

DOI: 10.3724/SP.J.1005.2008.01228

遗传与基因表达数据的整合——eQTL 的方法及应用

刘刚, 彭惠茹, 倪中福, 秦丹丹, 宋方威, 宋广树, 孙其信

中国农业大学植物遗传育种系, 农业生物技术国家重点实验室, 教育部作物杂种优势研究与利用重点实验室, 北京市作物遗传改良重点实验室, 农业部作物基因组与遗传改良重点实验室, 北京 100193

摘要: 高通量的基因型分析和芯片技术的发展使人们能够进一步研究哪些遗传差异最终影响基因的表达。通过表达数量性状座位(eQTL)作图方法可对基因表达水平的遗传基础进行解析。与传统的 QTL 分析方法一样, eQTL 的主要目标是鉴别表达性状座位所在的染色体区域。但由于表达谱数据成千上万, 而传统的 QTL 分析方法最多分析几十个性状, 因此需要考虑这类实验设计的特点以及统计分析方法。本文详细介绍了 eQTL 定位过程及其研究方法, 重点从个体选择、基因芯片实验设计、基因表达数据的获得与标准化、作图方法及结果分析等方面进行了综述, 指出了当前 eQTL 研究存在的问题和局限性。最后介绍了 eQTL 研究在估计基因表达遗传率、挖掘候选基因、构建基因调控网络、理解基因间及基因与环境的互作的应用进展。

关键词: eQTL; 个体选择; 基因芯片实验设计; 作图方法

Integrating genetic and gene expression data: methods and applications of eQTL mapping

LIU Gang, PENG Hui-Ru, NI Zhong-Fu, QIN Dan-Dan, SONG Fang-Wei, SONG Guang-Shu, SUN Qi-Xin

Department of Plant Genetics & Breeding and State Key Laboratory for Agrobiotechnology, Key Laboratory of Crop Heterosis and Utilization (MOE), Key Laboratory of Crop Genomics and Genetic Improvement (MOA) and Beijing Key Laboratory of Crop Genetic Improvement, China Agricultural University, Beijing 100193, China

Abstract: The availability of high-throughput genotyping technologies and microarray assays has allowed researchers to investigate genetic variations that influence levels of gene expression. Expression Quantitative Trait Locus (eQTL) mapping methods have been used to identify the genetic basis of gene expression. Similar to traditional QTL studies, the main goal of eQTL is to identify the genomic locations to which the expression traits are linked. Although microarrays provide the expression data of thousands of transcripts, standard QTL mapping methods, which are able to handle at most tens of traits, cannot be applied directly. As a result, it is necessary to consider the statistical principles involved in the design and analysis of these experiments. In this paper, we reviewed individual selection, experimental design of microarray, normalization of gene expression data, mapping methods, and explaining of results and proposed potential methodological problems for such analyses. Finally, we discussed the applications of this integrative genomic approach to estimate heritability of transcripts, identify candidate genes, construct gene networks, and understand interactions between genes, genes and environments.

Keywords: eQTL; individual selection; experimental design of microarray; mapping methods

收稿日期: 2007-12-17; 修回日期: 2008-03-24

基金项目: 国家重点基础研究发展规划(973 计划)项目(编号: 2007CB109003)资助[Supported by the National Basic Research Program of China (973 Program) (No. 2007CB109003)]

作者简介: 刘刚(1979-), 男, 湖北仙桃人, 硕士, 专业方向: 小麦遗传育种。E-mail: llyf923@163.com

通讯作者: 彭惠茹(1968-), 女, 内蒙古赤峰人, 博士, 副教授, 硕士生导师, 研究方向: 小麦遗传育种。E-mail: penghuihu@cau.edu.cn

20 世纪 80 年代以来, 分子标记技术的飞速发展, 使得性状与 DNA 序列信息能够联系起来, 结合遗传作图统计方法, 可以在全基因组水平上定位复杂性状的 QTL, 为人们认识复杂性状的遗传基础、分离克隆基因和辅助选择育种均提供了很大帮助。但是到目前为止, 用该方法克隆数量性状基因仍然非常困难, 同时也忽略了基因表达水平的差异对性状的影响。而基因芯片技术的诞生, 使得从全基因组水平上进行高通量的基因表达谱分析成为可能, 通过比较两个或多个处理间表达谱的差异获得了大量差异表达基因, 这些差异表达基因涉及了各种代谢途径并最终控制复杂性状的建成。因此, 如何进一步建立性状表型观测值、DNA 序列与基因表达谱之间的联系, 整合遗传作图信息和基因表达数据, 对于我们深入认识高等生物许多重要而复杂的性状(如作物的产量、人类的疾病等)的分子机理和调控网络至关重要(图 1)^[11]。

遗传和表达数据的整合通常有两种策略: 一是结合性状 QTL 定位和亲本间差异表达基因的信息, 筛选定位在 QTL 区域内控制性状的候选基因^[2]; 二是 Janson 和 Nap^[3]提出的表达 QTL (expression QTL, eQTL) 作图方法, 即将来自分离群体的各基因型的表达数据作为一个数量性状, 利用传统的 QTL 分析方法进行分析。目前, eQTL 作图方法不仅成功地应用到酵母^[4,5]、鼠类^[6~11]、人^[12~13]的相关研究中, 而且应用到了植物, 如拟南芥^[14,15]、玉米^[6,16]、小麦^[17]、大麦^[18]等物种的相关研究中, 成为国际上一个新的研究热点。本文主要从个体选择、基因芯片实验设计、基因表达数据的获得与标准化、eQTL 作图方法及结果的分析方法等方面对 eQTL 作图进行了阐述。

