浙 江 林 学 院 学 报 2008，**25**(1)：104－108
Journal of Zhejiang Forestry College

数量性状位点(QTLs)内候选基因的生物信息学分析方法

吴家胜[1]，汪旭升[2]
(1. 浙江林学院 林业与生物技术学院，浙江 临安 311300；2. 浙江大学 生物信息学研究所，浙江 杭州 310029)

摘要：几乎所有的农作物都开展了 QTL(quantitative trait loci)定位研究，定位的方法也很多，如区间作图、复合区间作图、基于混合线性模型的复合区间作图和贝叶斯作图等，但这些研究方法均有它们自身的缺陷，定位的 QTL 在基因组上区间仍然很大，精度不高。为了更好地理解 QTLs 的分子生物学基础。文章简要地介绍了利用生物信息学方法分析 QTL 区间内候选基因的方法，并着重从遗传、基因组结构、表达和功能等方面来剖析 QTL 内的候选基因，为将来更好地利用 QTL 提供了借鉴。图 1 参 22

关键词：遗传学；数量性状位点；候选基因；生物信息学；综述
中图分类号：S718.46；Q343.1　　**文献标志码**：A　　**文章编号**：1000-5692(2008)01-0104-05

Methods of bioinformatic analysis for candidate genes underlying quantitative trait loci (QTLs)

WU Jia-sheng[1], WANG Xu-sheng[2]
(1. School of Forestry and Biotechnology, Zhejiang Forestry College, Lin'an 311300, Zhejiang, China;
2. Bioinformatics Institute, Zhejiang University, Hangzhou 310029, Zhejiang, China)

Abstract: Almost all crops have been studied on QTL (quantitative trait loci) mapping with many QTL mapping methods, such as IM (interval mapping), CIM (composite interval mapping), MCIM (mixed-model based composite interval mapping) and Bayesian QTL mapping having been developed. However, these methods have had shortcomings, namey, the genomic region of the QTL detected, within which there were probably hundreds of candidate genes was still too large. In this paper, a better understanding of the molecular functions of QTLs was obtained by first briefly reviewing methods of analyzing the candidate gene within QTL intervals based on bioinformatics. Then, in order to provide a new analytical method for better use of QTLs in the future, the genetics, genome organization, gene expression and function of candidate genes located on QTL region were analyzed. [Ch, 1 fig. 22 ref.]

Key words: genetics; QTL (quantitative trait loci); candidate gene; bioinformatics; review

自从 Lander 等[1]发展了以连锁图谱上 2 个相邻的遗传标记为基础的数量性状位点(QTLs, quantitative trait loci)定位方法以后，QTL 定位方法得到了长足的发展。Zeng[2]探讨了将多元回归用于 QTL 作图的理论问题，提出了多元线性回归与区间作图法相结合的复合区间作图法(CIM，composite interval mapping)。基于混合线性模型的复合区间作图(MCIM，mixed-model based composite interval mapping)方法的提出[3,4]，为分析各种复杂的 QTL 提供了有力的工具。随着 QTL 定位方法的不断发展，QTL 定位方法在很多物种的各种性状上得到了广泛的应用。在过去几十年里，QTL 定位在植物上已经找到数以百计的抗病、抗虫、抗旱、株高等复杂性状在染色体上的位置，但是这个区段往往都比较大，通常认为在这区段内的基因都为候选基因。即使是精细作图可以减少一些候选基因的数量，但是在一定的区域内的基因仍然比较多。如果要最终利用 QTL 信息，我们还得找到在这些 QTL 区域内的基因的信息。

收稿日期：2007-04-20；修回日期：2007-09-20
基金项目：国家自然科学基金资助项目(30671704)
作者简介：吴家胜，副教授，博士，从事森林培育和数量遗传学研究。E-mail：wujs@zjfc.edu.cn

基因定位在目前是一个研究热点，尤其对复杂性状的 QTL 定位还刚刚开始。随着基因定位工作的开展，对 QTL 的分子生物学基础有了进一步的理解。但是，真正已经定位并克隆的 QTL 仍然比较少，如番茄 *Lycopersicon esculentum* 中控制大小和酸度的主效基因、水稻 *Oryza sativa* 中控制抽穗期和粒质量的基因等[5-7]。目前，无论采用上述的任何一种方法，定位的 QTL 在基因组上的区间仍然很大，可能有几百个基因。虽然目前也有几种方法来缩小 QTL 的区间，如精细作图、染色体单片段替换等，但是这些方法在各方面消耗较大。随着基因组序列呈指数上升，如水稻 2 个亚种粳稻和籼稻已于 2002 年完成草图序列[8,9]，其中粳稻的精细图也于 2005 年完成[10]。此外，玉米 *Zea mays*，小麦 *Triticum sestivum* 和大麦 *Hordeum vulgare* 等进行了大批量的 EST（expressed sequence tag）测序[11-13]，都对基因组上 QTL 的研究起到很大的作用。本文试图系统地阐明从 QTL 到基因的生物信息学的研究方法（图 1）。从 QTL 到基因需要很多步骤。①要把 QTL 定位到染色体上。②从多个紧密连锁的 QTLs 中分离一个 QTL 出来。③找到和评估候选基因。④测试这些基因的功能。

