内含子序列在甘蓝 MAGIC 亲本间亲缘关系分析中的应用

安光辉　严承欢　张维奕　彭丽莹　陈炯炯

华中农业大学园艺林学学院/园艺植物生物学教育部重点实验室,武汉 430070

摘要　为探究甘蓝多亲本高级世代互交系(multiparent advanced generation inter-cross,MAGIC)亲本间的亲缘关系和遗传多样性,以便对 MAGIC 杂交亲本的选择提供依据,采用生物信息学的方法对单拷贝基因进行全基因组鉴定,并采用其内含子序列分析对 7 个甘蓝亚种共 12 个 MAGIC 亲本构建系统发育树。全基因组鉴定共获得 7 930 个单拷贝基因,每条染色体上随机挑选 1 条单拷贝基因,采用 PCR 方法进行内含子扩增并测序,最终获得 3 427 bp 内含子长序列。运用邻接法构建系统发育树,可将甘蓝类蔬菜分为 3 大类型:花椰菜与西兰花类(Clade Ⅰ)、羽衣甘蓝类(Clade Ⅱ)以及结球甘蓝、抱子甘蓝、茎蓝与芥蓝类(Clade Ⅲ),表明内含子序列适于甘蓝亚种亲缘关系聚类;同时也表明,芥蓝属于甘蓝亚种,西兰花与花椰菜具有较近的亲缘关系;利用单拷贝基因内含子序列分析的方法可以较好地显示甘蓝 MAGIC 亲本间的亲缘关系,为 MAGIC 群体的构建提供参考。

关键词　甘蓝;单拷贝基因;内含子;系统进化树

中图分类号　S 635.9　　**文献标识码**　A　　**文章编号**　1000-2421(2017)01-0016-06

甘蓝(*Brassica oleracea*)为十字花科芸薹属 C 基因组物种,包括了食用叶的结球甘蓝(*B. oleracea* var. *capitata* L.)、抱子甘蓝(*B. oleracea* var. *gemmifera* Znk.)和羽衣甘蓝(*B. oleracea* var. *acephala* DC.);食用茎薹的茎蓝(*B. oleracea* var. *caulorapa* DC.)和芥蓝(*B. oleracea* var. *alboglabra* Bailey);食用花球的花椰菜(*B. oleracea* var. *botrytis* DC.)和西兰花(*B. oleracea* var. *italica* P.)等 7 大类型亚种[1]。甘蓝亚种性状丰富,如叶、茎、花的颜色,叶形,花球的发育等,是植物发育遗传研究的理想材料。

在物种性状的研究中,人工构建非自然群体(如 F2、BC1、RIL 等)是遗传作图以及性状分析的重要方法。随着群体构建方法的发展,Mackay 等[2]于 2007 年首次提出了多亲本高级世代互交系(multiparent advanced generation inter-cross, MAGIC)的概念,MAGIC 群体的构建可以提高作图的精度,为研究数量性状基因的克隆提供极大的便利。Cavanagh 等[3]认为在数量性状基因的定位、分析、作图等方面,MAGIC 群体较重组自交系、近等基因系、染色体代换系更有特色优势。由于 MAGIC 群体构建前期涉及到多个亲本,所以在考虑亲本性状差异的同时,参考亲本间的亲缘关系,使遗传标记和性状差异之间获得平衡,对 MAGIC 亲本配组的选择有重要意义。

长期以来,研究者利用形态学标记和分子标记研究物种亲缘关系,田源等[4]用 RAPD 标记对 30 份甘蓝材料进行亲缘关系和遗传多样性分析;王冬梅[5]对 183 份甘蓝材料的植物学性状进行调查分析,并使用 EST-SSR 标记研究了材料之间亲缘关系;李飞飞等[6]利用 SSR 及 ISSR 分子标记研究苜蓿属及其近缘植物的亲缘关系。虽然这些方法普遍用于分析物种亲缘关系,但存在形态标尺鉴定困难和分子标记信息量少的不足。随着分子生物学的兴起和发展,生物的亲缘和进化研究已经从宏观领域深入到分子水平,同源基因的核苷酸序列或同源蛋白质的氨基酸序列比对是研究生物进化的有效方

收稿日期:2016-07-05
基金项目:湖北省自然科学基金项目(2013CFB204)
安光辉,硕士研究生. 研究方向:植物功能基因组学. E-mail:anguanghui369@163.com
通信作者:严承欢,博士研究生. 研究方向:植物功能基因组学. E-mail:yan1014936712@163.com