1 eQTL 的研究方法

eQTL 定位与传统的 QTL (causative QTL, cQTL)^[19]一样, 一般经过构建作图群体和遗传连锁图谱(若分子标记连锁图谱未知)、测量数量性状(基因表达谱分析)、QTL 作图等步骤, 但由于 eQTL 定位中的性状为分离群体各基因型的基因表达谱, 所以二者在研究方法和对结果的生物学意义分析上有很多不同, 其具体分析过程见图 2, 以下重点介绍 eQTL 作图与传统 QTL 在研究方法上的不同之处。

1.1 个体选择

传统 QTL 作图所用的遗传群体一般较大(尤其是在对植物的研究中), 然而, 在 eQTL 定位中由于基因表达谱分析的成本高昂, 不可能对作图群体的所有个体都进行表达谱的分析, 只能选择部分基因型进行分析, 以降低 eQTL 研究成本。这样, 就面临一个怎样选择个体才能达到 QTL 作图的最大效率的问题。

1.1.1 基于性状和标记的选择

Wang 和 Nettleton^[20]提出了基于性状和标记基因型选择的方法, 其目的是使所检测的基因的表达与性状表现相关联的功效达到最大。假设控制某一性状的主效 QTL 只具有两种基因型, 按该座位的基因型, 把群体中所有个体分为两类, 从每一类中分别选择表型值最大和最小的个体进行 eQTL 分析。例如, 在包含 200 个个体的分离群体中, 控制某一性状的主效 QTL 只有两种基因型(X 和 O), 把基因型为 X 的个体分为一类, 从这一类中按该性状表型值, 选择最大值和最小值各 10 个个体。同样, 把该

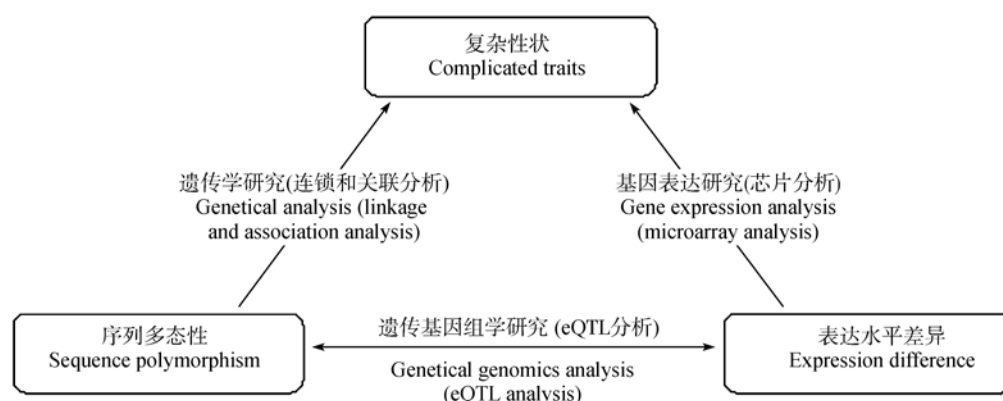


图 1 性状、DNA 序列和基因表达信息的整合

Fig.1 The integration of complex traits, sequence polymorphisms and variations of transcript levels

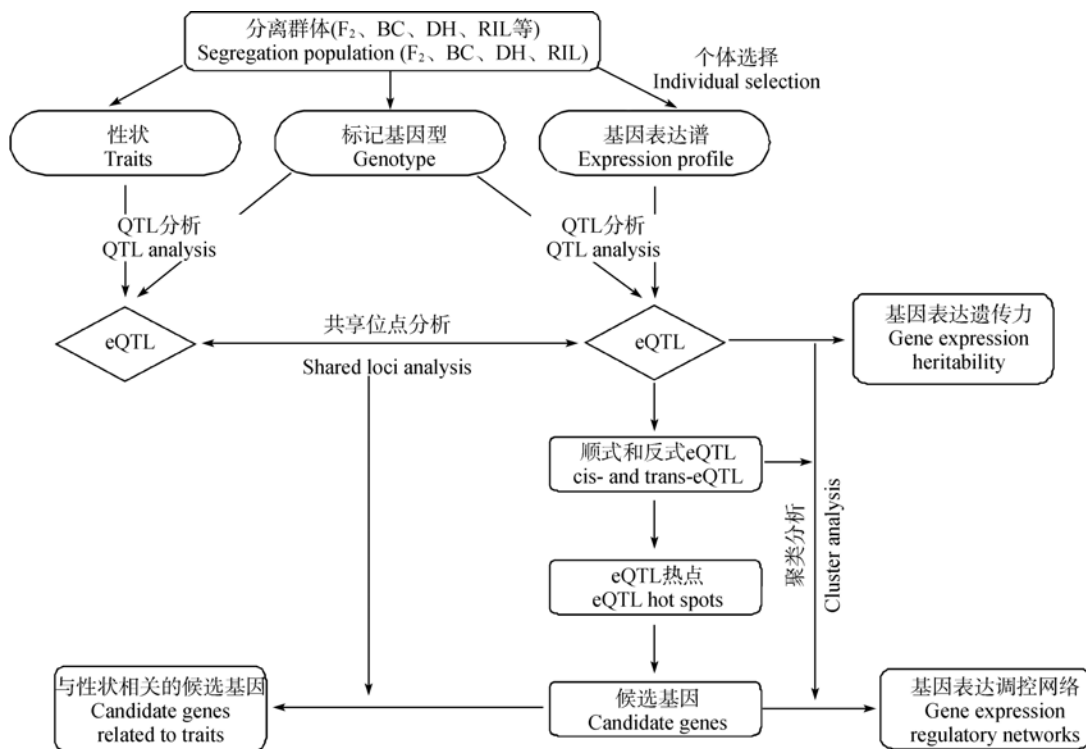


图 2 遗传与基因表达数据的整合路径

Fig. 2 Flow chart of the integrative strategy of genetic and gene expression data