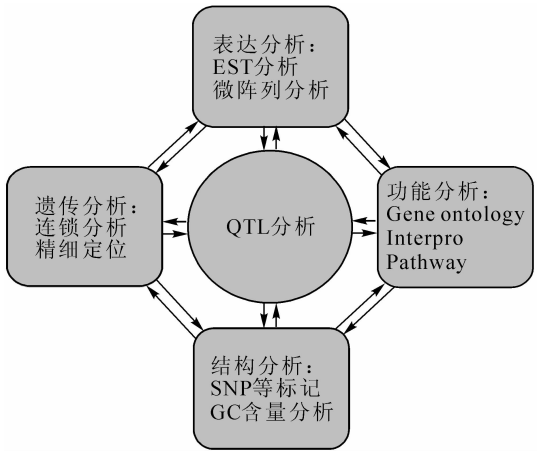

图 1　QTL 内基因的生物信息学分析方法图示

Figure 1　Summary of bioinformatic analyzing methods of genes underlying quantitative trait loci (QTL)

1　遗传分析

1.1　QTL 位置、遗传图谱、物理图谱和基因组序列

以往我们寻找 QTL 位点的方法有很多的不确定性，现在我们用的最好的途径就是定义一个核心区域。我们一般把 LOD（log of odds）值 >3 的区域称为核心区域，有时也把 LOD >2 或者 >1 称为核心区域。由于 QTL 定位是基于遗传图谱，定位出来的 QTL 也只是遗传图谱上的一个片段，而在基因组上的物理位置是未知，更不要说其基因的结构，所以最好尽可能地饱和遗传图谱。首先，要把遗传图谱上标记定位到物理图谱。把一个标记定位到物理图谱上一般有如下几种方法：①通过已有信息，如水稻在 RGP（rice genome research project）网站（http：//rgp. dna. affrc. go. jp/publicdata/geneticmap2000/index. html）提供了 3 267 个 RFLP（restriction fragment length polymorphism）在粳稻克隆上的标记位置和 TIGR（the institute for genomic research）网站（http：//www. tigr. org/tdb/e2k1 /osa1/mappedbacends /BACends. htm）提供了 2 240 个 SSR（single sequence repeat）的标记位置。②通过标记的一对引物，用电子 PCR（polymerase chain reaction）的方法[14]定位到基因组的位置上。③通过标记的序列信息，用 BLASTN 的方法[15]与整个基因组进行联配，从而获得这标记在基因组上的位置。一旦通过标记定位了一个 QTL 区间之后，就要提取这个区间内所有的基因及基因的序列。目前，在一定染色体区间内显示基因组信息，一般是通过 Gbrowser 来显示，如人类基因组和水稻基因组可以分别在 UCSC 的网站（http：//genome. ucsc. edu/）和 Gramene 网站（http：//www. gramene. org）。

既然在基因组上定位到了 QTL 位点，下一步的目标就检查在这个区间内克隆的位置和方向及缺口区域（gap）与重复的情况。在这个区间内可能会存在缺口区域，如果在所研究的区间内有缺口区域存在，那么我们认为这段为未知的区间。由于目前在一些基因组的测序是 clone-by-clone 的方法进行测序列的，如水稻的日本晴基因组，因此，在这个区间的克隆会存在较大的重复。通过 Gbrowser 这个软件可以明显找到那些基因是重复的，在 QTL 区间内知道了克隆的次序和方向之后，接下来就找到在这些克隆上的已知功能的和未知功能的基因，从而来评估在这些区间内的基因。目前，TIGR 和 Genbank 都对现有的一些基因组进行了注释，为分析基因的功能提供了很大的方便。

1.2　饱和遗传图谱和精细定位

利用各种标记，增加 QTL 区间内的标记数目，从而饱和该区段内的标记数，为精细作图和图位克隆提供新机会。利用不同分子标记的遗传图谱，根据它们在基因组上的位置，来饱和整个遗传图

