

摘 要

在动物遗传育种领域中,迄今为止主要的研究对象多为正态连续分布的生产性状,并获得了显著的遗传进展。对于这类性状,在观察值服从正态分布的假设前提下,用线性方法进行遗传分析是较为合理的选择,并可以获得合理的结果。而对于多基因离散性状或阈性状的遗传分析,由于这类性状的分布为非正态的,线性模型的功效就显得不足,甚至会导致错误的结果。动物育种中的离散性状,在多基因控制的理论前提下,其观察值的分布是离散不连续的,观察值不能表达为遗传效应与环境效应的线性组合。近年来,随着统计基因组学的飞速发展,使人们从分子水平上了解离散性状的遗传机理成为可能。但目前用于 QTL 检测与定位的统计方法大多都是基于对连续数量性状的标记-QTL 连锁分析。而对于不服从正态分布的离散性状来说,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,因此,采用常规的线性方法对离散性状进行标记-QTL 连锁分析的功效不高。相对于常规的数量性状,多基因控制下离散性状或阈性状的 QTL 定位就显得更具有挑战性。

广义线性模型 (GLM) 是一般正态线性模型 (LM) 的直接推广,它适用于连续性状和离散性状的遗传分析。GLM 无需常规线性模型对随机误差分布的正态性假设,并可通过连接函数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,以实现参数的无偏估计。另外,广义线性模型的参数估计量具有大样本正态分布,因而具有良好的统计性质。正因如此,广义线性模型成为了当前统计学领域中的主流方法,其应用范围涉及众多研究领域。

本研究在广义线性模型的框架内,对多基因离散性状的遗传参数估计和 QTL 定位进行了模拟研究,并与常规的线性方法进行了效率比较,参数的估计方法采用最大似然法,分析群体设计为同胞家系,群体大小为 500 头,并就影响遗传参数估计和 QTL 定位的各种因素进行了研究。研究结果表明:(1) 在遗传力和个体育种值的估计准确度方面, GLM 方法具有一定的优势;(2) 性状的遗传力和表型发生率对遗传力估计及个体育种值预测的效果均有直接影响;(3) GLM 方法具有良好的统计性质,其应用可扩展到多阈值性状和更为一般的家系群体;(4) GLM 框架内的阈模型方法能有效地估计出离散性状 QTL 参数,在检验功效方面,不同参数组合下的 QTL 检验功效也较常规线性方法高;(5) 类似于常规的数量性状,性状的遗传力、QTL 效应等对离散性状的 QTL 定位有明显的影响;(6) 双侧翼标记间距对 QTL 定位的准确性和功效也有较大程度的影响,较高密度的标记图谱对离散性状 QTL 连锁分析是非常必要的;(7) 潜在的阈值和阈值数对似然统计量 LR 值和 QTL 定位的效率有直接的影响。潜在阈值的大小则决定着性状各表型的发生率,因此,性状表型发生率的高低影响着离散性状 QTL 定位的功效。潜在的阈值数决定着性状的表型分类数。(8) 广义线性模型框架内的离散性状 QTL 定位方法 (GLM) 同样适用于一般群体、多标记和其它表型分布性状的情况。

总的来看, GLM 方法在离散性状的遗传分析方面具有一定的优势,可作为家畜离散性状遗传分析的主要方法之一。

关键词: 广义线性模型 (GLM), 线性模型 (LM), 离散性状, 遗传分析, QTL 定位

ABSTRACT

Linear mixed models provide a powerful means of genetic analyses. However, for many traits of economic importance the assumptions of linear responses, constant variance, and normality are questionable. So methods for continuous data are not appropriate for discrete data. Many economically important characters in livestock vary in a discontinuous manner but are not inherited in a simple Mendelian fashion. These traits are presumably controlled by polygenes with expressions modified by environmental effects like quantitative traits. Owing to the recent advances in marker technology, the molecular dissection of heritable quantitative traits into their individual Mendelian components or quantitative trait loci (QTL) is now possible, and novel statistical techniques to detect and map individual genes affecting quantitative traits have been developed and subsequently refined. However, most QTL mapping methods share a common assumption that the phenotype follows a normal distribution and hence are not appropriate for categorical traits. Thus, new methods are needed for QTL mapping for categorical traits.

Generalized linear models (GLM) have attracted considerable attention over the years. GLM is extend from ordinary linear mixed models, and can deal with discrete or continuous data conveniently. It differs from linear mixed model in that it focuses more on the inverse link function to model the relationship between the linear predictor and the conditional mean. GLM provide a flexible way to model production traits which do not satisfy the assumption of a linear mixed model.

In this study, methodology of genetic analyses and QTL mapping for discrete traits in livestock basing on the framework of a generalized linear model was present. The parameters were estimated using Maximum Likelihood method. The efficiency and power were compared with the linear model. The factors of influencing genetic analyses and QTL detection efficiency were simulated in this study too. Daughter design with multiple families was applied, and the number of segregating population is 500. From the results of simulation study, following conclusions can be derived. (1) GLM has a great advantage in heritability and breeding value estimation, and has nice accuracy for discrete traits. (2) The accuracy of estimation of heritability depended on the true heritability and incidence of the trait directly. (3) GLM can be extended for different situation such as general populations. (4) The GLM method has a certain advantage in location estimation and power of QTL mapping, and has nice efficiency and accuracy for ordinal traits. (5) The accuracy of QTL mapping largely depended on the effect of putative quantitative trait loci and the value of heritability. (6) The accuracy and power of QTL mapping depended on the distance between the two flanking markers. (7) The threshold value and the number of underlying liability can influence the power and accuracy of QTL mapping, too. (8) The GLM method for QTL mapping can be extended for different situation (e.g. general populations, multiple markers and different distribution of phenotype).

Key words: generalized linear models, linear model, discrete traits, genetic analyses, QTL mapping

国家重点基础研究发展规划（973）项目

（G200001603）

农业动物遗传资源中优质基因的分离和利用机理

国家高技术研究发展（863）计划

（2001AA243011）

奶牛低脂高蛋白新品系的选育

国家自然科学基金重点项目

（30430500）

仔猪 *E. coli* 腹泻抗性的分子遗传机理研究

表目录

表 1-1 动物育种中的分类性状及其分布	4
表 1-2 典型连接函数及方差函数	6
表 1-3 牛、猪、鸡、山羊和绵羊基因组上座位数统计结果	10
表 1-4 部分家畜基因组或染色体上的遗传图谱	11
表 1-5 不同设计的期望 QTL 方差及获得相同统计效力的样本大小	13
表 1-6 统计效率为 90% 时 BC 与 F2 设计所需的后代样本大小	14
表 1-7 不同设计试验效力和所需基因型测定数目	18
表 2-1 模拟参数的组合	29
表 2-2 三种方法在不同参数组合下的估计遗传力及标准误	35
表 2-3 三种方法估计遗传力的均方误及偏差	39
表 2-4 不同估计方法下的估计育种值排序与真实育种值排序之间的秩相关及排序的相对准确性	41
表 3-1 女儿设计后代的基因型及相对概率	49
表 3-2 后代 QTL 基因型的条件概率 ($\Pr(G = k m)$)	51
表 3-3 单阈值二分类性状在不同参数组合下的 QTL 参数估计效果 (表型发生率 $\pi_1 = 0.2$)	56
表 3-4 单阈值二分类性状在不同参数组合下的 QTL 参数估计效果 (表型发生率 $\pi_1 = 0.5$)	57
表 3-5 四分类等级性状在不同参数组合下的 QTL 参数估计效果 ($\pi = 0.20/0.60/0.15/0.05$)	58
表 3-6 四分类等级性状在不同参数组合下的 QTL 参数估计效果 ($\pi = 0.20/0.30/0.30/0.20$)	59
表 3-7 单阈值二分类性状在不同参数组合下 QTL 定位功效	62
表 3-8 多阈值等级性状在不同参数组合下 QTL 定位功效	63
表 3-9 不同模拟条件下的显著性阈值	64

图目录

图 1-1 广义线性模型的组成成分	8
图 1-2 三种图距函数的差异	11
图 1-3 QTL 检测的试验设计	12
图 2-1 不同性状遗传力真值下的遗传力估计效果	34
图 2-2 不同表型发生率下的遗传力估计效果（基础群大小为 240 的情况）	36
图 2-3 不同基础群大小下的遗传力估计效果	37
图 2-4 不同表型发生率下三种方法的估计遗传力差异（基础群大小为 240 的情况）	38
图 2-5 不同遗传力真值和表型发生率下三种方法的遗传力估计偏差(基础群大小为 240)	40
图 2-6 不同遗传力真值和表型发生率下三种方法的遗传力估计均方误比较(基础群大小为 240)	40
图 3-1 不同标记间距下的平均似然率曲线（4 分类等级性状）	55
图 3-2 不同表型发生率下的平均似然率曲线（4 分类等级性状, $h^2 = 0.4, \Delta_q = 0.5, r = 0.3$ ）	61
图 3-3 不同参数组合下 GLM 与 LM 方法的检验功效	61

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得中国农业大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

研究生签名：

殷宗俊

时间： 2005 年 6 月 20 日

关于论文使用授权的说明

本人完全了解中国农业大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。同意中国农业大学可以用不同方式在不同媒体上发表、传播学位论文的全部或部分内容。

(保密的学位论文在解密后应遵守此协议)

研究生签名：

殷宗俊

时间： 2005 年 6 月 20 日

导师签名：

张中印

时间： 2005 年 6 月 24 日

第一章 文献综述

1.1 基因组学对数量性状遗传模型的影响

数量遗传学自其诞生到现在,已经历了近 100 年的历史,形成了一整套理论体系,并在动物育种中发挥了巨大作用。Nilsson-Ehle (1909) 提出的数量性状遗传模型—多基因假说 (Polygene Hypothesis) 消除了孟德尔学派与统计学派的分歧,为数量遗传学的诞生铺平了道路;Fisher (1918) 那篇划时代的论文《根据孟德尔遗传假设的亲属间相关研究》则进一步系统论述了数量遗传学的研究对象和方法,标志着数量遗传学作为一个新的遗传学分支踏进了历史舞台;重复率、遗传力和遗传相关等遗传参数的提出以及统计理论的发展、计算机技术的引入,都推动着数量遗传学在动植物育种中起着主导作用。同时,在数量遗传学理论上发展起来的现代动植物育种方法,又以其巨大的成就强有力地支持了数量遗传学。随着分子生物技术的不断发展,如今,遗传学已进入基因组时代,人类的基因组测序工作已经完成,各种家畜的基因组测序工作正在展开,基因的奥秘正在被逐渐揭示,人们可以实现从分子水平去了解数量性状的遗传变异 (张勤, 2001)。

经典数量遗传学的理论核心是用统计学方法对造成数量性状变异性的因素进行剖析,其基本模型是 $y = u + g + e + g \times e$, 其中 u 、 g 、 e 分别表示群体均值、多基因遗传效应和环境效应, $g \times e$ 是遗传与环境的互作效应。在这一遗传模型中, g 是影响表型 y 的所有基因作用的总和,由于这些基因是未知的,分析中只能将它作为随机效应来处理。根据这个模型可进行各效应和方差组分的估计。经典数量遗传模型的理论基础是微效多基因 (infinitesimal gene) 假设,但这一假设从理论上说是不完善的,因为没有理由认为影响数量性状的所有基因都是微效的 (张勤, 2001),而近年来的大量研究也证明许多数量性状确实存在具有较大效应的主效基因 (major gene) 或数量性状基因座 (Quantitative Trait Locus, QTL)。同时,在家畜育种中还有许多用现有的数量遗传学理论和方法未能很好解决的问题,例如对抗病性状、胴体性状及低遗传力性状的遗传改良一直未能取得理想的结果。

经典数量遗传学把控制某一性状的多个基因作为一个整体进行研究,这种方法无法鉴别出单个数量基因以及与之有关的染色体片段,更难确定数量性状座位在染色体上的位置及与其它基因的关系,数量性状依然是个“黑箱”。如果能够将 QTL 定位在某个具体位置,能对其进行直接操作,将无疑使动物的育种更具有针对性,也更加高效。然而遗憾的是,复杂的数学方法和手段终究不能使抽象的数量基因具体化、物化到分子水平的 DNA 序列。毫无疑问,数量遗传学在这一方面的升华需要新的推动 (丁向东, 2004)。随着分子生物技术的飞速发展, DNA 测序技术不断发展完善,分子遗传标记日益丰富,遗传图谱的密度越来越大,都为 QTL 定位提供了先决条件。同时在分子遗传学方面,2001 年人类基因组计划公共协作组和 Celera 公司分别在 *Nature* 和 *Science* 杂志上发表了各自的人类基因组草图结果,这标志着遗传学已进入基因组时代。基因组时代给人们带来了大量新的概念,其中影响较大的有 DNA 水平的多态性、QTL、候选基因和基因定位等,这些概念在很大程度上弥补了微效多基因模型的缺陷。利用 DNA 标记可以对影响数量性状的单个基因 (即主效基因) 或 QTL 进行检测及定位,随着影响数量性状的主基因或 QTL 被不断检出,数量性状的遗传模型也发生了巨大变化,由传统的数量性状微效多基因遗传模型发展到目前的主

基因-多基因混合遗传模型, 将分子标记或主基因 (Gm) 的信息加入到性状遗传模型中以提高遗传分析的准确性, 加快遗传改良速度。

$$y = u + Gm + g + e + g \times e$$

从一定意义上说, 现代基因组学对数量遗传学的影响可用上述两种遗传模型的相对效率来度量, 显然, 这种影响主要取决于在总的遗传变异中由 Gm 所引起的变异比重以及对 Gm 基因型推断的准确性程度。因此, 如何解读 DNA 序列, 有效识别和分离影响性状的功能基因、主效基因及 QTL 成为现代分子数量遗传学的研究目标 (吴常信, 1997; 张勤, 2001; Andersson 等, 2004)。

1.2 遗传评估模型与标记辅助选择

遗传评定的准确性直接影响家畜重要经济性状的遗传改良。育种值估计是家畜遗传评定的核心内容, 个体的遗传评定, 就是要了解其加性遗传效应, 即育种值的大小, 由于育种值是观察不到的, 只能根据表型值用特殊的统计学方法进行估计。育种值的估计在家畜育种中占重要地位, 因而受到人们的重视。如何尽可能准确地估计育种值是关键, 传统的育种值估计方法主要是选择指数法, 但这一方法的假定在实际育种资料中很难满足, 相应地发展了动物模型 BLUP (best linear unbiased prediction) 法 (Henderson, 1984)。至今, 在家畜育种中人们普遍采用动物模型 BLUP 方法进行个体遗传评定, 这一方法主要是基于数量性状的多微效基因遗传模型, 利用表型信息和系谱信息估计个体的育种值。BLUP 方法可以克服传统选择指数法的不足, 可适应于不同来源、不同世代及不同环境下种畜个体的遗传评定。目前, 动物模型 BLUP 法已成为种畜育种值估计的国际通用标准方法, 与其它评估模型相比, 减少了诸多限制, 并通过分子血缘相关阵 (numerator relationship matrix) 充分利用各种来源信息, 提高了个体遗传评估的准确性 (张沅, 2001; 张沅、张勤, 1993)。利用这一方法, 许多数量性状都取得了显著的遗传进展, 但是, 育种学家们仍然有不满意的地方, 如一些低遗传力性状、限性性状以及一些难以测量或不能早期度量的性状, 取得的遗传进展非常有限或效率很低 (Fernando, 1989; Lande, 1990)。

近年来, 分子遗传学和分子生物信息技术的飞速发展使得在 DNA 水平上研究数量性状的遗传机理成为可能。用 DNA 标记可以对影响数量性状的单个基因 (即主效基因) 或数量性状基因座进行检测及定位。随着影响数量性状的主基因或 QTL 被不断检出, 数量性状的遗传模型也发生了巨大变化, 由传统的数量性状多微效基因遗传模型发展到目前的主基因-多基因混合遗传模型。将分子标记或主基因的信息加入到遗传模型中可以提高遗传分析的准确性。在家畜育种方案中, DNA 标记可以用来辅助亲本的遗传选择, 即标记辅助选择 (marker assisted selection, MAS)。人们提出了多种 MAS 的实施方案 (Gomez-Raya, 1997; Lande, 1990), 其中最为常用的 MAS 方案是基于主基因-多基因混合模型下的标记辅助 BLUP (marker assisted best linear unbiased prediction, MBLUP) 方法。MBLUP 方法实际上是常规动物模型 BLUP 方法的扩展, 也就是在动物模型中将基因的效应分为已定位的 QTL 的效应和未知基因 (剩余多基因) 的效应, 并借助分子标记信息来推断 QTL 基因型或 QTL 等位基因的概率, 个体总的育种值是 QTL 育种值和剩余多基因育种值之和。由于 DNA 标记信息的获取不易受环境、性别和发育阶段等因素的影响, 利用标记信息可以较准确地估计 QTL 的育种值, 总的育种值估计的准确性也会进一步提高, 因而以此为基础的标记辅助选择可大大提高选种的准确性和育种效率 (Soller, 1994)。大量研究表明,

MAS 可提高对数量性状的遗传选择效率, 相对于常规的动物模型 BLUP 方法, 在不同的情况下, MAS 可提高选择效率 1%~60%。许多学者通过随机模拟技术对 MAS 和常规 BLUP 选择效果进行了比较。Meuwissen 等 (1992) 采用确定模型 (deterministic model) 对奶牛产奶性状的标记辅助选择进行了模拟研究, 结果表明, 在开放和闭锁两种育种体系中, 选择进展分别比常规 BLUP 提高 9.5~25.8% 和 7.7~22.4%。Ruane 等 (1996) 研究表明, 在一个奶牛育种核心群的第 1、2、3、6 世代中, 利用标记辅助选择可使产奶性状的遗传进展提高 3%、9%、12% 和 6%; QTL 有利等位基因初始频率较低时 (如 0.1), 选择进展可分别达到 9%、19%、24% 和 15%。不同的研究结果不尽相同, 这取决于研究所采用的 QTL 遗传模型、群体结构、家系大小、家系数目、选择世代数、标记的特性、标记和 QTL 连锁程度等各项假设。但大多研究的共同结论是: 标记辅助选择的效率优于常规 BLUP 方法, 特别是对于低遗传力性状、限性性状和不宜直接测定的性状, 利用 MAS 选择可能会具有更大的价值 (Liu, 2001; Johan, 1994; Van Arendonk, 1994)。

标记辅助选择是提高选择效率的有效方法, 然而在实际育种工作中实用的例子并不多, 主要原因在于 MAS 在操作过程中所涉及的内容很广泛, 而且也很复杂。要使标记辅助选择能够在育种实践中得到广泛应用, 首先要构建高分辨率、高密度的标记连锁图谱; 其次是要能准确、高效地检测和定位影响数量性状的 QTL, 这就需要发展一些高效的遗传统计方法; 再者是要制定出完善的育种方案。

1.3 多基因离散性状的遗传分析

在动物遗传育种领域中, 迄今为止主要的研究对象是正态连续分布的生产性状, 并获得了显著的遗传进展。对于这类性状, 在观察值服从正态分布的假设前提下, 用线性方法进行遗传评定及参数估计等遗传分析是较为合理的选择, 并可以获得合理的结果。后续的遗传操作 (如个体育种值估计、遗传参数估计、QTL 定位和 MAS 等) 也都是基于线性模型的基础上进行的。而对于多基因控制的离散性状 (categorical traits) 或阈性状 (threshold traits) 的遗传分析, 由于这类性状的分布为非正态, 线性模型的功效就显得不足, 甚至会导出错误的结果。动物育种中的离散性状, 在多基因控制的理论前提下, 其观察值的分布是离散不连续的, 观察值不能表达为遗传效应与环境效应的线性组合, 从而限制了线性方法在离散性状遗传分析上的应用。

现代动物育种中, 人们对离散性状的重视程度不断增大, 一方面是因为许多重要的经济性状如繁殖性状、与疾病和生存有关的性状都属于离散性状; 再者是由于许多选择试验都证明了对离散性状的选择是有效的 (Tempelman, 1996; Rekaya, 2001)。特别是随着家畜育种技术的进一步发展, 一些当前育种目标中的主选性状不断趋于选择极限, 进一步选择的进展甚微, 从而一些重要的次级性状受到育种学家们的极大重视, 特别是与经济效益密切相关的离散性状, 如能将其纳入到家畜的育种目标中将会获得巨大的育种效益。在此前提下, 家畜离散性状的遗传分析方法显得日益迫切。

1.3.1 家畜主要离散性状及其分布

动物育种中的离散性状 (discrete traits), 有时也可称为阈性状 (threshold traits) 或拟连续性

状 (quasi-continuous traits), 在多基因控制的理论下, 其观察值是离散不连续的, 可能是一种分布 (如 0-1 性状的二项分布), 也可能是多种分布的混合体。如母畜繁殖性状中的产羔数、受胎率、乳头数、副乳头数、瞎乳头数、排卵率、受孕及分娩难易程度等; 胴体肉质性状中的肉色、肌间脂肪评分及胴体类型评分等; 与疾病抗性有关的乳房炎、抗线虫、抗腹泻、抗痒病、抗蹄病和抗应激等; 与生存有关的出生死亡等。这些性状的表型都是离散的, 为一个间断分布, 同时在受多基因控制的理论前提下, 又具有一个潜在的连续分布, 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。而离散的表型值与潜在的连续值是通过一系列固定的阈相联系, 这种潜在的连续分布可通过线性模型来描述, 但是潜在的连续值与可观察到的表型值是非线性的。

不同性状的表型分布不同, 动物育种中大多数的经济性状表型都是连续的, 且服从或近似服从正态分布。在动物育种中有几种重要的离散分布, 它们分别是二项分布 (binomial)、多项分布 (multinomial)、泊松分布 (poisson) 及负二项分布 (negative binomial)。如动物的生育能力、抗病和抗应激能力、成活或死亡等都服从二项分布或 0-1 分布, 而有关母畜排卵数、产羔数、乳头数等记数性状则服从 *multinomial* 或 *poisson* 分布。表 1-1 列出的是家畜主要的离散性状及其分布。

表 2-1 动物育种中的分类性状及其分布

Table 1-1 The categorical traits and distributions in the field of animal breeding

性状归属	分类性状	服从分布
type of traits	categorical traits	distribution
与繁殖有关	产仔数	Multinomial or Poisson
	排卵数	Multinomial or Poisson
	乳头数	Multinomial or Poisson
	生育力	Binomial
与生长有关	采食能力	Multinomial or Poisson
与胴体有关	胴体类型	Multinomial or Poisson
	肉质评分	Multinomial or Poisson
与生存有关	出生死亡	Binomial
与抗性有关	乳房炎	Binomial
	抗线虫	Binomial
	抗腹泻	Binomial
	抗蹄病	Binomial
	抗痒病	Binomial
	抗应激	Binomial or Poisson
	组织感染程度	Multinomial
与行为有关	分娩行为	Multinomial or Binomial
	异常行为	Multinomial or Binomial
	母性行为	Multinomial or Binomial

离散性状或阈性状的表观度量一般用有限的几个状态的相对发生率来表示, 对个体而言, 每个个体只能取其中的一个状态。但在一个群体中, 却可以取任一状态。如果给每一个状态确定一个取值, 则可用一随机变量来描述阈性状, 类似于常规的数量性状, 也可采用数值的方法来其进