法[7-8]。近年来，在研究物种亲缘关系的案例中，常运用叶绿体、线粒体和基因组核酸序列进行亲缘关系分析，因其能提供比 SSR 等分子标记更多的多态信息，更适于近缘物种的研究。例如，陈春梅等[9]利用茶树 cpDNA 序列对山茶属植物亲缘关系进行分析。对于同一物种不同亚种的亲缘关系分析，由于其叶绿体、线粒体以及基因的外显子差异较小，不能提供足够的多态性，因此，需采用几乎不受选择压力的内含子序列进行种内亚种的亲缘关系鉴定[10]。

本研究在甘蓝全基因组范围鉴定单拷贝基因，并从每条染色体中随机选出 1 条单拷贝基因，对 12 份具有代表性的 MAGIC 亲本材料中该基因的内含子进行 PCR 扩增并测序。根据序列分析，用邻接法构建 NJ 进化树并分析其亲缘关系；根据 MAGIC 亲本材料的亲缘关系结合其表型多样性，选择合适的 MAGIC 杂交配组，以期为构建表型更加丰富的甘蓝 MAGIC 群体提供指导。

1　材料与方法

1.1　材　料

本试验采用 12 份甘蓝 MAGIC 亲本，羽衣甘蓝 3 份：绿色羽衣甘蓝（ace 101）、白鸥羽衣甘蓝（ace 102）和红鸥羽衣甘蓝（ace 103）；芥蓝 2 份：矮脚香芥蓝（alb 101）和红色芥蓝（alb 103）；花椰菜 2 份：金色花椰菜（bot 101）和青梗松花菜（bot 103）；结球甘蓝 2 份：紫色结球甘蓝（cap 101）和绿色结球甘蓝（cap 102）；抱子甘蓝 1 份：抱子甘蓝（gem 101）；西兰花 1 份：西兰花（ita 101）；茎蓝 1 份：绿色茎蓝（cau 102）。所有材料均于 2014 年 10 月 27 日播种于中国蔬菜改良中心华中分中心，2015 年 3 月 7 日采样进行分子生物学实验。

1.2　单拷贝基因的调取与内含子扩增

运用笔者所在实验室开发的程序进行 All-against-All BLASTP 分析甘蓝蛋白质序列（http://www.ocri-genomics.org/bolbase/），即以测序甘蓝基因组蛋白质数据为 query 序列，采用 BLASTP 调取该测序甘蓝预测蛋白质数据，E 值设为 1e－10。当某一基因只匹配到其本身则认为是单拷贝基因，反之则否。结合甘蓝 GFF3 文件，调取甘蓝全基因组的单拷贝基因。在甘蓝每条染色体上随机挑选 1 个符合上述标准的单拷贝基因，跨过该基因的内含子在外显子上设计 PCR 引物（表 1），对 12 份甘蓝 MAGIC 亲本进行 PCR 扩增，并对 PCR 产物进行测序。相关性分析采用统计软件 SPSS（Version 22）进行分析。

表 1　单拷贝基因内含子引物

Table 1　Specific primers for intron of single copy genes

引物名称 Primer name	正向引物 Primer sequence(forward)	反向引物 Primer sequence(reverse)	PCR 产物长度/bp PCR product length
C1-11.3	TCCATGATGCTGTCTCTGCT	TCAGACCCAAATGAAGGAGCT	613
C2-19.9	ACAGTTGAGTCGGAAATTCGT	TCTTGTGGGAGTTGCTTTTAGA	704
C3-1.8	TGCAGCGAGATGGAAGAAGA	AGTCCCCTCGATGTCACTTG	686
C4-40.2	CCCTGACGGAGATGCAAATC	CATGCGGGGTATGAGAGGAT	575
C5-13.3	ACCGTCCAATTTCACTGACA	AAAGCATCCCTCCCACCATT	646
C6-6.1	GGAACGCTTCGCTGTTACG	AGCTGTGGTATCTTGACGCA	602
C7-16.1	CTGAGCAGAGCCATCGAAAA	CAACTGCTTTTGGGAACGGA	638
C8-27.8	CTGAGGAATCGAACCGGAGA	ATACGCAGAGATCATCCCCG	668
C9-34.9	TCCACTTCGTTACTTGTGTGC	CCTCCCCATGAAACATCCGA	551

