文章编号：1004-0374(2009)01-0038-05

·评述与综述·

基因表达数量性状定位的研究进展

陈 颖[1†], 汪旭升[2,3†], 许玲莉[1], 沈 勤[1], 王晓冬[1], 陆 璐[1,3*]
（1 南通大学医学院，南通 226001；2 浙江大学生物信息学研究所，杭州 310029；
3 田纳西大学医学中心，孟菲斯38163，美国）

摘 要：近年来，随着人类和一系列模式生物全基因组测序工作的完成，阐明基因互作、调控网络及代谢途径的生物学功能，成为后基因组时代生物学研究的重点及热点。最近将数量性状定位(quantitative trait loci, QTL)和基因表达分析联合运用，产生了遗传基因组学或基因表达数量性状定位(expression QTL; eQTL)。本文简要地回顾了基因表达遗传变异的本质及eQTL分析的基本原理。在此基础上，结合我们当前的研究工作，重点介绍了eQTL分析方法在候选基因挖掘和基因调控网络构建中的运用，并结合单核苷酸多态性(SNP)对基因表达的影响等问题，讨论eQTL在实际研究分析中面临的困难，并探讨该领域的挑战和发展方向。

关键词：遗传基因组学；基因表达数量性状定位(eQTL)；基因表达；数量性状；基因调控网络
中图分类号：R786；Q343.1+7　　**文献标识码**：A

Advance in study of gene expression quantitative trait loci (eQTL)

CHEN Ying[1†], WANG Xu-sheng[2,3,†], XU Ling-li[1], SHEN Qin[1], WANG Xiao-dong[1], LU Lu[1,3*]
(1 Medical College of Nantong University, Nantong 226001, China; 2 Institute of Bioinformatics, Zhejiang University, Hangzhou 310029, China; 3 University of Tennessee, Health Science Center, Memphis 38163 TN, USA)

Abstract: In recent years, it has become one of the most critical issues and hot topics in the field of biology to elucidate the gene interaction, regulatory network and metabolic pathway in the post-genome era as the complete sequencing of the genomes of human and some model organisms was revealed. The combination of quantitative trait loci (QTL) mapping with high-throughput gene expression generated genetical genomics or gene expression QTL (eQTL). In this paper, we briefly reviewed the mechanism of genetic variation in gene expression and the rationale of the analysis of eQTL. Based on our current studies, we introduced the application of eQTL approach in the identification of candidate genes and constructing genetic regulatory network. In view of the potential impact of SNPs on gene expression, we discussed the problems in the application of eQTL and explored its challenges and the development trends.

Key words: genetical genomics; expression QTL(eQTL); gene expression; quantitative trait; gene regulatory network

在高等生物中，许多重要的农艺性状、生理性状及复杂疾病都是数量性状，如农作物的产量和人类的高血压、糖尿病等，这些复杂性状受多个基因和环境因素的控制。为了有效地研究多基因控制的复杂性状，数量性状基因定位 (quantitative trait loci, QTL) 分析技术在20世纪 90 年代应运而生，有效地将控制数量性状的众多主效基因定位在相应的染色体上。传统的QTL分析只对个别或几个复杂性状进行 QTL 定位，从而获得控制复杂性状的一个或几个染色体的区间，再通过精细定位等手段，

收稿日期：2008-09-18；**修回日期**：2008-11-06
基金项目：国家自然科学基金项目(30700517, 30771200, 30770666)；江苏省自然科学基金重点项目(BK2007703)；江苏省自然科学基金项目(BK2007065)
***通讯作者**：lulu521ut@gmail.com
†**相同贡献**

发现其候选基因。实际上，由于定位到的 QTL 置信区间内包含了大量的基因，因此要精确定位到某一个或几个主效基因很具挑战性。这种困扰很大程度上制约了复杂性状的研究。此外，这种相对独立的分析显然不能够充分解释复杂的生命现象。因此，如何发现控制复杂性状的基因间互作、基因调控网络和代谢途径成为当今研究的热点。

