多基因离散性状 QTL 连锁分析方法

殷宗俊[1,2],张 勤[2]

(1.安徽农业大学动物科技学院,合肥 230036;2.中国农业大学动物科技学院,北京 100094)

摘 要:动物中有许多重要的离散性状,与常规的数量性状类似,其遗传基础受多基因控制并受到环境因子的修饰。由于多基因离散性状的表型特殊性,利用常规的 QTL 连锁分析方法很难获得理想的统计效果,相应地发展了许多基于广义线性模型框架内的非线性方法。文章就目前离散性状的 QTL 连锁分析方法作简要综述,并对可预期的改进方法进行了展望。

关键词:离散性状;QTL;连锁分析;广义线性方法

中图分类号:Q348 **文献标识码**:A **文章编号**:0253－9772(2006)05－0578－05

Methodology of QTL's Linkage Analysis for Discrete Traits Controlled by Polygenes

YIN Zong-Jun[1,2], ZHANG Qin[2]

(1. *College of Animal Science and Technology, AnHui Agricultural University, Hefei* 230036, *China*;
2. *College of Animal Science and Technology, China Agricultural University, Beijing* 100094, *China*)

Abstract: Many characters of biological interest and economical importance that are not inherited in a simple Mendelian fashion vary in a discrete form. These traits are called complex discrete traits. A complex discrete trait is presumably controlled by several genes, and characteristic of the trait is modified by environmental effects. Although methods of mapping QTL for continuous quantitative characters have been well developed, such methods for discrete characters are not available yet. So many non-linear methods were developed under the framework of the generalized linear model. In this paper, we reviewed the recent progress and improvement of the methods for QTL mapping in this field.

Key words: discrete traits; QTL; linkage analysis; generalized linear methods

基因组研究的飞速发展,大量分子标记的涌现,使人们在分子水平上了解数量性状的遗传机理得以实现。当前分子遗传标记的主要用途是构建连锁图谱和进行基因检测,对于数量性状而言,主要是检测与定位影响数量性状的基因座(Quantitative Trait Loci, QTL),QTL 定位即是利用基因组中遗传标记的分离来估计与其连锁的 QTL 可能存在位置和效应大小,QTL 定位效率的高低主要取决于所采用的实验设计和统计方法是否恰当。近年来,一些用于 QTL 检测与定位的统计方法和实验设计相继提出[1],并得到了不断的改进完善,QTL 检测与定位的效率大大提高,定位的精度也在不断上升。但现有方法的提出多数是基于对连续数量性状的标记-QTL 连锁分析,在性状表型服从正态分布及方差不依赖于均值(方差固定)的假设前提下,现有的多种 QTL 定位方法都能给出参数的一致性估计[2,3]。而

收稿日期:2005－05－08;**修回日期**:2005－07－10

基金项目:国家重点基础研究发展规划(编号:G2000016103)和安徽省教育厅项目(编号:2002jq126,2004kj151)资助[Supported by the National Key Basic Research Program of China (No. G200001603) and Educational Committee of Anhui Province (No. 2002jq126,2004kj151)]

作者简介:殷宗俊(1967—),男,博士,副教授,研究方向:动物遗传育种。E-mail: yinzongjun@yahoo.com.cn

对于不服从正态分布的离散性状(discrete traits)来说,由于这类性状的表型特殊性,表型值与遗传信息的相关程度低[4,5],因此,采用线性方法对多基因控制的离散性状进行标记——QTL 连锁分析的功效不高,甚至会导出错误的结果。尽管如此,目前在动植物离散性状 QTL 研究方面主要还是采用一些常规的线性方法。相对于常规的数量性状,多基因控制的离散性状或阈性状的 QTL 定位就显得更具有挑战性。