行遗传分析。对于较多状态的多阈值性状来说（如猪的产仔数、乳头数、排卵数，鸡的产蛋数等），其表型分布可看作近似的连续分布。由于表型上的特殊性，因此对离散性状的遗传分析也具有不同于常规数量性状的独特性。

1.3.2 离散性状的遗传分析思想

目前，对离散性状的遗传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，在假设表型数据为连续分布的前提下，采用线性的方法对性状进行遗传分析，称之为线性分析思想；另一种思路就是阈值概念（Wright, 1934; Falconer, 1996），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，而离散的表型值与潜在的连续值是通过一系列固定的阈相联系。离散性状非线性遗传分析方法的导出，主要基于贝叶斯概型和广义线性混合模型的发展应用。

对于离散性状，其观察值不能表达为遗传效应与环境效应的线性组合，从而限制了线性方法在阈性状遗传分析上的应用。另外，线性方法对于阈性状分析还有两个缺陷：一是线性方法的解不能保证阈性状发生的概率为正值，并且总和为 1；二是用线性方法只能在表观尺度上对分离性状的观察值进行描述，而在表观尺度上阈性状的遗传效应与环境效应是相关的。尽管如此，动物遗传育种中的离散性状至今仍然多用线性方法进行分析，Wright（1934）的阈值模型为正确地进行多基因离散性状的分析奠定了基础，但合理的方法是在近年来才逐渐建立起来的。

离散性状非线性遗传分析方法的导出，一方面基于贝叶斯概型，将关于参数的先验知识和来自数据的信息加以综合利用，将参数看作随机变量，导出其后验分布，并使后验概率密度函数最大化，从而求出相应的参数；另一方面就是基于广义线性模型（Generalized Linear Model, GLM）的应用（McCullagh, 1989），通过连接函数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，以实现参数的无偏估计。相对于贝叶斯概型，广义线性模型方法更受欢迎，其计算量远低于前者，并且有许多成熟的统计分析软件可以利用（如 SAS、SPSS），这也是广义线性模型成为当前动植物离散性状遗传分析主要手段的原因所在。

1.3.3 广义线性模型（GLM）及其在遗传分析中的应用

广义线性模型（GLM）是常规正态线性模型的直接推广（McCullagh, 1989），它适用于连续性状和离散性状的遗传分析，特别是后者，如属性数据，计算数据的统计分析。它与典型线性模型的区别是其随机误差的分布没有正态性要求，与非线性模型的最大区别则在于非线性模型没有明确的随机误差分布假定而广义线性模型的随机误差的分布是可以确定的；另外，广义线性模型的参数估计量具有大样本正态分布，因而具有良好的统计性质。

McCullagh 与 Nelder（1989）将广义线性模型定义为响应变量（Y）具有如下分布密度的指数分布族（exponential family）统称为广义线性模型：

$$f_Y(y, \theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\} \quad (1-1)$$

其拟然函数是

$$L = \log f = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \quad (1-2)$$

其中， $a(\phi)$ ， $b(\theta)$ ， $c(y,\phi)$ 均为已知函数， θ 为典范参数（canonical parameter）， ϕ 为离散参数（dispersion parameter）。响应变量（Y）的期望和方差可对（1-2）式求偏导数获得。可以看出，广义线性模型不同于一般线性模型，其响应变量的方差不是一个常数（ σ^2 ），而是一个依赖于均值的函数。

1.3.3.1 广义线性模型与一般线性模型的区别

统计回归模型中形如 $y = X\beta + \sum_{i=1}^k Z_i u_i + e$ ，其均值的期望为固定效应的函数（ $u_i = x_i' \beta$ ），响应变量的方差不依赖于均值（ $\text{var}(y_i) = \sigma^2$ ），且误差分布为正态的假设前提下，称之为一般线性模型，因为在模型中所有参数都是线性的。线性模型在多种场合下均发挥着重要的作用，当前绝大多数数量性状的统计分析都是采用这一模型。但是这一模型的缺陷也是显而易见的，特别是对于一些分类性状、阈性状、记数性状及属性资料等数据，由于上述多个假设都被打破，从而导致线性模型的不适应，或引起统计效率的显著下降，参数的估计严重偏差，效应的预测可能不在参数空间内。而广义线性模型则无需线性模型的这些假定条件，其对误差分布无需正态性假设，同时响应变量的方差定义为均值的函数。通过相应的连接函数（Link function）将响应变量观察值的期望与解释变量连接起来，从而使转化后的响应变量期望值线性化，以实现参数的无偏估计。常见分布的典型连接函数及方差函数如表 1-2 所示。

表 1-2 典型连接函数及方差函数
Table 1-2 Typical link and variance function

分布 Distribution	连接函数 Link function	反连接函数 Inverse link	方差函数 Variance function $V(\mu)$	性状类型 Trait type
Normal	$\eta = \mu$	$\mu = \eta$	1	Continuous
Poisson	$\log(\mu)$	$\exp(\eta)$	μ	Count
Binomial	$\log(\mu / (1-\mu))$	$\exp(\eta) / (1+\exp(\eta))$	$\mu(1-\mu)/n$	Proportion
Gamma	$1/\mu$	$1/\eta$	μ^{-2}	Posi. Conti.
Inverse-Gaussian	$1/\mu^{-2}$	$1/\eta^{-2}$	μ^{-3}	Posi. Conti.

1.3.3.2 广义线性模型的组成成分

广义线性模型的组成可由图 1-1 表示。与线性模型一样，广义线性模型也包括固定效应 β （如管理效应）、随机效应 $u \sim N(0, G)$ （如个体育种值），以及设计矩阵 X、Z 和观察值向量 y。

另外，广义线性模型还包括线性预测（linear predictor）， η ，连接或反连接函数， $h(\cdot)$ 。通

过反连接函数将条件均值 μ (conditional mean) 与线性预测 η 相联系, 并通过方差函数将协方差矩阵 R 与条件均值相联系, 从而实现误差方差依赖于条件均值。这也是广义线性模型与一般线性模型的主要区别。概括起来, 广义线性模型主要由三项构成要素, 它们分别为连接或反连接函数、线性预测 ($\eta = X\beta + Zu$) 和方差函数。

(1) 线性预测 (η) 对于一个线性混合模型, 其固定效应与随机效应可构成一个线性预测, $\eta = X\beta + Zu$, 线性混合模型中观察值向量可表示为这一线性预测向量与误差向量之和, 即:

$$y = \eta + e = X\beta + Zu + e \quad (1-3)$$

在给定 u 的条件下, 响应变量的条件分布服从于均值为 η 的正态分布 ($y|u \sim N(\eta, R)$)。

(2) 连接函数与反连接函数 用以连接观察值的线性预测 η 与条件均值 u 的函数。对于大多数性状来说, 反连接函数是一对一的关系。然而, 对于阈模型来说, u 为 $t \times 1$ 向量, t 为阈水平数, 对于生长曲线模型来说, u 为 $n \times 1$ 向量, η 为 $p \times 1$ 向量, 其中 n 为个体被测定的次数, p 为生长曲线中的参数个数。对于线性模型, 其反连接函数是广义线性模型的一个特例, $h(\eta) = \eta$, 属于一致性连接 (identity function)。0-1 二项分类性状的对数连接函数 (logit link) 是最为常用的, 其连接函数与反连接函数如下:

$$\eta_i = \ln\left(\frac{\mu_i}{1 - \mu_i}\right), \quad \mu_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}} \quad (1-4)$$

由 (1-4) 式可以看出, 对数连接不同于一致性连接, 其估计的均值 (μ_i) 范围在 0~1 之间, 同时, 效应的变化与线性预测值的改变不是线性的, 其变化曲线为 S 型而非直线。对于绝大多数单变量连接函数来说, 其变化趋势均为单调递增的, 也就是说, 随着线性预测值 η 的增加, 条件均值 μ 也是不断增大, 但其增加率则不是一个恒量。

连接函数与反连接函数的选择主要取决于误差分布, 同一分布可能有多种连接函数, 因此在选择连接函数时还应考虑到分析的简单性。

(3) 方差函数 方差函数 (Variance function) 主要用于恒量非系统变化程度 (non-systematic variability), 特别是在广义线性模型中, 残差的变化主要来源于两个方面, 一是来源于抽样分布, 二是来源于其它的额外变化, 如超分布 (over-dispersion) 等。由抽样分布引起的模型变异是直接的, 表 1 中给出了常见分布的抽样方差 ($v(\mu)$)。由超分布引起的模型变异可采用多种方法进行度量, 最为常用的方法是将残差的变化定义为 $\text{var}(y|u) = \phi v(\mu)$, 其中 ϕ 为超分布参数; 另一种常用的方法是将随机误差效应 e 加入到每个观察值的线性预测中, $e \sim N(0, \phi)$, 从而形成新的预测向量, 这种方法可产生较为理想的结果, 但是其计算工作量迅速增加。

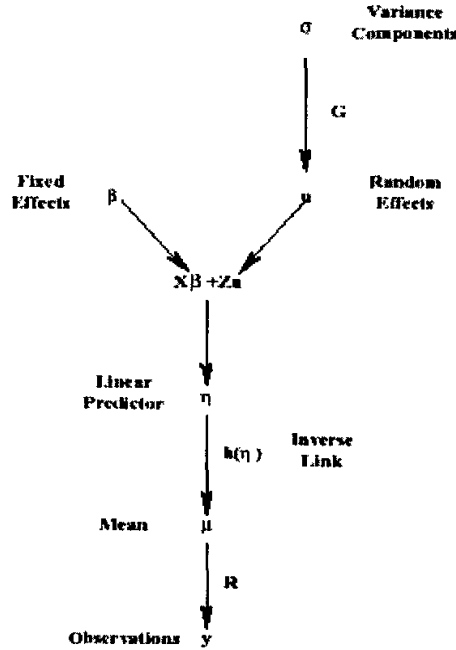


图 1-1 广义线性模型的组成成分

Figure 1-1 Symbolic representation of a generalized linear mixed model

1.3.3.3 模型参数的估计与预测方法

广义线性模型的参数估计方程可由多种方式导出，有传统的加权最小二乘法、似然法、Fisher 迹法和 Bayesian 法等，还有最近两年发展起来的一些新方法，如 MC-EM 算法、随机算法、复合似然法、求积法、BLUP 法等 (Nengjun 等, 2000; Rekaya 等, 2001; Engel 等, 1994; Thomson, 2004)。由拉普拉斯算子 (Laplacian) 拟似然法，可得到较为直观的固定效应与随机效应的估计方程：

$$\begin{pmatrix} X'H'R^{-1}HX & X'H'R^{-1}HZ \\ Z'H'R^{-1}HX & Z'H'R^{-1}HZ + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X'H'R^{-1}y^* \\ Z'H'R^{-1}y^* \end{pmatrix} \quad (1-5)$$

其中， $H = \partial\mu / \partial\eta'$ ， $R = \text{var}(y|u)$ ， $y^* = y - u + H\eta$ ， $\mu = h(\eta)$ 。与混合模型方程组一样，广义线性方程组的求解也需要通过多次迭代来完成。Fisher 迹法迭代估计方程如下：

$$\begin{aligned} [X'HR^{-1}HX](\beta^{[i+1]} - \beta^{[i]}) &= X'H'R^{-1}(y - u^{[i]}) \\ [X'HR^{-1}HX](\beta^{[i+1]}) &= X'H'R^{-1}(y - u^{[i]} + H\eta^{[i]}) \end{aligned} \quad (1-6)$$

观察值的方差构成主要有两个组分，一是随机效应方差，另一是误差分布方差。与混合线性模型方差组分估计类似，广义线性模型的方差组分估计可通过准似然函数 (quasi-likelihood) REML

法实现，具体估计方程如下：

$$ql(\beta, \sigma) = -\frac{1}{2} \ln |V| - \frac{1}{2} \ln |X'H'V^{-1}HX| - \frac{1}{2} (y^* - HX\beta)' V^{-1} (y^* - HX\beta) \quad (1-7)$$

式中， σ 为方差组分向量， $V = R + HZGZ'H'$ 。结合 (1-6) 式与 (1-7) 可对方差组分进行迭代求解。

广义线性模型中，误差分布方差组分的估计比一般线性模型复杂的多，这里不给出具体的求解过程，其相应的二次型为是 $(y - \hat{\mu})' R^{-1} \frac{\partial R}{\partial \phi} R^{-1} (y - \hat{\mu})$ 。

1.3.3.4 GLM 在家畜离散性状遗传分析中的应用概况

广义线性模型在动物离散性状的遗传分析中已达到了广泛应用 (Christoph, 2001)，其应用范围涉及个体的遗传评定、离散性状遗传参数估计及 QTL 定位等多个领域，并取得了理想的结果。广义线性模型是传统线性模型的扩展，GLM 有两个主要特性，一是通过连接函数实现响应变量的线性化，二是模型方差为均值的函数，它无需方差同质及残差项服从正态分布的假设，从而扩展了其适用的范围，这也是 GLM 区别于传统线性模型的关键所在。Gilmour 等 (1985, 1987) 首次较为系统地采用 GLM 对二分类数据 (dichotomous) 和顺序分类数据 (ordered categories) 进行了分析研究，这也是广义线性模型应用于分类数据分析的一个切入点，从而推动了广义线性模型在离散数据遗传分析中的广泛应用。在 GLM 中模型参数的估计，类似于线性模型，至于方差组分的估计可看作是 REML 的扩展 (Gianola, 1985; 1993)，但是在 GLM 中参数与方差的估计均需要多次迭代才能完成，其计算工作量较线性模型大的多。Meijering 和 Gianola (1985) 比较了分类性状在不同分析模型下，公牛育种值 (BLUP) 排序及截断选择响应 (truncation selection) 的差异，结果表明，当响应变量为二项分布时，GLM 分析具有最大的截断选择反应，与线性模型相比，选择效率可提高 12%。但对于多项分布的性状，GLM 与线性模型的估计和选择效率相当。Meuwissen 等 (1995) 和 Inge Riis Korsgaard 等 (2002) 的研究也得到了类似的结果。

目前广义线性模型主要用于离散性状的遗传分析，包括育种值估计 (效应预测)、方差组分及遗传参数的估计，而且分析的性状大多服从二项分布或泊松分布。除此之外，GLM 是否能渗入到动物遗传育种中的其它方面？答案应该是肯定的。近年来，随着统计基因组学的迅猛发展，对多基因控制的连续性性状实现 QTL (数量性状位点) 定位成为数量性状研究的热点，而对于分类性状由于其表型的不连续性，且不服从正态分布的假设，因此要实现分类性状 QTL 的准确有效定位，仍需要借助于非线性方法。这也将可能是今后一段时间内动物离散性状的一个研究重点。

1.4 高密度分子标记与遗传连锁图谱构建

标记-QTL 连锁分析要求有充分覆盖整个基因组的确定序列或多态标记座位的连锁图谱，从而能够在基因组的全范围内搜索 QTL。遗传图谱构建的理论基础是染色体的交换与重组，在细胞减数分裂时，非同源染色体上的基因相互独立、自由组合，同源染色体上的连锁基因产生交换与重组，交换的频率随基因间距离的增加而增大，因此可用重组率来揭示基因间的遗传图距。

现代的连锁图谱是随着遗传标记的发展而发展的，第一代遗传标记以限制性片段长度多态性

(RFLPs) 为代表; 其后发展起来的微卫星标记与 RFLPs 相比, 在基因组内存在的数量更多且多态性更高, 被称为第二代遗传标记; 单核苷酸多态性 (SNPs) 标记是最新发展起来的新一代遗传标记, SNPs 在基因组中的数量比微卫星大得多, 而且可以实现大规模自动化检测, 现在由于 DNA 芯片技术的发展, 使 SNPs 的应用前景更加广泛。发现新的标记是构建遗传图谱的主要工作, 随着分子生物学的发展和更多研究机构的参与, 新标记数量大幅度增加, Roslin 研究所 ARKdb 数据库 (<http://www.thearkdb.org/>) 2004 年的最新数据表明, 分布在猪, 鸡, 牛, 羊遗传图谱上的标记数分别已达到了 5109, 3400, 2209 和 14909。标记的不断丰富, 提高了遗传图谱的饱和度, 从而为 QTL 定位精确性的不断提高创造了前提条件。

DNA 分子标记技术的发展丰富了散布于基因组的标记资源, 也推动了人类、模式生物、以及动植物遗传标记图谱的构建。20 世纪 90 年代以来, 美、欧等发达国家先后启动了动物基因组研究计划: 欧洲经济共同体 (European Economic Community) 资助的欧洲 PiGMap 项目、美国农业部 (USDA) 资助启动的“国家动物基因组研究计划”(NAGRP, National Animal Genome Research)、美国农业部资助的肉畜研究中心 (USDA-MARC)、美国猪基因定位协作计划 (US Pig Gene Mapping Coordination Program)、英国 BBSRC (the Biotechnology and Biological Sciences Research Council)、英国的医学研究协会 (Medical Research Council, UK) 资助的鸡基因组定位计划等。同时, 广泛的国际合作也使得遗传连锁图谱不断更新。表 1-3 列出了部分家畜基因组上的总座位数、座位中的基因数和标记数、以及标记中微卫星数的统计结果。

表 1-3 牛、猪、鸡、山羊和绵羊基因组上座位数统计结果
Table 1-3 The number of loci scattering on genome of some animals

畜种	座位	基因	标记	微卫星
牛	4344	1554	2745	2402
猪*	2687	927	1739	1414
鸡*	2400	664	1706	1001
山羊	730	271	424	423
绵羊*	1722	370	1349	1118
马	1275	430	836	735

注: 表中数据引自 <http://locus.joucy.inra.fr>

其中, *数据引自: <http://www.thearkdb.org>

准确的、高密度的遗传标记图谱是实现精细定位 QTL 的前提。遗传图谱的质量取决于连锁群中遗传标记的数量、均匀分布程度和杂合度。基因间的交换受许多因素的影响, 如环境因素、性别、染色体异常、重组的基因控制等; 与此同时, 不同生物同一染色体的不同片段发生交换的频率也存在着差异, 因此得到的遗传图距也会存在差异。表 1-4 列出了部分家畜遗传图谱的构建情况。

表 1-4 部分家畜基因组或染色体上的遗传图谱

Table 1-4 Genetic map on genome or one chromosome of some animals

畜种	染色体号	座位数	覆盖基因组长度 (cM)	平均间距 (cM)	资料来源
牛	11	80	134.2	1.72	Casas 等, 2001
	27	39	75.5	1.98	Casas 等, 2001
猪	-	243	2561.9	11.4	Mikdwa 等, 2000
鸡	-	1889	3800	2.01	Groenen 等, 2000
	-	749	3569	4.8	Lee 等, 2002
绵羊	-	519	3063	5.9	De Gortari 等, 1998
	-	1062	3500	3.4	Maddox 等, 2001

在基因组内，标记与基因座位之间的遗传距离是以它们在配子形成过程中发生重组的概率，即重组率（recombination, r ）或图距（map distance, θ ）来度量。基因座位间重组率的取值在 0~0.5 之间，0 表示基因之间完全连锁，它们在配子形成过程中不发生重组，0.5 表示基因之间不连锁，它们在配子形成过程中完全自由组合；遗传图距的单位是摩根（M）或厘摩（cM），当两个座位之间的距离达到了可期望在一次减数分裂中它们之间发生 1 次染色单体互换时，称它们之间的距离是 1M。二者的关系可用图距函数来度量，但无论是何种图距函数，重组率是不可加的，而图距是可加的，即

$$\theta_{ac} = \theta_{ab} + \theta_{bc}, \quad r_{ac} \neq r_{ab} + r_{bc}$$

在基因位点间无干扰的理论假设下，Haldane（1919）根据 Poission 分布提出的作图函数为 $\theta = -\ln(1 - 2r)/2$ 。Kosambi 等（1994）在有干扰的情况下提出的经验转换公式如下：

$$\theta = 0.25 \ln\left(\frac{1 + 2r}{1 - 2r}\right)$$

这两种图距函数是有差异的，随着重组率的增大，这种差异也是不断加大的（如图 1-2 所示）。

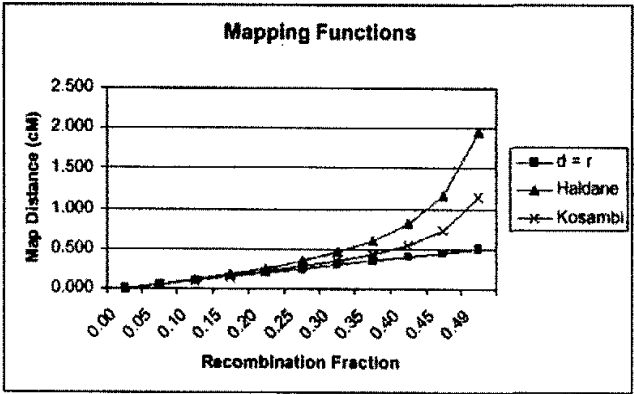


图 1-2 三种图距函数的差异

Fig. 1-2 The difference among three mapping functions

1.5 QTL 定位的试验设计与统计效率

数量性状基因座的定位是在质量性状基因定位的基础上发展起来的，但它与质量性状基因定位又有着本质的区别。QTL 定位主要是确定其在连锁图上的位置，定位的依据是遗传标记与数量性状之间的相关性，其基本步骤为：试验设计→数据收集→构建标记连锁图谱→标记 QTL 连锁分析或连锁不平衡分析及定位。影响 QTL 检测与定位的因素较多，贯穿于整个 QTL 检测与定位的全过程。其中，试验设计是 QTL 检测中至关重要的一环。

QTL 定位的试验设计主要是指资源群体的构建，资源群体是具有一定的系谱结构，具有丰富的 DNA 贮备，并对研究的目标性状作了详细表型测定的群体。由于在一般的家畜群体中标记与 QTL 可能处于连锁平衡状态，所以需要设计特定的资源群体，以产生足够的连锁不平衡。因此，试验设计的优劣，资源群体的结构直接影响着 QTL 检测的统计效力，也是 QTL 定位成功与否的关键所在。目前，用于动植物 QTL 检测与定位的试验设计概括起来主要有以下几种（图 1-3）。

畜禽 QTL 定位资源群体的设计大体上可分为两大类，一类是基于近交系或品系（远交系）杂交的试验设计，另一类是基于家系的试验设计。根据亲本交配方式的不同，基于近交系杂交、远交系杂交设计的有回交设计、 F_2 设计等。基于家系设计的有半同胞家系设计、全同胞家系设计、混合家系设计等。基于标记测定成本的考虑，Darvasi 和 Soller (1992,1994) 提出了选择性基因型测定 (selective genotyping) 和选择性混合 DNA 池 (selective DNA Pooling) 设计。