基因芯片技术的应用使得同时分析成千上万个基因的表达水平成为可能。2001年，Jansen 和 Nap[1] 提出将全基因组中的每个基因的 mRNA 表达量作为数量性状，对其进行 QTL 定位分析，即基因表达的数量性状定位分析技术(expression QTL, eQTL)，又叫遗传基因组学(genetical genomics)(图1)。一个 eQTL 就是染色体上的一个位点，这个位点可以包含一个或多个基因，这个（些）基因控制着某个基因表达的遗传变异。遗传基因组学综合运用基因组学、统计遗传学和生物信息学的方法，寻找控制这些基因表达的上游调控位点，发掘受该基因调节的下游基因及与该基因协同作用的基因，并进而建立基因调控网络，以阐明基因调控的机制，从而在表达及调控两个水平研究控制复杂性状的遗传基础。

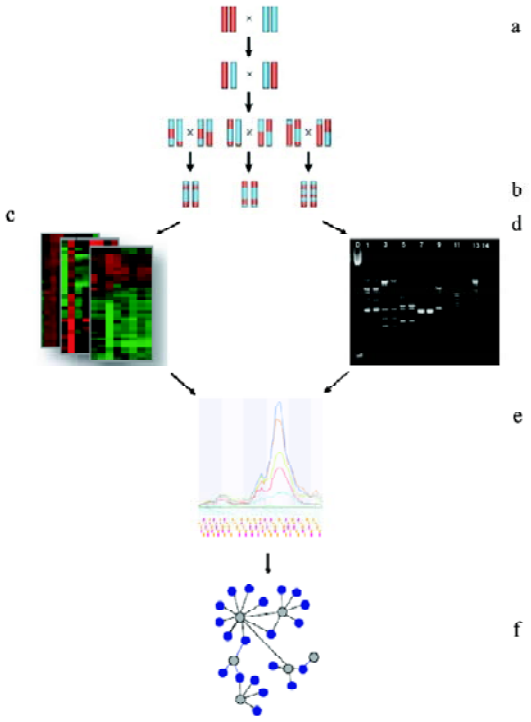

图1　基因表达数量性状定位分析技术的图示

a：亲本杂交；b：分离群体；c：每个个体的全基因组表达谱；d：分离群体内个体的分子标记；e：利用 QTL 定位分析方法定位转录子；f：基因调控网络的推断。

1　基因表达数量性状变异的遗传本质

基因表达数量性状定位的本质是定位控制 mRNA 表达变异的遗传因子。一个基因转录水平的调控受到多因素的影响，包括多个遗传和环境因子的影响。通过基因表达数量性状定位的方法能对基因组的几万个转录子同时进行 eQTL 分析。转录子与传统意义上的表型不同，每个转录子在基因组上的相应位置是已知的，因此可根据 eQTL 定位到的区间与转录子在染色体上的相对位置，将 eQTL 分为顺式作用eQTL (cis-eQTL)和反式作用eQTL(trans-eQTL)；若 eQTL 被定位到基因自身所在的区域，则为顺式作用；反之，若被定位到基因自身所在以外的其他区域，则为反式作用。

1.1　顺式作用eQTL　从全基因组水平定位产生的顺式作用 eQTL 可能有以下几种情况：(1) 由于基因自身启动子序列的多态性，从而造成了转录因子结合位点的变异或染色质结构的改变，进而引起基因表达的差异。产生这类变异的情况比较普遍，据我们对小鼠海马的研究发现，大约有 900 个 cis-eQTL 是由于这类变异造成的[2]；(2) 由于邻近基因在编码序列上的改变，反式作用于其定位基因，从而引起其基因表达的改变。从定位分析图上来看这是一个 cis-eQTL，但从本质上来讲，这其实是一种反式作用的 eQTL，仅仅根据分离群体的表达数据将很难将其与真正意义上的cis-eQTL区别开来，需要进一步的分子生物学实验，如建立荧光素酶报道系统、体外诱导后检测 mRNA 表达量等，深入了解其作用机制，从而判断其是否为真正意义上的 cis-eQTL；(3) 转录子自身在编码区域内序列的改变，引起其他基因在蛋白水平的改变，最后反馈影响其自身的表达变异。这实质上是一种自身反馈调节作用。在对酵母的研究中发现，*AMN1* 基因就是通过一系列的反式作用之后，引起自身的表达改变[3]。