谱。对于水稻而言，现有的分子标记已经相当丰富，应用广泛的有 SSR（simple sequence repeat），RLFP（restriction fragment length polymorphism），AFLP（amplified fragment length polymorphism）和 SNP（single nucleotide polymorphism）标记。日本水稻基因组计划 1998 年发表的水稻分子图谱已包括了 2 275个标记，其中 70% 为 EST，覆盖了 1 550 cM，平均图距为 1.5 cM，现在已经增加到 3 267 个标记。此外，几乎所有重要农作物，如小麦，玉米，大豆 *Glycine max*，马铃薯 *Solanum tuberosm*，棉花 *Gossypium hirsutum* 等的 RFLP 图谱均已构成。随着多种标记的不断添加与定位，分子标记图谱正日渐趋于饱和。高密度分子标记图谱对饱和遗传图谱及对分子标记辅助选择育种技术和基因图位克隆技术的发展起到了巨大的推动作用，尤其对 QTL 区间内的基因克隆提供了有用的资源。

2 区间内基因的结构分析

2.1 发展新的 SNP 和插入/缺失多态性(InDel)标记

通过上述的 QTL 两侧标记定位方法找到特定 QTL 位点内的基因后，对这些新基因进行不同种或亚种间基因组比较来发现一些新的标记。目前，基因组测序正在大规模地进行，像水稻和果蝇 *Drosophila melanogaster* 等都有不同的亚种或种已经测序。对这些种的等位基因进行联配，就可以比较清楚地比较基因的结构，如特定基因的 SNP 分布或 InDel（insertion，deltion）分布，有些 InDel 还可以发展成新的标记。如水稻可以通过 2 个亚种间的序列比较，获得一些新的 SNPs。这样可以饱和这个区段内的标记，并且把特定的基因功能与保守区域联系起来，将会对了解基因的构成有很大的帮助。而且有些变异可能就是造成表型变异的直接来源，如基因的外显子部分或启动子部分的变异。在序列数据中查找微卫星序列及发展 SSR 标记。到目前为止，我们找到了多态性数据，用 SNP 来扫描我们的位点可能已经够了。但是，在作图过程中可能需要将目标 QTL 定位缩短到 50 ~ 100 kb 区间。

2.2 分析 GC 含量

除了分析遗传图谱和物理图谱以后，我们还可以对这个位点内的重组率进行估算。通过计算这个位点的 GC（gas chromatograph）含量，可以间接地分析重组率。低的 GC 含量往往对应于低的重组率[16]。因此，在低 GC 的 QTL 区间内往往基因相对较少，而且往往变异相对较大。

3 QTL 区间内基因的表达分析

3.1 表达序列标签定位法

对 QTL 区间内基因的表达的分析，一种比较简单的方法就是通过定位 ESTs。目前已经有很多生物都已经测定了不同组织，不同处理，不同部位的 EST 序列，远远超过基因组的测序。通常的方法就是通过 BLASTN 联配的方法来定位区间内基因的 EST 表达情况，同时，也可以通过 EST 数量来测定这些基因的表达量的情况。我们还可以根据 EST 序列与 UniGene 的记录来证实特定组织的某个基因的表达。目前，EST 方法已经非常有效地应用到候选基因的确认和分离的分析中。如：EST 处于基因的编码区，那么 EST 是与目标基因共分离的。以此为基础，综合分析已有的遗传图谱、物理图谱和基因图谱，可将基因在染色体上准确定位。

3.2 微阵列(microarray)法

最近研究表明，在很多生物中调控的变异是复杂性状变异的主要原因[17]。而通过基因芯片的表达研究能够揭示大量基因的调控的变异。而单独的微阵列分析能够分析不同组织和不同处理的变异，但是不能跟特定的表型联系在一起。通过组合 QTL 分析和微阵列技术，能找到特定处理、特定的组织和特定的表型下的位置基因。Wayne 等[18]曾利用组合 QTL 作图和微阵列的分析找到了控制果蝇卵巢管数目的 34 个候选基因。

4 区间内基因的功能分析

4.1 Gene ontology (GO)[19]的分析方法

GO（http：//www.godatabase.org/dev/）是一组术语，由描述分子功能、生物学过程、基因产物相

关的细胞组成的分等级机构词汇(控制性的结构词汇)组成。用 GO 的方法对基因组注解的已知或预测的基因进行分析，可以很快得到结果。它是一个模拟生物学数据库、序列信息中心和其他基因组数据的提供者，GO 数据库是构成 GO 功能的支柱。AmiGO 浏览器可以直接访问它的数据库。因此，可以利用 GO 数据库和 AmiGO 浏览器，对 QTL 区间内的已知基因和未知功能的基因的产物进行分类。