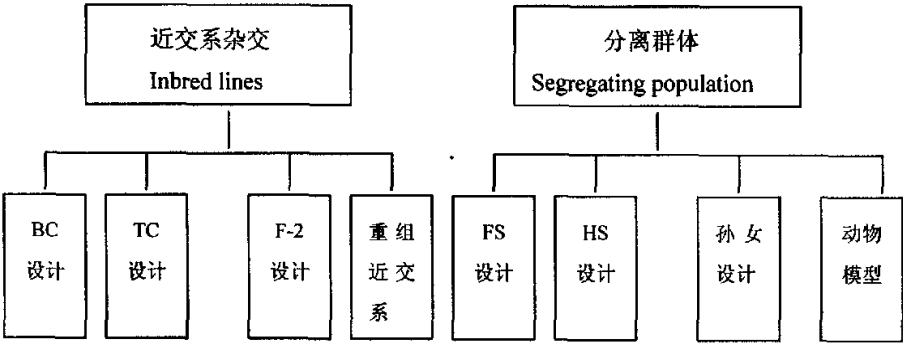


图 1-3 QTL 检测的试验设计

Fig. 1-3 Experimental designs for QTL detection

1.5.1 基于近交系、远交系杂交的试验设计

1.5.1.1 近交系杂交设计

通过近交系间的杂交来产生分离群体，主要用于植物、试验动物、小家畜及禽类的 QTL 定位。因近交系群体内标记与 QTL 基因型分别接近纯合，故可望产生标记与数量性状位点双杂合的 F_1 代，减少背景遗传方差，同时可望产生更大程度的连锁不平衡，从而保证一定的检测力 (Alfonso and Haley, 1998)。近交系中个体标记连锁相已知，所以可通过连锁分析估计标记间的重组率。其不足之处在于 (1) 紧密连锁的标记在分离群体中的重组事件很少，因此，高密度的连锁图谱只能提供有限的信息，除非分离群体中群体规模很大；(2) 许多研究材料是自然群体

的随机抽样，而非近交系的杂交后代。对 F1 实施不同的交配制度产生不同的分离群体，从而形成了不同的近交系统设计类型。利用回交后代的标记基因型信息和数量性状的表型信息为回交（BC）设计，F₁ 代个体互交产生 F₂ 代个体，利用 F₂ 代个体的标记基因型信息和数量性状的表型信息为 F₂ 设计，因此 F₂ 设计与回交设计的主要区别在于 F₁ 代的交配体系。此外还有重组近交系（Recombinant Inbred Lines）（将 F₂ 代个体再行近交以重新构建近交系）、重复回交（Repeated Back Cross）（回交后代再与某个亲本回交）、非亲本测交（TC）、四元杂交（Four-way Cross）等设计。

从有效信息来源上看，BC 与 TC 设计中，每个家系仅一个 F1 亲本提供了 QTL 检测的有效信息，而 F2 设计中两个杂合 F1 亲本都能提供有效信息。但 BC 设计的分离群体中每个个体都对 QTL 的检测提供信息，而 F2 群体中只有纯合型个体才对 QTL 加性效应的检测提供有效信息。在分离群体的基因型丰度方面，TC 与 F2 设计要优于 BC 设计。

Weller 等（2001）比较了不同试验设计下的相对效率（表 1-5），研究表明，在标记与 QTL 完全连锁的情况下，为获得相同的统计效力，F2 设计所需要的样本数最少，为 BC 设计的 1/2，为孙女设计或全同胞设计的 1/16。Soller 等（1978）推导出了基于 BC 或 F-2 设计的效率公式：

$$n = \frac{2(z_{\alpha/2} + z_{\beta})^2}{(\delta_n / \sigma)^2}$$

这个公式反应了统计效率参数（ z_{β} ）与标记内测定样本大小的关系，对于这个公式稍加变换可得： $z_{\beta} = \sqrt{n/2} \times \delta_n / \sigma - z_{\alpha/2}$ ，式中 z_{β} 为标准正态分布上 II 类错误的临界点，其上侧概率即为统计效力（Power）； $z_{\alpha/2}$ 为标准正态分布上 I 类错误的上尾临界点，如要达到 95% 的显著水平，则其上尾概率 $\alpha/2 = 0.025$ ， $z_{\alpha/2} = 1.96$ ； δ_n 为标记类型之间的期望差异，对于 BC 设计 $E(M_1M_2 - M_2M_2) = (1-2r)(d+a)$ ，对于 F2 设计 $E(M_1M_1 - M_2M_2) = 2a(1-2r)$ ； σ 为剩余标准差； n 为标记类型内的样本数。由统计效率的估计公式可以看出，QTL 检测的统计效率受诸多因素的影响。首先，样本大小和 QTL 的分离效应直接影响统计效率，样本愈大，QTL 的分离效应（ δ_n / σ ）愈大，则 z_{β} 也就愈大，统计效率就愈高；其次，可接受的 I 类错误（ α ）也直接影响 z_{β} 的大小， α 愈大，则 $z_{\alpha/2}$ 就愈小，相对来说 z_{β} 也就愈大，统计效率也会提高；再者，QTL 与标记之间的重组率（ r ）和所采用的试验设计均会影响 QTL 检测的统计效率（表 1-6）。

表 1-5 不同设计的期望 QTL 方差及获得相同统计效力的样本大小

Table1-5 Expected variance due to the QTL, and required sample size for different designs

设计类型 (Design)	F-2 (d=a)	F-2 (d=0)	BC (d=0)	全同胞 Full-sibs (d=0)	半同胞 Half-sibs (d=0)	孙女设计 Granddaughters (d=0)
QTL 方差 QTL variance	$3 a^2/2$	$a^2/2$	$a^2/4$	$a^2/8$	$a^4/8$	$a^2/32$
样本大小 Required sample size	1/3	1	2	16	4	16

注：a 为 QTL 的替代效应（ $a=0.5$ ）；d 为显性效应；样本大小与 F-2 设计相比。

表 1-6 统计效率为 90%时 BC 与 F2 设计所需的后代样本大小

Table1-6 The expectation of the contrasts, and required numbers of progeny to obtain statistical power of 0.9 for the BC and F2 designs. ($2a=0.282\sigma$, $\alpha=0.05$, $r=0$)

设计 design	对照 contrast (δ)	样本大小 sample size	显性程度		
			$d=-a$	$d=0$	$d=a$
Backcross	$(a-d)(1-2r)$	$2n$	525	2100	
F-2	$2a(1-2r)$	$4n$	1050	1050	1050

1.5.1.2 远交系杂交设计

常用于畜禽数量性状基因座的检测与定位，因畜禽群体中，近交系难以获得和维持，近交程度不够高时，亲本基因型得不到理想的纯合程度，从而很难利用近交系间的杂交来进行 QTL 的检测定位。这种设计实际上是利用两个极端品种或品系间的杂交，产生在所研究位点上相对杂合的 F1 代，之后可采用不同的交配体系，构建分离群体。如 Andersson 等（1994）利用欧美猪种与中国猪交配及利用野猪与培育品种杂交来进行 QTL 的定位是较为成功的实例。因为中国猪种在生长、繁殖、及胴体肉质性状方面与欧美猪种有着显著的差异，这样交配产生的群体有望在标记及数量性状基因位点等方面产生杂合子，从而取得类似近交系杂交的效果。

远交系内标记基因型相对复杂，通常假定两个远交群体内的 QTL 座位等位基因差异很小，而群体间的差异很大，从而形成极端差异。在此假设下，所有 F1 代在所研究性状上的 QTL 基因型相对杂合，对 F1 实施不同的交配，从而产生相应的分离群体，如 BC 群体、F-2 群体。由于杂交设计的假定条件在实际群体中很难满足，其统计效率和 QTL 的发现率也不如其它设计，因此，通过这种设计来进行 QTL 检测定位的成功率不是太高。

标记的多态信息含量（PIC）可作为比较两种设计效率的标准。PIC 是指当一个亲本为杂合，另一个亲本为任意基因型时，通过后代基因型能推断亲代基因型的概率。当两近交亲本具有不同的纯合标记基因型时，远交系的 F₂ 设计并不比近交系的 F₂ 设计效率更高，但当两近交系有一样的纯合标记基因型时（设此概率为 f ），远交系 F₂ 设计与近交系 F₂ 设计相对效率由 $R_y = (1-f)/PIC_K$ 决定。如果 $PIC_K > 1-f$ ，远交系 F₂ 设计效率更高； $PIC_K < 1-f$ 时，近交系 F₂ 设计效率更高； $PIC_K = 1-f$ 时，二者效率相等。因此，提高标记的 PIC 对提高远交系 F₂ 设计的效率至关重要，制造和维持高度近交的远交系需要很高的投入，畜禽中高度近交的远交系更是很少，远交系相比较而言可操作性更强。

1.5.2 基于家系的试验设计

这类试验设计是直接用常规的育种群作为资源群体，也称为远交群体（Outbred population）设计，是家畜 QTL 检测与定位的主要手段。在家畜中普遍存在的是分离群体，在数量性状基因定位方面，不同于近交系杂交。首先，近交群体检测数量性状基因位点是以群体间的差异为基础，而在远交群体中是以群体内的差异为基础。其次，近交群体中所有个体具有近乎一致的基因型和连锁相，由此构建的试验群体可造成遗传标记与 QTL 之间最大的连锁不平衡，且每个个体都能

为 QTL 的检测提供信息；而在远交群体中由于在各个家系内标记之间及标记与 QTL 之间的连锁状态是不同，具有不同的连锁相，同一位点存在着多个等位基因（复等位基因），只有部分个体提供信息（Y.Da, 1999; Zhang Qin, 1998），因而在分离群体中进行标记-QTL 的连锁分析时，其资源群体要以家系为单位。

远交群体中 QTL 的检测效力同样受多种因素的影响：（1）标记信息量的影响。单标记的信息含量可通过其多态信息含量（PIC, Polymorphism Information Content）衡量。当考虑多个标记时，还需要估计标记间的连锁相。（2）QTL 座位上的杂合程度的影响。当等位基因的频率过高或过低时（即杂合度较低）时，在该座位上发生分离的家系数会减小，从而影响 QTL 的检测。（3）QTL 基因频率和效应的影响。两等位基因的 QTL 的加性遗传方差为 $2p(1-p)a^2$ （其中 p 是基因频率， a 是替代效应），对于相同的加性遗传方差，基因频率接近 0.5 时，更可能检测到替代效应相对较小的 QTL。另外，替代效应越大，QTL 也越容易检测；（4）定位分析采用的群体规模也会影响 QTL 的检测。因此，可考虑采用后裔测定（Lander and Botstein, 1989）、多 QTL 定位等方法来降低环境效应的干扰或遗传噪音。

1.5.2.1 同胞对分析（Sib-pair analysis）

主要用于多个小全同胞家系的 QTL 检测与定位。该方法以分析全同胞对之间标记等位基因的 IBD（identical by descent）数目为基础（Fulker and Cardon, 1994），IBD 值可通过单标记或侧翼标记来推断。对于单标记连锁分析，估计模型为 $y_j = \alpha + \beta\pi_j$ ，其中 y_j 为两个同胞对表型值差异的平方； π_j 是同胞对之间标记等位基因 IBD 的函数，有三个取值 0, 1/2, 1，表示同胞对之间的 IBD 数目； $\alpha = \sigma_e^2 + 2\sigma_a^2$ ，这里 σ_a^2 为 QTL 的加性遗传方差； $\beta = -2(1-2r)^2\sigma_a^2$ 。对于双侧翼标记的 QTL 连锁分析，可根据 QTL 上的 IBD 对标记座位 IBD 的回归求得 QTL 上的 IBD 值为 $\pi_q = \alpha + \beta_1\pi_1 + \beta_2\pi_2$ ，这里， π_q 为 QTL 上的 IBD 值； π_1 和 π_2 为两个侧翼标记上的 IBD 值。

配子对设计在统计分析上不同于近交系杂交，其表型值的期望已表示为 QTL 遗传方差的函数，而非 a 和 d 的函数： $E(y_j|\pi_j) = (\sigma_e^2 + 2\sigma_a^2) - 2(1-2r)^2\sigma_a^2\pi_j$ 。大多数情况下，QTL 的加性效应远低于剩余方差，其统计效率也低于近交系杂交与同胞设计。另外，IBD 数目的推算也会受到各种因素的影响，直接影响着试验的统计效率。一般来说，在同胞对分析中要获得合理的统计效率，需要较大的测定样本，这无疑将增大试验的成本。

1.5.2.2 半同胞家系（Half-sib families）设计

主要用于半同胞家系的 QTL 检测与定位。对于较大的家系，同胞设计是理想的 QTL 检测手段。设计中以父系半同胞为资源群体，利用父系及其半同胞后代的标记信息和表型信息进行 QTL 分析，是肉牛和奶牛、羊等单胎动物 QTL 分析的常用方法。对于某些限性性状如产奶性状、产仔性状等，为了克服公畜不能提供表型信息的限制，从而相继产生了一些改进的试验设计，如孙女设计（grand-daughter design）和女儿设计（daughter design）等（Weller, 1990）。女儿设计是对双杂合公畜的女儿按标记类型进行分类，而孙女设计是对双杂子公畜的儿子给予标记，全面分析孙女的表型记录，在检测 QTL 存在时，孙女根据儿子的标记等位基因进行分类，然后进行均

值显著性检验。在孙女设计下,双杂合公畜 ($MmQq$) (M, m 为标记等位基因, Q, q 为 QTL 等位基因) 儿子的所有九种基因型中,只有五种基因型可以提供连锁信息,它们是 $MMQq, MmQQ, MmQq, Mmqq, mmQq$, 而其余四种 $MMQQ, MMqq, mmQQ, mmqq$ 在孙女中不能观测到重组,因而不能提供任何信息。

通常在半同胞分析中,只对公畜或女儿进行标记,而或略女儿母亲的标记信息,导致半同胞女儿中有部分为非信息个体,不能提供有效的标记信息,引起统计效率和定位精度的下降,同时 QTL 位置和效应的估计出现较大的偏差。Gomez-Raya 等 (2001) 研究表明,在考虑母亲标记信息的情况下,连锁分析的效率提高,重组率、LOD 值等参数接近无偏。因此,在采用同胞设计进行 QTL 分析时,应尽量利用全部后代的信息,以减小估计偏差。

在统计效率方面,孙女设计要高于女儿设计,对于同一性状 ($h^2 = 0.2$),要获得相同的统计效率,女儿设计的样本为孙女设计的 4 倍 (Weller and Soller, 1990),但在相同样本大小和 QTL 效应条件下,随着性状遗传力的提高,孙女设计的统计效率呈不断下降趋势,而女儿设计的统计效率则不断上升,因此孙女设计适合于低遗传力性状的 QTL 检测,而对于遗传力高的性状最好采用女儿设计。另外,无论是女儿设计还是孙女设计,随着家系群体的增大,其统计效率都是不断提高。宋九洲等 (1998) 通过反映连锁信息含量的期望 LOD 值 (ELOD) 和说明重组率估计精确性的 Fisher 信息测度,进一步对女儿设计和孙女设计的连锁分析效率进行比较,结果表明,在连锁紧密时,孙女设计的连锁信息要比女儿设计高,在精确性方面,孙女设计要比女儿设计高两倍。重组率大于 0.2 时,孙女设计在 ELOD 方面高于女儿设计,但在 Fisher 信息测度方面,孙女设计不如女儿设计,但这一不成为问题的主要方面。

1.5.2.3 全同胞家系 (Full-sib Families) 设计

主要用于大的全同胞家系的 QTL 检测与定位,如猪、猫、狗等多胎动物 (Lynch and Walsh, 1998)。将若干全同胞家系作为资源群体,利用亲本与其全同胞后代的标记信息和数量性状的表型信息来进行 QTL 的检测定位。

与半同胞家系一样,研究群体中可能有多个 QTL 的分离,另有部分个体为非信息个体,不能提供标记的有效信息。最好的情况是双亲具有 3~4 个不同的标记等位基因,这样后代可依据标记基因型划分为四个亚群,对这些亚群可采用 ANOVA 来检验其差异显著性。利用全同胞设计可以估计每个亲本的基因替代效应,也可用来估计总的遗传方差。

在人类中,只存在一些小的全同胞家系,由于缺乏足够的自由度,这种巢式设计的全同胞分析是不适用的,可采用同胞对设计来检测和定位 QTL。

1.5.2.4 配合动物模型分析 (Animal model analysis)

主要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。这里不再将 QTL 效应看作固定效应,而是作为一个随机效应来进行方差组分的估计。特别是在家畜群体中,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。

建立在动物模型基础上的 QTL 分析方法主要有 BLUP 法 (Fernando, 1989; Van Arendonk, 1994), Bayes 估计 (Hoeschele, 1993), 基于 Bayes 原理基础上的 Gibbs 抽样技术 (Hoeschele, 1996; Uimari, 1997), 以及基于混合线性模型的 ML 法、REML 法等。总的来看, QTL 随机效应模型主要有两种形式, 一是由 Fernando 和 Grossman (1989) 提出的随机配子效应模型, 另一种为 Goddard 等 (1992) 所采用的随机 QTL 效应模型。Van Arendonk 等 (1994) 对 Fernando 的随机配子模型作了修改, 以使之能处理多个标记的情况。

1.5.3 选择性基因型测定和选择性混合 DNA 池设计

无论是近交设计还是远交群体设计都需要大量的标记基因型测定, 尤其是当 QTL 效应较小时, 进行 QTL 的检测需要较大的群体, 标记成本增大。为了提高标记信息的统计效率, Darvasi 等 (1992) 和 Wang 等 (1994) 分别提出了选择性基因型测定与选择性样本 DNA 池设计。

选择性基因型测定 (Selective genotyping) 是从所研究的群体中选择高、低两种极端表型的个体构成一个子群体, 仅对这些子群体内的个体予以标记测定, 并用于 QTL 分析。虽然极端表型子群体的样本数远远少于原群体, 但其 QTL 定位的灵敏度和精确度却不亚于原群体, 因为在极端差异的子群体中, QTL 的效应通常被扩大。而且大多连锁信息位于极端表型个体上, 所以利用选择基因型测定可以提高 QTL 定位效率。特别是在所研究的性状较少时, 或所研究的性状为相关性状, 同时有大量的现有表型记录, 或表型记录测定成本较低的情况下, 选择性基因型测定的应用价值较高。但该方法也存在着一些缺陷, 首先它需要大量的表型测定, 对于表型测定存在困难或表型测定费用较高的性状, 应用选择性基因型测定的价值不大; 再者, 选择性基因型测定每次只能考虑一个性状, 因此, 当要研究多个性状的分离时, 这种方法也会受到一定限制, 最好与样本 DNA 池法配合使用。选择性基因型测定的统计效率可由下式给出:

$$z_{\beta} = \frac{N_g^{1/2} D_t}{2\sigma_t} - z_{\alpha/2}$$

式中, $z_{\alpha/2}$ 和 z_{β} 是 I 类与 II 类错误的标准正态分布临界点; D_t 为两尾样本性状的表型差异; σ_t 为两尾样本基因型内的表型标准差; N_g 为标记基因型的个体数。 D_t / σ_t 可近似表示为: $D_t / \sigma_t \approx \delta_n (1 + z_p i_p)^{1/2}$, 其中 δ_n 为随机标记下两种基因型之差的期望数; z_p 是上尾概率为 P 时的临界点值, $p = N_g / (2N_z)$, N_z 为表型测定的个体数; i_p 为选择强度, $i_p = \phi / p$, ϕ 为标准正态分布在 P 点处的密度值。

样本 DNA 池 (Sample pooling) 设计也称为“分离体分组混合分析法” (bulk segregant analysis, BSA), 由 Michelmore 等 (1991) 首次提出, 最初用于人类疾病性状的基因分析, 之后逐步移植到家畜数量性状的 QTL 分析 (Darvasi 等, 1994; Lipkin 等, 1998)。样本 DNA 池有两种不同的构建方式: (1) 根据表型差异构池。该方法是依据目标性状的表型差异, 将个体分为极端的两组, 如有病和无病个体、高产与低产个体等。之后, 将各组个体的 DNA 进行等量混合, 形成两个 DNA 池, 由于这两个 DNA 池之间除了目标基因座所在区域的 DNA 组成存在差异之外, 来自基因组其它部分的 DNA 组成是完全相同的。所以通过检验两池间的 DNA 多态性, 在两池间表现出差异的分子标记即被认为与 QTL 连锁。(2) 根据侧翼标记基因型差异构池。这种方法适合于目标基

因已定位在分子连锁图上,但其两侧标记与目标基因之间相距还较远,需进一步寻找更为紧密连锁标记的情况。

目前,样本 DNA 池主要是通过个体的表型差异来构建的。与选择性基因型测定一样,样本 DNA 池设计适合于所研究的目标性状较少,同时表型测定较为方便且成本较低的情况。样本 DNA 池设计的局限性在于以下三个方面:首先,个体 DNA 等量混合存在技术误差;其次, DNA 池设计在分析上主要是通过 PCR 扩增电泳条带的强度(亮度)来估计标记基因的频率,这种基因频率的估计方法也存在着较大的技术误差;再者,对 QTL 进一步定位的统计方法有限。样本 DNA 池设计的统计效率可由下式给出 (Darvasi and Soller, 1994):

$$Z_{\beta} = \frac{0.5 - \Phi(Z_p + \delta_n/2)/2p}{[0.25/(2pN_p) + V_{\pi}/2]^{1/2}} - Z_{\alpha/2}$$

式中, $\Phi(\bullet)$ 是标准正态分布函数; N_p 为表型测定个体数; V_{π} 为试验误差的方差。Darvasi 等还比较了在相同统计效率下完全基因型设计、选择性基因型设计及样本 DNA 池设计的基因型测定数,由表 1-7 不难看出,选择性基因型测定和选择性混合 DNA 池设计都大大降低了基因型测定的数目。尤其是选择性混合 DNA 池设计,比选择性基因型测定更适合多个性状的检测分析。选择性基因型测定设计在同时分析多个性状时,将按照性状值对个体进行归类,这样,几乎所有的个体都将进行标记测定,因此与杂交设计、远交群体家系设计等完全基因型测定的设计等同,选择性基因型测定丝毫起不到降低试验费用的作用,而选择性混合 DNA 池设计仅是增加了 DNA 池而已。如表 1-7 所示,即使对 10 个性状同时进行分析,基因型测定的数目只需要 20 个,当每个标记有 4 个重复时也仅为 80 个,并且统计效力并未下降太多,与完全基因型测定的 500 次相比,选择性混合 DNA 池设计的效率非常明显。

表 1-7 不同设计试验效力和所需基因型测定数目

Table 1-7 Example of experimental power and number of genotypings required of different experimental

试验设计类型	统计效力	基因型测定数目
完全基因型测定	0.80	500
选择性基因型测定 ^a	0.77	250
选择性混合 DNA 池		
$\sigma_x=0.0$	0.74	2
$\sigma_x=0.04$	0.50	2
$\sigma_x=0.04^b$	0.67	8

注: 群体大小 $N=500$, 基因型测定群体效应 $\delta=0.125$ 和显著性水平 $\alpha=0.05$, σ_x : 试验误差的标准差^a 每尾 25%,
^b 每个标记 4 个重复。Note: presented for the case of total population size $N=500$, genotypic group effect $\delta=0.125$ and type I error $\alpha=0.05$. σ_x : degree of experimental error standard deviation^a Twenty- five percent at each tail. ^b Analyzed with four replications per marker