座位基因型为O的个体分为一类,从这一类中按该性状表型值,选择最大值和最小值各10个个体。这样,我们就从这200个个体的群体中选择到包含40个个体的亚群体去进行基因表达谱分析。这是一种揭示特定数量性状遗传结构的理想方法,它不仅可以用来直接检测与目标性状有关的基因,还能用来阐明导致该性状变异的遗传路径中的主要调控因子。但是这种选择方法也有一定的局限性,它只适用于DH和RIL等异质纯合的群体(任何QTL座位只可能有两种基因型),对于利用像F₂这样的群体进行eQTL作图的研究中,每一座位存在3种基因型,该选择方法则不适用^[21]。

1.1.2 基于标记基因型遗传差异的选择

Jin等^[22]提出了一种在标记基因型基础上增加亚群体的遗传差异的基因型选择的方法。通过比较作图群体所有个体的标记基因型,并且采用最小阶矩偏差(Minimum moment aberration, MMA)设计来选择个体进行表达谱分析。这一过程可以对全基因组的标记基因型进行选择,也可以对某一特定区域的标记基因型进行选择。亚群体的样本容量和相似性测试考虑的标记数决定所要选择的个体。Jin等给出了计算个体间相似性的公式(1)(δ_{ij} 指两个体间相

同基因型的标记位点数, $P = n(n-1)/2$ 指个体配对的总数, n 指所选择的个体数)。为了便于比较不同大小的亚群体的相似值, Jin等通过公式(2)把 K 标准化(M 指 K 的最大可能值; R 是 K 的最大与最小可能值的差值)。从这两个公式我们可以知道, K 值越小, 则 S 值越大, 表明选择个体间的遗传差异越大。

$$K = \frac{1}{P} \sum_{i=1}^{n-1} \sum_{j=i+1}^n \delta_{ij} \quad (1)$$

$$S = \sqrt{n} \frac{(M-K)}{R} \quad (2)$$

这种基于MMA标准的选择方法简单易行,但是它依赖于无缺失的数据。对于缺失基因型, Jin等也提出可利用Haldane作图函数和两侧标记信息添上缺失数据。该选择方法还有另外一个局限性,它倾向于选择不同等位基因纯合的个体,所以,增加遗传差异只能提高检测加性效应的能力,却很难估计非加性效应^[22,23]。

1.1.3 基于实验目的的选择

最好的个体选择标准可能是针对目标座位选择符合实验目的的等位基因组合的个体。在这种情况下, Bueno等^[24]提出基于实验目的的选择方法。例如,利用一个F₂群体去估计反式eQTL的显性效应,则选

择的亚群体的A、H和B的基因型比例应该为1:2:1,如果要同时估计加性和显性效应,则选择的亚群体的A、H和B的基因型最优比例为0.293:0.414:0.293。对于多位上位性效应的估计则通过选择不同个体配对的基因芯片实验设计,得到想要的等位基因组合来估计这些遗传效应。

以上的个体选择方法都只注重对eQTL效应的检测,而忽略了对eQTL位置的检测精度。为提高eQTL位置的检测精度,Leon和Rosa^[25]提出了增加亚群体中重组事件的选择标准。模拟实验也表明,基于重组率的选择标准能充分提高估计QTL位置的精度。Jannink^[26]和Xu等^[27]提出了一个更为全面的基于重组率的选择方法,他们不仅考虑增加亚群体中的重组事件,也考虑了重组在全基因组中的平均分布。他们的方法能显著提高QTL位置估计的精度,当遗传图谱的距离较短时尤其如此。

1.2 基因表达分析的方法

进行基因表达分析的方法有多种,如差异显示反转录PCR技术(Differential display of reverse transcriptional PCR, DDRT-PCR)、代表性差异分析技术(Representational difference analysis, RDA)、抑制消减杂交技术(Suppression subtractive hybridization, SSH)、基因表达系列分析(Serial analysis of gene expression, SAGE)、限制性显示技术(Restriction Display PCR, RD-PCR)、cDNA-AFLP、实时定量PCR(Quantitative real-time PCR, qRT-PCR), Northern blotting 和基因芯片技术(Microarray)等,其中,cDNA-AFLP^[14]、qRT-PCR^[28]和基因芯片技术^[4-18]已被应用于eQTL的研究。但基因芯片技术以其高通量、高灵敏度的优点,成为eQTL研究的主要方法。

1.2.1 基因芯片实验设计

以往利用基因芯片技术进行大规模的基因表达水平的检测,主要是针对不同基因型、组织、发育阶段、分化阶段、病变、胁迫的表达特性进行综合分析。而在eQTL的研究中,将每个基因的表达水平作为数量性状,利用QTL作图方法来研究基因组和转录组之间的关系。因此eQTL研究除了考虑QTL分析应注意的问题——群体的遗传结构外,还要考虑转录谱研究中应注意的问题:什么样的基因芯片平台最适合特定的实验要求?多少张芯片能达到期望的实验效果?什么样的实验设计能更好地实现实验目的?在eQTL的研究中应用较多的是双色荧光

芯片,因为利用它能进行较为灵活的实验设计^[29]。基于双色荧光芯片,一些学者提出了适用于eQTL研究的基因表达分析的实验设计方法。

Piepho^[30]在利用基因芯片技术进行杂种优势机理的研究中,提出了一种能够有效地估计显性效应的双色荧光芯片实验设计方法,即将纯合和杂合的两个基因型在同一张芯片上共杂交来估计显性效应。在eQTL的研究中,F₂群体存在3种基因型A、H和B,因此在进行配对共杂交时,采取A和H、B和H的配对法更有利于估计显性效应。

