sup>,然而,应用同样的方法对水稻 626 个多基因家族进化规律进行分析,确定了 377 个基因置换事件<sup>[12]</sup>,导致这一结果出现的原因可能是拟南芥重复基因对比水稻分化得快,或者是由于拟南芥多次全基因组加倍后大量的基因丢失使得重复基因对避免了基因置换的发生。关于多个重要基因家族的进化研究表明,基因置换可能会对一些基因家族进化产生不同程度的影响,如对花生抗病基因<sup>[13]</sup>、组蛋白和 rRNA<sup>[14-15]</sup> 的比较分析表明,基因家族进化中都存在基因置换事件。上述研究结果暗示了物种全基因组加倍创造的大量重复基因片段在进化过程中受基因置换的影响。然而,在全基因组水平上对重复基因间基因置换发生的规模、形式和规律还处于探索阶段。

禾本科植物共有 1 个多倍体祖先物种<sup>[3,16]</sup>,且在大约 7 000 万 a 前发生过 1 次多倍化事件。谷子(*Setaria italica*)和玉米(*Zea mays*)作为禾本科植物家族的成员,不仅是重要的食物和饲料作物,而且

可以作为 C<sub>4</sub> 生物燃料。最近它们的全基因组测序工作相继完成<sup>[17-19]</sup>,为比较基因组学研究提供了良好数据材料。关于谷子和玉米全基因组加倍已有一些研究<sup>[3,17-19]</sup>,但基因组加倍之后产生的重复基因间是否发生过基因置换未见报道。为此,以谷子和玉米全基因组为研究对象,利用多重序列比对工具 McScan<sup>[11]</sup>和共线性方法 ColinearScan<sup>[20]</sup>确定基因组内由全基因组加倍产生的重复基因,基于比较基因组学系统发育分析重复基因间可能的基因置换事件,揭示谷子和玉米重复基因间发生基因置换的规模、形式,旨在为摸清多倍化产生的部分同源染色体对间重复基因进化机制提供理论依据。

## 1 材料和方法

### 1.1 物种基因组数据

禾本科物种玉米和谷子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,其中包括每个物种的全基因组 DNA 序列、蛋白质序列,以及基因序列的注释文件。

### 1.2 基因共线性推断

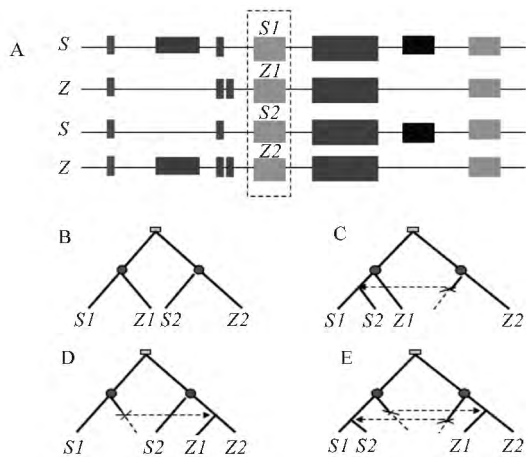
首先,对种内和种间基因组进行 Blast 比对分析得到基因同源性;然后,利用多重序列比对工具 McScan 寻找基因组内和基因组间同源共线 DNA 片段,同源共线区域中每对匹配基因的打分为  $\min(-\lg^{E\text{-value}}, 50)$ ,失配得分为 -1,打分路径中匹配基因对与邻近匹配基因对之间的距离超过 10 000 bp 时得分为 -1,并要求共线性区域的打分大于 300 时保留;最后利用加入统计估计的共线性分析工具 ColinearScan 对获得的基因组共线区域进行统计,估计其显著性,将显著性大于  $1e-10$  的区域去掉,确定基因组内重复基因和基因组间的同源共线信息。

### 1.3 基因置换推断

同源基因四联子(quartet)是由谷子的 1 对旁系同源基因 S1 和 S2,以及它们各自对应应在玉米中的直系同源基因(orthologs) Z1 和 Z2 共同构成的 1 组同源基因(图 1A)。

根据系统发育推断基因置换:对得到的同源基因四联子,应用 ClustalW<sup>[21]</sup>进行多重序列比对。比对后序列的空位数占总序列长度的 50% 以上,氨基酸一致性小于 40% 时删除,然后进行基因置换分析。(1)推断全基因置换,由于旁系同源基因产生早于物种分化,因此两物种间直系同源基因对 S1-Z1 和 S2-Z2 之间相似性应该高于它们旁系同源基因对(图 1B),然而基因对由于置换事件的发生导致它