1.3　序列分析及系统进化树构建

通过软件 Geneious 4.85[11] 对单拷贝基因内含子序列进行比对，利用软件 MEGA 5.1[12] 构建材料之间的 NJ 系统树[13]，采用置信度（Bootstrap）1 000 检验。

2　结果与分析

2.1　甘蓝基因组单拷贝基因

采用 All-against-All BLASTP 分析甘蓝蛋白质序列，E 值设为 1e－10，当有且只有最佳匹配为

自己本身时认为该序列是单拷贝基因。在甘蓝全基因组数据中,总计调取7 930条单拷贝基因,每条染色体上的单拷贝基因数变异幅度为614～1 152,单拷贝基因密度为13.94～23.39条/Mb。相关性分析可得,单拷贝基因数目与染色体长度正相关,相关系数r为0.678(图1),即单拷贝基因在染色体组上

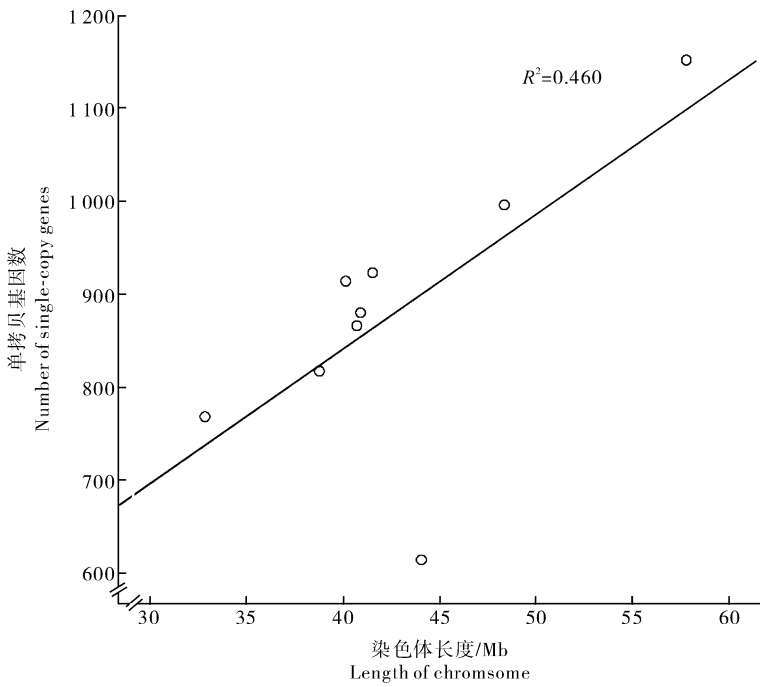

图1 单拷贝基因数目与染色体长度正相关

Fig.1 Strong correlation of the number of single-copy genes and length of chromosomes

趋于均匀分布。

2.2 内含子序列多态性

在甘蓝基因组中选取9个单拷贝基因,在基因内含子两侧设计引物扩增内含子序列并测序(序列两端测序效果较差的区域不采用),每个基因的内含子获得173～527 bp的序列,共计3 427 bp(表2,序列信息未显示)。在3 427个碱基中共有155个多态性位点,平均每22个碱基有1个差异位点。选择单拷贝基因的内含子序列进行分析,由于内含子区域不存在选择压力,可以认为是中性选择,导致了不同单拷贝基因内含子SNP位点的变异幅度的差异[14]。虽然,每个内含子所含SNP位点数目的变

表2 各内含子序列的扩增

Table 2 Summary of amplified introns sequences

染色体号 Chromosome No.	基因号 Gene No.	内含子号 Intron No.	可靠序列长度/bp Clean sequence length	多态性位点数 The number of polymorphic sites
1	Bol039501	Intron 1	173	10
2	Bol019012	Intron 1	222	3
3	Bol008767	Intron 1	472	8
4	Bol021703	Intron 1	370	7
5	Bol020912	Intron 1	493	3
6	Bol032871	Intron 1	352	30
7	Bol041764	Intron 1	451	54
8	Bol025058	Intron 1	527	11
9	Bol043349	Intron 1	367	29
总计 Total			3 427	155

