

玉米亚基因组中代谢基因的分化研究

张 宪¹, 王金朋², 雷天宇¹, 金殿川^{1,2*}

(1. 河北联合大学 理学院, 河北 唐山 063009; 2. 河北联合大学 生命科学学院, 河北 唐山 063009)

摘要: 为了确定玉米代谢通路中 2 个亚基因组基因的分化情况, 对玉米和高粱进行了共线性分析, 确定了玉米的 2 个亚基因组 Maize1 和 Maize2 基因组成, 统计了玉米所有 407 个代谢通路中的亚基因组基因保留情况, 并对它们的基因长度进行了双样本均值分析(Z -检验)和通路中 2 个亚基因组基因保留情况进行 χ^2 检验。结果表明, Maize1、Maize2 的基因长度没有显著的差异; 基因保留具有显著差异的通路大约占总代谢通路的 4.4%, 作为亚基因组, Maize2 在全局上受到更大基因丢失的影响, 有更多的 DNA 丢失, 而 Maize1 相对保留了更多的代谢基因, 拥有更大的表达量。

关键词: 玉米; 亚基因组; 代谢通路; 基因丢失; 分化

中图分类号: S513 文献标志码: A 文章编号: 1004-3268(2014)02-0019-05

Differentiation of Metabolic Genes in Subgenomes of Maize

ZHANG Xian¹, WANG Jin-peng², LEI Tian-yu¹, JIN Dian-chuan^{1,2*}

(1. College of Science, Hebei United University, Tangshan 063009, China;

2. College of Life Science, Hebei United University, Tangshan 063009, China)

Abstract: In order to understand the differential difference of 2 subgenomes in maize, the colinearity analysis of maize and sorghum was conducted to determine the gene composition of subgenomes Maize1 and Maize2. The number of subgenomic genes in all the 407 metabolic pathways of maize was counted up. The double-sample mean (Z -test) analysis for the gene length and the chi-square test for the retention of two subgenomic genes in the pathways were carried out. The result showed that there was no significant difference of gene length between Maize1 and Maize2; the pathways with remarkable difference in gene retention were about 4.4% in the total metabolic pathways, the Maize2 (as subgenome) had been affected more by gene loss in the global and lost more DNA, while the Maize1 relatively reserved more metabolic genes and had more expression level.

Key words: maize (*Zea mays* L.); subgenomes; metabolic pathway; gene loss; differentiation

玉米是世界上广为种植的粮饲兼用作物, 其基因组是继水稻和高粱基因组测序完成后第 3 个成功测序的农作物, 这为物种的比较基因组学研究提供了良好的机会^[1-3]。全基因组倍增或多倍化过程被认为是植物尤其是禾本科作物物种形成和进化过程中非常普遍和重要的事件, 几乎所有的开花植物在进化过程中均经历了一次或多次染色体加倍过程^[4]。倍增产生的大量重复基因为物种的遗传创新提供了丰富的材料来源^[4-5]。基因组加倍后, 再经历

二倍化过程(diploidization), 进化成当代的二倍化物种, 这一过程伴随着大量基因丢失(gene loss)、染色体重排(chromosome recombination)、基因倒位(gene inversion), 成为物种演化最重要的动力源泉^[5-6]。不同物种基因组的同源染色体在进化过程中基因丢失率不一致, 且存在阶段性非正常遗传重组现象^[4,6-7]。不同基因组重复基因丢失和保留与基因在物种基因组中的功能紧密相关^[8-10]。玉米的一个重要特征是其存在由 2 个异源多倍体进化而来的

收稿日期: 2013-09-13

基金项目: 国家自然科学基金项目(31100913); 河北联合大学青年自然科学基金项目(201230)。

作者简介: 张 宪(1987-), 女, 河北晋州人, 在读硕士研究生, 研究方向: 应用数理统计。E-mail: heuuzx2867@gmail.com

* 通讯作者: 金殿川(1970-), 男, 河北唐山人, 副教授, 硕士, 主要从事生物信息学研究。E-mail: jindianchuan@aliyun.com

亚基因组,它们经历了不同程度的基因丢失^[11-12]。玉米物种中如今已识别的代谢通路有 407 个,所涉及的功能基因多达 4 082 个,玉米代谢通路中 2 个亚基因组基因丢失和保留是否一致已成为学术界关注的热点。为此,本研究通过对玉米和高粱基因组进行比较基因组学分析,确定代谢通路中代谢基因所属亚基因组,进而统计分析其基因丢失和保留的差异,揭示可能的分化规律,为玉米基因组研究提供理论基础和材料来源。

