

水稻全基因组 R 基因鉴定及候选 RGA 标记开发

汪旭升 吴为人* 金谷雷 朱 军

(浙江大学农业与生物技术学院生物信息学研究所, 杭州 310029; 福建农林大学作物科学学院, 福州 350002.

* 联系人, E-mail: wuwr@zju.edu.cn)

摘要 用 45 个已知功能的植物抗病(R)基因序列对粳稻全基因组序列进行搜索, 共找出 2119 个 R 基因同源序列或类似物(RGA), 表明 RGA 在水稻基因组中成簇存在, 呈非随机分布. 采用隐马尔柯夫模型(HMM), 将这些 RGA 按其功能域分成了 21 类. 将粳稻的 RGA 与籼稻的基因组序列进行比较, 共找到 702 个两亚种间等位的 RGA, 并发现其中有 671 个(占 95.6%)RGA 的基因组序列(包括编码区和非编码区)在两亚种间存在长度差异(InDel), 表明水稻 RGA 在两亚种间存在很高的多态性. 通过在 InDel 两侧设计引物并进行 e-PCR 验证, 共开发出 402 个基于 PCR 的、表现为共显性的候选 RGA 标记. 这些候选标记在两亚种间的长度差异在 1~742 bp 之间, 平均为 10.26 bp. 有关数据均可从我们的网站(<http://ibi.zju.edu.cn/RGAs/index.html>)上获得.

关键词 水稻 抗病基因 RGA 多态性 分子标记

植物抗病(R)基因是决定寄主植物对病原菌专化性识别并激发抗病反应的基因, 与病原菌的无毒基因互补. 经典遗传学认为, 植物与病原菌间相互作用的遗传机制是“基因对基因”, 并提出了配体-受体模型来解释这一学说^[1,2]. 自 1992 年以来, 利用图位克隆和转座子标签法, 已经在水稻、拟南芥、玉米、烟草、亚麻等植物中克隆了 40 多个 R 基因^[3]. 研究发现, 这些 R 基因存在一些共同的结构. 根据其蛋白结构及在细胞中的位置, R 基因大致可分为 5 类^[4,5]: () NBS-LRR, 是含有核苷酸结合位点(NBS)和富亮氨酸重复(LRR)的胞内受体蛋白基因, 包括 2 个亚类: (1) TIR-NBS-LRR, 以拟南芥的 *RPP5* 基因、烟草的 *N* 基因及亚麻的 *L6* 和 *M* 基因为代表; (2) CC-NBS-LRR, 以拟南芥的 *RPS2* 和 *RPM1* 基因、番茄的 *I2* 基因及大麦的 *Mla1* 基因为代表. () 细胞间的苏氨酸/丝氨酸蛋白激酶(PK)基因, 包括番茄的 *Pto* 基因和大麦的 *Rpg1* 基因. () LRR-TM, 是 N 端存在一个胞外 LRR, C 端具有由疏水氨基酸组成的跨膜区的受体蛋白基因, 包括番茄抗叶霉病的基因 *Cf-2*, *Cf-4*, *Cf-5* 和 *Cf-9* 等. () PK-LRR-TM, 除含有 LRR-TM 结构外, 还具有 PK 结构, 包括水稻的 *Xa-21* 基因和拟南芥的 *FLS2* 基因. () SA-CC, 包括拟南芥的 *RPW8.2* 和 *RPW8.1* 基因. 此外, 还有玉米的 *Hm1* 及 *Hs1^{Pro-1}* 和 *Asc* 等其他结构域的 R 基因.

R 基因是一个庞大的基因家族. 尽管已经克隆了 40 多个 R 基因, 但对 R 基因的了解还非常有限. 目前, 对拟南芥和水稻的基因组测序工作皆已基本完成.