1　多基因离散性状遗传模型

多基因控制的离散性状,也称阈性状(threshold traits),在多基因控制的理论假设下,其观察值(响应变量)是离散不连续的,可能是一种分布或多种分布的混合体。离散性状的表型为一个间断分布,同时在受多基因控制的假设前提下,又具有一个潜在的连续分布(liability),它表示造成性状遗传变异的基础,一般是正态的或经过变换后成为正态的连续分布。而离散的表型值(Y)与潜在的连续变量(z)是通过一系列固定的阈(υ)相联系。在性状与 QTL 存在相关的前提下,潜在值可通过下面的线性混合模型来描述,但是潜在的连续值与可观察到的表型值是非线性的。

$$z = x'\beta + g_{QTL} + g_{poly} + e$$

其中 $x'\beta$ 为对应个体的所有固定效应(包括群体的均值);g_{QTL} 为个体 QTL 基因型的效应;g_{poly} 为个体的随机多基因效应,e 为随机环境效应。性状的表型值由潜在变量中特定的阈值(υ)决定,高于或低于 υ 时会表现出不同的表型状态。对于 C 个多表型分类性状来说,有多个阈值($\upsilon_1, \upsilon_2, \cdots \upsilon_{c-1}$),且有 $\upsilon_1 < \upsilon_2 < \cdots \upsilon_{c-1}$,个体潜在变量的取值在不同的阈值区间内,表现出不同的表型等级状态。

$$\upsilon_{k-1} < z < \upsilon_k \Leftrightarrow Y = k, k = 1, 2 \cdots, c, \upsilon_0 = -\infty, \upsilon_c = +\infty$$

对于上面的分类性状,其表型呈现出二项分布或多项分布,线性分析的基本假定得不到满足,通过常规线性方法来构建全似然函数进行分析的功效也将受到影响,因此借助于阈模型或广义线性模型(GLM)来进行分析是较为合理的选择。

2　多基因离散性状 QTL 连锁分析方法

利用连锁图谱,根据经典遗传学中基因的重组交换原理,可以通过相关统计方法分析 DNA 标记变异和性状值变异二者间的关联,从而判断标记是否与影响性状的基因连锁,以及标记和 QTL 之间的遗传距离。连锁分析是基因定位的重要策略之一,其基本原理是:在家系中,位于同一条染色体上的两个座位(QTL 与遗传标记)在减数分裂的过程中会发生交换与重组。因此,由标记与 QTL 间的重组率可估算出两者间的距离及连锁程度。目前,对多基因离散性状的 QTL 连锁分析有两种完全不同的思想,这两种思想的主要区别在于是否具有潜在连续分布的假设,一种是不考虑离散数据的实际离散性,采用线性的方法进行标记-性状之间的关联分析,称之为线性分析思想;另一种思路就是阈值概念[4],认为在性状表型离散分布的背后存在着一个潜在的连续分布,潜在的连续变量受 QTL 与多基因效应的共同影响,在此基础上的 QTL 连锁分析方法称之为非线性分析思想。

2.1　QTL 连锁分析的线性方法

随着统计基因组学的迅猛发展,一些用于 QTL 检测与定位的统计方法相继提出,并得到了不断的改进完善。根据在统计分析中同时采用的标记数目的不同,连锁分析可以分为:(1)以单一标记为基础的零区间定位;(2)以多标记为基础的区间定位和复合区间定位[6]。区间定位利用染色体上所有可提供信息的标记,在标记连锁群内搜索 QTL 最可能存在的位置。相对于单标记分析,区间定位更大程度地利用了有效标记信息,提高了 QTL 的检测力和定位准确性。但是,当多个 QTL 连锁时,区间定位很难正确估计 QTL 位置或对各 QTL 进行独立的参数估计,因为连锁 QTL 的存在可能会导致出现"鬼影"QTL(ghost QTL)。在此基础上,ZENG[7] 提出了多 QTL 模型和复合区间定位方法,其中复合区间定位法由于直观性好、计算上易于程序化而被广泛应用,已逐步取代了区间定位法。Rebai 等[8] 的研究表明,在回交及 F_2 群体中,利用常规的线性回归方法可以有效的进行等级离散性状的 QTL 连锁分析,但是在统计效率方面则损失较大。