1 材料和方法

1.1 材料

玉米 (*Zea mays*) B73^[13] 和高粱 (*Sorghum bicolor*) 全基因组序列数据从植物基因组倍增数据库 Plant Genome Duplication Database (PGDD) (<http://chibba.agtec.uga.edu/duplication/index/files>) 下载得到,其中包含物种的 DNA 序列、氨基酸序列以及基因组的注释文件;玉米代谢通路中有关代谢基因数据从公共数据库 GRAMENE (<http://pathway.gramene.org/MAIZE/organism-summary>) 下载得到,包括所有的代谢基因、酶以及代谢通路信息。

1.2 共线性分析方法

玉米和高粱基因组内均含有大量由多倍化产生的重复基因,它们以不同的规模排列在物种不同染色体上,基于基因的同源性以及基因在染色体上的位置分布,若将任意 2 条染色体基因按顺序排列在二维平面图中,连续出现的同源基因在图中将以连续的点构成线形,这一染色体间的共线片段称为 1 个模块 (block),即共线性区域 (synteny)。这里对得到的每个模块进行了一次经验打分,打分规则采取的是一个动态规划的算法,即:

$$S(v) = M(v) + \max_m \{S(m) + G(m, v), 0\}$$

式中: S 为模块中共线基因对在基因链中得分; M 为基因对匹配时得分; G 为两基因对间的空位罚分; m 和 v 是在某个共线区域中的 2 个基因对,且 m 在 v 的前面出现。

这里运用 Perl 语言编程结合集成共线性分析软件 Mcscan 对玉米和高粱染色体进行共线性分析,设定的评分方案是对每 1 对匹配基因对打分为 $(-\log_{10} E_{\text{value}}, 50)$ 最小值,失配罚分为 -1,对任意同源基因对之间的距离设定为最大 1 000 bp;最后将得分 > 300 ,并且共线性区域的期望值 $E_{\text{value}} < 1 \times 10^{-5}$ 作为重要的分界标准,确定物种中由基因倍

增产生的旁系同源基因对 (paralogs),以及两物种之间的直系同源基因对 (orthologs)。玉米和高粱两物种的共线性示意图如下:

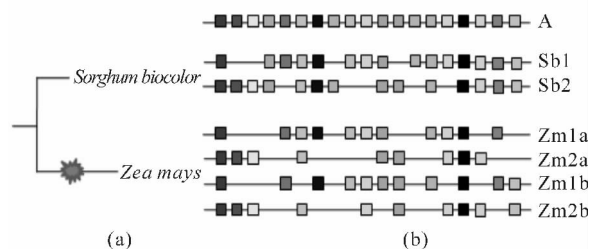


图 1 玉米和高粱物种间的共线性示意图

图 1(a)表示玉米和高粱物种的系统发育树,图中星形表示玉米在与高粱分化后的某一时刻发生过一次全基因组倍增;(b)中黑色的横线表示两物种中共线的 DNA 片段,横线上不同颜色的正方形代表物种中不同的基因。A 是假设的祖先物种中的共线基因;Sb1、Sb2 表示高粱物种内的共线基因模块,相对于祖先共线基因,旁系同源基因间经历了不同的基因丢失情况;Zm1a、Zm2a、Zm1b、Zm2b 表示玉米与高粱同源的共线模块。

2 结果与分析

2.1 玉米、高粱染色体的共线性分析

构建基因组水平点阵图是从整体上观察物种间共线情况以及发现可能的基因倍增的一种直观表示方法。根据 1.2 构建点图的方法得到玉米和高粱基因组水平上的点阵图 (图 2),图中横、纵轴分别代表玉米和高粱物种的 10 条染色体序列,深黑色的点表示玉米和高粱直系同源基因对,连续相邻同源基因对构成了直系同源的共线模块,用黑色斜线表示;浅