水稻籼、粳 2 个亚种的基因组草图已经完成^[6,7], 而且粳稻的 1 号、4 号和 10 号 3 条染色体已经发布了精细图^[8~10]. 这些成果为在整个基因组水平上研究 R 基因提供了契机. 利用拟南芥基因组的测序结果, Meyers 等人^[11]分析了 NBS-LRR 型 R 基因在拟南芥基因组中的分布. 对水稻基因组序列的初步分析显示, 水稻基因组中存在大量的 R 基因^[6~7], 且往往成簇存在^[12~14]. Monosi 等人^[15]发现在水稻中存在近 500 个 NBS-LRR 的基因, 但没有发现 TIR 的基因. Chelkowski 等人^[16]和 Koczyk 等人^[17]利用 18 个已知的 R 基因分别对拟南芥和粳稻基因组序列进行分析, 发现拟南芥和水稻分别存在 549 和 1744 个 R 基因, 其中水稻的 R 基因中有 597 个属于 NBS-LRR 类型. 可以看出, 目前这些基于基因组序列的研究主要都集中在对 NBS-LRR 型 R 基因的分析上.

分子标记是现代遗传学研究的有力工具. 自 1980 年首次提出分子标记的概念以来^[18], 分子标记已广泛应用于遗传图谱构建、基因定位、基因克隆、基因组比较、遗传多样性分析、标记辅助育种等领域的研究^[19~21]. 利用 R 基因类似物(RGA)作探针, 可以检测相应座位上的限制性片段长度多态性(RFLP), 从而开发成为 RFLP 标记, 常称为 RGA 标记. 由于 RGA 本身可能就是潜在的 R 基因, 而且 R 基因常常成簇分布于基因组中, 因此 RGA 标记对于 R 基因的克隆和标记辅助选择可能具有特别的应用价值. 但是, 由于 RGA 标记是基于 RFLP 分析技术的, 操作上比较麻烦, 因此目前应用得并不多.

本研究利用已公布的籼稻和粳稻的基因组序列,采用生物信息学的方法,通过收集目前所有已知的R基因序列,对水稻基因组中RGA的数目、分布和类型进行了更为详尽的分析,以期在全基因组水平上加深对水稻R基因的了解.同时,对RGA在水稻2个亚种间进行了遗传多态性(SNP和InDel)比较、引物设计和e-PCR验证,为开发方便实用的基于PCR技术的新型RGA标记奠定基础.

1 材料与方法

() 基因组及蛋白质序列的来源. 从TIGR网站(<http://www.tigr.org/>)和北京基因组学研究所网站(<http://rise.genomic.org.cn/>)分别下载粳稻(Nipponbare)和籼稻(93-11)的基因组及蛋白质序列,它们的更新时间皆为2004年4月.所有数据的处理和分析皆在IBM P650的服务器上完成,使用IBM AIX的Unix操作系统.

() 水稻RGA的搜索. 搜集已报道的45个R基因,然后用它们对粳稻蛋白质数据库进行BLASTP^[22]搜索(参数 $E < +10^{-10}$,最小长度为该基因的80%),获得所有粳稻的RGA序列.去除粳稻数据库中存在的克隆重复,建立一个粳稻RGA蛋白数据库.同时,根据每个BLASTP搜索中匹配最好的结果,得到这些粳稻RGA的核苷酸序列.为了进一步验证得到的RGA,我们进行候选RGA序列与TIGR发布的CDS序列进行比较,去除不符的序列.

() 水稻RGA的结构分类及其在染色体上的分布. 利用Hmmer程序^[23]中的hmmsearch部分,采用前面建立的功能域列表,在新建立的数据库中搜索基因序列所包含的功能域.利用pepcoil程序^[24]分析序列中CC结构的可能性,数值大于90%的认为具备该结构.运用TM-HMMer^[25]分析TM跨膜功能域.依据功能域分布的情况,对RGA进行分类,分别统计各类RGA在不同染色体上的分布情况.

() 2个亚种间RGA多态性的鉴定和开发. 将所有粳稻的RGA序列分别与籼稻基因组数据库进行TBLASTN^[22]联配,以确定籼稻中对应的等位基因.为消除非等位联配,在TBLASTN搜索中采用了严格的判别标准,将E值设为 1×10^{-20} .对初筛到的籼、粳稻RGA等位基因,进一步用sim4程序^[26]进行联配分析,去除匹配率 $\leq 85\%$ 且2条序列同时间断200 bp以