1.6 QTL 检测与定位方法

Sax (1923) 和 Thoday (1961) 先后对 QTL 定位的基本思想进行了阐述: 如果在控制某性状的基因附近存在多态性 DNA 标记, 那么标记的传递在一定程度上反应了基因的传递情况, 标记的分离可用于检测和估计与之连锁的 QTL, 从而实现 QTL 的遗传定位。因此, 采用合理的试验设计, 利用经典遗传学中基因的重组交换原理, 通过分析 DNA 标记变异和性状表现变异二者间的关联可实现 QTL 定位, 包括估计 QTL 的位置和效应。

分子标记技术和相关统计方法的发展, 有力地促进了 QTL 定位研究的发展。目前借助分子生物学技术进行 QTL 检测的方法主要有两类, 一类是标记-QTL 连锁分析 (marker-QTL linkage analysis), 也称基因组扫描 (genome scanning); 另一类是候选基因分析 (candidate gene approach)。

1.6.1 候选基因分析

候选基因分析是从一些可能是 QTL 的基因中筛选 QTL。对 QTL 最有效的利用是直接测定 QTL 基因型, 而不是通过与之连锁的标记去推测, 因而在基因定位的基础上, 我们希望能够找到相关的基因或引起基因功能发生改变的突变。候选基因是已知在性状的发展或生理过程中具有某种生物学功能并经过测序的基因, 可以是结构基因或者是在调控或生化路径中影响表达的基因。候选基因分析的基本方法可概况如下: (1) 选择候选基因。根据掌握的生物学和生理学知识选择; 另一途径是选择在人类或其它物种中发现的具有较大效应的突变基因, 因为这些基因可能对我们所研究的家畜的相应性状也有重要作用。此外, 还可以根据标记-QTL 连锁分析的结果, 在可能存在 QTL 的区域去选择已经定位的基因, 这些基因将有较大的可能性成为 QTL; (2) 获得用于扩增基因的引物序列。可从现有的基因数据库中去寻找特定基因的序列信息, 如果数据库中没有所研究的畜种或与之相近的物种的相应基因的序列信息, 可从几个其它物种的序列信息中找出一个比较一致的序列, 再根据它来设计引物。如果这些物种的序列没有一致性, 则需要用特定的技术为所研究的物种测定这个基因的序列并设计引物; (3) 揭示候选基因内的多态性。在一个基因内部通常会有一些中性的多态位点, 这些位点一般会与决定候选基因功能的功能位点之间存在很强的连锁不平衡, 因而这个基因内的任何多态位点和它的调控序列对于候选基因分析都将很有用的。用于研究基因内的多态性的技术很多, 目前常用的主要是基因 PCR 技术的方法, 如 RFLP 和 SSCP 等; (4) 建立对所揭示的多态位点进行大规模的测定基因型的实用方法。目前常用的主要是基因 PCR 技术的方法; (5) 选择用于进行候选基因分析的群体。这个群体要用来获得 DNA 样品和表型信息并研究候选基因与性状的关系。原则上, 任何群体都可用于候选基因分析, 可以是正在进行常规育种的商业化育种群, 也可是一个未经选择的地方品种群体。但如果用两个近交系杂交所得到的群体则有一定缺点, 因为在这些群体中, 基因之间易处于高度连锁不平衡, 这样就很难确定所发现的性状与候选基因之间的关联是由于候选基因本身引起的, 还是由于另一个与候选基因连锁的基因引起的; (6) 研究候选基因与性状的关系。候选基因内的各个多态位点与性状的关系, 也可研究整个候选区基因与性状的关系; (7) 证实所发现的候选基因与性状的关系。考虑到与之连锁的 QTL, 可在同一群体的多个世代中证实, 在同一品种的多个群体中证实, 以及在不同的品种中证实。

从目前所报道的畜禽数量性状主效基因来看,最具代表性的是影响猪产仔数的ESR基因(雌激素受体基因)。ESR基因是根据其在繁殖性状生理或生化过程中的作用来选择的。类固醇激素和它们的受体在繁殖的生理或生化过程中扮演着重要的角色,在作用的初期,可在靶组织细胞中检测到与激素特异性结合的受体蛋白。雌激素与妊娠有着密切的关系,雌激素通过与其受体结合来发挥作用,而且,该蛋白上的突变与早产和人的乳腺癌有关,该基因的转基因小鼠的生殖系统有明显的变化。基于上述现象,雌激素受体基因被作为猪窝产仔数的候选基因。Rothschild等(1996)用梅山猪与大白猪杂交的合成系进行了验证,结果表明,估计的ESR基因的加性效应为1.15头,在大白猪中,ESR基因对头胎总产仔数的加性效应为0.42头。在猪上,应激易感性、脂肪、采食、生长、肉质、窝产仔数以及毛色等性状的候选基因的生产应用试验表明,候选基因分析是一种有效的方法。

候选基因分析的优势主要表现在如下几个方面:(1)统计检验的效率较高。如果候选基因确实就是QTL,则检验出它是QTL的可能性较大;(2)应用范围广。只需要一个世代的群体,因而可用于任何可测定表型值的群体;(3)成本低,易实施。由于是对单个基因进行分析,而不是对整个基因组进行扫描;(4)便于应用。可直接用于遗传评定,可直接利用克隆和转基因等技术进行畜群改良。但是,候选基因分析只能用已知生物学功能的基因作为候选基因,因此无法找出那些生物学功能未知的QTL。另外,候选基因分析对QTL的验证比较困难,除非候选基因的效应非常大,否则很难断定候选基因就是QTL,很有可能我们所发现的候选基因的效应并不是它本身的效应,而是它与之相近的QTL之间的连锁不平衡所造成的。候选基因与QTL的关系有三种可能:候选基因就是QTL本身、候选基因与QTL紧密连锁、候选基因与QTL连锁不紧密或不连锁。候选基因与性状的关联分析,可为证明候选基因是否为QTL提供进一步的证据。一般来讲,证明候选基因就是QTL本身比较困难,但二者的高度关联至少可以证明候选基因可能与QTL紧密连锁。即使候选基因不是QTL本身,但如果二者紧密连锁,候选基因也可直接在群体中应用。除连锁之外,一些其它因素也会影响候选基因和性状间的关联分析。因此,候选基因方法尚需同QTL的位置信息结合,作为基因组扫描的第二阶段(Haley等,1998),从而在一定程度上消除连锁之外因素的影响。

1.6.2 标记-QTL 连锁分析

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(linkage analysis)是基因定位的主要策略之一。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,染色体上的两个座位间相距越远,发生重组的机率越高,两个座位在一起传给后代的机会就越多。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。通过对覆盖密度适当的遗传图谱中的遗传标记在群体中进行分型,以此找到与QTL紧密连锁的某一标记,从而确定该基因在染色体上的粗略位置,在此基础上再在该染色体区域使用覆盖密度更高的标记作深入的连锁分析或连锁不平衡分析,就可将QTL确定在较小的区域,这样,再利用物理图谱提供的该区域的染色体片断将基因精细定位或克隆。

通过标记-QTL 连锁分析,许多畜禽重要性状的 QTL 得到定位 (Georges 等, 1995; Zhang 等, 1998; Thomsen 等, 2001; Ashwell 等, 2001)。但目前连锁分析一般将 QTL 定位在一个约 10-20cM 甚至更大的片断上,这离目前分子克隆技术可以操作的定位要求还差很远。QTL 定位的关键是重组事件的发生次数,对于连锁分析,影响重组事件的两个主要因素为标记密度和群体规模。

目前,遗传连锁图谱中标记密度已达到 0.5~3cM,且随着新的 DNA 标记类型(如 SNPs)的出现,标记密度还可大幅度提高,随着新的测定技术(如 DNA 芯片)的发展,测定标记的速度加快,成本降低。因此一种观点乐观地认为,利用高密度的遗传连锁图谱即可作到精细定位,显然这是不全面的,高密度的遗传连锁图谱仅是精细定位的前提。但对于连锁分析,标记密度越高,并不意味着定位越准确。在用多元回归模型分析时,过高的标记密度将会增大 QTL 加性效应和显性效应估计值的方差,进而加大了统计推断的误差 (Rodolphe and Lefort, 1993)。Darvasi 等人 (1993) 的研究结果表明对于中等效应的 QTL,即使有较高的标记密度,由于缺乏重组事件的发生,也只能定位在一个较大的区间。因此对于连锁分析而言,遗传连锁图谱的密度不是越高越好,除非能观测到有效数目的重组事件,否则也只能是浪费。另外,也可以通过扩大资源群体来增加重组事件, Darvasi 和 Soller (1997) 提出了 QTL 定位置信区间与群体规模的关系式为 $CI_{95} \approx 3000/kN\alpha^2$, 其中 CI_{95} 为 95% 的 QTL 定位置信区间; k 为每一个体可提供信息的亲本数 (回交设计或半同胞设计: $k=1$, F_2 设计: $k=2$); N 为资源群体中有效个体数; α 是 QTL 基因替代效应。由上式可以看出,扩大定位群体规模不能有效的减小 QTL 定位的置信区间。

1.6.2.1 QTL 定位的线性方法

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, QTL 检测定位的效率与精度也大大提高。

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 (Weller 等, 1986); (2)以多标记为基础的区间定位 (Lander and Bostein, 1989) 和复合区间定位 (Jansen, 1994; Zeng, 1993, 1994)。以单一标记为基础的零区间定位方法,其效率随着标记与 QTL 之间实际距离的增加而下降,为了提高效率,每一染色体上就必须覆盖有更多的遗传标记。同时,在单标记分析中,由于 t 检验和回归分析无法区分 QTL 效应和标记与 QTL 的连锁程度,因而只能检测而不能定位 QTL。区间定位对其作了改进,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,找到 QTL 最可能的位置。较单标记而言,区间定位更大程度的利用了有效标记信息,从而降低了定位偏差,提高了 QTL 的检测力和定位准确性。但是,当多个 QTL 连锁时,区间定位不能正确估计 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) 或检测不出 QTL (连锁 QTL 间效应相互抵消) 的结果。解决方法之一是,检测到一个 QTL 后,用个体的标记基因型对表型进行校正,得到一个新的表型值,然后再进行区间定位,当 QTL 为加性作用时,该方法效果较好。对区间定位最有效的改进方法是将其与多元回归分析相结合。Zeng (1993) 的研究表明,多元回归中每个 QTL 的效应只被其两侧相邻标记所吸收,且与其它区间的 QTL 无关。因此可以用被检区间以外的部分或全部标记作为回归模型中的余因子来消除其它 QTL 或遗传背景对被检

区间的影响。在此基础上, Zeng (1994) 提出了多 QTL 模型 (multiple-QTL model) 和复合区间定位方法 (composite interval mapping), 其中复合区间定位方法由于直观性好、计算上易于自动化而被普遍接受和广泛应用, 已逐步取代了区间定位法。

远交群体的结构比较复杂, 有多个结构简单的小家系组成的群体, 也有包含多世代、具有复杂亲缘关系、缺失数据 (missing data) 的复杂群体。鉴于此, 相继发展了基于最小二乘的方法 (Haley and Knott, 1992; Martínez and Curnow, 1992)、基于最大似然分析的方法 (Lander and Bostein, 1989; Knott and Haley, 1992; Zeng, 1994; Georges 等, 1995; Mackinnon and Weller, 1995; Hoeschele 等, 1997)、非参数估计方法 (e.g. Kruglyak and Lander, 1995), 以及贝叶斯分析 (Hoeschele and Vanraden, 1993; Thaller and Hoeschele, 1996) 等一系列的 QTL 定位分析方法。

基于最小二乘的方法 如①单标记基因型的方差分析; ②多标记基因型的方差分析; ③对 QTL 条件概率的回归; ④ *Haley-Knott* 回归; ⑤性状表型值对标记型的回归等多种方法。Whittaker 等 (1996) 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。最小二乘法计算速度快, 容易做成标准统计软件包, 可以作为定位分析的第一步。

基于最大似然分析的方法 QTL 等位基因效应或基因型效应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 (Grignola 等, 1996; Grignola 等, 1997) 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。基于最大似然的方差组份分析方法可用于个体分析、大家系或是多个家系联合分析, 因为不要求 QTL 等位基因的数目和各等位基因的频率已知 (Xu and Atchley, 1995; Grignola and Hoeschel, 1997), 所以稳健性较好。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。多数情况下, 回归定位对 QTL 位置和效应的估计结果与最大似然法所得到的结果相近, 如果能将回归方法和基因型概率算法结合起来, 将媲美于 ML。但最大似然法能够利用整个系谱中的信息, 可处理更一般化的、相对复杂的系谱。

1.6.2.2 QTL 定位的非线性方法

在性状表型服从正态分布及方差不依赖于均值的假设前提下, 上面提及的各种 QTL 线性定位方法都能给出参数的一致性估计。而对于离散性状或阈性状, 由于其分布是非正态的, 多呈现二项 (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能效不高。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 (Tilquin 等, 2001)。相对于常规的数量性状, 离散性状的 QTL 定位就显得更具有挑战性。阈值模型 (Falconer 等, 1996) 的提出开辟了合理的离散性状遗传分析方法, 阈模型在思想上是基于潜在连续分布 (underlying liability) 的假设, 而在分析方法上则是基于广义线性模型 (Generalized Linear Model, GLM)。

目前, 用于离散性状 QTL 定位的非线性方法主要有三种: Bayesian 法、GLM 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, 特别是对于二分类性状的分析, 并取得了可喜的成果 (Nengjun 等, 2000; Heather 等, 2001; Christoph 等, 2001; Rao 等,

2001), 但这些非线性方法在动物分类性状的QTL定位中则少见报导。

广义线性方法 目前, 广义线性方法在离散性状QTL定位中的应用主要是在近交系杂交群体中单阈值二分类性状的研究方面。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 (1995) 利用阈模型首次对二分类性状的QTL参数进行了估计。在此基础上, 相继发展了许多有效的方法 (Xu & Atchley, 1996; Visscher等, 1996; Rebai, 1997; Xu等, 1998)。

上述的这些方法主要是基于近交系杂交的设计上发展的, 对于家畜来说有很多不适应的地方。Yi & Xu等 (1999) 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等人 (2000, 2001) 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分类离散性状 (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 (Rao & Xu, 1998; Spyrides-Cunha 等, 2000)。

贝叶斯分法 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 (1993) 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两等位基因时、单标记、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。

非参数方法 QTL定位中通常假设, 性状表型值服从正态分布, 在这一假设条件下, 可以通过简单的参数检验来检测QTL (例如, 单标记时采用 t 检验, 区间定位时采用LOD值)。对于多基因控制的离散性状来说, 解决这一问题的方法一是通过合适的数学转换, 将这些数据转化为近似正态分布, 但很难找到这种数学转化方法。另外一种途径是采用非参数方法定位QTL。Kruglyak等 (1995) 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。

1.6.3 显著性阈值与置信区间

QTL定位的统计分析中, 一个重要问题是如何确定合适的显著性阈值来判断是否存在QTL。QTL检测时, 经常涉及到多次假设检验的问题, 如多标记检测、多QTL检测、多性状QTL定位等。Lander and Botstein (1989) 和Kruglyak and Lander (1995) 介绍了不同标记密度下多标记检验时的校正方法。当标记密度足够低, 多次检验间相互独立时, 可采用Boferroni校正方法 (Simes, 1986)。

由于 QTL 定位中采用的检验统计量分布未知, 因此, 只能采用统计量的近似分布 (如似然率近似服从卡方分布)。但诸如样本大小、基因组长度、标记密度、QTL 效应以及数据缺失等因素均会影响统计量的分布。Churchill 和 Derge (1994) 提出用数据重排的方法打乱标记和观察值之间的关联, 得到符合零假设的数据。通过多次数据重排得到检验统计量的经验分布, 并利用经验分布估计显著性阈值。进行数据重排可得到两种水平的显著阈值: 比较水平阈值 (Comparisonwise threshold) 和试验水平阈值 (Experimentwise threshold)。

QTL 定位中, 得到的估计值的分布通常未知, 因而无法估计其置信区间。QTL 定位置信区间的确定对决定下一步的精细定位或者 QTL 在育种工作中的应用至关重要。Lander and Botstein (1989) 提出 LOD drop-off 方法, 即在估计的 QTL 位置左右各降低一个或两个 LOD 单位所得到的区间长度, 分别相当于 96.8 和 99.8% 的置信度 (Mangin 等, 1994)。Mangin 等 (1994) 和 Visscher 等 (1996) 证明, 在小群体或中等大小群体中, 由于检验统计量不服从卡方分布, LOD-drop off 方法得到的置信区间有一定的偏差。

另外一类方法是通过抽样来估计参数的置信区间, 如: 参数 Bootstrap 抽样、非参数 Bootstrap 抽样和 Jackknife 抽样。参数 Bootstrap 抽样中, 首先分析数据得到参数估计值, 然后, 根据参数的估计值和理论分布产生与实际数据具有相同样本大小的新样本, 并估计参数。多次抽样可得到参数的经验分布, 进而估计参数的置信区间。参数 Bootstrap 抽样要求参数的估计值和理论分布正确, 否则, 估计的置信区间与实际值会有很大差别。区别于参数 Bootstrap, 非参数 Bootstrap 是对实际数据进行有放回抽样, 得到与实际数据具有相同样本大小的新样本, 然后根据参数的经验分布估计置信区间。非参数 Bootstrap 抽样方法并不是严格的非参数 Bootstrap 抽样, 因为用样本估计参数时仍然需要假设参数的分布。Jackknife 抽样是将实际数据去掉一个记录, 然后对实际数据进行抽样获得新样本, 因此, Jackknife 抽样只能得到 N 个新样本 (其中 N 为实际数据的样本含量)。

Visscher 等 (1996) 通过模拟对 LOD drop-off 和 Bootstrap 两种方法进行了比较, 并对相关问题如影响置信区间的因素进行了阐述。置信区间和检验统计量间存在强烈的负相关, 置信区间的平均长度取决于群体大小和 QTL 效应, 而受标记区间的影响较小。

1.6.4 QTL 定位方法的改进与精细定位

连锁分析通常利用当前的 2~3 个世代的资料追踪染色体片段的遗传, 并通过发现群体中的重组事件来估计 QTL 与标记最可能的相对位置。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 (Darvasi 等, 1993)。所以, 连锁分析一般将 QTL 定位在一个约 5-20cM 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, 因为在 1cM 的基因组区间上, 平均会有 20~30 个基因 (Haley, 1998)。QTL 定位的目的是识别性状基因, 而家畜中识别基因的主要方法一位置候选克隆要求将 QTL 定位在更小的片段上。因此, 依据基因的图谱位置 (map location) 实现 QTL 克隆分离的定位克隆不仅要求目标基因的正确定位, 而且要求定位的解析度精细到目前分子克隆技术可以操作的程度 (例如 <0.1cM)。

目前, 在 QTL 精细定位方面主要有三种策略可供选择, 它们分别是 (1) 连锁不平衡 (linkage

disequilibrium, LD) 定位 (Kruglyak, 1997); (2) 传递不平衡检验定位 (transmission disequilibrium test, TDT) (Spielman 等, 1993) 和同源相同 (identical by descent, IBD) 定位 (Charlier 等, 1996)。三种方法均是基于连锁不平衡原理, 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 5~20cM 区间的特定区域内; ②在这个区域内有高密度的标记 (间隔 0.25~1cM)。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 (张勤, 2001)。

同源相同 (identical by descent, 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段, 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, 可作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, Michel Gorges 等人 (1999) 利用 IBD 方法将影响奶牛产奶性状的一个 QTL 定位在奶牛第 14 号染色体 5cM 的区间。Meuwissen 等人 (2002) 利用 IBD 将控制牛双胎率的 QTL 定位在不到 1cM 的区间内。此外与泌乳量有关的 *DGAT*, *GHR* 基因也分别用 IBD 定位得到了很好的结果 (Grisart 等, 2002; Winter 等, 2002)。

Spielman 等 (1993) 提出的基于家系分析的传递连锁不平衡 (Transmission disequilibrium test, 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, 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, 亦即精细定位。

1.7 研究的主要内容

现代动物育种中, 人们对离散性状的重视程度不断增大, 特别是随着家畜育种技术的进一步发展, 一些当前育种目标中的主选性状不断趋于选择极限, 进一步选择的进展甚微, 从而一些重要的次级性状受到育种学家们的极大重视, 特别是与经济效益密切相关的离散性状, 如能将其纳入到家畜的育种目标中将会获得巨大的育种效益。

离散性状和数量性状有着相同的遗传基础, 同样受多基因控制, 只是离散性状表现为不连续而数量性状呈连续分布。它们都具有一定的生物学意义和经济价值, 在畜禽生产中具有重要的意义, 同样受到动物遗传学家的重视。对于连续的数量性状, 在观察值服从正态分布的假设前提下, 用线性方法进行遗传评定、参数估计和 QTL 定位等遗传分析是较为合理的选择, 并可以获得合理的结果。而对于离散性状或阈值性状的遗传分析, 由于这类性状的表型分布为非正态, 线性模型的功效就显得不足, 甚至会导致错误的结果, 从而限制了线性方法在离散性状遗传分析上的应用。在此前提下, 家畜离散性状的遗传分析方法显得日益迫切。

广义线性模型 (GLM) 是一般正态线性模型的直接推广, 它适用于连续性状和离散性状的遗传分析, 特别是后者, 如属性数据、计算数据的统计分析。GLM 在分析方法上克服了线性模型对响应变量的正态假设, 另外, 广义线性模型的参数估计量具有大样本正态分布, 因而具有良好

的统计性质。因此，广义线性模型成为了当前统计学领域中的主流方法，其应用范围涉及众多研究领域。

本研究将广义线性模型引入家畜离散性状（阈性状）的遗传分析，在广义线性模型的框架内对家畜离散性状进行遗传分析，并对多基因离散性状的 QTL 定位进行模拟研究。通过对不同方法的效率比较，探索合理的离散性状遗传分析和 QTL 连锁分析方法。

第二章 家畜离散性状遗传分析的广义线性方法

2.1 前言

在动物遗传育种领域中,迄今为止主要的研究对象多为正态连续分布的生产性状,并获得了显著的遗传进展。对于这类性状,在观察值服从正态分布的假设前提下,用线性方法进行遗传评定及参数估计等遗传分析是较为合理的选择,并可以获得合理的结果。而对于离散性状(categorical traits)或阈值性状(threshold traits)的遗传分析,由于这类性状的分布为非正态的,线性模型的功效就显得不足,甚至会导出错误的结果。动物育种中的离散性状,在多基因控制的理论前提下,其观察值的分布是离散不连续的,观察值不能表达为遗传效应与环境效应的线性组合,从而限制了线性方法在离散性状遗传分析上的应用。另外,线性方法只能在表观尺度上对离散性状的观察值进行描述,而在表观尺度上离散性状的遗传效应与环境效应是相关的。尽管如此,动物遗传育种中的离散性状至今仍多用线性方法进行分析。

Wright (1934)的阈值模型为正确地进行多基因离散性状的分析奠定了基础,但合理的方法是在近年来才逐渐建立起来的。离散性状非线性遗传分析方法的导出,一方面是基于贝叶斯概型,将关于参数的先验知识和来自数据的信息加以综合利用,将参数看作随机变量,导出其后验分布,并使后验概率密度函数最大化,从而估计出相应的参数;另一方面就是基于广义线性混合模型(generalized linear mixed model, GLMM)的应用,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,以实现参数的无偏估计。