Fu和Jansen^[29]提出了一种适用于RIL和DH群体来估计eQTL的加性效应的实验设计方法——远缘配对设计(Distant pair design),是指把全基因组或基因组某一特定区域内遗传差异较大的两种基因型进行配对与芯片杂交。该设计方法的基本原则是尽量使芯片内的两样本的遗传差异比芯片间大,且在芯片数目一定的情况下,增加生物学重复要比增加技术重复更有效^[23]。

以上两种设计方法都只侧重显性或加性效应的估计,而我们往往要同时估计加性、显性效应及互作效应。针对这一问题,Bueno等^[24]提出了相应的解决方法,还专门讨论了针对不同实验目的的基因芯片设计的优化方法。

1.2.2 基因表达数据的获得及校正

提取样品的RNA经反转录成荧光素标记的cDNA,在液相中与基因芯片上的探针杂交,洗膜后用图像扫描仪捕获芯片上的荧光信号,由此获得的图像就是基因芯片的原始数据。此后还需用图像分析软件从中提取各点的吸光度值、面积和吸光度比等数据,并转化成基因表达矩阵。为避免芯片实验中的系统差异引起的芯片间数据比较的困难,首先需将原始数据进行标准化(Normalization)^[31]。常用的是“看家基因”法:预先选择一组表达水平不变的看家基因,计算出这组基因平均ratio值为1时的标准化系数,然后用其对全部的数据进行标准化^[32]。此外,整体平均值法和信号强度依赖法也常用于芯片数据的标准化。原始数据经标准化后,就可以用于统计分析了。在eQTL研究中,是将每个基因在分离群体中的表达水平作为一个数量性状(e-trait)来分析。

1.3 eQTL 作图方法

在eQTL定位中,常需要对成千上万个表达性状

同时进行分析,但是发展多性状QTL定位模型几乎不可能,因为太多的性状使得模型过于复杂。以下两种eQTL作图策略可以简化模型:(1)转录本为基础的作图方法:独立地分析每一个表达性状(e-trait),最后获得全部表达性状的eQTL;(2)标记为基础的作图方法:先对每个标记基因型进行分组,然后检测该标记的不同基因型间表达量的差异,判断它与eQTL是否连锁,最后完成对全基因组标记的扫描^[33]。前一种方法只需了解传统的单个性状的定位方法即可。单标记回归分析(Single marker regression analysis)^[6]、区间作图(Interval mapping, IM)^[34]、复合区间作图(Composite interval mapping, CIM)^[35,36]、多区间作图(Multiple interval mapping, MIM)^[37]等方法在这里均适用,只是需用它们对成百上千个表达性状进行逐一分析。为减少工作量,实际应用中可以依照一定标准对表达性状进行选择(如前文所述),最终只对选中的目标性状进行定位。第二种方法的基础是差异基因表达分析的方法,通常的差异表达分析包含两个样本——对照和处理,而在表达-标记关联分析中,需将携带某一标记的一种基因型作为对照与该标记的其他基因型(处理)进行表达谱的多重比较。如F₂群体,有3种基因型,需进行两两差异表达分析:一种产生于纯合子之间,另一种产生于纯合子(两个)与杂合子之间。

eQTL定位方法与过去的QTL定位方法的主要不同在于需要同时对大量的“表达性状”(即单个基因的表达水平)进行QTL定位分析。因此不仅要在基因组位置上进行多重测验,而且在不同性状上也要进行多重测验,在统计学上就面临着I型错误的概率增高的问题。公认的控制多重比较带来的高假阳性率的方法就是控制假阳性率(False discovery rate, FDR)。目前统计上常用的方法有:(1)利用Bonferroni方法等对两两 t 测验的 P 值进行校正,将调整后的 P 值与FDR比较,确定是否存在显著差异^[38]或用更严格的 q 值代替 P 值来控制假阳性率^[39];(2)通过重复排列和重置抽样(Permutation)对 P 值进行纠正^[15,40];(3)通过eQTL定位中LOD峰值对应的 P 值来控制基因组位置上的假阳性发生率,同样可用重置抽样(Permutation)对 P 值进行纠正^[40],因为这里只考虑一个LOD峰值,所以当在一个转录本定位在两个以上位置时,该方法会得出错误结论。近几年,学者们提出了一些更加完善的统计方法来解决eQTL定位中存在的问题,如经验Bayesian评判法^[41]以及分级混合

法^[42],对大量表达性状的自动化分析的eQTL策略^[43],将针对转录本的分析 and 针对标记的分析结合在一起的分析混合模型(Mixture over markers, MOM)方法^[44],以及基于标记基因型、表达量以及标记基因型与表达量互作线性模型的贝叶斯压缩估计方法^[45]等等。