上的结果.接着运用diffseq程序^[27]分析SNP和InDel在RGA的基因组序列中的分布情况,将存在InDel的作为候选的RGA标记.以粳稻的基因组序列为模板,在InDel位置的两侧各取100 bp的序列,连接成一条200 bp长的模板序列,然后利用ePrimer3程序¹⁾在模板序列上设计引物.ePrimer3程序一般给出5对候选引物,我们选取其中设计最合适的一对,并要求正、反向引物分别位于InDel的左、右侧,且扩增出的目标片段长度不大于1000 bp.最后,通过电子PCR(e-PCR)^[28]进行验证.对得到的水稻RGA标记进行命名,规则为以OSR开头,后跟4个数字,例如:OSR0255.上述步骤主要通过编写perl脚本程序来实现.

2 结果与分析

2.1 粳稻中RGA的数目、密度及其在染色体上的分布

通过对粳稻蛋白质数据库的搜索,共获得2119个RGA(表1).它们在各染色体上的数量变化在113(3号染色体)~333个(1号染色体)之间,平均为176个,以1,2,11号染色体最多.单条染色体上RGA的平均密度变化在0.66~2.42或2.68~9.44个/Mb之间.无论是遗传图密度还是物理图密度,都是以11号染色体最多,3号染色体最少.根据TIGR发布的拼接好的水稻基因组序列,分析RGA在染色体上的分布情况,发现大部分RGA都以成簇形式存在(多数情况下每簇包含2~12个RGA),如在1号染色体的AP003209克隆上发现有10个RGA.

表1 粳稻中RGA在各染色体上的数量和密度

染色体	染色体长度		RGA 数目	RGA 平均密度	
	/cM	/Mb		/cM ⁻¹	/Mb ⁻¹
1	181.8	44.3	333	1.83	7.51
2	157.9	39.9	217	1.37	5.44
3	166.4	41.1	110	0.66	2.68
4	129.6	38.2	195	1.50	5.10
5	122.3	33.2	134	1.09	4.04
6	126.3	31.7	190	1.50	5.99
7	118.6	35.0	125	1.05	3.57
8	121.2	27.6	158	1.30	5.72
9	93.5	21.6	133	1.42	6.16
10	83.8	25.7	113	1.35	4.40
11	117.9	30.2	285	2.42	9.44
12	109.5	30.6	126	1.15	4.12
全基因组	1528.8	399.1	2119	1.39	5.31

2.2 水稻RGA的结构分类

通常将R基因分为5大类,其中NBS-LRR是最

1) <http://www.hgmp.mrc.ac.uk/>

多的一类^[4,5]. 我们根据R基因的结构与功能域, 将水稻RGA进行了更细致的分类, 共分为 21 类(图 1). 其中PK类RGA(*Pto*, *Fen*, *Lr10*)数目最多, 占 26.7% (566/2119). 第 2 大类是TM-LRR, 占总数的 20.5% (435/2119). 需要指出的是, 在本研究中, 具有NBS或LRR功能域的RGA被分成了 9 类, 即TM-LRR, PK-LRR, NBS-LRR, CC-NBS-LRR, CC-LRR, CC-NBS, PK-NBS-LRR, PK-NBS和CC-PK-LRR, 因此每一类都不是最多的, 但若将它们皆计为NBS-LRR类型, 则其数量占水稻RGA的半数以上(1091/2119). 第 1 个被克隆的玉米抗圆斑病基因*Hm1* 所代表的毒素还原酶类RGA共发现了 77 个. 这类基因还与CC结构相结合成为CC-Hm1 类, 共发现 3 个该类型的成员. PK-NBS, CC-PK-LRR, TIR, Hs1 和Pad4 这几类RGA在水稻中皆只存在一个成员, 对这些基因进行结构分析后显示, 其中大部分是假基因或没有功能的基因. Pan等人^[29]研究认为, 在双子叶和单子叶植物的分化过程中, NBS-LRR分化成TIR-NBS-LRR和CC-NBS-LRR共 2 大类. 本研究显示, 水稻中不存在TIR-NBS-LRR类的RGA, 这与甜菜相似^[30], 但在拟南芥中已发现 117 个这类基因^[19]. 本研究发现的水稻RGA的有关数据可以从我们的网站(<http://ibi.zju.edu.cn/RGAs/index.html>)获得.