异幅度并不与内含子长度呈现相关性(表 2)，但从分子进化的角度这种中性选择更有利于反映真实的甘蓝亚种间的亲缘关系。

2.3　系统发育树构建

基于 3 427 bp 内含子序列通过软件 MEGA5.1 构建 12 个甘蓝 MAGIC 亲本的 NJ 系统发育树，以白菜(*B. rapa*)的同源序列为外类群(outgroup)，如图 2。可将甘蓝类蔬菜清晰分为 3 大类型：花椰菜与西兰花类(Clade Ⅰ)、羽衣甘蓝类(Clade Ⅱ)以及结球甘蓝、抱子甘蓝、苤蓝与芥蓝类(Clade Ⅲ)。Clade Ⅰ由金色花椰菜(bot 101)、青梗松花菜(bot 103)和西兰花(ita 101)构成(Bootstrap 值 97)，被称为花椰菜与西兰花类，在该类型中金色花椰菜与西兰花亲缘关系较青梗松花菜更近，表明花椰菜与西兰花有较近的亲缘关系，且两者间关系复杂，可能出现亚种间亲缘关系较亚种内更近的现象。Clade Ⅱ由羽衣甘蓝类构成(Bootstrap 值 91)，包括绿色羽衣(ace 101)、红鸥羽衣(ace 102)以及白鸥羽衣(ace 103)。绿色羽衣甘蓝较鸥系列羽衣甘蓝支长较长，可见其亲缘关系较红鸥羽衣和白鸥羽衣更远。Clade Ⅲ由剩余其他材料构成 (Bootstrap 值 71)，包含了紫色结球甘蓝(cap 101)和绿色结球甘蓝(cap 102)、抱子甘蓝(gem 101)、绿色苤蓝(cau 102)、矮脚香芥蓝(alb 101)和红色芥蓝(alb 103)，说明这 4 个亚种类型材料亲缘关系较近，且遗传背景与亲缘关系复杂，在育种过程可能经历过亚种间的基因转移。在 Clade Ⅲ中的拓扑结构表明：芥蓝属于 Clade Ⅲ，相对于 Clade Ⅰ与 Clade Ⅱ，其与结球甘蓝、抱子甘蓝以及苤蓝亲缘关系更近，因此芥蓝属于甘蓝的一个亚种。如果单独对各染色体上的内含子进行序列分析，结球甘蓝和抱子甘蓝、苤蓝在基因 Bol039501 的 Intron1 序列完全相同，而结球甘蓝在 Bol008767 的 Intron 1 上的序列和芥蓝一致，但与抱子甘蓝、苤蓝存在 4 个 SNP。可见，不同甘蓝亚种间遗传关系复杂多变，可能是由于在育种过程中一个亚种导入了其他亚种的基因组序列所致，这

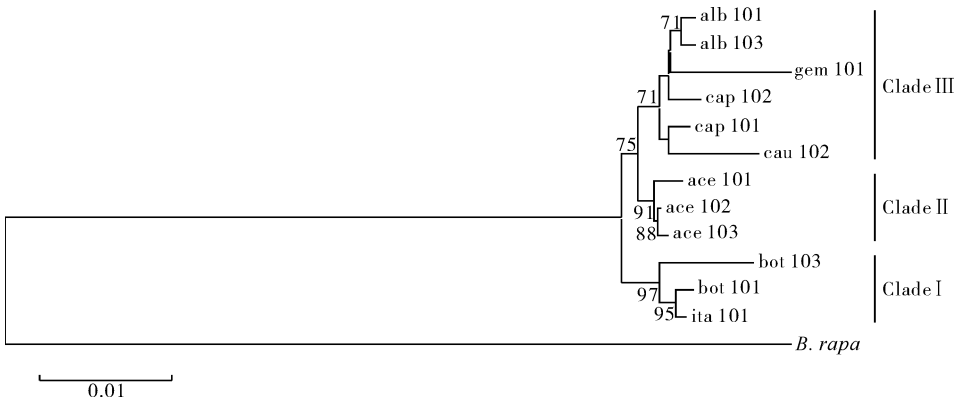

图 2　基于内含子序列的(NJ)系统发育树

Fig.2　The neighbor-joining phylogenetic tree based on intron sequence

也能部分地解释这 4 个甘蓝亚种亲缘关系较近的原因。

3　讨　论

本试验通过生物信息学的方法在全基因组鉴定了甘蓝的单拷贝基因，在每条染色体上随机挑选 1 条单拷贝基因，采用 PCR 扩增并测序获得 12 个 MAGIC 材料的内含子序列，将 9 个不同染色体上的内含子序列合并为 1 条长为 3 427 bp 的序列进行 NJ 系统发育树的构建，对甘蓝类物种的 7 个亚种的亲缘关系进行了分析，并证明了芥蓝与甘蓝亚种、花椰菜与西兰花在进化中具有较近的亲缘关系。