1.4 eQTL 定位结果分析

定位得到的差异显著 eQTL 可以分为顺式 eQTL(Cis-acting eQTL)和反式 eQTL(Trans-acting eQTL)两种,顺式 eQTL 指的是某个基因的 eQTL 定位到该基因所在的基因组区域,表明可能是该基因本身的多态性引起的表达水平的差异;反式 eQTL 指 eQTL 座位的某一基因的多态性导致与该 eQTL 座位不在同一基因组区域的基因表达水平的差异^[23]。顺式eQTL 能直接提供候选基因的信息,而反式eQTL的信息不仅能与其他方法结合获得调控网络,还有利于减少网络的候选节点。另外,一旦通过定位获得成千上万个eQTL,人们首先感兴趣的是那些在染色体上成簇分布的区域,这些区域称为eQTL热点(Hot spots)^[46],确定eQTL热点最简单的方法是直接计算定位的转录本数目。但有些区域可能是“ghost”热点,应该进行统计学检验,目前对转录本数目进行检验的方法主要有:根据 n 个转录本是否定位在热点区域的概率进行泊松测验,或对eQTL进行全基因组随机排列,得到每个区域eQTL数目的极值分布,据此计算显著的eQTL热点区域应满足的数目。另外,Kendzioriski等^[46]通过定位偏好(用LOD值评价)来确定显著的eQTL热点区域。对于eQTL热点候选基因的验证可采用芯片的基因集合丰度分析(Gene set enrichment analysis, GSEA),对热点区域候选基因进行功能分类,再将相关生物学功能基因的比例与期望值进行卡方测验,但一般置信度较低^[47]。eQTL热点为我们提供了候选调控基因信息,根据常用的基因网络分析方法,可进一步明确候选基因间的交互作用。另外,考虑到每个基因在网络上的表达变化不同,贡献大小可能不一致,参与的途径可能不止一条等问题,因此,拓扑分析(Topological analysis of pathway phenotype association, TAPPA)^[48]和影响分析(Impact analysis, IA)^[49]等统计分析方法也相继用于构建基因的调控网络。另外,利用相关分析对基因表达性状QTL和表型性状QTL定位结果进行综合分析,并结合结构方程模型^[50]、系统动态模型^[51]、贝叶斯网络模型^[52]等统计

方法,不仅能挖掘与性状相关的候选基因,还能系统描述复杂性状形成的分子机理和调控网络。

2 eQTL 的应用

虽然 eQTL 这一研究方法提出才短短几年时间,但其研究进展迅速。研究者用该方法在估计基因表达遗传率、挖掘候选基因、构建基因调控网络等许多方面开展了大量工作,使得对复杂性状的分子机理及调控网络的认识更加深入。

2.1 估计基因表达遗传率

Cheung等^[53]对人的淋巴母细胞进行基因表达谱分析,结果表明基因的转录水平是可遗传的。利用传统遗传学的广义遗传率公式可估计基因表达的遗传率。Monks等^[12]对人的淋巴细胞系中基因表达的研究发现,31%的差异表达基因是可遗传的,遗传率均值为34%。酵母的研究表明,3546个基因表达的遗传率大于69%^[43]。Chesler等^[9]通过对小鼠的研究发现,有608个基因表达的遗传率超过33%(其中最高可达78%),平均11%。Vuylsteke等^[4]对拟南芥的研究发现,有792个差异表达基因是可遗传的,其遗传力最高为93%,平均为30%。

2.2 确定 eQTL 热点和筛选候选基因

Morley等^[13]用Affymetrix芯片对来自于14个法语人类多态性研究中心(CEPH)尤他州家族的成员的免疫B细胞中的3554个基因进行表达谱的分析,在第14和20条染色体上检测到两个eQTL热点。

Brem等^[4]采用寡核苷酸微阵列,鉴定由实验酵母菌株和野生酵母菌株杂交产生的40个单倍体分离后代的基因型,构建了一个包含3312个分子标记的图谱,覆盖基因组的99%。同时采用包含6215个可读框的cDNA微阵列,对亲本和分离后代进行表达谱分析,进行eQTL定位,共发现8个eQTL聚集的区域(eQTL热点),分别有7~94个功能相关的基因定位在这些区域。并且结合已有的研究结果推测,定位在6个eQTL热点中的部分基因分别被基因*CST13*、*LEU2*、*MAT*、*URA3*、*HAP1*、*SIR3*调控。Yvert等^[5]采用相同的酵母菌株进行杂交,得到86个单倍体分离后代群体,并以此为材料,运用相同的实验和分析方法,不仅证实了Brem等^[4]的结果,而且还检测到13个调控1265个基因表达的eQTL热点,许多功能相似的基因定位在同一eQTL上。同时,他们通过图位克隆和功能鉴定发现,基因

*GPA1*和*AMN1*的多态性可能影响了与信息素应答和细胞分裂有关基因的表达水平。

Schadt等^[6]利用寡核苷酸芯片,对小鼠标准自交系C57BL/6J和DBA/2J的111个F₂个体的肝组织进行分析,检测23574个基因的表达情况,并采用区间作图法进行eQTL定位,发现在第2、6、7、9、10、16和17号染色体上存在eQTL热点区域,并找到了一个与小鼠肥胖有关的候选基因*Mup1*。Ghazalpour等^[54]根据Schadt等^[6]的数据,鉴定了13个与肥胖相关的基因调控途径,发现位于第3、6、16、19号染色体上的4个热点区域参与这些途径的调控,并发现了可能参与这些途径的新基因。Bystrykh等^[7]利用来源于DBA/2(D2)和C57BL/6(B6)两个小鼠近交系杂交的30个重组近交系,分离出造血干细胞,进行表达谱分析,在*Scp2*区间确定了8个可能影响造血干细胞增殖的候选基因。Chesler等^[9]采用相同的小鼠重组自交系群体的前脑组织,在第1、2、6、10、11、14和19染色体上检测到7个eQTL热点。

在植物中,Shi等^[16]用一个包含有40个株系的玉米重组近交系群体进行eQTL分析,在玉米染色体bin1.07、1.12、3.05、8.03和9.04上检测到5个eQTL热点,并找到102个与细胞壁降解能力有关的候选基因。Jordan等^[17]从来源于RL4452和‘AC Domain’的小麦DH群体的118个株系中选出41个用Affymetrix芯片进行表达谱分析,在染色体1D和4B上检测到2个eQTL热点。Potokina等^[18]用139个重组近交系的大麦群体对16000个基因进行eQTL分析,确定了14个eQTL热点。