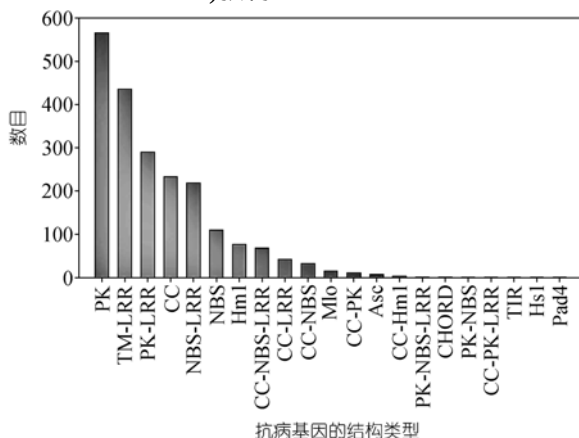


图 1 水稻中 RGA 的类型及其数量分布

PK, 苏氨酸-丝氨酸蛋白激酶; TM, 跨膜蛋白; LRR, 富亮氨酸重复; CC, 卷曲螺旋结构; NBS, 核苷酸结合位点; Hm1, 玉米 *Hm1* 基因; CHORD, 富半胱氨酸-组氨酸结构域; TIR, 白细胞介素-1 受体; Mlo, Asc, Hs1, Pad4 分别代表各自基因的特有结构域

2.3 RGA 在水稻亚种间的多态性

通过用粳稻 RGA 序列对籼稻基因组序列进行 TBLASTN 联配, 在籼稻上找到 1860 个 R 基因的同源序列. 经过人工分析后去除重复的或匹配不好(匹

配序列长度 < 80 bp, 一致性 < 40%)及与 TIGR 数据库中 CDS 序列不符的同源序列, 得到 861 个同源序列. 进一步去除位于不同染色体的非等位基因, 最终获得了 702 个在籼、粳间等位的 RGA. 用 sim4 程序对这 702 个 RGA 序列进行分析, 发现有 671 个(占 95.6%)在籼、粳间存在 InDel 的现象, 说明在籼粳亚种间 RGA 存在很高的长度多态性. 用 ePrimer3 程序在 InDel 两侧设计 PCR 引物, 并进行 e-PCR 验证, 选出能够获得惟一预期扩增产物的引物对, 最终得到 402 个候选的水稻亚种间 RGA 标记. 有 269 个多态的 RGA 未能开发成候选标记, 其原因可能是: () 一些 RGA 间的结构相似性, 使得引物的特异性不强, 不能得到惟一的扩增产物; () 我们将 e-PCR 产物的长度限制在 1000 bp 以内, 有些 RGA 的扩增产物可能过大而不能入选; () 引物是依据粳稻 Nipponbare 的基因组序列设计的, 有些在籼稻 93-11 上未能完全匹配. 这些候选 RGA 标记的有关信息(包括标记的引物、序列、所在粳稻 Nipponbare 的 BAC 克隆和籼稻 93-11 的 Scaffold 等)都在我们的网站(<http://ibi.zju.edu.cn/RGAs/index.html>)上发布. 根据 TIGR 发布的拼接好的水稻基因组序列, 对候选 RGA 标记进行了物理定位. 结果显示, 跟全体 RGA 的情况一致, 候选 RGA 标记在基因组中的分布也是非随机的, 表现为“成簇”分布的现象(图 2). 有些染色体区域(如 1 号染色体的长臂)出现大片的空缺.

候选 RGA 标记的等位基因间长度差异(LD)变化在 1~742 bp 之间, 平均长度为 10.15 bp, 呈指数分布, 大部分(68.16%) < 5 bp; 24.88%落在 5~30 bp 之间; 仅 6.96% > 30 bp(图 3). 值得指出的是, 我们发现 14 个 RGA 在 2 个亚种间的长度差异超过 1 kb, 其插入序列都具有独立完整的基因结构. 同源性分析显示, 这 14 个插入序列的基因功能与受体蛋白密切相关. 由于 R 基因本身就是一类受体蛋白, 因此这种在 R 基因中插入与受体蛋白相关的基因的现象是否隐含着某种重要的生物学机制, 是一个令人感兴趣的问题.