广义线性模型是一般正态线性模型的直接推广(McCullagh, 1989),它适用于连续性状和离散性状的遗传分析,特别是后者,如属性数据,计算数据的统计分析。它与典型线性模型的区别是其对随机误差的分布无正态要求。另外,广义线性模型的参数估计量具有大样本正态分布,因而具有良好的统计性质。因此,广义线性模型成为了当前统计学领域中的主流方法,其应用范围涉及众多研究领域。广义线性模型在动物离散性状的遗传分析中已达到了广泛应用(Christoph, 2001),包括个体的遗传评定、离散性状遗传参数估计及QTL定位等多个领域,并取得了理想的结果(Gianola, 1993; Meijering, 1985; Meuwissen, 1995; Inge Riis Korsgaard, 2002)。

现代动物育种中,人们对离散性状的重视程度不断增大,一方面是因为许多重要的经济性状如记数繁殖性状、与疾病、应激和生存有关的性状都属于离散性状;再者是由于许多选择试验(Tempelman, 1996; Rekaya, 2001)都证明了对离散性状的选择是有效的。特别是随着家畜育种技术的进一步发展,一些当前育种目标中的主选性状不断趋于选择极限,进一步选择的进展甚微,从而一些重要的次级性状受到育种学家们的极大重视,特别是与经济效益密切相关的分类性状,如能将其纳入到家畜的育种目标中将会获得更大的育种效益。在此前提下,家畜离散性状的遗传分析方法显得日益迫切,本研究模拟分析了常规线性方法(LM)、阈值性状方法(TM)和广义线性方法(GLM)在离散性状遗传分析中的相对效率及精度,旨在探索合理的离散性状遗传分析方法。

2.2 材料与方法

2.2.1 研究设计与数据模拟

本研究模拟的分析群体是全同胞-半同胞混合家系。考虑单个二分类性状（单阈值性状），根据 Wright（1934）的阈模型，阈性状 k 个非连续的表现状态分布，具有一个潜在的不可观察的连续分布（潜在变量）。潜在变量的遗传基础与常规的数量性状类似，所以在阈性状的模拟中，首先用混合遗传模型模拟潜在变量的分布，然后根据阈性状各类型的发生率（或潜在阈值），确定性状表型状态。

2.2.1.1 性状的遗传模型与表型状态的确定

单阈值两个表型状态的性状（如家畜的受孕与不受孕，单胎与双胞胎，对某种疾病的感染与不感染、发病与不发病等），其潜在连续变量（ z ）可用如下的线性模型描述：

$$z = \mu + u + e$$

其中 μ 为群体的均值， u 为个体加性遗传效应（多基因效应）， e 为随机环境效应。

在基础群中，假定个体育种值效应 u 服从正态分布 $N(0, \sigma_u^2)$ （ σ_u^2 为加性遗传方差）， e 服从 $N(0, \sigma_e^2)$ （ σ_e^2 为随机环境效应方差，设定为 1）。在给定性状遗传力 h^2 和 σ_e^2 的情况下， σ_u^2 的推算公式如下：

$$\sigma_u^2 = h^2 \sigma_e^2 / (1 - h^2)$$

对于只有两个类别的二分类性状来说，只有一个潜在的阈值（ τ ），当潜在变量（ z ）的取值大于该阈值时，性状表现（ y ）为一种状态，否则表现为另一种状态。

$$y = \begin{cases} 1 & z \geq \tau \\ 0 & z < \tau \end{cases}$$

基础群由 ns 头公畜和 nd 头母畜组成，基础群个体间无亲缘关系。假定基础群中状态 1 的发生率为 π ，并由此确定阈值 τ 。

后代的多基因效应为

$$u_i = 0.5u_i^s + 0.5u_i^d + m$$

其中 u_i^s 和 u_i^d 分别为个体 i 父亲和母亲的多基因效应， m 为孟德尔抽样离差，服从分布 $N(0, \sigma_m^2)$ ，并有

$$\sigma_m^2 = 0.25[(1 - f_s) + (1 - f_d)]\sigma_u^2$$

f_s 和 f_d 分别为父亲和母亲的近交系数。后代个体潜在变量和离散表型的产生同基础群，假定上下代间 σ_e^2 保持不变。

2.2.1.2 参数设置与组合

基础群大小（公畜 *ns*，母畜 *nd*）设定为三个水平（20，100），（40，200），（80，400），基础群中公母畜个体随机交配，每头公畜与 5 头不同的母畜交配，每头母畜产 10 个后代，这样每头公畜的后代数 50 头。群体中状态 1 的发生率（ π_1 ）设定三个水平 0.10，0.20，0.40，性状的遗传力（ h^2 ）也设定为三个水平 0.10，0.20，0.40（表 2-1）。

表 2-1 模拟参数的组合
Table 2-1 Cases simulated.

参数组合 parameter case	遗传力 heritability	基础群大小(头) basic population	发生率 π_1 incidence
1	0.10	120	0.10
2	0.10	120	0.20
3	0.10	120	0.40
4	0.10	240	0.10
5	0.10	240	0.20
6	0.10	240	0.40
7	0.10	480	0.10
8	0.10	480	0.20
9	0.10	480	0.40
10	0.20	120	0.10
11	0.20	120	0.20
12	0.20	120	0.40
13	0.20	240	0.10
14	0.20	240	0.20
15	0.20	240	0.40
16	0.20	480	0.10
17	0.20	480	0.20
18	0.20	480	0.40
19	0.40	120	0.10
20	0.40	120	0.20
21	0.40	120	0.40
22	0.40	240	0.10
23	0.40	240	0.20
24	0.40	240	0.40
25	0.40	480	0.10
26	0.40	480	0.20
27	0.40	480	0.40

这样可产生 27 种群体参数的组合, 每一参数组合的群体重复模拟产生 100 次, 利用不同的方法分别算出离散性状的方差组分及遗传力参数, 之后进行 100 次模拟结果的平均和相关统计量的计算。

2.2.2 分析方法

离散性状的遗传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, 在假设表型数据为连续分布的前提下, 采用线性的方法对性状进行遗传分析, 称之为线性分析思想; 另一种思路就是阈值概念 (Falconer, 1996), 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, 而离散的表型值与潜在的连续值是通过一系列固定的阈相联系。

2.2.2.1 线性方法 (LM)

在不考虑离散数据的实际离散性, 假设表型数据为连续分布的前提下, 对于离散性状或阈值性状可采用常规的线性模型进行遗传分析, 具体的分析模型如公式 2-1 所示。

$$Y = Xb + Zu + e \quad (2-1)$$

其中 y , b , u 和 e 分别为观察值向量、固定效应向量、随机效应向量 (个体育种值)、随机残差; X 和 Z 是相应的设计矩阵。 $V(u) = \sigma_u^2 A$, $V(e) = \sigma_e^2 I$, $Cov(u, e') = 0$ 。对应的混合模型方程组如下:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \hat{\alpha}A^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

其中, $\hat{\alpha}$ 是需要估计的方差参数 σ_u^2 与 σ_e^2 的函数。利用 REML 方法, 方差组分的 EM 算法迭代估计方程如下 (张沅等, 1993):

$$\begin{aligned} \sigma_u^2 &= [\hat{u}'A^{-1}\hat{u} + tr(A^{-1}C)\hat{\sigma}_e^2]/q \\ \sigma_e^2 &= [y'y - \hat{b}'X'y - \hat{u}'Z'y]/(N - r(X)) \end{aligned} \quad (2-2)$$

对于阈值性状用混合模型方程估计的育种值不再是最佳线性无偏估计, 因此用 REML 方法得到的方差组分估计值也不再具有 REML 估计值的性质。这时的 BLUP 方法和 REML 方法为 QBLUP (伪最佳线性无偏预测) 和 QREML (伪约束最大似然) 方法 (McCullagh & Nelder, 1989)。

2.2.2.2 阈值性状方法 (TM)

利用 χ^2 分析, 阈值性状遗传力的估计公式如下 (Falconer, 1996; Gianola, 1982 Foulley, 1992)。

$$h^2 = \frac{\chi^2 - (S-1)}{r_A[k - (S-1)]} \quad (2-3)$$

式中, r_A 是全群平均的亲缘相关 (由公式 2-4 计算); S 是全群总的公畜数; $k = N - \frac{1}{N} \sum_i n_i^2$;

χ^2 值可利用 $C \times R$ 列联表的 χ^2 独立性检验的简易计算公式求得。

$$\chi^2 = N \left[\sum_i \sum_j \frac{p_{ij}^2}{T_{R_i} T_{C_j}} - 1 \right]$$

$$r_A = \frac{\sum_i \sum_j n_{ij}^2 + \sum_i n_{i*}^2 - 2N}{4(\sum_i n_{i*}^2 - N)} \quad (2-4)$$

上式中, n_{ij} 为第 i 头公畜与配的第 j 头母畜的后代数, n_{i*} 为第 i 头公畜的后代数, N 为总的后代数, p_{ij} 为第 i 头公畜在第 j 个阈值水平下的后代数, T_{R_i}, T_{C_j} 分别为第 i 头公畜和第 j 个阈值水平的后代数。

2.2.2.3 广义线性模型方法 (GLMM)

广义线性模型定义为响应变量 (Y) 具有如下分布密度的指数分布族 (McCullagh & Nelder, 1989)

$$f_Y(y, \theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\} \quad (2-5)$$

相应的拟然函数为: $L = \log f = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$ (2-6)

式中, $a(\phi)$, $b(\theta)$, $c(y, \phi)$ 均为已知函数, θ 为典范参数 (canonical parameter), ϕ 为离差参数 (dispersion parameter)。响应变量 (Y) 的期望和方差可对 (2-6) 式求偏导数获得, 求解方程 $E(\partial L / \partial \theta) = 0$, $E(-\partial^2 L / \partial \theta^2) = E((\partial L / \partial \theta)^2)$ 得

$$E(Y_i, \theta_i, \phi) = \mu_i = b'(\theta_i)$$

$$Var(Y_i, \theta_i, \phi) = b''(\theta_i) a(\phi) = a(\phi) \left(\frac{\partial \mu_i}{\partial \theta_i} \right) \quad (2-7)$$

由 (2-7) 式可看出, 广义线性模型不再等同于一般线性模型, 其响应变量的方差不是一个常数 (σ^2), 而是一个依赖于均值的函数。在广义线性模型中, 通过反连接函数将条件均值 μ (conditional mean) 与线性预测 η (linear predictor) 相联系, 并通过方差函数将协方差矩阵 R

与条件均值相联系, 从而实现误差方差依赖于条件均值。这也是广义线性模型与一般线性模型的主要区别。概括起来, 广义线性模型主要由三项构成要素, 它们分别为线性预测 ($\eta = X\beta + Zu$)、连接或反连接函数和方差函数。常见分布的典型连接函数及方差函数参见表 1-2。

广义线性模型的参数估计方程可由多种方式导出, 由拉普拉斯算子 (Laplacian) 拟似然法, 可得到较为直观的固定效应与随机效应的估计方程:

$$\begin{pmatrix} X'H'R^{-1}HX & X'H'R^{-1}HZ \\ Z'H'R^{-1}HX & Z'H'R^{-1}HZ + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X'H'R^{-1}y^* \\ Z'H'R^{-1}y^* \end{pmatrix} \quad (2-8)$$

其中: $H = \partial\mu/\partial\eta'$, $R = \text{var}(y|u)$, $y^* = y - u + H\eta$, $\mu = h(\eta)$ 。这里, $h(\cdot)$ 为反联结函数, 要注意的是不同表型分布的反联结函数是不同的, 如二项分布一般采用 *logist* 或 *probit* 联结, 而对于泊松分布的性状表型一般可采用 *log* 联结。

方程 (2-8) 在形式上与混合线性方程组相类似, 对其稍加变换可以看的更为清楚, 这里记: $W = H'R^{-1}H$, $hry = H'R^{-1}y^*$, 则 (2-8) 可变形如下,

$$\begin{pmatrix} X'WX & X'WZ \\ Z'WX & Z'WZ + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X'hry \\ Z'hry \end{pmatrix} \quad (2-9)$$

与线性混合模型方程组一样, 广义线性方程组的求解需要通过多次迭代来完成。本研究采用 *Fisher* 迹法进行固定效应与随机效应的估计, *Fisher* 迹法的迭代估计方程为:

$$\begin{aligned} [X'HR^{-1}HX](\beta^{[l+1]} - \beta^{[l]}) &= X'H'R^{-1}(y - u^{[l]}) \\ [X'HR^{-1}HX](\beta^{[l+1]}) &= X'H'R^{-1}(y - u^{[l]} + H\eta^{[l]}) \end{aligned} \quad (2-10)$$

观察值的方差构成主要有两个组分, 一是随机效应方差, 另一是误差分布方差。与混合线性模型方差组分估计类似, 广义线性模型的方差组分估计可通过准似然函数 (quasi likelihood) REML 方法实现, 准似然函数如公式 (2-11) 所示。

$$ql(\beta, \sigma) = -\frac{1}{2} \ln|V| - \frac{1}{2} \ln|X'H'V^{-1}HX| - \frac{1}{2} (y^* - HX\beta)'V^{-1}(y^* - HX\beta) \quad (2-11)$$

式中, σ 为方差组分向量, $V = R + HZGZ'H'$ 。类似于线性模型的方差组分估计, 通过 REML 法可得到方差组分的近似估计。

$$0.5 \left\{ \text{Diag} \left[\frac{1}{\sigma_i^4} (q_i - 2\text{tr}(C^{ii})/\sigma_i^2) \right] + \left[\frac{1}{\sigma_i^4 \sigma_j^4} \text{tr}(C^{ij}C^{jj}) \right] \right\} \hat{\sigma} = 0.5 \left\{ \frac{\hat{u}_i' \hat{u}_i}{\sigma_i^4} \right\} \quad (2-12)$$

由公式 (2-10) 与 (2-12) 可对方差组分迭代求解 (McCullagh, 1989; Breslow 等, 1993; Gianola 等, 1983; Ramon 等, 1996; Donoghue, 2004)。

2.2.3 估计与预测效果的度量

遗传力估计的准确度 主要通过估计遗传力的偏差 (BIAS) 和均方误 (MSE) 来度量。

$$MSE = \frac{\sum_{i=1}^n (\hat{h}_i^2 - h^2)^2 / n}{h^2}$$

$$BIAS = \frac{\sum_{i=1}^n (\hat{h}_i^2 - h^2) / n}{h^2} \quad (2-13)$$

n 是每一参数组合重复的次数, $\hat{h}_i^2$ 为第 i 次重复的估计遗传力 ($i=1,2,3,\dots,n$), h^2 是遗传力真值。二者反映了遗传力估计的精度和准确性 (Gamal, 1999)。

育种值排序间的秩相关检验 对估计的个体育种值进行排序, 并与个体的真实育种值排序之间进行秩相关检验。秩相关 (rank correlation), 也称为等级相关和顺序相关, 是指以等级尺度或排队顺序度量的两个变量之间的相关, 秩相关程度越高, 则说明遗传评估的准确性就越高。常用的度量指标是 Spearman 秩相关等系数:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (2-14)$$

其中 $d_i = X_i - Y_i$, 为相对应的两个等级变量之差。

2.3 结果与分析

2.3.1 遗传力估计

各参数组合下，三种方法遗传力估计及估计标准误的结果见表 2-2 所示。可以看出，不同的估计方法及参数组合对估计遗传力的影响均较大。总体上看，对于离散性状采用 GLMM 方法的估计遗传力结果最接近真值，相对而言，LM 方法估计遗传力的效果不太理想，离真值的偏差也较大。用 TM 方法估计遗传力时，资料的变异程度对估计的结果影响较大， χ^2 值或者组间差异的变异直接影响遗传力估计的准确度，本研究中 TM 方法的估计遗传力结果多数介于 LM 与 GLMM 方法之间。

2.3.1.1 性状遗传力对估计结果的影响

性状遗传力和表型发生率较低时，三种方法的估计结果都有一定的低估现象，但是随着性状遗传力的提高，特别是当 $h^2 > 0.2$ 时，GLMM 方法在多数参数组合下出现高估现象；而 LM 和 TM 方法在不同的参数组合下大多没有这种高估的情况。在相同的表型发生率下，随着性状遗传力真值的提高，LM 与 TM 方法的遗传力估计的偏差有增大的趋势，特别是在性状表型发生率较高的情况下，而 GLMM 方法的估计偏差则相对稳定（表 2-2，图 2-1）。

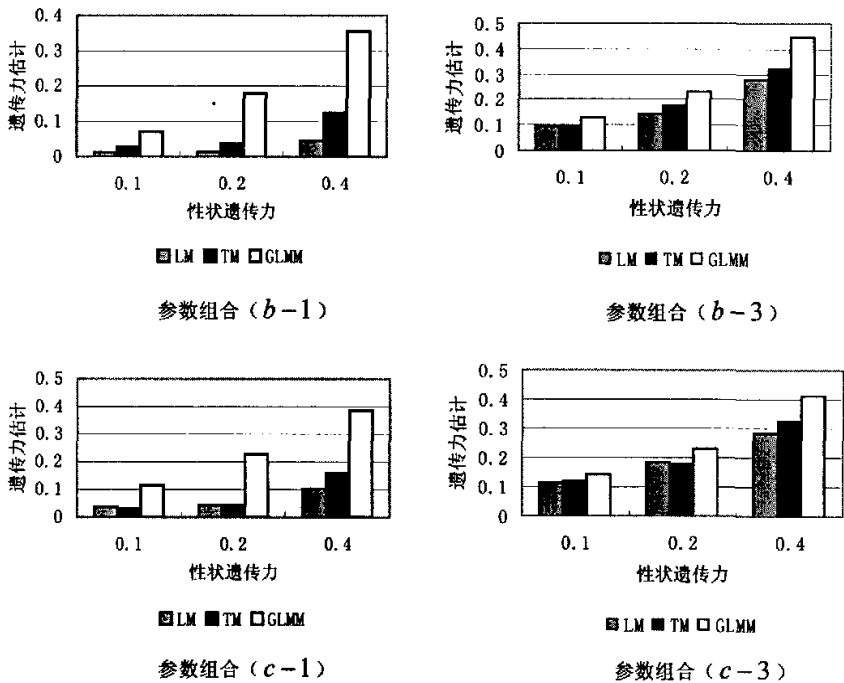


图 2-1 不同性状遗传力真值下的遗传力估计效果

注：b， c 代表基础群大小分别为 240 和 480；1， 2， 3 分别代表表型发生率为 0.1、0.2 和 0.4 情况

表 2-2 三种方法在不同参数组合下的估计遗传力及标准误

Table 2-2 Comparison of heritability estimates computed by three methods: linear model (LM); threshold method (TM) and generalized linear mixed model (GLMM).

参数组合 Case	LM		TM		GLMM	
	$\hat{h}^2$	$\sigma_{\hat{h}^2}$	$\hat{h}^2$	$\sigma_{\hat{h}^2}$	$\hat{h}^2$	$\sigma_{\hat{h}^2}$
1	0.0001	0.0000	0.0223	0.0078	0.0706	0.0115
2	0.0625	0.0036	0.0342	0.0186	0.1235	0.0892
3	0.1367	0.0412	0.1320	0.0369	0.1262	0.0756
4	0.0116	0.0024	0.0274	0.0024	0.0851	0.0082
5	0.0318	0.0089	0.0623	0.0176	0.1018	0.0801
6	0.0927	0.0112	0.0943	0.0212	0.1301	0.0769
7	0.0362	0.0041	0.0303	0.0033	0.1127	0.0432
8	0.0501	0.0108	0.0563	0.0902	0.1158	0.0667
9	0.1100	0.0318	0.1150	0.0275	0.1429	0.0858
10	0.0220	0.0076	0.0412	0.0152	0.1400	0.0667
11	0.0718	0.0130	0.0578	0.0326	0.1812	0.1020
12	0.1981	0.0528	0.1559	0.0779	0.2305	0.1101
13	0.0121	0.0074	0.0356	0.0057	0.1782	0.1164
14	0.1009	0.0524	0.0866	0.0236	0.2055	0.1032
15	0.1423	0.0861	0.1754	0.0527	0.2264	0.1204
16	0.0414	0.0066	0.0423	0.0096	0.2104	0.1018
17	0.1012	0.0526	0.0972	0.0774	0.2166	0.0892
18	0.1851	0.0811	0.1765	0.0420	0.2270	0.0796
19	0.1140	0.0133	0.1126	0.0381	0.3301	0.1327
20	0.2022	0.0769	0.1920	0.0765	0.3812	0.1415
21	0.3218	0.1012	0.2861	0.1325	0.4101	0.1567
22	0.0441	0.0239	0.1236	0.0253	0.3548	0.1671
23	0.1216	0.0550	0.1923	0.0369	0.4210	0.1855
24	0.2751	0.1006	0.3161	0.0837	0.4462	0.1549
25	0.1010	0.0544	0.1567	0.0239	0.3865	0.1212
26	0.1925	0.0827	0.2502	0.0566	0.4070	0.1347
27	0.2837	0.1535	0.3215	0.1876	0.4100	0.1413

注: 1. 遗传力估计值为 100 次重复的均值; 2. 各参数组合中的参数取值在表 2-1 中给出。

Note: 1. Each value of heritability estimates is the average of 100 replications. 2. Factors influencing the estimates for each case are defined in Table 2-1.