2.3 构建调控网络

生物体中的基因往往不是单独起作用的,而是一组调控因子调控一系列基因的表达,从而形成了网络状的相互作用关系^[55]。虽然芯片数据也能提供某一通路中相关基因的信息,但是不能鉴别不同通路的主调控基因,也不能预测由于某一系统混乱所导致的结果。eQTL分析则有利于对生物系统基因间互作的更高层次的理解,Janson和Nap^[3]首次把eQTL热点与基因调控网络的构建联系起来,eQTL热点的检测为构建基因调控网络提供了线索。

基于贝叶斯网络分析方法,Zhu等^[56]提出了一个结合分离群体的相对表达水平和基因型数据来构建基因调控网络的方法。他们用这种方法对Schadt

等^[6]的eQTL结果进行分析,构建了多个基因调控网络。Schadt等^[57]利用相同的小鼠群体进行基因表达与肥胖性状之间的因果关系研究,构建了与肥胖相关的基因调控网络。Li等^[52]提出了一个结合eQTL作图、SNP分析和贝叶斯网络分析的计算方法来构建基因调控网络,得到了66个候选调控网络。

eQTL研究不仅有助于我们估计基因表达水平的遗传率、挖掘候选基因和构建基因调控网络,而且为我们正确理解基因与基因及基因与环境间的互作提供了条件,从而使我们能更深入认识生物复杂性状的分子机理和遗传基础。

3 eQTL研究的局限性

由于基因表达谱分析的成本高昂,一般不会对整个分离群体进行基因表达谱分析,而只是选择部分个体进行分析。但是,这种人工抽样可能会导致偏分离,同时样本容量小使得实际分析的群体并不能代表理论群体,不利于QTL定位。目前的QTL作图方法只能定位到染色体区域某一置信区间内,这一区间可能存在很多个基因,不利于对eQTL定位结果的分析。另外,基因表达谱分析过程中出现的生物学误差和系统误差都可能限制了eQTL研究方法发挥更大作用。

4 结语

eQTL的研究方法首次在全基因组水平上建立了基因组与转录组间较为精确的关系并引入表达性状(expression trait, e-trait)的概念,扩展了传统的表现型的含义。即利用eQTL作图的方法,以全新的视角解析基因与性状的关系:综合考虑多种基因变异,多种表型、多种环境和表观遗传变异,在生物系统的不同层面之间建立相互联系。目前eQTL研究已经在候选基因的挖掘、基因调控网络的构建等方面取得了较好的进展。我们相信,在不久的将来,随着QTL作图方法、基因组学、转录组学、蛋白质组学和代谢组学的不断发展完善,eQTL研究将在整合不同层次信息,正确理解基因与基因、基因与环境互作以及复杂性状形成的分子机理等方面发挥越来越重要的作用。

参考文献(References):

- [1] Li J, Burmeister M. Genetical genomics: combining genetics with gene expression analysis. *Hum Mol Genet*, 2005, 14(S1): R163–R169. [\[DOI\]](#)
- [2] Yagil C, Hubner N, Monti J, Schulz H, Sapojnikov M, Luft FC, Ganten D, Yagil Y. Identification of hypertension related genes through an integrated genomic transcriptional approach. *Circ Res*, 2005, 96(6): 617–625. [\[DOI\]](#)
- [3] Jansen RC, Nap JP. Genetical genomics: the added value from segregation. *Trends Genet*, 2001, 17(7): 388–391. [\[DOI\]](#)
- [4] Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 2002, 296(5568): 752–755. [\[DOI\]](#)
- [5] Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L. Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat Genet*, 2003, 35(1): 57–64. [\[DOI\]](#)
- [6] Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colnayo V, Ruff TG, Milligan SB, Lamb JR, Cavet Guy, Linsley PS, Mao M, Stoughton RB, Friend SH. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 2003, 422(6929): 297–302. [\[DOI\]](#)
- [7] Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T, Su AI, Vellenga E, Wang J, Manly KF, Lu L, Chesler EJ, Alberts R, Jansen RC, Williams RW, Cooke MP, Haan G. Uncovering regulatory pathways that affect hematopoietic stem cell function using ‘genetical genomics’. *Nat Genet*, 2005, 37(3): 225–232. [\[DOI\]](#)
- [8] Hubner N, Wallace C, Zimdahl H, Petretto E, Schulz H, Maciver F, Mueller M, Hummel O, Monti J, Zidek V, Musilova A, Kren V, Causton H, Game L, Born G, Schmidt S, Müller A, Cook SA, Kurtz TW, Whittaker J, Pravenec M, Aitman TJ. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat Genet*, 2005, 37(3): 243–253. [\[DOI\]](#)
- [9] Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA, Threadgill DW, Manly KF, Williams RW. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat Genet*, 2005, 37(3): 233–242. [\[DOI\]](#)
- [10] Vazquez-Chona FR, Khan AN, Chan CK, Moore AN, Dash PK, Hernandez MR, Lu L, Chesler EJ, Manly KF, Williams RW, Geisert EE. Genetic networks controlling retinal injury. *Mol Vision*, 2005, 11: 958–970.
- [11] Lum PY, Chen Y, Zhu J, Lamb J, Melmed S, Wang S, Drake TA, Lusis AJ, Schadt EE. Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. *J Neurochem*, 2006, 97(Suppl. 1): 50–62. [\[DOI\]](#)
- [12] Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, Edwards S, Phillips JW, Sachs A, Schadt EE. Genetic inheritance of gene expression in human cell lines. *Am J Hum Genet*, 2004, 75(6): 1094–1105. [\[DOI\]](#)
- [13] Morley M, Molony CM, Weber TM, Davlin JL, Ewens KG, Spielman RS, Cheung VG. Genetic analysis of genome-wide variation in human gene expression. *Nature*, 2004,