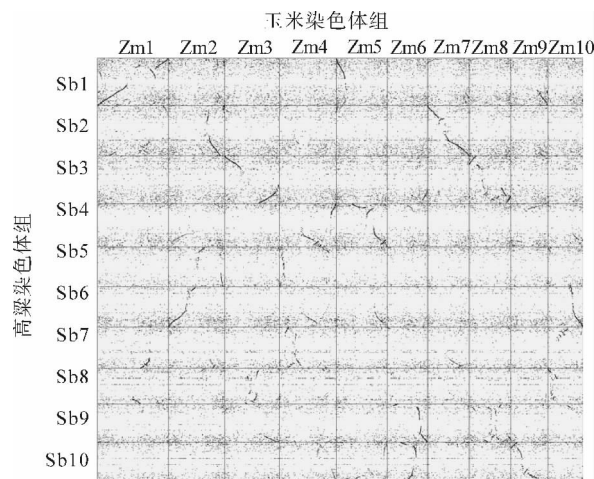


图 2 玉米和高粱基因组结构点阵图

灰色点表示的是物种基因组中旁系同源基因对,连续相邻同源基因对构成了旁系同源的共线模块,用灰色斜线表示,这很大程度上是由物种内基因倍增产生的重复基因构成的共线基因。

从玉米和高粱的比较作图中可以看出,两物种间存在广泛的共线性,且在共线区域高粱染色体基因同时与不同染色体上的 2 个玉米基因保持同源性,其从属于玉米上不同的亚基因组,从而可以根据高粱和玉米的共线同源关系得到玉米的 2 个亚基因组 Maize1 和 Maize2 的基因组成。为了进一步观察玉米和高粱基因的共线情况,分别对玉米、高粱以及它们之间的共线性区域进行了搜

索,在共线性数据文件中发现高粱、玉米各自染色体内部具有的共线模块分别为 170 个、332 个,所含旁系同源基因对个数均为 3 505 个,即玉米染色体相对于高粱来说含有更多的共线模块,但模块长度都比较短,长度大于 50 对同源基因的模块仅 0.3%左右;玉米和高粱物种间相对于它们自身具有更多的共线模块,存在 577 个模块,包含 20 023 个直系同源基因对,最长的模块存在于 Sb3 与 Zm3 上,共有 684 对直系同源基因,其长度主要集中在 10~50 对同源基因,占到总模块的 67.6%,大于 50 对同源基因的模块也有 96 个(约 16.6%),结果如图 3 所示。

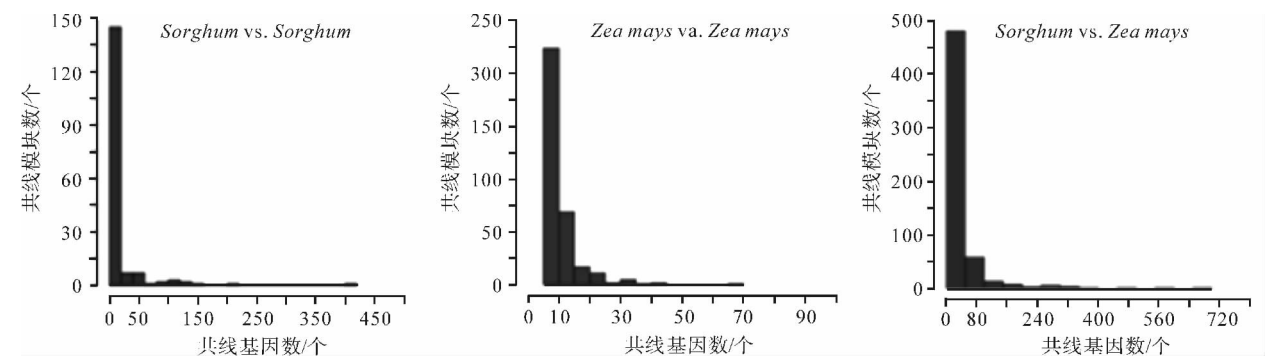


图 3 高粱、玉米及其之间共线模块分布

2.2 玉米代谢基因所属亚基因组的划分

根据 Schnable JC^[8] 等对 2 个亚基因组的定义,对所有的代谢基因进行了所属亚基因组(Maize1、Maize2)的划分。总体上两亚基因组在代谢通路中基本保持了与整个染色体组一致的丢失偏向性,在已识别代谢通路中存在的 4 082 个基因,属于 Maize1、Maize2 的基因分别有 2 476 个、1 551 个,即 Maize1 相对于 Maize2 保留了更多的代谢基因,说明 Maize1 中保留了更多的表达基因,对玉米总表达量的贡献更多。另外,对于剩余未确定归属的基因有 55 个,占总代谢基因的 1.3%,这对于研究代谢通路中 2 个亚基因组基因的分化影响不大。为了进一步观察代谢基因中 2 个亚基因组基因是否具有长度上的偏向性,研究中根据基因的起始位置和终止位置计算其长度,再依据统计上的双样本均值分析方法,对代谢基因中属于 Maize1、Maize2 的基因长度进行 Z-检验,分析中假设 Maize1、Maize2 基因长度均值差为 0,检验结果见表 1。