2.3.1.2 表型发生率对遗传力估计结果的影响

随着性状表型发生率的提高，三种方法的估计遗传力结果都有上升的趋势。相对于 GLMM 方法，性状表型发生率的提高对 LM 和 TM 的估计结果影响较大，遗传力估计的提高速度较快（表 2-2，图 2-2）。在一定程度上也说明 LM 和 TM 方法对性状表型发生率的依赖性较强，当性状表型发生率较低时（10%-20%），利用这两种方法估计遗传力的效果较差；而当表型发生率较高时（40%），LM 和 TM 两种方法与 GLMM 方法估计的效果相当，特别是在性状遗传力较低的情况下。相对而言，GLMM 方法对性状发生率的敏感性不是太强。

从本研究的结果看，在不同的表型发生率下，GLMM 的估计遗传力结果最接近真值，并且随着性状发生率的提高，GLMM 的遗传力估计出现偏高的趋势。

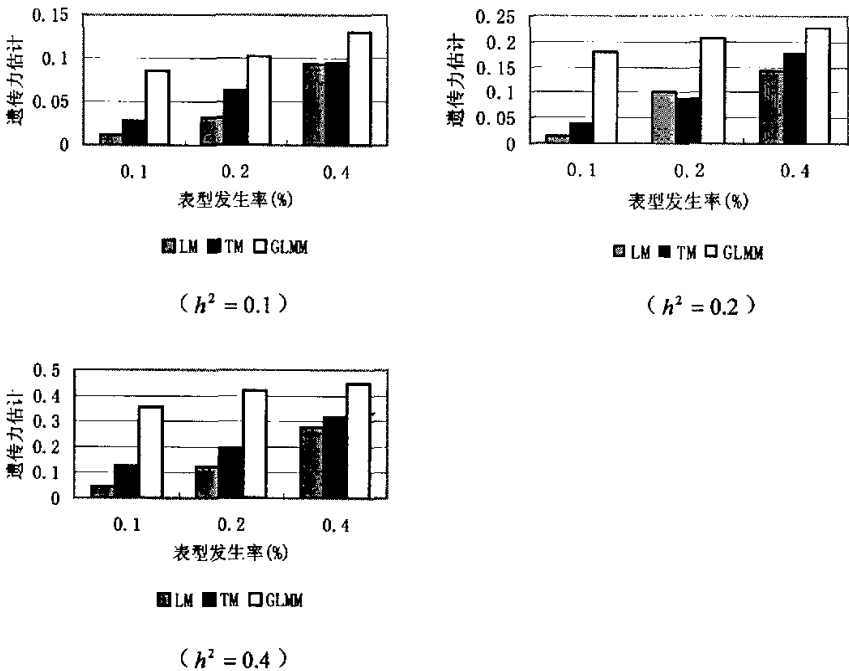


图 2-2 不同表型发生率下的遗传力估计效果（基础群大小为 240 的情况）

Figure 2-2 The estimation effect of heritability under different categories of incidence (population size is 240)

2.3.1.3 群体大小对遗传力估计结果的影响

一定大小的群体结构是遗传力估计的基础，群体过小会导致遗传力估计出现偏差。从本研究的结果来看，在多数组合下，随着基础群体的增大（120→480），LM、TM 和 GLMM 三种估计方法的效果均有一定程度的上升，但是影响的程度不是太明显。特别是在高遗传力（ $h^2 = 0.4$ ）或较高的表型发生率（ $\pi = 0.4$ ）情况下，基础群大小对各种方法遗传力估计的效果基本上没有影响，而且都能给出相对准确的遗传力估计（图 2-3）。

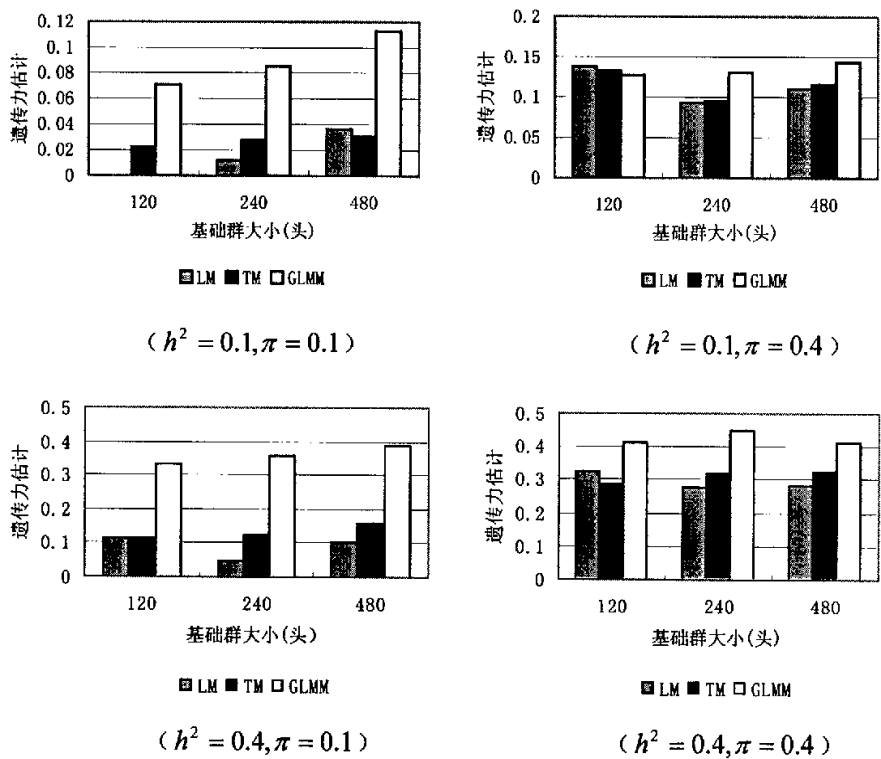


图 2-3 不同基础群大小下的遗传力估计效果

Figure 2-3 The evaluation effect of heritability from different population size

2.3.1.4 不同估计方法的比较

总体上看，在多数参数组合下，GLMM 方法的估计遗传力结果最接近真值，估计偏差也相对较小，显示了 GLMM 的良好统计特性。在 GLMM 分析中，连接函数采用对数连接 ($\mu_i = e^\eta / (1 + e^\eta)$)，方差函数为 $v(\mu_i) = \mu_i(1 - \mu_i)/n$ ，通过连接函数将响应变量观察值的期望与解释变量连接起来，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。

相对而言，利用常规的线性方法 (LM) 估计多基因离散性状遗传力的效果不太理想，离真值的偏差也相对较大，特别是在表型发生率较低的情况下。用 TM 方法估计遗传力时，资料的变异程度对估计的结果影响较大，本研究中 TM 方法的估计遗传力结果多数介于 LM 与 GLMM 之间。

LM 和 TM 方法的估计效果受表型发生率和性状遗传力的影响较大，而 GLMM 方法即使在低遗传力和低表型发生率的情况下也能获得遗传力的有效估计。随着性状遗传力和表型发生率的增大，LM 和 TM 方法的遗传力估计效果均有较大幅度的增加 (图 2-4)。

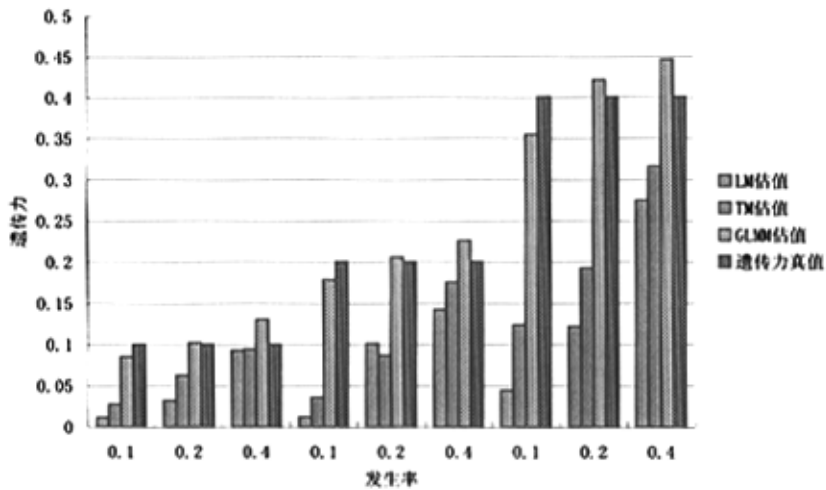


图 2-4 不同表型发生率下三种方法的估计遗传力差异（基础群大小为 240 的情况）

Figure2-4 Comparison of heritability estimates computed by three methods with difference incidence of categories. (the base population is 240)

2.3.2 遗传力估计的准确度

不同方法估计遗传力的均方误（MSE）和估计偏差（BIAS）见表 2-3 所示。MSE 和 BIAS 由 100 次重复模拟获得（参见公式 2-13），二者分别反映了不同方法遗传力估计的准确性和精度。

在多数参数组合下，GLMM 方法的估计偏差相对较小，也最接近真值，特别是在性状的表型发生率较低的情况下（如 $\pi_1 = 0.10$ ），GLMM 方法的相对优势较高。随着表型发生率的提高，由于 LM 与 TM 方法的遗传力估计准确度提高，因此 GLMM 方法的相对优势也有一定的下降。

伴随着性状表型发生率的增大，TM 与 LM 方法的遗传力估计偏差与估计均方误都呈下降趋势，而 GLMM 方法的遗传力估计偏差和估计均方误却有增大的趋势（图 2-5，图 2-6），这可能是由于 GLMM 方法对遗传力的高估引起。在性状表型发生率较高时（40%），TM 与 LM 方法的遗传力估计准确度与 GLMM 方法的相差不大，但在性状表型发生率较低的情况下，GLMM 方法的估计准确度明显高于前面两种方法。

在其它参数相同的条件下，性状遗传力的高低对 LM、TM 和 GLMM 三种方法的估计偏差及趋势影响不明显，但随着性状遗传力的增大（0.1→0.4），LM 与 TM 方法的估计均方误有明显增大的趋势。

可以看出，相对于 LM 和 TM 方法，GLMM 方法可以较为准确的估计出离散性状的遗传参数，但其相对优势受性状遗传力和表型发生率等多种因素的影响。

表 2-3 三种方法估计遗传力的均方误差及偏差

Table 2-3 Comparison of mean square error and bias of heritability estimates among three methods: linear model (LM); threshold method (TM) and generalized linear mixed model (GLMM).

参数组合 Case	LM		TM		GLMM	
	MSE	BIAS	MSE	BIAS	MSE	BIAS
1	0.0998	-0.9280	0.0604	-0.8123	0.0086	-0.3014
2	0.0141	-0.3655	0.0433	-0.6554	0.0055	0.2350
3	0.0134	0.3682	0.0102	0.3321	0.0069	0.2512
4	0.0781	-0.8812	0.0527	-0.7260	0.0022	-0.1528
5	0.0465	-0.7123	0.0142	-0.3817	0.0005	0.0116
6	0.0015	-0.0682	0.0003	-0.0570	0.0091	0.3201
7	0.0407	-0.6253	0.0486	-0.7152	0.0016	0.1273
8	0.0249	-0.4791	0.0191	-0.4327	0.0025	0.1528
9	0.0012	0.1023	0.0023	0.1520	0.0184	0.4290
10	0.1584	-0.7682	0.1261	-0.7940	0.0180	-0.3231
11	0.0822	-0.6476	0.1011	-0.7011	0.0018	-0.0952
12	0.0018	-0.0095	0.0097	-0.2412	0.0047	0.1525
13	0.1765	-0.9355	0.1351	-0.8229	0.0024	-0.0829
14	0.0491	-0.4911	0.0643	-0.5674	0.0015	0.0277
15	0.0166	-0.2885	0.0030	-0.1230	0.0035	0.1226
16	0.1258	-0.7931	0.1243	-0.7885	0.0054	0.0656
17	0.0488	-0.4934	0.0528	-0.5143	0.0014	0.0853
18	0.0011	-0.0745	0.0028	-0.1175	0.0036	0.1425
19	0.2045	-0.7520	0.2065	-0.7185	0.0122	-0.1747
20	0.0978	-0.5121	0.1082	-0.5239	0.0009	-0.0470
21	0.0153	-0.1955	0.0324	-0.2846	0.0025	0.0252
22	0.3167	-0.7897	0.1910	-0.6628	0.0052	-0.1146
23	0.1938	-0.6960	0.1078	-0.5092	0.0011	0.0653
24	0.0390	-0.3122	0.0176	-0.2107	0.0054	0.1169
25	0.2235	-0.7473	0.1480	-0.6082	0.0045	-0.0346
26	0.1076	-0.4875	0.0561	-0.3645	0.0005	0.0175
27	0.0338	-0.2773	0.0154	-0.2252	0.0025	0.0258

注：各参数组合中的参数取值在表 2-1 中给出。

Note: Factors influencing the estimates for each case are defined in Table 2-1.

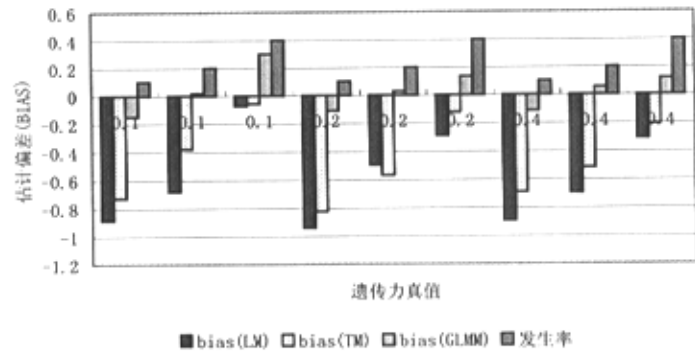


图 2-5 不同遗传力真值和表型发生率下三种方法的遗传力估计偏差(基础群大小为 240)

Figure 2-5 The bias of heritability estimates among three methods under the different true heritability and incidence of categories. (the base population is 240)

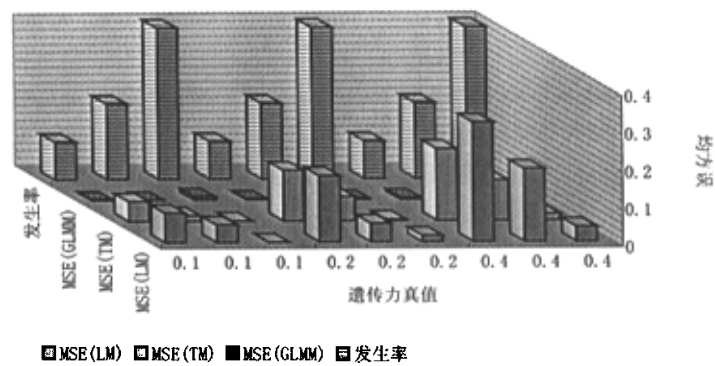


图 2-6 不同遗传力真值和表型发生率下三种方法的遗传力估计均方误比较(基础群大小为 240)

Figure 2-6 Comparison of mean square error of heritability estimates among three methods under the different true heritability and incidence of categories. (the base population is 240)

2.3.3 育种值排序间的秩相关

对每一参数组合重复模拟 100 次，并利用 LM 和 GLMM 两种方法分别估计基础群中的公畜个体育种值，并对公畜个体按平均育种值排序，计算所得的育种值排序之间的 Spearman 秩相关见表 2-4 所示。在 GLMM 分析中，连接函数采用对数连接 ($\mu_i = e^{\eta} / (1 + e^{\eta})$)，方差函数为 $v(\mu_i) = \mu_i(1 - \mu_i) / n$ ，通过 Fisher 迹法对公畜的个体育种值进行估计。排序的相对准确性定义为 GLMM-真值的秩相关与 LM-真值的秩相关之比，其值的高低反映不同遗传评定方法的相对优势和劣势。本研究采用的是全同胞-半同胞平衡设计，即每头公畜的后代数相同（50 个后代）。由于随机环境效应的作用，这里育种值的高低不能直接决定个体的表型状态。

表 2-4 不同估计方法下的估计育种值排序与真实育种值排序之间的秩相关及排序的相对准确性
Table 2-4 Rank correlations between true order and prediction order of breeding values using different methods (LM and GLMM).

参数组合 case	公畜育种值排序间的秩相关 rank correlations			相对准确性 relative accuracy
	LM -真值	GLMM -真值	LM-GLMM	
1	0.5324	0.8572	0.7921	1.6101
2	0.6874	0.8948	0.8275	1.3017
3	0.7332	0.9215	0.9240	1.2568
4	0.5823	0.8624	0.8044	1.4810
5	0.6832	0.8962	0.8331	1.3117
6	0.7157	0.9205	0.9210	1.2862
7	0.5623	0.8452	0.8435	1.5031
8	0.6928	0.9013	0.9061	1.3009
9	0.7036	0.9281	0.9427	1.3191
10	0.6092	0.8921	0.8733	1.4644
11	0.7135	0.9068	0.8924	1.2709
12	0.7610	0.9016	0.9061	1.1848
13	0.6184	0.9040	0.8832	1.4618
14	0.7126	0.9023	0.8951	1.2662
15	0.8132	0.9160	0.9511	1.1264
16	0.6863	0.9122	0.8917	1.3291
17	0.7128	0.9102	0.9241	1.2769
18	0.8549	0.9114	0.9237	1.0661
19	0.6914	0.9025	0.9012	1.3053
20	0.7431	0.9221	0.9114	1.2408
21	0.8762	0.9316	0.9621	1.0632
22	0.7228	0.9101	0.9109	1.2591
23	0.7544	0.9218	0.9226	1.2288
24	0.8845	0.9305	0.9617	1.0520
25	0.7336	0.9155	0.8957	1.2480
26	0.7642	0.9201	0.9175	1.2040
27	0.8928	0.9146	0.9623	1.0244

育种值排序的秩相关结果可以看出, LM 与 GLMM 方法的估计结果与真实育种值之间都有一定的差异, 这种差异主要是由于本研究中可利用的表型信息较少的原故, 但在实际的遗传评定中我们关心的不是个体的育种值本身, 而是个体育种值在群体中的排序结果。秩相关程度高说明估计的结果与实际育种值排序较吻合, 育种值预测的准确性也就较高。在实际应用中, 由于资料来源的复杂性, 还可以在模型中增加一些处理来对环境与管理等影响因子加以校正, 以达到更有合理的结果。

从个体育种值的排序秩相关检验结果来看, 在不同的参数组合下, GLMM 方法的估计效果都不同程度地优于 LM 方法, 说明 GLMM 方法在估计多基因离散性状育种值方面具有一定的优势。

从估计育种值排序的相对准确性来看, 常规 LM 与 GLMM 两种遗传分析方法的效果主要取决于性状的遗传力和表型发生率的高低(表 2-4)。随着性状遗传力和表型发生率的不断提高, GLMM 的相对准确性不断下降, 如当 $h^2 = 0.4, \pi = 0.4$ 时(参数组合 27), GLMM 方法的育种值估计效果与 LM 方法相当(相对准确性只有 1.0244)。在遗传力较低和表型发生率较小时, GLMM 方法的相对准确性最高(如参数组合 1 的相对准确性达到 1.6101), 说明对低遗传力和低表型发生率的离散性状的遗传分析方面, GLMM 方法的优势明显高于常规的线性方法, 遗传评定的相对准确性增加最大。

总的来看, GLMM 方法的遗传分析效果较常规的 LM 方法有一定的优势, 但其优势的大小取决于多种因素, 在育种实践过程中应根据现场和所评定性状的实际情况合理选择。

2.4 讨论与结论

目前, 对多基因控制的离散性状的遗传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, 在假设表型数据为连续分布的前提下, 采用线性的方法对性状进行遗传分析, 称之为线性分析思想; 另一种思路就是阈值概念(Falconer, 1996), 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, 而离散的表型值与潜在的连续值是通过一系列固定的阈相联系。离散性状非线性遗传分析方法的导出, 主要基于广义线性混合模型的发展应用。广义线性混合模型是一般正态线性模型的直接推广(Mccullagh, 1989), 它适用于连续性状和离散性状的遗传分析; 另外, 广义线性模型的参数估计量具有大样本正态分布, 因而具有良好的统计性质。

2.4.1 GLMM 方法的优势和影响因素

本研究对单阈值二项分类性状的遗传力估计结果显示, 在遗传力的估计准确度方面, GLMM 方法具有一定的优势, 这与前人的研究结果基本一致的(Rekaya, 2001; Gianola, 1983; Inge Riis Krosgaard, 2002; Meuwissen 等, 1995)。Meuwissen 等(1995)离散性状选择效率的模拟研究表明, 在一个封闭的核心群内对离散性状实施选择, 当考虑的性状为多等级性状时, 利用 GLMM 方法来替代 LM 方法可多增加 1~2% 的遗传进展; 但是当考虑的性状为二分类性状时, 利用 GLMM 方法则可多获得约 85% 的遗传改进, 因为在这种情况下, LM 方法可利用的表

型信息减少,而 GLMM 方法则可通过连接函数来实现响应变量的线性化。相对而言,LM 方法估计遗传力的效果较差;TM 方法的遗传力估计效果介于 LM 与 GLMM 之间。从公畜个体育种值的排序结果来看,在不同的参数组合下,GLMM 方法的育种值估计效果都不同程度地优于 LM 方法,能有效的给出公畜育种值的较为合理的排序,说明 GLMM 方法在进行多基因离散性状遗传分析上具有一定的优势,特别是对于只有两种表型的二分类性状。

性状的遗传力和性状的表型发生率对遗传力估计和个体育种值预测的效果均有直接的影响,随着性状遗传力真值的提高,GLMM 方法出现一定程度的高估现象,LM 与 TM 方法估计均方误(MSE)则有明显增大的趋势;随着性状发生率的提高,三种方法的遗传分析效果都有上升的趋势,特别是对 LM 和 TM 的估计结果影响较大。而随着性状表型发生率的提高,GLMM 方法的遗传力估计偏差及均方误则具有增大的趋势,这主要是由于 GLMM 方法对遗传力的高估引起。因此,在育种实际中应根据性状的性质(如遗传力和表型发生率)来合理的选择所要采用的遗传分析方法。对于具有较高遗传力和表型发生率的离散性状,可选用 LM 或阈模型方法进行分析,因为毕竟 GLMM 方法的计算成本较高。

2.4.2 GLMM 方法的扩展

虽然本文仅研究了单阈值的二分类性状,但对多于2个类型的多阈性状而言,GLMM方法的分析也是有效的(Tempelman, 1996; Gianola, 1983),只是在连接函数及方差函数的选择上有所差异,这主要取决于离散性状的表型分布(不同分布的连接函数和方差函数参见表1-2),例如,对于呈现 $poisson$ 分布的产仔数等性状可采用 log 连接函数来实现响应变量期望值的线性化,从而得到有效估计的目标(Thomson 等, 2003; Gianola, 1983)。动物育种中大多数的经济性状表型都是连续的,且服从或近似服从正态分布,而对于分类性状,其表型值呈现出离散分布,在动物育种中有几种重要的离散分布性状,它们分别是二项分布(binomial)、多项分布(multinomial)、泊松分布($poisson$)以及负二项分布(negative binomial)。如家畜中的生育能力、抗病或抗应激能力、成活或死亡等都服从二项分布或0-1分布,而有关母畜排卵数、产仔数、乳头数等记数性状则服从多项分布或者泊松分布。大量研究表明(Foulley, 1993; Olesen, 1994; 盛志康等, 1999),对于分类较多的多阈值性状,可以将其看作连续性状来处理,利用常规的线性方法可以获得与GLMM方法相当的参数估计效率和选择进展,特别是在阈值超过4个以上时。

2.4.3 GLMM 方法的应用前景

目前,广义线性模型主要用于离散性状的遗传分析,包括育种值估计(效应预测)、方差组分及遗传参数的估计等,而且分析的性状大多服从二项分布或者服从泊松分布,如家畜的疾病易感性、产羔数、分娩难易程度和肢蹄结构等性状。除此之外,GLMM 是否能渗入到动物遗传育种中的其它方面?答案应该是肯定的。近年来,随着统计基因组学的迅猛发展,对多基因控制的连续性状实现 QTL(数量性状位点)定位成为数量性状研究的热点,而对于具有相同遗传基础的多基因离散性状由于其表型的不连续性,且不服从正态分布的假设,利用传统的线性方法很难实现这类性状 QTL 的准确有效定位,仍需要借助于一些特殊的非线性方法。目前,用于

多基因控制的离散性状 QTL 定位的统计分析方法主要有以下几种: Bayesian 法、GLMM、STM (Structural threshold model) 和非参数方法等, 这些方法在人类及植物离散性状的 QTL 定位中已发挥了重要的作用。相对于其它几种方法, GLMM 方法更具优势, 因此, 将 GLMM 方法引入离散性状的 QTL 定位研究将可能是今后一段时间内动物多基因离散性状的一个研究重点。

2.4.4 可获得的初步结论

从本研究对二分类性状的结果来看, 可获得以下主要结论: 相对于常规的线性方法, 利用 GLMM 方法对二分类性状进行遗传分析具有一定的优势, 在遗传力的估计精度和公畜育种值预测准确性方面均明显的高于 LM 方法。但是, GLMM 方法的这些优势在很大程度上受到性状遗传力和表型发生率的影响, 随着性状遗传力和表型发生率的增大, GLMM 的优势不断下降。