3 讨论

本研究通过序列同源性比较结合功能域位点分析, 共发现了 2119 个 RGA, 说明水稻基因组中 R 基因的数量是十分丰富的, 是一个非常庞大的基因家族. 当然, 在这些 RGA 中, 除了包含 R 基因外, 还可能包含没有功能的基因或假基因. 为了鉴定其中哪些是真正的 R 基因, 我们把所有的 RGA 同已发布的

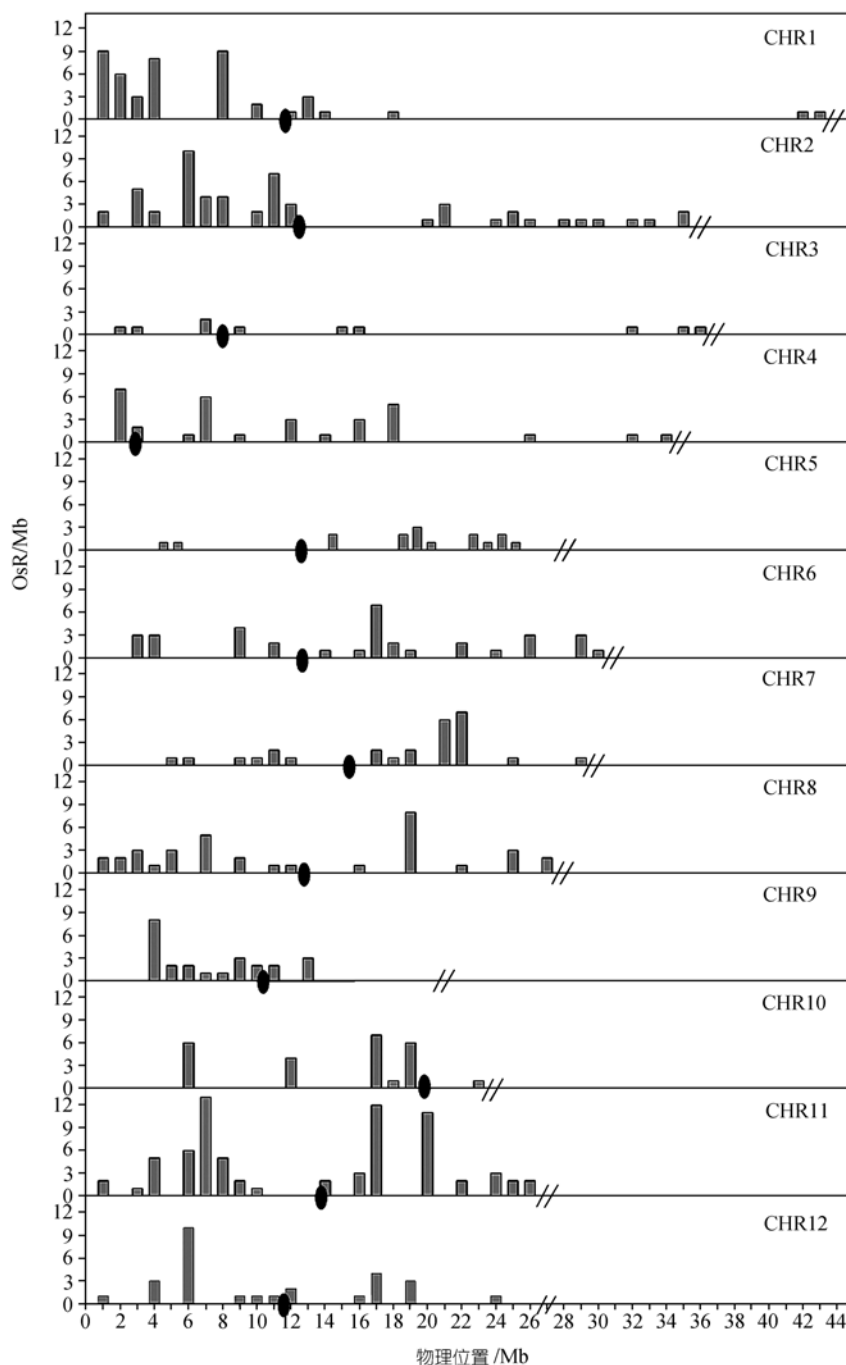


图2 候选 RGA 标记在水稻基因组上的分布

横坐标是物理图位置, 纵坐标是每 Mb 所含 RGA 的个数. 两斜杠表示染色体终止的位置, ● 表示着丝粒的位置

32127 个水稻全长cDNA^[31]进行BLAST分析, 结果有 1851 个RGA能够很好地与cDNA匹配, 因而可以认为它们可能是真正的R基因(当然不排除其中有些可能是可表达的假基因). 剩余的 268 个RGA可能存在 3 种情况: () 是cDNA数据库中未包括的基因, 因为水稻中报道有约 5 万个基因; () 是不表达的假基因;

() 是特定病原菌诱导表达的基因. 随着水稻全长cDNA数据库的不断充实和完善, 这部分RGA的身份将得到进一步的鉴别. 将来的挑战是对水稻中所有R基因的功能阐明. 利用生物信息学的方法在全基因组范围内获取RGA的有关信息, 将大大促进对R基因的功能研究.