表 1 结果显示,由于 $Z < Z_{\alpha/2}$,所以接受原假设, Maize1、Maize2 基因长度并没有显著的差异,说明在基因长度上它们具有一定偏向性但不明显,其中 Maize2 的基因相对短一些。这个实际上说明作为

亚基因组, Maize2 在全局上受到更大的影响,有更多的 DNA 丢失,而且这种影响也反映到了基因的碱基构成上。

表 1 Z-检验:双样本均值分析

项目	Maize1	Maize2
平均	4 520.093 7	4 303.445 519
已知协方差	54 530 675.94	34 389 739.24
观测值	2 476	1 551
假设平均差	0	
Z	1.030 533 191	
$P(Z \leq z)$ 单尾	0.151 379 89	
Z 单尾临界	1.644 853 627	
$P(Z \leq z)$ 单尾	0.302 759 78	
Z 双尾临界	1.959 963 985	

2.3 玉米代谢通路中 2 个亚基因组基因的分化差异

为了进一步分析玉米 2 个亚基因组基因在代谢通路中的分化差异,对所有代谢通路所含基因按照 Maize1、Maize2 所属情况进行了分类汇总,并对分类结果做了 2 个 χ^2 检验,1 个是对理论值取简单平均;1 个是考虑基因组相对丢失水平的平均。表 2、3 分别列出了处理结果中有关亚基因组基因存在一丢失情况具有显著差异的代谢通路处理信息。

表 2 玉米代谢通路中 Maize1、Maize2 基因数目汇总 个

通路 ID	Maize1	Maize2	NA	总计
CITRULBIO-PWY	19	28		47
GLUTATHIONESYN-PWY	2	7	1	10
GLYCOCAT-PWY	26	6		32
PROSYN-PWY	15	21		36
PWY0-166	35	42	3	80
PWY-2841	35	40		75
PWY-2902	192	83		275
PWY-401	17	3		20
PWY-4983	2	6		8
PWY-5143	13	2	1	16
PWY-5687	18	27	2	47
PWY-5723	63	24	4	91
PWY-5941	17	3		20
PWY-5995	18	3	1	22
PWY-6029	3	9		12
PWY-6035	3	9		12
PWY-6475	2	7		9
XYLCAT-PWY	3	7		10

表 3 玉米代谢通路亚基因组基因保留-丢失差异性 χ^2 检验

通路 ID	理论平均	P 值	相对均值 1	相对均值 2	P 值
CITRULBIO-PWY	23.5	0.189	28.81	18.19	0.003
GLUTATHIONESYN-PWY	4.5	0.096	5.52	3.48	0.016
GLYCOCAT-PWY	16	0.000	19.61	12.39	0.020
PROSYN-PWY	18	0.317	22.06	13.94	0.016
PWY0-166	38.5	0.425	47.19	29.81	0.004
PWY-2841	37.5	0.564	45.97	29.03	0.009
PWY-2902	137.5	0.000	168.54	106.46	0.004
PWY-401	10	0.002	12.26	7.74	0.029
PWY-4983	4	0.157	4.90	3.10	0.035
PWY-5143	7.5	0.005	9.19	5.81	0.044
PWY-5687	22.5	0.180	27.58	17.42	0.003
PWY-5723	43.5	0.000	53.32	33.68	0.033
PWY-5941	10	0.002	12.26	7.74	0.029
PWY-5995	10.5	0.001	12.87	8.13	0.022
PWY-6029	6	0.083	7.35	4.65	0.010
PWY-6035	6	0.083	7.35	4.65	0.010
PWY-6475	4.5	0.096	5.52	3.48	0.016
XYLCAT-PWY	5	0.206	6.13	3.87	0.042