同时应该考虑到这一结论是在本研究的特定试验设计中获得的, 对于其它群体、其它分布的性状仍需要进行深入的研究。

第三章 家畜离散性状 QTL 定位方法研究

3.1 前言

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遗传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 最为可能位置和效应大小,QTL 定位效率的高低主要取决于所采用的试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相继提出(Christoph, 2001; Weller, 2001; Zhang Qin, 1998; Lander, 1989; Knott, 1992; Haley, 1992; Gomez-Raya, 1997; Thaller, 1996),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。

目前用于 QTL 检测与定位的统计方法大多都是基于区间定位,在染色体的一个区间里寻找 QTL 最为可能的位置,较为常用的有最小二乘法、最大似然法以及贝叶斯方法等。这些方法的提出主要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,以上提及的各种 QTL 定位方法都能给出参数的一致性估计(Thaller, 1996; Weller, 2001; Yi, 1999; XU, 1996, 1998; Lunetta, 2000)。而对于不服从正态分布的多基因离散性状(discrete traits)来说,由于这类性状的表型特殊性,多呈现出二项(Binary)分布、多项(Multinomial)分布或者泊松(Poisson)分布,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低(Falconer, 1996; Gamal, 1999; Gianola, 1983),因此,采用线性方法对多基因离散性状(或阈值性状)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 定位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值性状的 QTL 定位就显得更具有挑战性。阈值模型(Wright, 1934; Falconer, 1996)的提出开辟了合理的离散性状(阈值性状)遗传分析方法,阈模型是基于潜在连续分布(underlying liability)的假设基础之上,认为在离散表型的背后存在一个潜在的无法观察的连续分布,离散的表型值与潜在的连续值是通过一系列固定的阈相联系。阈值模型在分析方法上则是基于广义线性模型(generalized linear model, GLM)(McCullagh, 1989),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,以实现参数的无偏估计。

阈模型在家畜的遗传评定中已发挥了重要的作用(Tempelman, 1996; Rekaya, 2001; Gamal, 1999; Gianola, 1983),在植物及人类的抗性性状 QTL 定位中也有大量的研究报道。Hackett 和 Weller (1995)利用阈模型首次对近交系杂交设计下二分类性状的 QTL 参数进行了估计,在此基础上,相继发展了许多有效的方法(Xu & Atchley, 1996; Thaller, 1996; Heather, 2001; Yi, 1999; Visscher 等, 1996; Rebai, 1997; Xu 等, 1998; Inge Riis Korsgaard, 2002)。但大多方法主要是基于近交系杂交设计(如 BC、F₂ 群体),由于家畜的群体结构较为复杂,需要估计的参数较多等原故,阈模型方法在家畜群体 QTL 定位方面的应用相对较少。Yi & Xu 等(1999)

提出了适应于多个大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 (2000, 2001) 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 (Rao & Xu, 1998; Spyrides-Cunha 等, 2000)。

现代动物育种中, 人们对离散性状的重视程度不断增大, 一方面是因为许多重要的经济性状如家畜对疾病的抵抗力, 单胎畜种的产仔数以及与生存有关的性状都属于离散性状; 再者是由于许多选择试验都证明了对离散性状的选择是有效的 (Tempelman, 1996; Rekaya, 2001)。特别是随着家畜育种技术的进一步发展, 一些当前育种目标中的主选性状不断趋于选择极限, 进一步选择的进展甚微, 从而一些重要的次级性状受到育种学家们的极大重视, 特别是与经济效益密切相关的分类性状, 如能将其纳入到家畜的育种目标中将会获得巨大的育种效益。在此前提下, 对家畜离散性状遗传机理的认知以及相应的遗传分析方法显得日益迫切, 本研究主要针对离散性状的 QTL 连锁分析方法进行了探讨, 在广义线性模型的框架内, 采用最大似然区间定位法对阈模型与一般线性模型的定位效率进行比较, 并对影响离散性状 QTL 定位效率的主要因素进行模拟研究, 以寻求高效的离散性状 QTL 定位方法。

3.2 材料与方法

3.2.1 研究设计与数据模拟

用于 QTL 连锁分析的试验设计为半同胞家系设计,产生的资源群体为半同胞群体。分别考虑了单阈值(二分类)和 3 阈值(等级)离散性状,类似于第二章的数据模拟方法,根据 Wright (1934) 的阈模型,阈性状 k 个非连续的表现状态分布,具有一个潜在的不可观察的连续分布(潜在变量)。潜在变量的遗传基础与常规的数量性状类似,所以在阈性状的模拟中,首先用混合遗传模型模拟潜在变量的分布,然后根据阈性状各类型的发生率(取决于阈值大小),确定性状表型状态。

3.2.1.1. 试验设计与资源群体的产生

利用在标记和 QTL 位点上均为杂合的公畜与来自随机平衡群体的母畜群(一般假定处于连锁平衡状态,且母畜间以及母畜与公畜间均无亲缘关系)交配产生半同胞信息个体。同时,本研究只考虑在标记与 QTL 位点上只有两个等位基因的基本情况,公畜的连锁相已知,当然公畜的连锁相可以通过后代的基因型进行推断。用 Monte Carlo 方法,产生连锁群中每个女儿从公畜获得标记与 QTL 位点上的基因,然后根据母畜群体中的标记位点和 QTL 位点的基因频率产生每个后代的基因型。模拟产生 10 头公畜,每头公畜分别与 50 头母畜交配,每头母畜有一个信息后代,这样资源群体的大小为 500 头。

3.2.1.2. 性状的遗传模型与表型状态的确定

根据潜在变量模型,离散的表型是由特定的阈值决定的。在 QTL 与潜在变量相关的前提下,潜在值可用下面的混合模型描述

$$z = x'\beta + g_{QTL} + g_{poly} + e \quad (3-1)$$

其中 z 为个体无法观察的潜在变量取值,为一连续变量; $x'\beta$ 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。 $g_{poly} \sim N(0, \sigma_{poly}^2)$ (σ_{poly}^2 为多基因加性遗传方差), $e \sim N(0, \sigma_e^2)$ (σ_e^2 为随机环境效应方差)。

性状的表型值由潜在变量中特定的阈值(ν)决定,高于或低于 ν 时会表现出不同的表型状态。设 y_i 为个体的离散表型取值,对于单阈值二分类性状来说 y_i 有两种可能的取值

$$y_i = \begin{cases} 1 & z_i \geq \nu \\ 0 & z_i < \nu \end{cases} \quad (3-2)$$

但对于 C 个多分类性状来说, 则有多个阈值 ($v_1, v_2, \dots, v_{c-1}$), 且有 $v_1 < v_2 < \dots < v_{c-1}$, 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。

$$v_{k-1} < z < v_k \Leftrightarrow y = k, k = 1, 2, \dots, c, \quad v_0 = -\infty, v_c = +\infty \quad (3-3)$$

对于上面的分类性状, 其表型呈现出二项分布或多项分布, 因此线性分析的基本假定不能够满足, 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。

在 QTL 与潜在变量相关的前提下, 潜在变量的遗传模型见 (3-1) 式。为了便于模拟研究, 这里没有将其它固定效应考虑在模型中。 σ_g^2 表示总的遗传方差 (包括随机多基因方差和 QTL 的方差贡献), Δ_q 为 QTL 加性遗传方差 (σ_q^2) 在总遗传方差 (σ_g^2) 中的比例。在给定性状遗传力 h^2 、 Δ_q 和 σ_e^2 的情况下, 可推出所需的方差组分:

$$\sigma_p^2 = \sigma_e^2 / (1 - h^2), \quad \sigma_g^2 = \sigma_p^2 * h^2, \quad \sigma_q^2 = \sigma_g^2 * \Delta_q, \quad \sigma_{poly}^2 = \sigma_g^2 - \sigma_q^2$$

QTL 的基因型值可根据模型 $\sigma_q^2 = 2p(1-p)a^2$ 推算 (Falconer, 1996), p 和 $(1-p)$ 分别为 QTL 的两个等位基因在群体中的频率, 本研究 p 的取值为 0.5; a 为基因替代的平均效应。在不考虑显性作用的前提下 ($d = 0$), $a = \sqrt{\sigma_q^2 / 2p(1-p)}$, QTL 的 3 种基因型 QQ 、 Qq 和 qq 的基因型值分别为 a 、0 和 $-a$ 。后代的多基因遗传效应为

$$g_{poly} = 0.5g_{poly}^s + 0.5g_{poly}^d + m \quad (3-4)$$

其中 m 为孟德尔抽样离差, 服从分布 $N(0, \sigma_m^2)$, $\sigma_m^2 = 0.25[(1-f_s) + (1-f_d)]\sigma_{poly}^2$, f_s 和 f_d 分别为父亲和母亲的近交系数。假定在每世代中 σ_e^2 保持不变。

3.2.1.3. 参数设置

为了便于探讨方法的优劣, 本研究中只考虑了双侧翼标记的情况, 双侧标记间 ($M_1 - M_2$) 及左侧标记与 QTL 之间 ($M_1 - QTL$) 的重组率 (r_{12}, r_1) 设定为三个水平 (0.30, 0.15), (0.20, 0.10), (0.10, 0.05), 它们带着 3 种标记密度; QTL 方差对总遗传方差的贡献 (Δ_q) 设定为三个水平 0.10, 0.30, 0.50; 性状的遗传力设定为 2 个水平 0.10, 0.40。

对于二分类性状, 群体中状态 1 的发生率 (π_1) 设定为两个水平 0.2, 0.50; 对于 3 阈值等级性状, 群体中 4 个等级状态的发生率 ($\pi_c, c = 1, 2, 3, 4$) 也分别设定为两个水平 ($\pi_1 = 0.20, \pi_2 = 0.60, \pi_3 = 0.15, \pi_4 = 0.05$) 和 ($\pi_1 = 0.20, \pi_2 = 0.30, \pi_3 = 0.30, \pi_4 = 0.2$)。

每一参数组合的群体重复模拟产生 1000 次, 分别对 QTL 位置参数和效应参数进行估计, 之后进行 1000 次模拟结果的平均。

3.2.2 连锁分析方法

标记—QTL 连锁分析采用两种不同的分析思想进行，第一种是常规的线性方法（LM），将离散的表型看作连续性状来处理（Weller 等，1995，2001）；第二种是采用广义线性模型框架内的阈模型方法（GLM），它是基于潜在连续分布假设基础之上的一种离散性状分析方法（Xu 等，1998；Rao & Xu，1998；Spyrides-Cunha 等，2000）。通过一系列固定的阈将离散的表型值与潜在的连续值相联系。

3.2.2.1. 线性方法（LM）

这种方法不考虑性状表型的实际离散性，将离散的表型看作连续变量来对待。采用的分析模型为（Weller 等，1995，2001）：

$$Y_{sjk} = \mu + a_s + Z_{j1}a_{s1} + (1 - Z_{j1})a_{s2} + e_{sjk}$$
 (3-5)

其中 Y_{sjk} 为第 s 头公畜第 j 个标记基因型内第 k 个后代的表型值； a_s 为公畜的多基因效应； a_{sk} 为 QTL 两个等位基因的加性效应（ $k = 1, 2$ ）； Z_{jk} 表示在标记基因型 j 内后代继承杂合子公畜第 k 个等位基因的条件概率（由表 3-1 给出），这一条件概率主要取决于重组率。上述模型可以再参数化为

$$Y_{sjk} = \mu + S_s + Z_{j1}\alpha_s + e_{sjk}$$
 (3-6)

这里 $S_s = a_s + a_{s2}$ ，为公畜组效应； $\alpha_s = a_{s1} - a_{s2}$ ，为加性 QTL 效应； μ 为群体均值，为了分析上的简便设定 $\mu = 0$ 。对于这一线性模型，可以利用常规的最大似然法进行相关参数的估计。

表 3-1 女儿设计后代的基因型及相对概率
Table 3-1 Genotype and relative frequency of progeny in the daughter design

公畜单倍型	后代基因型	基因型值	相对频率	标记基因型
sire gamete genotype	progeny genotype	genotypic value	relative frequency	marker genotype
M1QM2	Q-	$a_{s1} + a_\gamma$	$1 - r_1r_2 / (1 - r)$	M1M2
M1qM2	q-	$a_{s2} + a_\gamma$	$r_1r_2 / (1 - r)$	
M1Qm2	Q-	$a_{s1} + a_\gamma$	$(1 - r_1)r_2 / r$	M1m2
M1qm2	q-	$a_{s2} + a_\gamma$	$r_1(1 - r_2) / r$	
m1QM2	Q-	$a_{s1} + a_\gamma$	$r_1(1 - r_2) / r$	m1M2
m1qM2	q-	$a_{s2} + a_\gamma$	$(1 - r_1)r_2 / r$	
m1Qm2	Q-	$a_{s1} + a_\gamma$	$r_1r_2 / (1 - r)$	m1m2
m1qm2	q-	$a_{s2} + a_\gamma$	$1 - r_1r_2 / (1 - r)$	

3.2.2.2. 广义线性方法 (GLM)

多基因控制的等级离散性状, 其表现多呈非连续性变异, 但是又不服从简单的孟德尔分离比例, 遗传基础受多基因控制, 与常规的数量性状类似。在多基因控制的理论前提下, 一般认为在性状离散表型的背后存在一个潜在的连续变量 (z), 潜在变量可由公式 (3-1) 中的模型描述, 多等级的离散表型 (y) 是由一系列固定的阈值决定 (见公式 3-2 和 3-3)。

对于半同胞女儿设计, 在 QTL 位点上只有两个等位基因的假设前提下, 女儿的潜在变量为三元混合分布, 其分布密度为

$$f(z_j) = \sum_{i=1}^3 fre_{j(i)} \cdot f_{(i)}(z_j)$$

其中 $fre_{j(i)} = \Pr(G=i|m)$, 表示个体 j 在给定标记基因型下 QTL 基因型为 $i(i=QQ, Qq, qq)$ 的条件概率, 可通过给定的重组率进行推算 (表 3-2 给出); $f_{(i)}(z_j)$ 为三种基因型个体潜在变量的正态分布密度函数, 具有同质方差和不同的均值 ($a, d, -a$)。

假设有 N 个标记后代, Y_j, G_j 分别表示个体 j 的表型值和基因型值, 可以将 Y_j 转化为多变量向量, $W_j = [w_{j1} \ w_{j2} \ \dots \ w_{jc}]^T$, 其中

$$w_{jk} = \begin{cases} 1 & Y_j = k \\ 0 & Y_j \neq k \end{cases}, \quad k=1, 2, \dots, c$$

在给定个体 j 基因型值的条件下, W_j 将服从多项分布, 并有 $\Pr(w_{jk}=1|G_j) = P_{jk}$ 。这样个体 j 表型值向量 W_j 的似然函数为 $f_j(G_j) = \prod_{k=1}^c P_{jk}^{w_{jk}}$ 。由于个体的基因型值 G_j 无法观察, 因此需要借助 W_j 期望似然 (即用 3 种可能的基因型概率进行加权), 这样

$$f_j = \sum_{i=1}^3 pr(G_j = G_i) f_j(G_j = G_i) = \sum_{i=1}^3 pr(G_j = G_i) \prod_{k=1}^c P_{jk}^{w_{jk}} \quad (3-7)$$

其中, $G_1 = a, G_2 = d, G_3 = -a$, 分别为 3 种基因型值。在 (3-7) 式中我们需要解决两个方面的问题, 一是要弄清个体基因型值与 P_{jk} 之间的内在联系, 再者是如何去推断个体基因型的归属问题。

在 e 服从正态分布的假设前提下, P_{jk} 可以定义为

$$P_{jk} = \Phi(v_k - x' \beta - g_j - G_j) - \Phi(v_{k-1} - x' \beta - g_j - G_j) \quad (3-8)$$

其中, g_j 表示个体 j 的多基因效应, $\Phi(\cdot)$ 为标准正态分布函数。由于这种计算 P_{jk} 的方法要涉及到数值积分, 为了避免这点可以利用 Logistic 模型来近似 (Bonny, 1986),

$$\Phi(\xi) \approx \psi(c\xi) = \frac{\exp(c\xi)}{1 + \exp(c\xi)}, \quad \text{这里 } c = \pi / \sqrt{3}。 \text{ 这样转化后的 } P_{jk} \text{ 值为}$$

$$P_{jk} \approx \frac{\exp\{c[v_k - (x'\beta + g_j + G_j)]\}}{1 + \exp\{c[v_k - (x'\beta + g_j + G_j)]\}} - \frac{\exp\{c[v_{k-1} - (x'\beta + g_j + G_j)]\}}{1 + \exp\{c[v_{k-1} - (x'\beta + g_j + G_j)]\}} \quad (3-9)$$

个体 j 归属于三种 QTL 基因型的概率可通过已知的侧翼标记基因型来推断, 在给定侧翼标记基因型下, QTL 的条件概率是 $fre_{j(i)} = \Pr(G = i|M)$, 其中 M 为标记基因型的单倍型, 可以是双侧翼标记, 也可以是多个标记。前面已经提及这一条件概率, 它可以通过给定的重组率来推算 (表 3-2)。这样后代 j 的似然函数可表示为

$$f_j = \sum_{i=1}^3 pr(G_j = G_i|M) f_j(G_j = G_i) = \sum_{i=1}^3 pr(G_j = G_i|M) \prod_{k=1}^c P_{jk}^{w_k}$$

后代群体的似然函数 (N 个后代) 可表示为

$$L = \prod_{j=1}^N f_j \quad (3-10)$$

对于 (3-10) 式先求对数似然, 在给定参数 (v, u, β, g, G) 初始值的情况下, 引入隐型变量 P_{jk} , 通过 EM 算法可得到所求参数的最大似然值 (Nelder 等, 1965; McCullagh, 1989; XU, 1996; Rao 等, 1998)。并可获得此时的最大对数似然函数值 ($\ln L_{(\max)}$), 利用似然率 ($LR = 2(\ln L_{\max} - \ln L_o)$) 进行 QTL 位置的检验。

表 3-2 后代 QTL 基因型的条件概率 ($\Pr(G = k|m)$)

Table 3-2 The conditional probabilities of a QTL genotype with given marker genotype

公畜基因型	后代标记型	QTL 基因型		
		QQ	Qq	qq
M1M1M2M2	M1M1M2M2	$[(1-r1)(1-r2)/(1-r)]p$	$1 - r1r2/(1-r)(1-p) + [r1r2/(1-r)]p$	$(r1r2/(1-r))(1-p)$
	M1M1M2m2	$(1-r1)p$	$(1-r1)(1-p) + r1p$	$r1(1-p)$
	M1M1m2M2	$[r2(1-r1)/r]p$	$[r2(1-r1)/r](1-p) + [r1(1-r2)/r]p$	$[r1(1-r2)/r](1-p)$
	M1m1M2M2	$(1-r2)p$	$(1-r2)(1-p) + r2p$	$r2(1-p)$
	M1Qm2/m1qm2	$0.5p$	0.5	$0.5(1-p)$
	M1m1m2M2	$r2p$	$r2(1-p) + (1-r2)p$	$(1-r2)(1-p)$
	m1m1M2M2	$[r1(1-r2)/r]p$	$[r1(1-r2)/r](1-p) + [r2(1-r1)/r]p$	$[r2(1-r1)/r](1-p)$
	m1m1m2M2	$r1p$	$r1(1-p) + (1-r1)p$	$(1-r1)(1-p)$
m1m1m2m2	m1m1m2m2	$[r1r2/(1-r)]p$	$[r1r2/(1-r)](1-p) + [1-r1r2/(1-r)]p$	$[1-r1r2/(1-r)](1-p)$

注: r 为两标记间的重组率; $r1, r2$ 分别为 QTL 与左、右侧标记间的重组率, $r = r1 + r2 - 2r1r2$;

p 为母畜群中 Q 的频率

Note: r is the recombination fraction between the two markers; $r1$ and $r2$ is the recombination fraction between the putative QTL and each marker, respectively; p is the frequency of alleles in dam population.