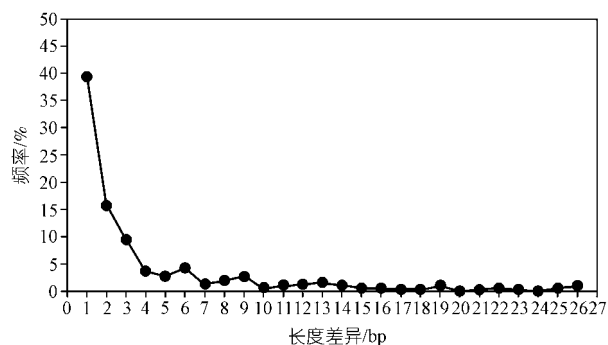


图 3 候选 RGA 标记在 2 个亚种间的长度差异(LD)的频率分布

其中 LD > 26 的 30 个标记未标出

传统的 RGA 标记是一种 RFLP 标记, 应用上不方便, 而且由于开发上成本较高, 所以数量上十分有限. 本研究利用水稻籼、粳亚种的基因组测序数据和生物信息学手段, 开发出了基于 PCR 技术的候选 RGA 标记, 这将使 RGA 长度多态性成为一种实用的分子标记. 我们用相似的方法已成功开发出水稻内含子长度多态性(ILP)标记(结果未显示). 经实验分析, 发现水稻 ILP 标记具有明显的亚种特异性. 据此推测, 本研究基于籼、粳亚种间序列比较而开发的候选 RGA 标记也将具有较高的亚种特异性. 该特性可使 RGA 标记在水稻亚种间杂交育种和亚种间杂种优势利用方面具有重要的应用价值. 另外, 已知 RGA 在水稻基因组中呈簇状非随机分布, 而本研究开发出来的候选 RGA 标记在基因组上的分布对全体 RGA 的分布具有很好的代表性(图 3). 而且, RGA 标记本身就是候选的 R 基因. 因此, 利用 RGA 标记将有助于快速定位 R 基因, 加快 R 基因定位和图位克隆的进程.

致谢 本工作为国家高技术研究发展计划(批准号: 2003AA207160, 2002AA234031)和福建省自然科学基金(批准号: B9910011)资助项目.