们之间的相似性发生变化(图 1C-E),利用基因树的拓扑结构变迁推断可能的基因置换,对于发生置换的基因树都进行置换检验(bootstrap),获得置换基因的置信度。(2)部分基因置换事件的发生利用系统发育树和动态规划相结合来推断<sup>[22]</sup>。序列相似性用同源基因间的同义核苷酸置换率( $K_s$ )度量,利用物种进化分析软件 PAML<sup>[23]</sup>中的 Nei-Gojobori 方法计算<sup>[24]</sup>。



A 中不同颜色的矩形表示同源染色体片段上的基因,相同颜色表示同源基因;B-E 中矩形表示物种基因组加倍,圆表示物种分化。B. 假如没有基因置换发生;C. S2 被 S1 置换;D. Z1 被 Z2 置换;E. 2 个物种都发生置换

图 1 同源基因四联子和基因置换推断

## 2 结果与分析

### 2.1 谷子和玉米中的基因置换

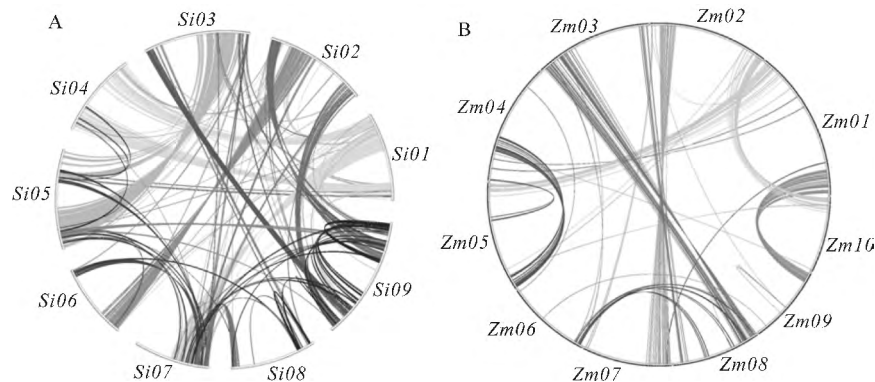
根据基因共线性分析,并将共线性基因 blocks 小于 10、显著性小于  $1e-3$  去掉,获得了谷子中 2 748 对、玉米中 1 835 对重复基因对,两物种之间有 12 647 对直系同源基因对,这些重复基因不均匀

地分布在谷子和玉米的每条染色体上(图 2)。基于基因组内和基因组间的同源共线信息,在谷子和玉米之间共发现 1 128 个同源基因四联子。去掉序列较为分化的同源基因四联子,即对得到的同源基因四联子进行多重序列比对后,若序列的空位数目占总序列长度 50% 以上,氨基酸一致性小于 40% 时删除掉,得到 1 014 个同源基因四联子,用于进行基因置换的系统发育分析。

通过对获得的同源基因四联子构树,根据树的拓扑结构变迁,推断谷子中有 18.9% (192 对) 重复基因在谷子、玉米分化之后发生过基因置换,其中包括 2.2% 的全基因置换和 16.7% 的部分基因置换,发生置换的基因对的置信度在 80% 以上。在玉米中发现了相对较少的重复基因对发生基因置换,玉米重复基因的 11.9% (121 对) 在进化过程中受基因置换的影响,其中 2.8% 发生了全基因置换,9.1% 发生了部分基因置换,发生基因置换的置信度在 80% 以上。

### 2.2 局部重复基因对置换事件

为进一步认识重复基因间的置换事件,以及它们间发生的不同形式的置换,对局部的重复基因对进行了比较分析。如谷子中的 1 对旁系同源基因对 *Silg213901* 和 *Si4g371972* 之间的  $K_s$  为 0.009 1,小于它们与各自在玉米中的直系同源基因 *Zm4g240610886* 和 *Zm9g27092199* 间的  $K_s$  (0.027 5 和 0.121 2),这表明谷子中的这对旁系同源基因在物种分化之后发生了全基因置换,但这一四联子中玉米的旁系同源基因对间并未发生基因置换,因为它们间的  $K_s$  为 0.123 1,大于它们与各自直系同源基因对间的距离,置换检验这一四联子构建的树,结果显示其置信度为 100%。将这一四联子利用 MEGA 进行构树(图 3A),显示谷子旁系同源基因对之间的相似