4.2　InterPro 数据库[20]的分析方法

InterPro (http：//www. ebi. ac. uk/interpro/)是将几大数据库(PROSITE，PRINTS，Pfam 和 ProDom)的蛋白质结构的信息统一起来的一个蛋白质结构域和功能位点的数据库，因此，只要访问这个站点，就可以获得所有这些数据库相关的信息。来自这些数据库的合并的注释构成了 InterPro 的核心，每个条目包括蛋白质的功能描述和文献参考，以及与相关数据库的链接，包括前面所提到的 GO 记录。同时，InterPro 还提供了一个使数据库本地化后的检索工具 Interproscan。此外，PDB (protein data bank)，SCOP (structural classification of proteins)，CATH [Class (C)，Architecture (A)]，Topology (T) and Homologous superfamily (H)，HOMSTRAD (homologous structure alignment database)，CAMPASS (CAMbridge database of protein alignments organised as structural superfamilies)等为蛋白质结构数据库，它们运用不同的原理来识别结构相似的蛋白质超家族，由于结构域在进化过程中比序列保守，一些通过核苷酸序列识别不到的蛋白质超家族在这些数据库中可以识别到。而且，这些数据库大部分是建立在结构域基础上的，结构域范围很容易在整个蛋白质结构中找到。通过用 InterPro 方法，可对 QTL 区间所有基因的功能域进行分析，并对不同功能域的基因进行分类。

4.3　Pathway 的分析方法

通常基因之间的关系相当复杂，遗传学家需要扩大路径并且鉴定复杂的路径之间的交叉。根据文献查找某个性状的生理途径，搜索是否有基因位于这个 QTL 区段内，如果有，则认为这个基因可能为该 QTL 内的一个候选基因。Pahtway 的分析方法是对基因组的路径进行分析的生物信息学方法，并建立 Pathway 的数据库。这些数据库包含基因组功能的详细的注释，比如说对新陈代谢和信号路径的描述。Pathwayassist 和 GeneWays 就是 2 个很好的分析软件。Pathwayassist (ariadne genomics，rockville，MD)允许用户利用 ResNet 数据库开发基因间的互作网络。ResNet (tm)是一个庞大的分子网络数据库，这个数据库中包含有 10 万个结果的规则、相互作用、1. 5 万个蛋白质、细胞过程和小分子的修饰。这个软件可以帮助我们从实验的结果中分析路径之间的前后联系，鉴别基因、小分子、细胞结构和过程之间的联系，建造便于出版的图表和网络图。用 Pathwayassist 这个工具，可以分析 microarray 和 SAGE 数据，对蛋白质分类和排序，绘制路径图，输入、输出和过滤数据，并且用文献中的新信息修正我们的路径。GeneWays 是一个用语言算法提取文献中的分子和分子过程中的联系，并且把它们之间的联系变成路径[21]，包括提取、分析、形象化和整体化分子路径数据的过程。Pathwayassist 和 GeneWays 是分析基因之间联系的很好的方法。当然，没有一种方法是完美的，每一种方法都有其自身的局限性，只要我们把握了特点，充分利用它们的优点，就可以为区间内基因功能分析提供新的思路。Yonan 等[22]利用 Pathway 的方法对 孤独症的 QTL 内候选基因进行了深入的研究。

5　QTL 内基因的分析的应用及前景

虽然不同的 QTL 定位方法有时定位的结果不一样，但是我们相信 QTL 作图和突变研究是揭示复杂性状的唯一方法。QTL 定位为人们了解染色体片段、基因结构及基因的分布提供了一个很好的工具，同时基因组测序反过来对 QTL 定位所得到的基因的分析提供了强大的数据支持。随着基因组序列的完善，为剖析 QTL 的基因结构提供了一种可能。通过突变的方法来找到基因的生理生化途径，我们会找到一些基因是编码的基因，有些是不编码的基因(启动子、内含子或基因间的区域)序列。不管怎么样，在将来近几十年，QTL 方法仍将是揭示复杂性状的重要方法。对目标性状 QTL 的准确定位是实施分子育种的前提，QTL 定位研究的目的之一就是尽量地挖掘有利用价值的等位基因，从而提供标记辅助选择应用于育种实践，以培育优质、高产和多抗新品种。

生物信息学为分析 QTL 提供新工具和思路，为解释表型变异成为可能。当前，各种基因组序列

不断增加，遗传图谱也日趋饱和，为基因的定位、表达研究提供了崭新工具。相信在人类基因组计划的推动下，在农作物基因组研究的影响下，QTL 内基因的分析研究必将迎来一个迅速发展的时代。

参考文献：

[1] LANDER E S, BOTSTEIN D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps [J]. *Genetics*, 1989, **121**: 185 – 199.