- 430(7001): 743–747.[\[DOI\]](#)
- [14] Vuylsteke M, Daele HVD, Vercauteren A, Zabeau M, Kuiper M. Genetic dissection of transcriptional regulation by cDNA-AFLP. *Plant J*, 2006, 45(3): 439–446.[\[DOI\]](#)
- [15] West MAL, Kim K, Kliebenstein DJ, Leeuwen HV, Michelmore RW, Doerge RW, Clair DAS. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. *Genetics*, 2007, 175(3): 1441–1450.[\[DOI\]](#)
- [16] Shi C, Uzarowska A, Ouzunova M, Landbeck M, Wenzel G, Lübberstedt T. Identification of candidate genes associated with cell wall digestibility and eQTL (expression quantitative trait loci) analysis in a Flint × Flint maize recombinant inbred line population. *BMC Genomics*, 2007, 8: 22.[\[DOI\]](#)
- [17] Jordan MC, Somers DJ, Banks TW. Identifying regions of the wheat genome controlling seed development by mapping expression quantitative trait loci. *Plant Biotechnol J*, 2007(5): 442–453.[\[DOI\]](#)
- [18] Potokina E, Druka A, Luo ZW, Wise R, Waugh R, Mike K. Gene expression quantitative trait locus analysis of 16000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. *Plant J*, 2008, 53(1): 90–101.
- [19] Drake TA, Schadt EE, Lusis AJ. Integrating genetic and gene expression data: application to cardiovascular and metabolic traits in mice. *Mamm Genome*, 2006, 17(6): 466–479.[\[DOI\]](#)
- [20] Wang D, Nettleton D. Identifying genes associated with a quantitative trait or quantitative trait locus via selective transcriptional profiling. *Biometrics*, 2006, 62(2): 504–514.[\[DOI\]](#)
- [21] Nettleton D, Wang D. Selective transcriptional profiling for trait-based eQTL mapping. *Anim Genet*, 2006, 37(1): 13–17.[\[DOI\]](#)
- [22] Jin CF, Lan H, Attie AD, Churchill GA, Bulutuglo D, Yandell BS. Selective phenotyping for increased efficiency in genetic mapping studies. *Genetics*, 2004, 168(4): 2285–2293.[\[DOI\]](#)
- [23] Rosa GJM, Leon N, Rosa AJM. Review of microarray experimental design strategies for genetical genomics studies. *Physiol Genomics*, 2006, 28 (1): 15–23.[\[DOI\]](#)
- [24] Bueno JSD, Gilmour SG, Rosa GJM. Design of microarray experiments for genetical genomics studies. *Genetics*, 2006, 174(2): 945–957.[\[DOI\]](#)
- [25] Leon N, Rosa GJM. Optimization of selective phenotyping for QTL mapping. *Plant and Animal Genome XIII Meetings*, 2005, 860, 283.
- [26] Jannink JL. Selective phenotyping to accurately map quantitative trait loci. *Crop Sci*, 2005, 45(3): 901–908.[\[DOI\]](#)
- [27] Xu Z, Zou F, Vision TJ. Improving QTL mapping resolution in experimental crosses by the use of genotypically selected samples. *Genetics*, 2005, 170(1): 401–408.[\[DOI\]](#)
- [28] Yamashita S, Wakazono K, Nomoto T, Tsujino Y, Kuramoto T, Ushijima T. Expression quantitative trait loci analysis of 13 genes in the rat prostate. *Genetics*, 2005, 171(3): 1231–1238.[\[DOI\]](#)
- [29] Fu J, Jansen RC. Optimal design and analysis of genetic studies on gene expression. *Genetics*, 2006, 172(3): 1993–1999.[\[DOI\]](#)
- [30] Piepho HP. Optimal allocation in designs for assessing heterosis from cDNA gene expression data. *Genetics*, 2005, 171(1): 359–364.[\[DOI\]](#)
- [31] LI Yao. Microarray and Functional Genomics. Beijing: Chemical Industry Press, 2004, 77–100.
李瑶. 基因芯片与功能基因组. 北京: 化学工业出版社, 2004, 77–100.
- [32] Bilban M, Buehler LK, Head S, Desoye G, Quaranta V. Normalizing DNA microarray data. *Curr Issues Mol Biol*, 2002, 4(2): 57–64.
- [33] Jia ZY, Xu SZ. Mapping quantitative trait loci for expression abundance. *Genetics*, 2007, 176(1): 611–623. [\[DOI\]](#)
- [34] Land ES, Botstein D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 1989, 121(1): 185–199.
- [35] Zeng ZB. Theoretical basis for separation of multiple linked gene effects in mapping of quantitative trait loci. *Proc Natl Acad Sci USA*, 1993, 90(23): 10972–10976.[\[DOI\]](#)
- [36] Zeng ZB. Precision mapping of quantitative trait loci. *Genetics*, 1994, 136(4): 1457–1468
- [37] Kao CH, Zeng ZB, Teasdale RD. Multiple interval mapping for quantitative trait loci. *Genetics*, 1999, 152(3): 1203–1216.
- [38] Dudoit S, Yang YH, Speed TP, Callow MJ. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 2002, 12(1): 111–139.
- [39] Storey JD, Tibshirani R. Statistical significance for genome-wide studies. *Proc Natl Acad Sci USA*, 2003, 100(16): 9440–9445.[\[DOI\]](#)
- [40] Brem RB, Kruglyak L. The landscape of genetic complexity across 5700 gene expression traits in yeast. *Proc Natl Acad Sci USA*, 2005, 102(5): 1572–1577.[\[DOI\]](#)
- [41] 1235endziorski CM, Newton MA, Lan H, Gould MN. On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statist Med*, 2003, 22(24): 3899–3914.[\[DOI\]](#)
- [42] Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 2004, 5(2): 155–176.[\[DOI\]](#)
- [43] Carlborg O, Koning DJD, Manly KF, Chesler E, Williams RW, Haley CS. Methodological aspects of the genetic dissection of gene expression. *Bioinformatics*, 2005, 21(10): 2383–2393.[\[DOI\]](#)
- [44] Kendziorski C, Chen M, Yuan M, Lan H, Attie A. Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics*, 2006, 62(1): 19–27.[\[DOI\]](#)