A. 谷子中重复基因; B. 玉米中重复基因。圆周按逆时针依次排列的是谷子的 9 条染色体、玉米的 10 条染色体,每条灰色的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,且玉米基因中只有禾本科共同加倍产生的重复基因

图 2 谷子和玉米中的重复基因

性高于它们与各自直系同源基因对的相似性。同样地在玉米中也存在这样的基因对,如旁系同源基因对 *Zm1g281320337* 和 *Zm5g5636846*  $K_s$  为 0.035 9, 小于它们各自与直系同源基因 *Si2g1861079* 和 *Si9g4451048* 间的  $K_s$  (0.256 3 和 0.048 7), 这表明玉米的这一旁系同源基因对发生过全基因置换。构建这一四联子的系统发育树(图 3B), 发现玉米的 1 对旁系同源基因(*Zm1g281320337* 和 *Zm5g5636846*)相似性高于他们各自与直系同源基因对的相似性, 推断玉米染色体片段上的基因 *Zm5g5636846* 是由对应同源染色体片段上的 *Zm1g281320337* 置换而来。

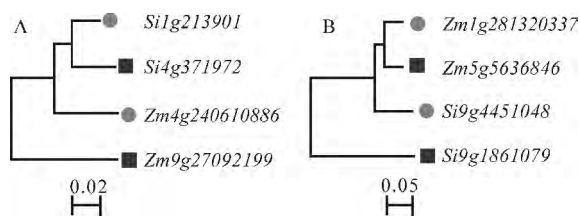


图 3 同源基因四联子系统发育树

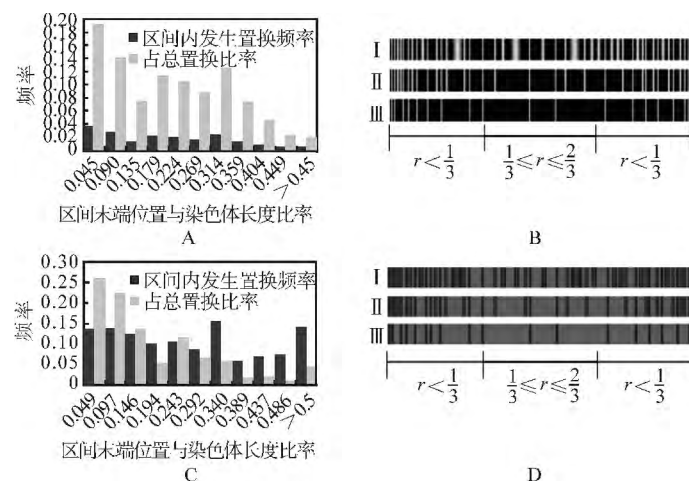
对于有些旁系同源基因对是否发生置换, 基于基因间的相似性并不能确定, 因此, 利用基于动态规划的方法, 在核苷酸水平上对同源基因四联子进行了系统发育分析, 发现物种重复基因对间存在部分基因置换。如谷子旁系同源对 *Si1g40501504* 和 *Si4g3847281* 之间发生了部分基因置换事件, 基因 *Si4g3847281* 作为供体, 它的第 1 393—1 523 bp 和 449—578 bp 2 个 DNA 片段置换了 *Si1g40501504* 上的核苷酸序列, 其置信度都在 80% 以上; 玉米旁系同源对 *Zm7g167980519* 和 *Zm9g126933531* 发生了 2 个

DNA 片段的部分基因置换, 分别是 *Zm7g167980519* 基因第 1 055—1 091 bp 的片段和 849—902 bp 的片段置换 *Zm9g126933531* 序列上的遗传信息。

### 2.3 基因置换与物理位置

为了推断基因置换是否和基因在染色体上的物理位置有关, 这里分别对谷子和玉米中的基因置换事件和基因在染色体上的位置分布进行了相关性研究。在谷子中发生基因置换的重复基因距离染色体末端的平均距离约 8 Mb, 小于未发生基因置换基因对距离染色体末端的距离(约 10 Mb,  $P < 0.05$ ); 在玉米中发生基因置换的重复基因距离染色体末端的平均距离约 30 Mb, 小于未发生基因置换基因对距离染色体末端的距离(约 35 Mb,  $P < 0.05$ ), 这一数据结果表明重复基因间发生置换的基因比未发生置换的基因更靠近染色体末端。