前人对甘蓝变种间亲缘关系的研究大多局限于植物学性状和 SSR 等传统标记，而且往往因为标记的不同使亲缘分析结果存在一定的差异。这些差异源于分析方法的本质区别，传统的方法基于表型以及传统分子标记的聚类分析是通过表型或标记的相似性来构建亲缘关系聚类图。植物形态学标记只反映材料之间的生长状态和表型性状的差异，这些差异并不稳定，易受到环境因素的影响，如温度、肥力、光照等，而造成性状鉴定误差。随着对植物发育、基因调控等方向研究的深入，人们认识到一个基因的

突变可能造成植物形态的巨大改变，同时影响植物多个表型性状。因此，表型性状（尤其是经过人工选择的性状）不能真实反映不同物种的亲缘关系。同样地，传统分子标记，以SSR分子标记为例，只显示简单重复序列长度的差异，但无法检测碱基差异。如果标记数量有限，SSR则不一定能真实地反映材料之间的亲缘关系，尤其对于多态性低的物种内材料。可见，在对于物种内材料亲缘关系的研究中，植物形态学以及传统分子标记均存在一些不足。

基因内含子序列能够同时表现同源序列长度和碱基差异，且内含子作为非编码区，承受选择压小，能提供5倍于外显子的变异[15]，从而反映出更多的多态性，更好地揭示多态性不丰富的亚种间亲缘和进化，得出较为准确的亲缘关系。本研究采用生物信息学进行单拷贝基因的鉴定，通过相关性分析发现单拷贝基因在基因组上密度为13.94～23.39条/Mb，在染色体的尺度上可以看作随机分布，且单拷贝基因与染色体长度呈正相关。对单拷贝基因数目和分布的分析有助于了解内含子序列的分布和代表性，可以更有目的性地选择和扩增内含子序列。因此，内含子序列分析在鉴定甘蓝亚种间的亲缘关系中具有优势，有目的地选取较少的序列便可代表亚种类型的整体，将不同亚种区分开。在本研究系统发育树的构建中，Clade Ⅰ中包含花椰菜与西兰花2种类型甘蓝，表明花椰菜与西兰花有很近的亲缘关系，这与孙德岭等[16]使用AFLP标记得出的结果一致。而Clade Ⅲ包含紫色结球甘蓝（cap 101）和绿色结球甘蓝（cap 102）、抱子甘蓝（gem 101）、绿色苤蓝（cau 102）、矮脚香芥蓝（alb 101）和红色芥蓝（alb 103），这4个类型材料遗传关系复杂，但可以判断芥蓝属于甘蓝的一个亚种，且与结球甘蓝、抱子甘蓝、苤蓝的亲缘关系很近，这一结论与王冬梅[5]和周禹等[17]采用分子标记得到的结果一致。结球甘蓝、抱子甘蓝、苤蓝和芥蓝因亲缘关系较近被分为一组，出现这种情况的原因可能是甘蓝各亚种之间没有生殖隔离，在育种过程中，不同亚种材料间有基因交流（gene flow），最终导致不同亚种材料之间含有相同的基因序列。

甘蓝为古六倍体，经过三倍体化事件形成现代二倍体物种，其基因组复杂多变，遗传研究困难。随着甘蓝育种的快速发展，出现了越来越多的新品种，像“西兰苔”（芥蓝与西兰花的杂交种）、“紫衣骑士”（紫色抱子甘蓝）这样的新型甘蓝蔬菜品种层出不穷，甘蓝类蔬菜的遗传研究工作面临新的挑战。要解读同甘蓝一样拥有复杂基因组的亚种间亲缘关系，最理想的方法是进行全基因组重测序，构建“泛基因组”数据库，分析各亚种染色体不同区域的相关性，进而详细解析亚种间基因组的关系。

参 考 文 献

[1] 刘英，王超.简述甘蓝类植物的起源及分类[J].北方园艺，2006(4)：58-60.