我们对小鼠大脑进行基因水平上的表达定位分析时发现，假阳性率 (false discovery rate, FDR) 控制在 5% 以下时，大约 30% — 40% 的 eQTL 是 cis-eQTL[4]。当然在不同物种中 cis-eQTL 的数目也不完全相同，譬如人类大约有 70% 的 eQTL 是 cis-eQTL[5]；同样，在相同物种的不同组织中 cis-eQTL所占的比例也会不同。一般来说，连锁程度高的，即 LOD (log of odds)值高的 eQTL 通常是顺式作用 eQTL，而反式作用 eQTL 的 LOD 值则相对较低，因此影响基因本身转录的 DNA 变异较之影响其他基因转录的

DNA 变异更容易被检测到。

1.2 反式作用eQTL 反式作用eQTL不同于顺式作用eQTL，相对于定位的转录子而言，它定位于基因组其他位置，即不同染色体或相同染色体的不同区间(通常大于20Mb)。反式作用一般是通过改变基因的编码序列，从而引起蛋白结构的改变，最终影响定位转录子的基因表达差异。反式作用也可能是转录因子通过其他某种途径发挥作用。Morley等[6]在对人类类淋巴母细胞系(LCLs)进行研究时找到110个cis-eQTL和17个trans-eQTL，并提出受顺式作用调节的基因远多于受反式作用调节的基因。但实际上，由于主调控子的存在，trans-eQTL的数目可能要多于cis-eQTL。

1.3 基因表达的调控热点/主调控子 控制多个基因表达的调控基因(如转录因子)的遗传变异会影响几乎所有受其调控的基因表达。在基因表达数量性状定位中将这类遗传因子定义为主调控子(master regulator)，其调控区域为调控热点（图2)。部分反式作用eQTL有“热点”聚集现象，这些热点可能包含主效调控基因，而每个主效调控基因调节众多转录子的变异。我们在BXD小鼠前脑的基因表达数据中发现有7个反式作用eQTL影响着几百个基因的表达[4]。其中一个主调控子(位于1号染色体的末端)被命名为QTL富集区(QTL rich region, Qrr1)，我们目前正在对其进行深入的分析。该Qrr1主要是由于两个连锁单体（haplotype）造成，且大部分基因是与神经因子及主要的神经系统疾病相关。Brem等[7]对6个eQTL热点进行了分析，初步确定了转录因子Hap1是其中一个eQTL热点的主效调节基因。

主调控子在遗传基因组学中是一个非常普遍的现象，但也有研究未发现主调控子现象[8]。对于是否出现主调控子，目前部分研究认为可能是由于所使用的基因芯片、分子标记类型或者分析方法等的差异所造成的[9-11]。基因芯片的实验实际上是一系列的复杂过程，包括杂交、染色和扫描等过程。虽然在芯片处理过程中，标准化去除了大量的人为效应，如芯片批次、染色效应和性别效应等，但我们还是会发现一些基因表达量会随着芯片的批次和在芯片上的位置的改变而改变。

1.4 基因表达的遗传率 在经典的数量遗传学中，遗传率分为广义遗传率和狭义遗传率。广义遗传率是指基因型变量与表型变量的比率，包括加性效应和非加性效应（显性和上位性效应）遗传量。狭义遗传率是指基因加性作用所引起的变异占全部表型变异的比值。现研究表明许多基因的转录水平是可遗传的[7,12]。在基因表达遗传学中，基因表达水平被视为数量性状，因此将遗传率的概念引入基因表达遗传学，且主要是指狭义遗传率。Monks 等[8]对来自15个CEPH家系的LCLs进行研究，发现了2 430个差异表达的基因，其中762个基因（约31%）的表达呈高度遗传，中等遗传率约占34%。Brem和Kruglyak[13]在对酵母的研究中发现，具有最显著