- [45] Hoti F, Sillanpää MJ. Bayesian mapping of genotype \times expression interaction in quantitative and qualitative traits. *Heredity*, 2006, 97(1): 4–18. [\[DOI\]](#)
- [46] Kendziora C, Wang P. A review of statistical methods for expression quantitative trait loci mapping. *Mamm Genome*, 2006, 17(6): 509–517. [\[DOI\]](#)
- [47] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*, 2005, 102(43): 15545–15550. [\[DOI\]](#)
- [48] Gao SG, Wang XJ. TAPPA: topological analysis of pathway phenotype association. *Bioinformatics*, 2007, 23(22): 3100–3102. [\[DOI\]](#)
- [49] Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, Georgescu C, Romero R. A systems biology approach for pathway level analysis. *Genome Res*, 2007, 17(10): 1537–1545. [\[DOI\]](#)
- [50] Tomarken AJ, Waller NG. Structural equation modeling: strengths, limitations, and misconceptions. *Annu Rev Clin Psychol*, 2005, 1: 31–65. [\[DOI\]](#)
- [51] Baldwin NE, Chesler EJ, Kirov S, Langston MA, Snoddy JR, Williams RW, Zhang B. Computational, integrative, and comparative methods for the elucidation of genetic coexpression networks. *J Biomed Biotechnol*, 2005(2): 172–180. [\[DOI\]](#)
- [52] Li H, Lu L, Manly KF, Chesler EJ, Bao L, Wang J, Zhou M, Williams RW, Cui Y. Inferring gene transcriptional modulatory relations: a genetical genomics approach. *Hum Mol Genet*, 2005, 14(9): 1119–1125. [\[DOI\]](#)
- [53] Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, Spielman RS. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat Genet*, 2003, 33(3): 422–425. [\[DOI\]](#)
- [54] Ghazalpour A, Doss S, Sheth SS, Drake LAI, Schadt EE, Lusis AJ, Drake TA. Genomic analysis of metabolic pathway gene expression in mice. *Genome Biol*, 2005, 6(7): R59. [\[DOI\]](#)
- [55] Wyrick JJ, Young RA. Deciphering gene expression regulatory networks. *Curr Opin Genet Dev*, 2002, 12(2): 130–136. [\[DOI\]](#)
- [56] Zhu J, Lum PY, Lamb J, GuhaThakurta D, Edwards SW, Thieringer R, Berger JP, Wu MS, Thompson J, Sachs AB, Schadt EE. An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet Genome Res*, 2004, 105(24): 363–374. [\[DOI\]](#)
- [57] Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, GuhaThakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang L, Castle J, Zhu H, Kash SF, Drake TA, Sachs A, Lusis AJ. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet*, 2005, 37(7): 710–717. [\[DOI\]](#)

欢迎订阅 2009 年《林业科学》

《林业科学》是中国林学会主办、中国科协主管的林业综合性学术期刊，创刊于 1955 年。代表中国林业科学研究和林业科技期刊的最高水平。在三届“国家期刊奖”评选中，两次荣获中国期刊最高奖——“国家期刊奖”，一次名列“国家期刊奖提名奖”第一名。目前已被 CA、AJ、ZR、JST、CSCD 等国内外 20 多个重要检索系统与数据库收录。

主要刊登林业及相关领域的最新科研成果，及时反映国家林业建设重点和热点，评述学术动向，开展学术讨论，促进国内外学术交流。内容包括森林培育、森林生态、林木遗传育种、森林保护、森林经理、森林与环境、生物多样性保护、野生动植物保护与利用、园林植物与观赏园艺、经济林、水土保持与荒漠化治理、森林工程、木材科学与技术、林产化学加工工程、林业经济、林业可持续发展及林业宏观决策研究等方面。以学术论文、研究报告、综述为主，还设有学术问题讨论、研究简报、植物新品种与良种、科技动态、新书评介等栏目。读者对象为国内外从事林业及相关领域研究的科技人员、管理人员以及高等院校的师生。本刊接受英文稿件(附中文摘要)。

月刊，每月 25 日出版，大 16 开，每期 128 页。单价：25 元，全年：300 元。国内外公开发行。国内统一刊号：CN11-1908/S，邮发代号 82-6；国际刊号：ISSN1001-7488，发行代号：BM44。

可以直接向编辑部订阅。

联系地址：北京万寿山后中国林学会《林业科学》编辑部

邮编：100091 电话：(010)62889820 62888579

E-mail: linykx@forestry.ac.cn

期刊主页：http://lyke.chinajournal.net.cn

期刊博客：http://blog.sina.com.cn/linykx