远交群体的结构相对复杂,主要由多个结构简单的小家系组成,也有包含多世代、具有复杂亲缘关系、缺失数据的复杂群体。针对各种复杂群体的情况,相应发展了许多适用于远交群体的 QTL 连锁分析方法。所有的这些线性方法大致可区分为两大类,一类是基于最小二乘的方法,另一类是基于最大

似然分析的方法。Whittaker 等[9]提出直接对标记型进行回归可以得到同样的信息(QTL 位置和效应),而不用对区间上各点进行分析。在最大似然分析中,QTL 等位基因效应或基因型效应也可以作为随机效应或固定效应进行处理。当 QTL 作为随机效应处理时,REML 方法[10]假定 QTL 等位基因效应服从正态分布,QTL 等位基因效应方差协方差矩阵为 QTL 的 IBD 概率与 QTL 等位基因效应方差之积,其中 IBD 概率可通过系谱和标记信息进行估计。在远交群体中,最大似然分析在利用各种亲属信息推断标记和 QTL 的概率上优于回归方法。

2.2 QTL 连锁分析的非线性方法

在性状表型服从正态分布和方差不依赖于均值的假设前提下,上面提及的各种 QTL 线性分析方法都能给出参数的一致性估计。而对于多基因控制的离散性状,由于其分布是非正态的,一般呈现出二项(binary)、多项(Multinomial)分布或者泊松(Poisson)分布。由于离散性状的表型特殊性,可观察的表型值与基因效应间不是线性的关系,表型值与遗传信息的相关程度相应较低,因此,采用常规的线性方法对离散性状进行遗传分析的功效不高[11]。阈值模型[4]的提出开辟了合理的离散性状遗传分析方法,阈模型在分析方法上主要是基于广义线性模型(Generalized Linear Model, GLM),通过连接函数(link function)将观察值与遗传效应相联系,从而使转化后的响应变量期望值线性化,实现参数的无偏估计[12]。

目前,用于离散性状 QTL 定位的非线性方法主要有 3 种:GLM 方法、Bayesian 方法和非参数方法,这些方法在人类及植物分类性状的 QTL 定位中已发挥了重要的作用,特别是对于近交系设计下的二项分类性状,并取得了可喜的成果[1,13],但这些非线性方法在动物分类性状的 QTL 定位中则少见报道。

2.2.1 广义线性方法

目前,广义线性方法在离散性状 QTL 定位中的应用大多集中于近交系杂交群体中单阈值二分类性状的研究方面。主要是由于二分类性状只有两种表型取值,在计算上较为方便。在广义线性模型的框架内,Hackett 和 Weller[14]利用阈模型首次对二分类性状的 QTL 参数进行了估计。在此基础出上,相继发展了许多有效的方法[8,15]。但多数方法主要还是基于近交系杂交设计,而在家畜群体 QTL 定位方面的应用相对较少。YI 等[2]提出了适应于多个大全同胞家系设计的固定效应模型(fixed-model),用来定位分析家畜全群的二分类性状 QTL。研究表明,当全同胞家系的数量较小时这种方法的统计效率较高,而随着全同胞家系数的上升和需要估计的参数增多,该方法的估计效率降低。在此基础上,相应发展了 QTL 随机模型(random model),随机模型中将 QTL 位点的每个等位基因都看作随机变量,通过对 QTL 位点方差的估计和检验来取代固定效应模型中的各种参数估计。利用阈模型,Kadarmideen 等[16]建立了一套基于最大似然的广义区间作图(generalised interval mapping method)方法定位多个半同胞家系的二分类性状 QTL 位点,模型中的参数估计采用 Newton-Raphson 算法。由于阈模型同样适用于多等级性状(ordinal traits),因此,多等级性状的 QTL 定位方法可方便地从前面各种方法的扩展获得。