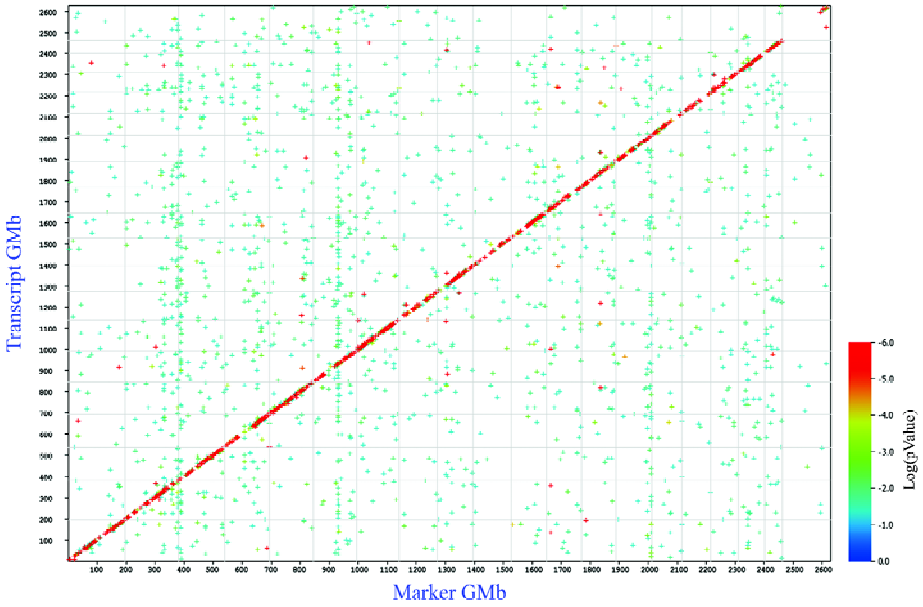

图2 LXS RI小鼠海马主调控子的例子(www.genenetwork.org)

图中X轴表示分子标记的位置；Y轴表示转录子的位置；黑色的框代表主调控子区域；红色对角线代表顺式调控子

QTL的1 038个基因的中等遗传率为27%。Vuylsteke等[14]在对拟南芥的研究中同样发现有 82%的差异表达基因认为是可遗传的，遗传率从 11%到93%不等，中等遗传率占30%。尽管各项研究的结果不尽相同，它们主要是由于组织来源、样本含量、统计方法、遗传多样性、环境因素等的差异所导致；但这些研究都无一例外地表明，基因表达水平完全可以作为可遗传的(数量)性状，对其进行遗传分析。

2　基因表达数量性状定位研究的运用

2.1　候选基因的挖掘　一旦在标记区间内发现存在具有统计学意义的eQTL，那么这个位点很可能包含了控制着某个基因表达变异的候选基因。传统QTL定位很难确定其定位区间内的候选基因，而eQTL不同于传统的QTL，如果定位到的eQTL具有顺式作用，那么往往认为引起该表达变异的基因就是其本身。最近，我们利用BXD和LXS两个重组近交系小鼠，分析与紧张焦虑相关的海马特异性表达的基因，结果发现40多个与紧张焦虑相关的基因具有cis-eQTL(未发表)。我们进一步利用等位基因特异性表达(ASE) 实验手段来验证这些 cis-eQTL，并构建含有这些基因启动子序列的载体，利用荧光素酶报告基因来检测启动子部分的序列变异对其基因表达变异的影响。由此可见，结合遗传基因组学的方法和分子生物学手段，不仅能够找到候选基因，而且能够发现引起基因表达变异的原因。

与cis-eQTL的分析相比，找到trans-eQTL内的候选基因具有更大的挑战性。目前主要有以下几种方法来剖析trans-eQTL内的候选基因：首先，在trans-eQTL区间内寻找自身为cis-eQTL的转录子，这些转录子很可能就是候选基因；其次，利用目前已发现的大量调控或生化途径，将转录子与trans-eQTL内的基因共定位于调控或生化网络将有助于发现目的基因。有研究报道通过特定的统计方法，如基于似然法因果模型选择法(LCMS)，来剖析候选基因[15]。此外，还可以通过差异表达、表型相关和序列多态性分析等生物信息学手段以及染色质免疫沉淀法、转基因技术、基因敲除等分子生物学方法对trans-eQTL内的候选基因进行进一步筛选，以最终找到目的基因。