发生置换的重复基因较未发生置换的基因距离染色体末端更近, 但并不能说明靠近末端的重复基因对置换具有偏好性, 为此将重复基因距离染色体末端的距离划分为 11 个区间, 每个区间分别是该区间的终止位置与染色体长的比率, 并统计每个区间内重复基因对发生置换的频率。统计结果显示, 谷子和玉米中靠近染色体末端的重复基因比其他重复基因间发生置换频率高(图 4A、C), 置换频率超过 1% 的重复基因都集中在染色体末端区段, 且这些区段发生置换的基因数占有置换基因数的 90% 以上(图 4B、D)。根据这些基因在染色体上的分布统计, 可以推断靠近染色体末端的重复基因相对于其他重复基因易受基因置换事件的影响。



A、C 分别为谷子、玉米重复基因置换频率分布; B、D 分别为谷子、玉米基因在染色体上分布示意图。I. 根据基因在区间上的数量统计谷子和玉米全基因组在染色体上的分布; II. 谷子和玉米中的重复基因在染色体上的分布; III. 谷子和玉米中发生置换的重复基因在染色体上的分布;  $r$  表示基因距染色体末端距离与染色体长的比率,  $r < 1/3$  的区域内, 物种所有发生置换的重复基因在 90% 以上, 且区间里发生置换与总重复基因数的比率  $> 1\%$ ;  $1/3 \leq r \leq 2/3$  区域内, 物种所有发生置换的重复基因不足 10%, 且区间里发生置换与总重复基因数的比率  $< 1\%$

图 4 重复基因置换频率分布和基因在染色体上的分布

### 3 结论与讨论

被子植物在进化过程都发生过 1 次或多次全基因组加倍事件<sup>[25]</sup>,这就为部分同源染色体片段间非正常遗传重组创造了条件。全基因组加倍之后产生的大量重复基因片段常常导致物种基因组极其不稳定,大量 DNA 片段丢失和染色体重排,从而出现新的“同源染色体对”保留了由全基因组加倍产生的重复基因片段。虽然 DNA 重排可以降低这一同源染色体对相互作用的概率,但是它们独立保持下来的过程中仍然发生着非正常的遗传重组,一个重要的特征就是重复基因间有基因置换事件发生。

本研究发现,谷子中有 192 对(18.9%)、玉米中有 121 对(11.9%)的重复基因在其进化过程中发生了基因置换,且重复基因对之间的置换常常是以部分基因置换为主,谷子发生置换的基因对中有 88.4%、玉米中有 76.5%是部分基因置换;基于动态规划和系统发育相结合的方法,确定了谷子、玉米多个重复基因对间发生过不止 1 个片段的基因置换;基因置换与重复基因在染色体上物理位置的相关性分析发现,基因置换与基因在染色体上位置显著相关,对不同染色体区域重复基因发生置换的频率进行统计分析表明,靠近染色体末端的重复基因对更易于发生置换。究其原因在于重组是以序列相似性为基础,那么靠近末端的基因易发生置换是合理的。一方面,靠近染色体末端的基因序列比其他位置上的基因保守,基因共线性常常是处在基因密度高的区域<sup>[26]</sup>,重组可能是为了保持序列相似性而去掉一些有害变异<sup>[27]</sup>;另一方面,靠近着丝粒区的重复元件比较多,降低了序列相似性。

基于系统发育分析在谷子和玉米重复基因中发现了基因置换事件,且有全基因和部分基因置换,初步研究了导致基因置换的因素,但仍然存在一些重要的问题尚不清楚,如基因置换如何影响序列碱基含量、物种进化速率、基因功能等一系列重要问题;若要全面了解物种重复基因置换事件及其对遗传进化带来的影响,有待今后更深入地研究。

#### 参考文献:

- [1] 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.
- [2] De B S, Maere S, Van de Peer Y. Genome duplication and the origin of angiosperms[J]. *Trends Ecol Evol*, 2005, 20: 591-597.