2.2.2 贝叶斯分析

由未知变量的后验联合分布及先验信息直接给出由观察数据得到的期望概率。Hoeschele 等[17]提出在远交群体中用贝叶斯方法进行连锁分析,之后不同学者发展了 QTL 多等位基因时、多标记、QTL 效应正态分布下多标记定位分析和两个 QTL 连锁等多种情况下的贝叶斯分析方法。由于参数的先验信息通常并不是很清楚,所以贝叶斯分析需要假设参数的统计分布,当具有大量记录时,似然函数可能会"掩盖"先验分布,这时贝叶斯估计趋近于最大似然估计。与最大似然法一样,贝叶斯分析可处理标记数据缺失的情况,计算强度较高。计算机技术的发展会逐步降低计算强度上的难题,从而使贝叶斯分析得以广泛的应用。

2.2.3 非参数方法

在性状表型值服从正态分布的假设条件下,可以通过简单的参数检验来检测 QTL 的存在(例如,单标记时采用 t 检验,区间定位时采用 *LOD* 值)。但对于离散性状,性状表型值并不服从正态分布。解决这一问题的是采用非参数方法来进行 QTL 检测。Kruglyak[18]介绍了通过扩展非参数 Wilcoxon rank-sum 检验得到服从标准正态分布的统计量 Z_w。相对于传统方法,非参数方法的效率相对较低,但在有些情况下,如服从指数分布时,其效率要高于 t 检验。参数方法可以直接估计 QTL 效应,而

非参数方法只能检测 QTL。

3　QTL 连锁分析方法的改进

进一步提高 QTL 检测与定位的精度,实现精细定位和高效定位,是今后 QTL 定位研究的重点,除了要不断改进实验设计外,合理优化的统计分析方法也是必不可少的,统计方法的好坏直接影响着 QTL 检测定位的准确性和功效。对于具有重要经济价值的离散性状或阈性状来说,由于其特殊的表现形式,在分析方法的选择方面更应多加考虑。

当标记与 QTL 紧密连锁时,在有限群体内就会很难发现其重组事件,因此紧密连锁的标记图谱只能对 QTL 定位提供很少的信息,除非每世代的群体规模很大。现有的研究结果来看,连锁分析一般将 QTL 定位在一个约 10 cM 甚至更大的片段上。连锁分析定位的精确度可满足实施标记辅助选择(MAS)的要求,但在如此大的片段上识别影响性状的基因则非常困难。QTL 定位的目的是识别性状基因,而动物中识别基因的主要方法(位置候选克隆)要求将 QTL 定位在更小的片段上。

连锁分析与连锁不平衡分析相结合应是今后动物离散性状 QTL 精细定位的主要手段,根据连锁不平衡的原理,目前,在 QTL 精细定位方面主要有 3 种策略可供选择,它们分别是(1)连锁不平衡(linkage disequilibrium, LD)定位[19];(2)传递不平衡检验定位(transmission disequilibrium test, TDT)[20]和同源相同(identical by descent, IBD)定位[21]。这些定位策略的共同前提是:(1)通过连锁分析已将 QTL 定位在一个 5～20 cM 区间的特定区域内;(2)在这个区域内有高密度的标记。其基本思想都是利用群体的连锁不平衡来寻找与 QTL 连锁最紧密的标记,或包含 QTL 的最小的标记区间,并且要利用群体历史上积累下来的重组事件。

IBD 定位主要利用系谱资料中观察到的非重组频率来确定连锁不平衡。其基本思想是在后代中找到来自共同祖先的某个片断,即 IBD 片段或 IBD 区域。IBD 方法最早用于人类,假设群体中祖先发生特定突变,利用 QTL 和紧密连锁标记间的连锁不平衡,在"得病"个体中寻找 IBD 区域,因为这种片段可能携带疾病基因。IBD 区域随着系谱中减数分裂数的增加而减小,通过紧密连锁的标记来检测这样的区域,可做到对基因的精细定位。目前,利用 IBD 对 QTL 精细定位的研究已有成功例子,Meuwissen 等人[22]利用 IBD 将控制牛双胎率的 QTL 定位在不到 1 cM 的区间内。此外与泌乳量有关的 *DGAT* 和 *GHR* 基因也分别用 IBD 定位得到了很好的结果[23]。