2.2　基因调控网络　eQTL技术结合了基因表达资料和表达水平的QTL分析，可用于分析基因之间的调控关系，进而构建基因调控网络。在eQTL研究中，由于顺式作用 eQTL定位于基因本身所在的区域，故顺式作用 eQTL 可以直接提供候选基因的信息；反式作用 eQTL 区间内可能包含多个候选基因，通过前述多种生物信息学分析方法和现代分子生物学方法可筛选和发现其上游调控基因。我们利用已发表的 eQTL 表达数据和 bayesian网络方法，对209个反式作用eQTL构建了66个候选调节网络，每个网络都是有向图，图中位于eQTL之间的基因是候选调控基因，而调控基因本身表达水平的调节基因又会被定位到其他区域[16]，由此形成了更为复杂的基因调控网络。当然，在eQTL定位中发现的上位性也是一种很好的潜在的调控网络。Brem和Kruglyak[13]在对酵母的实验研究中发现，在3 546个高遗传率的转录子中，约40%未发现QTL，约16%具有上位性QTL。与传统QTL上位性的研究相比，目前对eQTL上位性的分析还停留在初步认识阶段，因此如何分析eQTL的上位性及调控网络之间的关系，还有待进一步的研究。

3　基因表达数量性状定位研究中存在的问题

3.1　单核苷酸多态性(SNP)对基因表达的影响　SNP在基因组中广泛存在，目前利用多种手段发现小鼠C57BL/6J和DBA/2J两个品系之间存在1 400多万个SNPs。在基因组DNA中，位于编码区内的SNP(coding SNP, cSNP)比较少，因为外显子的变异率仅为其他序列的1/5，SNP更可能出现在非编码序列中，包括启动子、内含子和基因间的序列。调控序列主要分布在非编码区域，因此这部分序列的变异可能会造成表达的改变，但是如果序列的多态性(如SNP、INDEL)发生在探针序列上，则会导致假阳性cis-eQTL的产生。为了研究SNP对eQTL分析结果的影响，我们对SNP在探针不同位置上引起的eQTL假阳性率进行了评估，结果表明在探针中心位置的SNP对eQTL的定位影响大于SNP位于探针末端的影响。所有小鼠的基因芯片所用的探针都是基于C57BL/6J基因组序列的，因此在BXD重组近交系中发现的cis-eQTL中，C57BL/6J等位基因的表达量往往大于DBA/2J等位基因的表达量[17-19]。现有的多个SNP数据库，如SNPdb (http://www.ncbi.nlm.nih.gov/SNP/)，可以用来排除由于探针上SNP造成表达变异从而产生的假阳性cis-eQTL。

3.2　费用的昂贵　基因表达数量性状定位的研究需要足够重组信息量的QTL作图群体、足够精细的遗传图谱及表达性状资料。这需要大量的人力、物

力、财力的投入。基因芯片技术是研究高通量基因表达水平不可或缺的工具，然而其昂贵的技术成本使得许多实验室望而兴叹。作图群体的建立同样需要大量的投入，一个重组近交系小鼠的建立是经全同胞交配达 20 代以上培育而成，往往需要 7 － 9 年的时间。品系建立成功之后还需要大量工作，如品系间的基因分型、高密度遗传图谱等。虽然目前单个基因芯片的价格不算昂贵，但同时测量群体内的上百个个体的基因表达，仍需要非常大的投入。

3.3 证实的难度 除了 SNP 会导致假阳性 cis-eQTL，还有其他多种因素会影响 eQTL 的检测，如群体的大小、检测手段、统计方法的选择以及遗传异质性、等位基因频率等。这意味着所检测的 eQTL 还需要经过进一步的筛选，我们在研究中选择了置换试验 (permutation test) 和FDR方法筛选出在统计学上具有明显意义的上游基因调控位点[4]。虽然统计方法的应用能够提高检测的准确性，但通过eQTL分析所构建的基因调控网络仍需要分子生物学的实验进行验证，如 ASE 分析技术、建立荧光素酶报道系统及免疫共沉淀实验等。随着生物技术的进一步发展，我们对基因表达网络的理解会越来越深入。