参 考 文 献

- Flor H H. The complementary genic systems in flax and flax rust. *Adv Genet*, 1956, 8: 29~54
- Flor H H. Current status of the gene-for-gene concept. *Annu Rev Phytopathol*, 1971, 9: 275~296
- Dangl J L, Jones J D. Plant pathogens and integrated defense responses to infection. *Nature*, 2001, 411: 826~833[DOI]
- Hammond-Kosack K E, Jones J D G. Plant disease resistance genes. *Annu Rev Plant Physiol Plant Mol Biol*, 1997, 48: 575~607[DOI]
- Jones J D. Putting knowledge of plant disease resistance genes to work. *Curr Opin Plant Biol*, 2001, 4: 281~287[DOI]
- Goff S A, Ricke D, Lan T H, et al. A draft sequence of the rice genomes (*Oryza sativa* L. ssp. *japonica*). *Science*, 2002, 296: 92~100[DOI]
- Yu J, Hu S, Wang J, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, 2002, 296: 79~92[DOI]
- Sasaki T, Matsumoto T, Yamamoto K, et al. The genome sequence and structure of rice chromosome 1. *Nature*, 2002, 420: 312~316[DOI]
- Feng Q, Zhang Y, Hao P, et al. Sequence and analysis of rice chromosome 4. *Nature*, 2002, 420: 316~320
- The Rice Chromosome 10 Sequencing Consortium. In-depth view of structure, activity, and evolution of rice chromosome 10. *Science*, 2003, 300: 1566~1569
- Meyers B C, Kozik A, Griego A, et al. Genome-wide analysis of NBS-LRR encoding genes in *Arabidopsis*. *Plant Cell*, 2003, 15: 809~834[DOI]
- Meyers B C, Dickerman A W, Michelmore R W, et al. Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. *Plant J*, 1999, 20: 317~332
- Bai J, Pennill L A, Ning J, Lee S W, et al. Diversity in nucleotide binding site-leucine rich repeat genes in cereals. *Genome Res*, 2002, 12: 1871~1884[DOI]
- Meyers B C, Kozik A, Griego A, et al. Genome-wide analysis of NBS-LRR-Encoding genes in *Arabidopsis*. *Plant Cell*, 2003, 15: 809~834
- Monosi B, Wissner R J, Pennill L, et al. Full-genome analysis of resistance gene homologues in rice. *Theor Appl Genet*, 2004, 109: 1434~1447[DOI]
- Chelkowski J, Koczyk G. Resistance gene analogues of *Arabidopsis thaliana*: Recognition by structure. *J Appl Gen*, 2003, 44: 311~321
- Koczyk G, Chelkowski J. An assessment of the resistance gene analogues of *Oryza sativa* ssp. *japonica*: Their presence and structure. *Cell Mol Biol Lett*, 2003, 8: 963~972
- Botstein D, White R L, Skolnick M, et al. Construction of a genetic linkage map in the man using restriction fragment length polymorphisms. *Am J Hum Genet*, 1980, 32: 314~331
- McCouch S R, Chen X, Panaud O. Microsatellite mapping and applications of SSLP's in rice genetics and breeding. *Plant Mol Biol*, 1997, 35: 89~99 [DOI]
- Coburn J, Temnykh S, Paul E, et al. Design and application of microsatellite marker panels for semi-automated genotyping of rice (*Oryza sativa* L.). *Crop Sci*, 2002, 42: 2092~2099
- Ponce M R, Robles P, Micol J L. High throughput genetic mapping in *Arabidopsis thaliana*. *Mol Gen Genet*, 1999, 261: 408~415[DOI]
- Altschul S, Madden T, Schaffer A, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucl Acids Res*, 1997, 25: 3389~3402[DOI]
- Eddy S R. Profile hidden Markov models. *Bioinformatics*, 1998, 14: 755~763[DOI]
- Lupas A, van Dyke M, Stock J. Predicting coiled coils from protein sequences. *Science*, 1991, 252: 1162~1164
- Krogh A, Larsson B, von Heijne G, et al. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J Mol Biol*, 2001, 305: 567~580[DOI]
- Florea L, Hartzell G, Zhang Z, et al. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res*, 1998, 8: 967~974.
- Rice P, Longden I, Bleasby A J. Internet Resources: EMBOSS. *Trends Genet*, 2000, 16: 276~277[DOI]
- Schuler G D. Sequence mapping by electronic PCR. *Genome Res*, 1997, 7: 541~550
- Pan Q L, Wendel J, Fluhr R. Divergent evolution of plant NBS-LRR resistance gene homologues in dicot and cereal genomes. *J Mol Evol*, 2000, 50: 203~213
- Tian Y, Fan L, Thurauf T, et al. The absence of TIR-type resistance gene analogues in the sugar beet (*Beta vulgaris* L.) genome. *J Mol Evol*, 2004, 58: 40~53[DOI]
- Kikuchi S, Satoh K, Nagata T, et al. Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. *Science*, 2003, 301: 376~379[DOI]

(2005-02-05 收稿, 2005-04-25 收修改稿)