Spielman[24]提出的基于家系分析的传递连锁不平衡(TDT)分析可以完全消除群体层化现象的影响。TDT 定位的理论基础是:当标记和性状基因不连锁,且标记与性状不存在关联时,标记基因由亲本到后代的传递是随机的。因此,TDT 方法通过简单的卡方检验,在患病后代中对有无传递某一基因进行检验。TDT 方法最初用于分析当标记和性状存在关联时,检验标记和性状基因是否连锁的情况,但现在更为广泛的用于检验当标记和性状基因连锁时,二者是否存在关联。由于其简单易行和高效,TDT 方法得到了广泛应用和扩展,从最初仅能分析核心家系资料(双亲和一个发病后代),扩展到现在能够处理多个后代、同胞对、亲本信息缺失等多种情况。扩展系谱是人和畜禽中普遍存在、信息含量丰富的数据资料,它综合了父母、同胞及其他亲属的信息,通过 TDT 方法来分析扩展系谱资料,可最大限度地利用手中的数据。针对此,Martin[25]提出了 PDT(Pedigree disequilibrium test)方法,它能有效地对人类的扩展系谱进行分析。相对人类而言,扩展系谱在畜禽中数目多、信息量大、更易获得、更有利用价值。

在家畜多基因离散性状 QTL 连锁分析方法上,另一个可以预期的方法改进是 QTL 随机效应模型(random model)的合理使用。类似于常规的数量性状,在家畜群体中,由于双亲的 QTL 基因型一般未知,也不知道群体中 QTL 上等位基因的确切数目,另外,家系间基于标记基因型的 QTL 基因型的频率不同。鉴于这些复杂情况,将 QTL 的效应看作随机效应比较现实,不仅可以对 QTL 的效应和位置进行估计,而且还能有效地估计出 QTL 的方差贡献。这种 QTL 随机效应模型可用于一般系谱资料的 QTL 定位研究,利用动物模型分析 QTL 的方差组分,并实现对 QTL 的定位,这在分析策略上不同于前面各种定位方法。这样可以不再将 QTL 效应看作固定效应,而是看作一个随机效应来进行方差组分的估计,QTL 连锁分析的结果更具科学性。目前,在连续数量性状的 QTL 定位方面,建立在这种动物模型基础上的 QTL 连锁分析方法主要有 BLUP 法,

Bayes 估计,基于 Bayes 原理基础上的 Gibbs 抽样技术[17],以及基于混合线性模型的 ML 法、REML 法等。从理论上说,这些方法都可以扩展到多基因控制的离散性状 QTL 连锁分析研究,只是在理论推导和计算方面更加复杂了。可以预见,随着统计基因组学的飞速发展,也随着数理统计学、计算科学向生命科学中的不断渗透,这些障碍都将会被不断排除。

4 结 语

基因组研究的飞速发展,使人们从分子水平上了解数量性状的遗传机理成为可能,QTL 定位自 1961 年提出以来,一直是人们研究的热点,涌现出了大量的 QTL 定位分析方法,但是有关多基因离散性状 QTL 定位的研究不是很多。多基因控制的离散性状在畜禽生产中具有重要的经济意义,因此这类性状 QTL 定位方法显得日益迫切。目前在动物中,QTL 定位以连锁分析为主,但由于多基因离散性状的表型特殊性,通过常规连锁分析的效果并不理想。因此,借助于近年来发展起来的一系列 QTL 精细定位方法是非常必要的,可以预期的是:随着统计基因组学的深入发展,结合连锁不平衡分析以及 QTL 随机效应模型将是今后家畜多基因离散性状 QTL 精细定位的主要手段。

参 考 文 献(References):

[1] Christoph L. Mapping quantitative trait loci using generalized estimating equations. *Genetics*, 2001, 159:1325～1337.