4 基因表达数量性状定位研究的发展方向

20世纪，分子生物学家主要集中于对单个基因或单个蛋白质的研究，以及它们之间有限的交互作用。尽管这些研究工作对了解一个复杂的生物系统而言必不可少，但这些相对独立的数据无法诠释一个复杂的生物系统的生命本质。孟德尔遗传学研究单个性状与遗传变异的关联；复杂性状分析研究某数量性状与多个遗传变异的关系；而目前的系统遗传学则同时研究多个遗传变异、环境因子及多种表型间的复杂关系。现在人们又将 mRNA 表达量及蛋白质谱作为表型，融合到系统遗传学的研究中，加之过去几年里获得的大量基因型数据，使得同时在不同时间、不同部位的多个层次上分析基因表达和基因表达的调控成为了可能，从而系统了解基因转录的调控机制，最终构建基因表达的遗传网络。

遗传学的研究最终将把模式生物的研究与非模式生物联系起来，目前基因表达数量性状定位的研究同样主要是基于酵母、小鼠、大鼠和拟南芥等模式生物。在未来几年，如何进一步诠释模式生物的基因表达的遗传基础及如何将现有的研究结果扩展到非模式生物(如人类和水稻)的研究将成为研究的重点和难点。

[参 考 文 献]

[1] Jansen RC, Nap JP. Genetical genomics: the added value from segregation. Trends Genet, 2001, 17(7): 388-91

[2] Lu L, Cook MN, Bennett B, et al. Genetic dissection of transcriptional regulatory network in the hippocampus of LXS mice[C]//30th Annual Scientific Meeting of the Research Society on Alcoholism. Chicago, Illinois, USA, July 7-12, 2007

[3] Ronald J, Brem RB, Whittle J, et al. Local regulatory variation in *Saccharomyces cerevisiae*. PLoS Genet, 2005, 1(2): e25

[4] Chesler EJ, Lu L, Shou S, et al. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nat Genet, 2005, 37(3): 233-42

[5] Goring HH, Curran JE, Johnson MP, et al. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nat Genet, 2007, 39(10): 1208-16

[6] Morley M, Molony CM, Weber TM, et al. Genetic analysis of genome-wide variation in human gene expression. Nature, 2004, 430(7001): 743-7

[7] Brem RB, Yvert G, Clinton R, et al. Genetic dissection of transcriptional regulation in budding yeast. Science, 2002, 296(5568): 752-5

[8] Monks SA, Leonardson A, Zhu H, et al. Genetic inheritance of gene expression in human cell lines. Am J Hum Genet, 2004, 75(6): 1094-105

[9] Alberts R, Terpstra P, Bystrykh LV, et al. A statistical multiprobe model for analyzing cis and trans genes in genetical genomics experiments with short-oligonucleotide arrays. Genetics, 2005, 171(3): 1437-9

[10] Li J, Burmeister M. Genetical genomics: combining genetics with gene expression analysis. Hum Mol Genet, 2005, 14(Spec 2): R163-9

[11] Williams RW. Expression genetics and the phenotype revolution. Mamm Genome, 2006, 17(6): 496-502

[12] Schadt EE, Monks SA, Drake TA, et al. Genetics of gene expression surveyed in maize, mouse and man. Nature, 2003, 422(6929): 297-302

[13] Brem RB, Kruglyak L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc Natl Acad Sci USA, 2005, 102(5): 1572-7

[14] Vuylsteke M, Daele H, Vercauteren A, et al. Genetic dissection of transcriptional regulation by cDNA-AFLP. Plant J, 2006, 45(3): 439-46

[15] Schadt EE, Lamb J, Yang X, et al. An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet, 2005, 37(7): 710-7

[16] Li H, Lu L, Manly KF, et al. Inferring gene transcriptional modulatory relations: a genetical genomics approach. Hum Mol Genet, 2005, 14(9): 1119-25

[17] Doss S, Schadt EE, Drake TA, et al. Cis-acting expression quantitative trait loci in mice. Genome Res, 2005, 15(5): 681-91

[18] Manly KF, Wang J, Williams RW. Weighting by heritability for detection of quantitative trait loci with microarray estimates of gene expression. Genome Biol, 2005, 6(3): R27

[19] Peirce JL, Li HQ, Wang JT, et al. How replicable are mRNA expression QTL? Mamm Genome, 2006, 17(6): 643-56