表 2、3 中第 1 列列出了 χ^2 检验中 P 值小于 0.05,即在检验中拒绝原假设,而可以认为该代谢通路中两亚基因组基因存在、丢失具有显著差异的代谢通路 ID 号,表 2 中 Maize1、Maize2 所对应的列分别是相应代谢通路中所属两亚基因组基因个数,

NA 对应的是代谢通路中未确定其亚基因组归属的代谢基因个数,最后一列是相应代谢通路中代谢基因总个数。表 3 中第 2、3 列是对代谢通路中两亚基因组基因划分的理论平均值和 χ^2 检验结果,最后 3 列是表 2 中 Maize1、Maize2 基因个数考虑基因组相

对丢失水平下的平均值以及其 χ^2 检验 P 值。

从总体上看,亚基因组基因保留具有显著性差异的代谢通路占已识别代谢通路的 4.4%,且基本上偏向于 Maize1 基因保留下来,说明代谢通路中亚基因组基因大体上保持了一致性的丢失水平,但进化过程中,2 个相同功能的代谢基因,相对来说,属于 Maize1 的代谢基因更可能保留下来,从而具有更大的表达量。

3 结论

通过构建玉米和高粱全基因组水平的共线性点图,在 Schnable 等^[8]重构的 2 个祖先基因组基础上,精确定义了玉米的 2 个亚基因组。进一步对玉米 407 个代谢通路中所有代谢基因进行了 Maize1、Maize2 基因的划分,并对它们的基因长度进行了统计上的双样本均值分析(Z -检验)和通路中 2 个亚基因组基因保留情况的 χ^2 检验。结果表明:1) Maize1、Maize2 基因长度并没有显著差异,说明在基因长度上它们具有一定偏向性但不明显,其中 Maize2 的基因相对短一些,其在全局上有更多的 DNA 序列的丢失;2) Maize1、Maize2 基因保留具有显著差异的通路占总代谢通路的 4.4%,其中 Maize1 相对保留了更多基因,说明在进化过程中, Maize1 基因为玉米的总代谢贡献了更多的表达量, Maize2 基因则更可能丢失和沉默。但是,通路水平的 2 个亚基因组基因分化研究并不能从整体上说明它们在整个代谢网络中的分化差异,因此,在今后的研究工作中还需对其进一步分析。

参考文献:

- [1] 田清震,谢传晓,李新海,等. 玉米基因组学研究进展[J]. 玉米科学,2006,14(3):1-5.
- [2] Gaut B, d'Ennequin M, Peek A, *et al.* Maize as a model for the evolution of plant nuclear genomes[J]. PNAS, 2000,97(13):7008-7015.
- [3] 黎裕,王天宇. 玉米比较基因组学研究进展[J]. 生物技术通报,2004(1):23-26.
- [4] Wang X, Tang H, Paterson A H. Seventy million years of concerted evolution of a homoeologous chromosome pair, in parallel, in major *Poaceae* lineages[J]. Plant Cell, 2011,23(1):27-37.
- [5] Paterson A H, Bowers J E, Chapman B A. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics[J]. PNAS, 2004,101(26):9903-9908.
- [6] Wang X, Shi X, Hao B, *et al.* Duplication and DNA segmental loss in the rice genome: Implication for diploidization[J]. New Phytol, 2005,165(3):937-946.
- [7] Brunet F G, Roest Crollius H, Paris M, *et al.* Gene loss and evolutionary rates following whole-genome duplication in teleost fishes[J]. Mol Biol Evol, 2006,23(9):1808-1816.
- [8] 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.
- [9] Sankoff D, Zheng C. Fractionation, rearrangement and subgenome dominance [J]. Bioinformatics, 2012, 28(18):402-408.
- [10] Woodhouse M R, Schnable J C, Pedersen B S, *et al.* Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs[J]. Plos Biology, 2010,8(6):e1000409.
- [11] 樊龙江,郭兴益. 从水稻基因组序列中挖掘生物信息[J]. 浙江大学学报,2005,31(4):355-361.
- [12] 高清松,杨泽峰,徐辰武. 水稻基因组进化的研究进展[J]. 扬州大学学报,2009,30(2):34-44.
- [13] Patrick S S, Doreen W, Robert S F. The B73 maize genome: Complexity, diversity, and dynamics[J]. Science, 2009,326:1112-1114.