[2] MACKAY I，POWELL W.Methords for linkage disequilibrium mapping in crops[J].Trends in plant science，2007，12(2)：57-63.

[3] CAVANAGH C，MORELL M，MACKAY I，et al.From mutations to MAGIC：resource for gene discovery，validation and delivery in crop plants[J].Current opinion in plant biology，2007，11：1-7.

[4] 田源，王超.甘蓝类蔬菜亲缘关系的RAPD初步分析[J].中国蔬菜，2008(1)：20-22.

[5] 王冬梅.甘蓝类作物亲缘关系的SSR分析[D].北京：中国农业科学院，2011.

[6] 李飞飞，羊海军，崔大方.利用SSR及ISSR分子标记研究苜蓿属及其近缘植物的亲缘关系[J].中山大学学报（自然科学版），2014，53(1)：113-120.

[7] YANG Z，YODER A D.Estimation of the transition/transversion rate bias and species sampling [J].Journal molecular evolution，1999，48：274-283.

[8] YANG Z，O’ BRIEN J D，ZHENG X，et al.Tree and rate estimation by local evaluation of heterochronous nucleotide data [J].Bioinformatics，2007，23：169-176.

[9] 陈春梅，马春雷，马建强，等.茶树cpDNA测序及基于cpDNA序列的山茶属植物亲缘关系研究[J].茶叶科学，2014，34(4)：371-380.

[10] 王宁，陈润生.基于内含子和外显子的系统发育分析的比较[J].科学通报，1999，44(19)：2095-2101.

[11] DRUMMOND A，ASHTON B，BUXTON S，et al.Geneious v4.8.5[CP/OL].[2016-01-06].http://www.geneious.com/2010.

[12] TAMURA K，PETERSON D，PETERSON N，et al.MEGA 5：molecular evolutionary genetics analysis using maximum likelihood，evolutionary distance，and maximum parsimony methods [J].Mol Biol Evol，2011，28(10)：2731-2739.

[13] SAITOU N，NEI M.The neighbor-joining method：a new method for reconstructing phylogenetic trees [J].Mol Biol Evol，1987，4：406-425.

[14] HOFFMAN M M，BIRNEY E.Estimating the neutral rate of

nucleotide substitution using introns[J].Mol Biol Evol,2007,24(2):522-531.
[15] 杨永强,王巍杰,徐长波.单核苷酸多态性研究进展[J].化学与生物工程,2009,26(8):19-21.
[16] 孙德岭,赵前程,宋文芹,等.花椰菜类蔬菜自交系基因组间亲缘关系的 AFLP 分析[J].园艺学报,2002,29 (1):72-74.
[17] 周禹,李燕,孙勃,等.芥蓝与甘蓝其他变种分类关系的研究[J].园艺学报,2010,37(7):1161-1168.

Application of intron sequence in phylogenetic relationship of MAGIC parents of *Brassica oleracea*

AN Guanghui　YAN Chenghuan　ZHANG Weiyi　PENG Liying　CHEN Jiongjiong

College of Horticulture and Forestry Sciences/ Key Laboratory of Horticultural Plant Biology, Huazhong Agricultural University, Ministry of Education, Wuhan 430070, *China*

Abstract　In order to study genetic diversity and relationships of MAGIC parents of *Brassica oleracea*, and provide the basis of the hybridization experiments of different sub-species. The single-copy genes were identified with bioinformatics and some intron sequences of them were used to construct the phylogenetic tree of 12 parents from 7 subspecies. We identified 7 930 single-copy genes form whole *Brassica oleracea* genome and randomly chosen from them of different chromosomes to amplify and sequence their intron sequences, resulting in a total of 3 427 bp intron sequences. Phylogenetic analysis revealed that intron sequence method was suitable for relationship analysis in the different sub-species of *Brassica oleracea* L., which were obviously divided into three Clades: the Clade Ⅰ of cauliflowers and broccoli, the Clade Ⅱ of kales and the Clade Ⅲ of head cabbage, brussels sprouts, kohlrabi and cabbage mustard. What's more, cauliflower and broccoli are closely related and Chinese kale (*B. oleracea* var. *alboglabra* Bailey) belongs to *B. oleracea*. The intron sequences of single-copy genes will provide robust results on relationships of MAGIC parents and theoretical reference for hybridization of MAGIC parents in future.

Keywords　*Brassica oleracea*; single-copy genes; intron; phylogeny

(责任编辑:张志钰)