[2] ZENG Z B. Precision mapping of quantitative trait loci [J]. *Genetics*, 1994, **136**: 1 457 – 1 468.

[3] WANG D L, ZHU J, LI Z K, *et al*. Mapping QTL with epistatic effects and QTLxenvironment interactions by mixed linear model approaches [J]. *Theor Appl Genet*, 1999, **99**: 1 255 – 1 264.

[4] GAO Y M, ZHU J, SONG Y S, *et al*. Analysis of digenic epistatic effects and QE interaction effects QTL controlling grain weight in rice [J]. *J Zhejiang Univ Sci*, 2004, **5** (4): 371 – 377.

[5] FRARY A, NESBITT T C, GRANDILLO S, *et al*. fw2. 2: a quantitative trait locus key to the evolution of tomato fruit size [J]. *Science*, 2000, **289**: 85 – 88.

[6] YANO M, KATAYOSE Y, ASHIKARI M, *et al*. *Hd*1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the Arabidopsis flowering time gene *CONSTANS* [J]. *Plant Cell*, 2000, **12**: 2 473 – 2 483.

[7] KONISHI S, IZAWA T, LIN S Y, *et al*. An SNP caused loss of seed shattering during rice domestication [J]. *Science*, 2006, **312**: 1 392 – 1 396.

[8] GOFF S A, RICKE D, LAN T H, *et al*. A draft seqeunce of the rice genomes (*Oryza sativa* L. ssp. *japonica*) [J]. *Science*, 2002, **296**: 92 – 100.

[9] YU J, HU S N, WANG J, *et al*. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*) [J]. *Science*, 2002, **296**: 79 – 92.

[10] International Rice Genome Sequencing Project. The map-based sequence of the rice genome [J]. *Nature*, 2005, **436**: 793 – 800.

[11] GARDINER J, SCHRODER S, POLACCO M L, *et al*. Anchoring 9 371 maize expressed sequence tagged unigenes to the bacterial artificial chromosome contig map by two-dimensional overgo hybridization [J]. *Plant Physiol*, 2004, **134**: 1 317 – 1 326.

[12] LAZO G R, CHAO S, HUMMEL D D, *et al*. Development of an expressed sequence tag (EST) resource for wheat (*Triticum aestivum* L.): EST generation, unigene analysis, probe selection and bioinformatics for a 16 000-locus bin-delineated map [J]. *Genetics*, 2004, **168**: 585 – 593.

[13] ZHANG H, SREENIVASULU N, WESCHKE W, *et al*. Large-scale analysis of the barley transcriptome based on expressed sequence tags [J]. *Plant J*, 2004, **40**: 276 – 290.

[14] SCHULER G D. Sequence mapping by electronic PCR [J]. *Genome Res*, 1997, **7**: 541 – 550.

[15] ALTSCHUL S F, MADDENT T L, SCHAFFER A A, *et al*. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs [J]. *Nucleic Acids Res*, 1997, **25**: 3 389 – 3 402.

[16] YU A, ZHAO C, FAN Y, *et al*. Comparison of human genetic and sequence-based physical maps [J]. *Nature*, 2001, **409**: 951 – 953.

[17] MACKAY T F. The genetic architecture of quantitative traits [J]. *Annu Rev Genet*, 2001, **35**: 303 – 339.

[18] WAYNE M L, MCLNTYRE L M. Combining mapping and arraying: an approach to candidate gene identification [J]. *Proc Natl Acad Sci USA*, 2002, **99**: 14 903 – 14 906.

[19] ASHBUMER M, BALL C A, BLAKE J A, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Corsortium [J]. *Nature Genet*, 2000, **25**: 25 – 29.

[20] APWEILER R, ATTWOOD T K, BAIROCH A, *et al*. The Interpro databases, an intergrated documentation resource for protein families, domains and functional sites [J]. *Nucleic Acids Res*, 2001, **29**: 37 – 40.

[21] RZHETSKY A, KOIKE T, KALACHIKOV S, *et al*. A knowledge model for analysis and simulation of regulatory networks [J]. *Bioinformatics*, 2000, **16**: 1 120 – 1 128.

[22] YONAN A L, PALMER A A, SMITH K C, *et al*. Bioinformatic analysis of autism positional candidate genes using biological databases and computational gene network prediction [J]. *Genes Brain Behav*, 2003, **2**: 303 – 320.