[2] YI N, XU S. A random model approach to mapping quantitative trait loci for complex binary traits in outbred populations. *Genetics*, 1999, 153:1029～1040.

[3] Lunetta K L, Faraone S V, Biederman J, Laird N M. Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. *American Journal of Human Genetics*, 2000, 66: 605～614.

[4] Falconer D S, Mackay T F C. Introduction to Quantitative Genetics. 4th ed. Essex: Longman, 1996.

[5] Gamal A, Berger P J. Properties of threshold model predictions. *J Anim Sci*, 1999, 77: 582～590.

[6] Weller J I. Quantitative Trait loci Analysis in Animals. Trowbredge:Printed and Bound in the UK by Cromwell Press. 2001.

[7] ZENG Z B. Precision mapping of quantitative trait loci. *Genetics*, 1994, 136:1457～1468.

[8] Rebai A. Comparison of methods for regression interval mapping in QTL analysis with non-normal traits. *Genetical Research*, 1997, 69:69～74.

[9] Whittaker J C, Lewis C M. The effect of family structure on linkage tests using allelic association. *American Journal of Human Genetics*, 1998, 63: 889～897.

[10] Grignola F, Hoeschele I. Mapping linked quantitative trait loci via Residual Maximum Likelihood. *Genet Sel Evol*, 1997 29:539～544.

[11] Thomson P C. A generalized estimating equations approach to quantitative trait locus detection of non-normal traits. *Genetics Selection Evolution*, 2003, 35:257～280.

[12] Mccullagh P, Nelder J A. *Generalized linear models*. 2nd edn Chapman-Hall, London, 1989.

[13] Heather J, Cordell J A, Todd N J. Statistical modeling of inter-locus interactions in a complex disease. *Genetics*, 2001, 158: 357～367.

[14] Hackett C A, Weller J I. Genetic mapping of quantitative trait loci for traits with ordinal distributions. *Biometrics*, 1995, 51: 1252～1263.

[15] XU S. Further investigation of the regression method for mapping quantitative trait loci. *Heredity*, 1998, 80:364～373.

[16] Kadarmideen H N, Dekkers J C M. Generalized marker regression and interval QTL mapping methods for binary traits in half-sib family designs. *J Anim Breed Genet*, 2001, 118:297～309.

[17] Hoeschele I, P Vanranden. Bayesian analysis of linkage between genetic markers and quantitative trait loci. I. prior knowledge. *Theor Appl Genet*, 1993, 85:953～960.

[18] Kruglyak L, Lander E S. A nonparameteric approach for mapping quantitative trait loci. *Genetics*, 1995, 139: 1421～1428.

[19] Kruglyak L. What is significant in whole-genome linkage disequilibrium studies? *Am J Hum Genet*, 1997,61:810～812.

[20] DING X D, ZHANG Q, XU R H, WANG Y C. Pedigree transmission disequilibrium test for QTL mapping of threshold trait. *Chinese Science Bulletin*, 2004, 49: 1347～1353.

[21] Carlier C, Farnir F, Berzi P. Identify-by-descent mapping of recessive traits in livestock: application to map the bovine syndactyly locus to chromosome 15. *Genome Res*, 1996, 6:580～589.

[22] Meuwissen T H, Karlsen A, Lien S, Olsaker I, Goddard M E. Fine mapping of a quantitative trait locus for twinning rate using combined linkageand linkage disequilibrium mapping. *Genetics*, 2002,161:373～379.

[23] Winter A. Association of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-CoA: diacylglycerol acyltransferase (DGAT1) with variation at a quantitative trait locus for milk fat content. *Pro Natl Acad Sci USA*, 2002, 99: 9300～9305.

[24] Spielman R S, McGinnis R E, Ewens W J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet*, 1993, 52: 506～516.

[25] Martin E R, Bass M P, Kaplan N L. Correcting for a potential bias in the pedigree disequilibrium test. *Am J Hum Genet*, 2001, 68: 1065～1067.