- [3] Paterson A H, Bowers J E. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics[J]. *Proc Natl Acad Sci USA*, 2004, 101(26): 9903-9908.
- [4] Wang X, Shi X. Duplication and DNA segmental loss in the rice genome: Implications for diploidization[J]. *New Phytol*, 2005, 165(3): 937-946.
- [5] Wang X, Tang H. Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization[J]. *Genome Res*, 2009, 19: 1026-1032.
- [6] Puchta H, Dujon B, Hohn B, *et al.* Two different but related mechanisms are used in plants for the repair of genomic double-strand breaks by homologous recombination[J]. *Proc Natl Acad Sci USA*, 1996, 93: 5055-5060.
- [7] Khakhlova O, Bock R. Elimination of deleterious mutations in plastid genomes by gene conversion[J]. *Plant J*, 2006, 46: 85-94.
- [8] Datta A, Hendrix M, Lipsitch M, *et al.* Dual roles for DNA sequence identity and the mismatch repair system in the regulation of mitotic crossing-over in yeast[J]. *Proc Natl Acad Sci USA*, 1997, 94: 9757-9762.
- [9] Ezawa K, Oota S, Saitou N. Genome-wide search of gene conversions in duplicated genes of mouse and rat[J]. *Mol Biol Evol*, 2006, 23(5): 927-940.
- [10] Sawyer S. Statistical tests for detecting gene conversion[J]. *Mol Biol Evol*, 1989, 6: 526-538.
- [11] Zhang L, Vision T J, Gaut B S. Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana* [J]. *Mol Biol Evol*, 2002, 19: 1464-1473.
- [12] Xu S, Clark T. Gene conversion in the rice genome [J]. *BMC Genomics*, 2008, 9: 93.
- [13] Brown R D, Mattoccia E. On the role of RNA in gene amplification[J]. *Acta Endocrinol Suppl (Copenh)*, 1972, 168: 307-318.
- [14] Ohta T. Some models of gene conversion for treating the evolution of multigene families [J]. *Genetics*, 1984, 106(3): 517-528.
- [15] Bowers J E, Arias M A, Asher R, *et al.* Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses[J]. *Proc Natl Acad Sci USA*, 2011, 102: 13206-13211.
- [16] Salse J, Piegus B. New in silico insight into the synteny between rice (*Oryza sativa* L.) and maize (*Zea mays* L.) highlights reshuffling and identifies new duplica-

- tions in the rice genome[J]. Plant J, 2004, 38(3): 396-409.
- [17] Zhang G Y, Liu X, Wang J, *et al.* Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential[J]. Nature Biotechnology, 2012, 30(6): 549-556.
- [18] 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.
- [19] Schnable P S. The B73 maize genome: Complexity, diversity and dynamics[J]. Science, 2009, 326: 1112-1115.
- [20] 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.
- [21] Thompson J D, Higgins D G, Gibson T J. Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice[J]. Nucleic Acids Res, 1994, 22: 4673-4680.
- [22] Wang X, Tang H, Bowers J E, *et al.* Extensive concerted evolution of rice paralogs and the road to regaining independence [J]. Genetics, 2007, 177: 1753-1763.
- [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] Tang H, Bowers J E. Synteny and collinearity in plant genomes[J]. Science, 2008, 320: 486-488.
- [26] Carvalho A B. The advantages of recombination[J]. Nat Genet, 2003, 34: 128-129.
- [27] Yu J, Wang J, Lin W, *et al.* The genomes of *Oryza sativa*: A history of duplications[J]. PloS Biol, 2005, 3: 266-281.

## 欢迎订阅 2015 年《河南农业科学》

《河南农业科学》是河南省农业科学院主办的综合性农业科技期刊。多年来,深受省内外农业科技人员、农业院校师生等涉农读者的喜爱。本刊连续被评为全国中文核心期刊、中国科技核心期刊、中国科学引文数据库(CSCD)来源期刊、RCCSE 中国核心学术期刊(A<sup>+</sup>)、中国农业核心期刊。曾多次获得有关部门的奖励,被评为“全国优秀农业期刊”;连续荣获“河南省优秀科技期刊一等奖”“河南省自然科学期刊综合质量检测一级期刊”“河南省第一、二届自然科学二十佳期刊”。

栏目设置有:综述、作物栽培·遗传育种、农业资源与环境、植物保护、园艺·林学、畜牧·兽医、农产品加工·农业工程·农业信息技术。

本刊为月刊,国际标准 16 开本,160 页,彩色封面,每期定价 18.00 元,全年 216 元。各地邮局均可订阅,邮发代号:36—32。如错过订期,可直接与本刊编辑部联系订阅。

地址:郑州市花园路 116 号

邮编:450002

电话:0371—65739041

E-mail:hnnykx@163.com

传真:0371—65712747

网址: <http://www.hnnykx.org.cn>