对于二分类性状，上面的似然函数可简化为下式

$$L = \prod_{i=1}^N [\Pr(y_i = 1|m)]^{y_i} \cdot [\Pr(y_i = 0|m)]^{1-y_i} \quad (3-11)$$

其中， $\Pr(y_i = k|m)$ 为个体 i 在标记基因型 (m) 下表现为状态 k ($k = 0, 1$) 的条件概率

$$\Pr(y_i = 1|m) = \sum_{k=1}^3 fre_{i(k)} \cdot P_{(k)} \quad \Pr(y_i = 0|m) = \sum_{k=1}^3 fre_{i(k)} \cdot (1 - P_{(k)})$$

$P_{(k)} = \Pr(y_i = 1|G = k) = \Pr(z_i > v|G = k)$ ，表示个体的 QTL 基因型为 k 时表型值取状态 1 的条件概率。对 (3-11) 求自然对数，之后 $\ln L$ 分别对参数 a, d, v 求偏导数并设为 0，可得到参数 $P_{(k)}$ 的估计方程如下：

$$P_1 = \frac{\sum_{i=1}^N w_{i(1)} y_i}{\sum_{i=1}^N w_{i(1)}}, \quad P_2 = \frac{\sum_{i=1}^N w_{i(2)} y_i}{\sum_{i=1}^N w_{i(2)}}, \quad P_3 = \frac{\sum_{i=1}^N w_{i(3)} y_i}{\sum_{i=1}^N w_{i(3)}} \quad (3-12)$$

其中，

$$w_{i(1)} = \frac{fre_{i(1)} P_1^{y_i} (1 - P_1)^{1-y_i}}{\sum_{k=1}^3 fre_{i(k)} P_k^{y_i} (1 - P_k)^{1-y_i}}$$

$$w_{i(2)} = \frac{fre_{i(2)} P_2^{y_i} (1 - P_2)^{1-y_i}}{\sum_{k=1}^3 fre_{i(k)} P_k^{y_i} (1 - P_k)^{1-y_i}}$$

$$w_{i(3)} = \frac{fre_{i(3)} P_3^{y_i} (1 - P_3)^{1-y_i}}{\sum_{k=1}^3 fre_{i(k)} P_k^{y_i} (1 - P_k)^{1-y_i}} \quad (3-13)$$

在给定参数初值的情况下，可以利用上面的几组方程对 $P_{(k)}$ 进行迭代求解，得到 $P_{(k)}$ 的最大似然估计值，并可获得此时的最大对数似然函数值 ($\ln L_{(\max)}$)。利用似然比 ($LR = 2(\ln L_{\max} - \ln L_o)$) 进行 QTL 位置的检验。

3.2.3 显著性阈值的确定与检验功效

由于 QTL 定位中采用的检验统计量 (LR) 的分布未知，因此只能采用统计量的近似分布 (如似然率近似服从自由度为 1~2 的卡方分布)，但诸如样本大小、基因组长度、标记密度、QTL 效应以及数据缺失等因素均会影响统计量的分布。因此，本研究选用了在计算机水平上模拟产生显著性阈值的方法，在 H_0 假设下 (无 QTL 存在) 每个群体重复产生 1000 次，并计算

其最大似然统计量，在此基础上确定 0.05 水平的 LR 显著阈值。

检验功效 (Power) 定义为重复模拟过程中 QTL 检验统计量显著的次数占总模拟次数的百分比，Power 越高则说明方法较优。

置信区间 (CI) 估计采用 Lander and Botstein (1989) 提出的 LOD drop-off 方法，即在估计的 QTL 位置左右各降低一个或两个 LOD 单位所得到的区间范围，分别相当于 96.8% 和 99.8% 的置信度 (Mangin 等, 1994)。

3.3 结果与分析

3.1.1 QTL 参数的估计

估计的 QTL 参数主要有其效应参数 (a) 和位置参数 (r_1), 不同参数组合下的 QTL 参数估计结果见表 3-3、表 3-4、表 3-5 和表 3-6 所示。其中, 前两个表列出的是单阈值二分类性状在不同标记间距下的估计结果, 后两个表列出的是多阈值等级性状的结果。

3.3.1.1. 性状遗传力的影响

不同方法的 QTL 定位效果均受到阈值性状遗传力的影响, 性状的遗传力高低直接影响着 QTL 定位的效率和准确度。在本次模拟研究的各种参数组合下, 当模拟的性状遗传力较低时 ($h^2 = 0.1$), QTL 效应和位置参数的估计值与真值的偏差均较大, 随着性状遗传力的提高 ($h^2 = 0.4$), QTL 参数估计的准确性也不断上升, 不论是单阈值还是多阈值性状均出现这种趋势。

特别是在高遗传力和高的 QTL 方差贡献的情况下, 参数估计的结果最接近真值, 参数估计的标准差也相对较小。但是对于大多数的离散性状 (阈值性状) 来说, 其遗传力均不是太高 (多在 0.01~0.30 之间), 因此相对于常规的数量性状而言, 阈值性状的 QTL 连锁分析效果要差的多, 参数估计必然会出现一定程度的偏差。

3.3.1.2. QTL 效应的影响

QTL 的方差贡献对两种方法的 QTL 定位效果也有非常大的影响。QTL 方差贡献是 QTL 方差在整个遗传方差中所占的比率, 反映了 QTL 座位对性状表型的影响效应大小。因此, 与常规数量性状一样, 它会对连锁分析产生较大的影响, 随着 QTL 方差贡献的增大, 定位的准确性也会不断提高。

本研究设计了 3 种 QTL 效应大小等级, 由表 3-3~表 3-6 中的结果可以看出, 不管在何种参数组合下, GLM 或 LM 连锁分析的参数估计效果均有这种趋势。在 QTL 效应与性状遗传力较低的情况下 ($\Delta_q = 0.1$, $h^2 = 0.1$ 时), 参数的估计值与相应真值的偏差较大, 随着 QTL 效应和性状遗传力的增大, 参数的估计值与真值之间的偏差不断减小, 说明阈模型方法与常规的线性方法一样, QTL 定位的效率均受到其效应大小的影响。

3.3.1.3. 表型发生率的影响

性状的表型发生率对 QTL 参数估计的效果也有较大的影响。离散性状的区间定位比常规数量性状的效果要差的多, 其主要原因是由于离散的表型导致很多加性遗传信息的丢失, 而阈值的取值直接决定着遗传信息的丢失程度。性状表型发生率的大小主要取决于潜在变量中阈值的取值情况, 而性状的表型发生率又决定着潜在变量的遗传力。例如, 对于二分类性状, 当性状

的表型发生率为 50% 时, 潜在变量的遗传力达到最大, 而遗传力大小又直接影响着 QTL 定位的效果。因此, 随着性状表型发生率的增大 (10%→50%), QTL 定位的效率也相应增大, 参数估计值与真值之间的偏差减小。而当性状表型发生率较小时, 参数估计的偏差较大, 定位的准确性也下降。

相对而言, 性状的表型发生率对 GLM 方法的影响要比对 LM 方法的影响大。不论是单阈值还是多阈值离散性状, 随着表型发生率的增大, GLM 方法的 QTL 参数估计效果提高较快, 而 LM 方法的效果则不是太明显。虽然性状的表型发生率在自然群体中是很难把握的, 但可以通过人为精心的设计来调整资源群体中的表型发生率大小, 确保资源群体中的表型发生率较高, 以提高 QTL 定位的效力。

3.3.1.4. 标记间距的影响

在多标记连锁分析中, 标记间距反应了标记的密度。在双侧标记连锁分析中, 标记间距也可采用标记间的重组率来反映, 标记间的重组率愈小, 则标记间距也就愈小。本研究设计中设计了三种标记密度, 双侧翼标记间的重组率分别为 0.1、0.2 和 0.3。从 QTL 参数估计的偏差来看, 随着标记间距的减小, 参数估计值与真值之间的偏差也呈下降趋势, 说明在阈性状的连锁分析中标记密度的影响也是非常大的。

从似然率曲线变化趋势来看 (图 3-1), 标记间距越小, 曲线变化就越平稳, 但各重组率位置上的似然率 (LR) 则相对较大。似然率的增加必然导致 QTL 定位功率的增大, QTL 参数估计的准确性也会相应提高。

当 QTL 效应较大时 ($\Delta_q = 0.5$), 标记间距的大小对 QTL 位置参数估计和检测功效的影响相对较小, 因为较大 QTL 效应会带来 QTL 定位准确性的提高。在此情况下, 较小的标记密度也不会导致 QTL 定位效果的显著增加。但总的来看, 双侧翼标记间距的减小可以增加 QTL 定位的准确性。

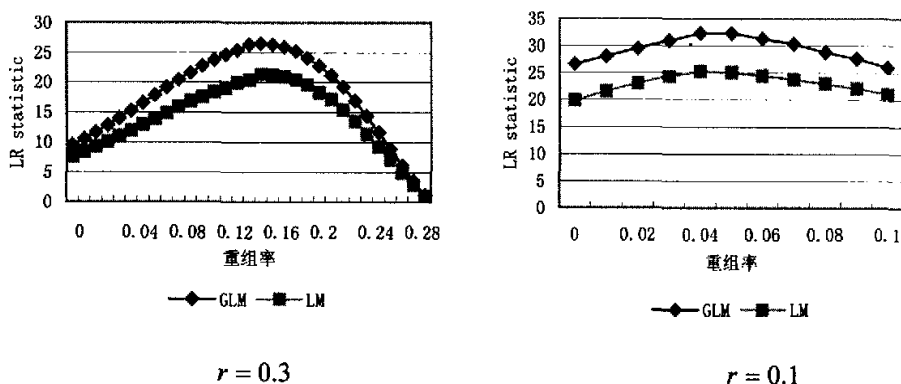


图 3-1 不同标记间距下的平均似然率曲线 (4 分类等级性状, $h^2 = 0.4, \Delta_q = 0.5, \pi = 0.2/0.3/0.3/0.2$)

Figure 3-1 The likelihood profile produced by averaging the analyses of 100 replicated simulations using two method under the different markers distance.

表 3-3 单阈值二分类性状在不同参数组合下的 QTL 参数估计效果 (表型发生率 $\pi_1 = 0.2$)

Table 3-3 Mean estimates of QTL parameters under different parameter case for binary traits ($\pi_1 = 0.2$).

重组率 (r)	遗传力	QTL 方差 贡献	QTL 参数	参数真值	估计均值 (标准误)	
					GLM	LM
$r = 0.3$	0.10	0.10	r_1	0.1500	0.141(0.021)	0.136(0.042)
			a	0.1414	0.1233(0.0876)	0.1349(0.1035)
		0.30	r_1	0.1500	0.143(0.036)	0.140(0.051)
			a	0.2449	0.1146(0.0923)	0.1275(0.1127)
		0.50	r_1	0.1500	0.143(0.036)	0.142(0.056)
			a	0.3162	0.1926(0.1245)	0.2147(0.1334)
	0.40	0.10	r_1	0.1500	0.140(0.051)	0.139(0.047)
			a	0.2828	0.2025(0.1343)	0.2163(0.1526)
		0.30	r_1	0.1500	0.144(0.062)	0.141(0.062)
			a	0.4899	0.3125(0.1212)	0.3236(0.1324)
		0.50	r_1	0.1500	0.143(0.058)	0.142(0.046)
			a	0.6325	0.5643(0.1426)	0.5762(0.1558)
$r = 0.2$	0.10	0.10	r_1	0.1000	0.084(0.024)	0.092(0.037)
			a	0.1414	0.1236(0.0835)	0.1210(0.0912)
		0.30	r_1	0.1000	0.091(0.031)	0.089(0.029)
			a	0.2449	0.1635(0.1124)	0.1727(0.1178)
		0.50	r_1	0.1000	0.092(0.013)	0.089(0.015)
			a	0.3162	0.2237(0.1313)	0.2416(0.1247)
	0.40	0.10	r_1	0.1000	0.092(0.028)	0.091(0.020)
			a	0.2828	0.2246(0.1227)	0.2115(0.0954)
		0.30	r_1	0.1000	0.094(0.024)	0.091(0.018)
			a	0.4899	0.3582(0.1524)	0.3742(0.1438)
		0.50	r_1	0.1000	0.093(0.016)	0.092(0.024)
			a	0.6325	0.5854(0.1235)	0.5827(0.1629)
$r = 0.1$	0.10	0.10	r_1	0.0500	0.045(0.007)	0.042(0.011)
			a	0.1414	0.1329(0.0926)	0.1212(0.0951)
		0.30	r_1	0.0500	0.045(0.008)	0.042(0.009)
			a	0.2449	0.1887(0.1228)	0.1851(0.1141)
		0.50	r_1	0.0500	0.047(0.013)	0.045(0.014)
			a	0.3162	0.2582(0.1223)	0.2413(0.1221)
	0.40	0.10	r_1	0.0500	0.042(0.015)	0.043(0.024)
			a	0.2828	0.2553(0.1167)	0.2419(0.1272)
		0.30	r_1	0.0500	0.045(0.017)	0.043(0.020)
			a	0.4899	0.3872(0.1263)	0.4038(0.1383)
		0.50	r_1	0.0500	0.046(0.015)	0.045(0.019)
			a	0.6325	0.5914(0.1448)	0.5725(0.1522)

注: r 为双侧翼标记间的重组率, QTL 与左侧标记间的重组率 (r_1) 为 r 的一半; a 为 QTL 的效应值; 括号内的数值为参数估计的标准误。下表相同。

表 3-4 单阈值二分类性状在不同参数组合下的 QTL 参数估计效果 (表型发生率 $\pi_1 = 0.5$)
Table 3-4 Mean estimates of QTL parameters under different parameter case for binary traits ($\pi_1 = 0.5$).

重组率 (r)	遗传力	QTL 方差 贡献	QTL 参数	参数真值	估计均值 (标准误)	
					GLM	LM
$r = 0.3$	0.10	0.10	r_1	0.1500	0.143(0.028)	0.135(0.036)
			a	0.1414	0.1277(0.1032)	0.1252(0.0760)
		0.30	r_1	0.1500	0.145(0.026)	0.142(0.048)
			a	0.2449	0.1426(0.0766)	0.1528(0.1230)
		0.50	r_1	0.1500	0.143(0.042)	0.141(0.042)
			a	0.3162	0.2323(0.1171)	0.2416(0.1228)
	0.40	0.10	r_1	0.1500	0.145(0.045)	0.133(0.037)
			a	0.2828	0.2346(0.1164)	0.1765(0.1375)
		0.30	r_1	0.1500	0.142(0.054)	0.143(0.058)
			a	0.4899	0.3973(0.1253)	0.4127(0.1260)
		0.50	r_1	0.1500	0.146(0.046)	0.136(0.053)
			a	0.6325	0.5762(0.1513)	0.5424(0.1427)
$r = 0.2$	0.10	0.10	r_1	0.1000	0.092(0.031)	0.087(0.028)
			a	0.1414	0.1243(0.0927)	0.1084(0.0864)
		0.30	r_1	0.1000	0.095(0.036)	0.093(0.031)
			a	0.2449	0.1921(0.1038)	0.1843(0.1163)
		0.50	r_1	0.1000	0.094(0.028)	0.091(0.032)
			a	0.3162	0.2761(0.1247)	0.2553(0.1166)
	0.40	0.10	r_1	0.1000	0.093(0.035)	0.116(0.041)
			a	0.2828	0.2435(0.1525)	0.2247(0.1102)
		0.30	r_1	0.1000	0.096(0.031)	0.094(0.023)
			a	0.4899	0.4053(0.1328)	0.4132(0.1376)
		0.50	r_1	0.1000	0.113(0.022)	0.087(0.034)
			a	0.6325	0.5529(0.1413)	0.5841(0.1572)
$r = 0.1$	0.10	0.10	r_1	0.0500	0.055(0.034)	0.056(0.021)
			a	0.1414	0.1332(0.1185)	0.1155(0.0763)
		0.30	r_1	0.0500	0.053(0.026)	0.039(0.012)
			a	0.2449	0.2046(0.0764)	0.1656(0.0679)
		0.50	r_1	0.0500	0.046(0.012)	0.041(0.020)
			a	0.3162	0.2582(0.1100)	0.2600(0.1202)
	0.40	0.10	r_1	0.0500	0.057(0.008)	0.044(0.017)
			a	0.2828	0.2547(0.1212)	0.2235(0.1163)
		0.30	r_1	0.0500	0.053(0.020)	0.058(0.018)
			a	0.4899	0.4382(0.1316)	0.3950(0.1158)
		0.50	r_1	0.0500	0.046(0.007)	0.062(0.021)
			a	0.6325	0.5878(0.1291)	0.5592(0.1431)

表 3-5 四分类等级性状在不同参数组合下的 QTL 参数估计效果 ($\pi = 0.20/0.60/0.15/0.05$)
Table 3-5 Mean estimates of QTL parameters and their stand errors under different parameter case for four ordinal traits ($\pi = 0.20/0.60/0.15/0.05$).

重组率 (r)	遗传力	QTL 方差 贡献	QTL 参数	参数真值	估计均值 (标准误)	
					GLM	LM
$r = 0.3$	0.10	0.10	r_1	0.1500	0.158(0.041)	0.136(0.033)
			a	0.1414	0.1162(0.0852)	0.1321(0.0924)
		0.30	r_1	0.1500	0.145(0.033)	0.143(0.038)
			a	0.2449	0.1549(0.1020)	0.1728(0.1106)
		0.50	r_1	0.1500	0.145(0.040)	0.142(0.038)
			a	0.3162	0.2638(0.1230)	0.2621(0.1326)
	0.40	0.10	r_1	0.1500	0.160(0.051)	0.140(0.055)
			a	0.2828	0.2461(0.1040)	0.2235(0.1315)
		0.30	r_1	0.1500	0.154(0.055)	0.141(0.060)
			a	0.4899	0.4366(0.1320)	0.4015(0.1336)
		0.50	r_1	0.1500	0.142(0.051)	0.160(0.038)
			a	0.6325	0.5964(0.1382)	0.5572(0.1451)
$r = 0.2$	0.10	0.10	r_1	0.1000	0.092(0.033)	0.089(0.037)
			a	0.1414	0.1236(0.0729)	0.1088(0.0884)
		0.30	r_1	0.1000	0.094(0.030)	0.093(0.033)
			a	0.2449	0.1844(0.1032)	0.1872(0.1241)
		0.50	r_1	0.1000	0.107(0.020)	0.092(0.021)
			a	0.3162	0.2761(0.1223)	0.2685(0.1310)
	0.40	0.10	r_1	0.1000	0.094(0.022)	0.088(0.035)
			a	0.2828	0.2438(0.1313)	0.2382(0.1088)
		0.30	r_1	0.1000	0.095(0.033)	0.093(0.020)
			a	0.4899	0.4258(0.1389)	0.4330(0.1400)
		0.50	r_1	0.1000	0.104(0.020)	0.092(0.032)
			a	0.6325	0.5872(0.1450)	0.5679(0.1525)
$r = 0.1$	0.10	0.10	r_1	0.0500	0.058(0.011)	0.040(0.020)
			a	0.1414	0.1320(0.1060)	0.1133(0.1105)
		0.30	r_1	0.0500	0.055(0.015)	0.041(0.022)
			a	0.2449	0.2124(0.1150)	0.1700(0.1030)
		0.50	r_1	0.0500	0.045(0.020)	0.041(0.015)
			a	0.3162	0.2685(0.1220)	0.2710(0.1413)
	0.40	0.10	r_1	0.0500	0.057(0.022)	0.041(0.020)
			a	0.2828	0.2432(0.1163)	0.2121(0.1170)
		0.30	r_1	0.0500	0.046(0.028)	0.045(0.025)
			a	0.4899	0.4463(0.1160)	0.3765(0.1238)
		0.50	r_1	0.0500	0.046(0.015)	0.043(0.021)
			a	0.6325	0.5528(0.1340)	0.5872(0.1421)

表 3-6 四分类等级性状在不同参数组合下的 QTL 参数估计效果 ($\pi = 0.20/0.30/0.30/0.20$)
Table 3-6 Mean estimates of QTL parameters and their stand errors under different parameter case for four ordinal traits ($\pi = 0.20/0.30/0.30/0.20$).

重组率 (r)	遗传力	QTL 方差 贡献	QTL 参数	参数真值	估计均值 (标准误)	
					GLM	LM
$r = 0.3$	0.10	0.10	r_1	0.1500	0.140(0.040)	0.138(0.035)
			a	0.1414	0.1076(0.0730)	0.1362(0.1021)
		0.30	r_1	0.1500	0.144(0.042)	0.140(0.030)
			a	0.2449	0.1728(0.1145)	0.1892(0.1220)
		0.50	r_1	0.1500	0.146(0.031)	0.142(0.044)
			a	0.3162	0.2828(0.1256)	0.2673(0.1473)
	0.40	0.10	r_1	0.1500	0.157(0.042)	0.137(0.050)
			a	0.2828	0.2461(0.1120)	0.2235(0.1235)
		0.30	r_1	0.1500	0.154(0.054)	0.140(0.055)
			a	0.4899	0.4366(0.1330)	0.4015(0.1424)
		0.50	r_1	0.1500	0.143(0.050)	0.158(0.044)
			a	0.6325	0.5964(0.1435)	0.5572(0.1378)
$r = 0.2$	0.10	0.10	r_1	0.1000	0.106(0.028)	0.112(0.030)
			a	0.1414	0.1182(0.0850)	0.1136(0.0737)
		0.30	r_1	0.1000	0.095(0.025)	0.092(0.042)
			a	0.2449	0.1626(0.1212)	0.1938(0.1180)
		0.50	r_1	0.1000	0.095(0.033)	0.090(0.020)
			a	0.3162	0.2861(0.1030)	0.2836(0.1220)
	0.40	0.10	r_1	0.1000	0.104(0.020)	0.085(0.033)
			a	0.2828	0.2360(0.1220)	0.2550(0.1100)
		0.30	r_1	0.1000	0.096(0.042)	0.092(0.035)
			a	0.4899	0.4336(0.1205)	0.4330(0.1370)
		0.50	r_1	0.1000	0.096(0.032)	0.093(0.030)
			a	0.6325	0.5535(0.1323)	0.5828(0.1408)
$r = 0.1$	0.10	0.10	r_1	0.0500	0.058(0.010)	0.039(0.021)
			a	0.1414	0.1220(0.0828)	0.1321(0.1030)
		0.30	r_1	0.0500	0.054(0.020)	0.040(0.022)
			a	0.2449	0.2210(0.0890)	0.1936(0.1155)
		0.50	r_1	0.0500	0.046(0.022)	0.040(0.021)
			a	0.3162	0.2835(0.1128)	0.2927(0.1362)
	0.40	0.10	r_1	0.0500	0.055(0.025)	0.059(0.028)
			a	0.2828	0.2550(0.1005)	0.2445(0.1220)
		0.30	r_1	0.0500	0.045(0.020)	0.043(0.031)
			a	0.4899	0.4500(0.0756)	0.4225(0.1129)
		0.50	r_1	0.0500	0.047(0.020)	0.043(0.021)
			a	0.6325	0.5760(0.1435)	0.5756(0.1512)

3.3.1.5. 阈值数的影响

潜在变量的阈值数决定着阈性状表型的分类情况, 阈值的大小直接影响着表型发生率的高低。本研究中考虑了单阈值和 3 阈值两种情况, 模拟研究的结果表明: 在 QTL 效应和位置参数估计准确度方面, 同一种方法的估计结果差异不大, 不管是单阈值性状或是 3 阈值性状都能得到 QTL 参数的合理估计值 (表 3-3~表 3-6)。相对而言, 在同一参数组合下, 阈模型方法的参数估计效果更好, 离真值的偏差相对较小。

阈值数对 QTL 检验功效的影响较大, 不管是阈模型方法还是常规的线性方法, 3 阈值性状的 QTL 检验功效都要高于单阈值性状, 特别是 LM 方法提高的幅度相对较大, 主要是由于随着阈值数的增加, 加性遗传信息的丢失相应减少。当阈值数足够大时, 离散性状 (阈性状) 可以近似看作连续性状来处理。

3.3.1.6. 不同定位方法的影响

从本次模拟研究的结果可以看出, 在相同参数组合下, 基于 GLM 框架内的阈模型方法与常规的 LM 方法在 QTL 效应参数 (α) 估计上没有显著的差异。同时, 在多数情况下, 阈模型方法和线性方法的 QTL 效应估计值都较真值偏低。在 QTL 位置参数估计方面, 阈模型方法要明显强于 LM 方法, 重组率 (r_1) 估计的偏差相对较小, 特别是在 QTL 效应较小、性状遗传力较低及双侧翼标记间距较大的情况下 GLM 方法的优势更为明显。

3.1.2 检验功效

检验功效定义为检验显著次数占模拟总数 (同一参数组合) 的百分率, 功效的高低决定于定位方法的优劣。不同方法在各参数组合下的检验功效列于表 3-7、表 3-8, 其中表 3-7 为单阈值二分类性状的检验功效, 表 3-8 为 3 阈值等级性状的检验功效。

模拟研究的结果表明, 阈模型方法获得的 LR (似然检验统计量) 值均高于相应组合的 LM 方法, 说明阈模型方法在检测离散性状的 QTL 方面比一般的线性方法效率要高。相同参数组合下的似然检验统计量愈高, 检验显著的概率相应就愈大。因此, 检验统计量的高低直接影响着检验功效大小。在相同参数组合下, 阈模型方法 QTL 检测的检验统计量在 $\alpha = 0.05$ 水平上的功效均高于 LM 方法 (图 3-1)。

检验统计量的高低还受到性状遗传力和 QTL 方差贡献的影响, 随着性状遗传力和 QTL 方差贡献的提高, 两种方法所得到的检验统计量都有增大的趋势。相对而言, 阈模型方法的增加较 LM 方法更快, 因此随着 QTL 效应和性状遗传力的增大, 阈模型方法的检验优势就更加明显。

离散性状不同状态的表型发生率 (π) 和状态数对 QTL 定位的功效也有很大的影响 (图 3-2)。本研究的结果显示: 相同参数组合下, 多阈值性状的检验统计量 (LR) 和功效要高于单阈值性状, 主要是由于单阈值性状遗传信息丢失相对较多的原故; 另外, 当性状表型发生率趋于中间型时, QTL 定位的功效也相应较高, 而当性状的表型发生率趋于极端时, QTL 定位的功效相应较低。因此, 在群体中该性状上不同表型状态的发生频率相差较大时, QTL 连锁分析的

效果相对较差，而各表型状态的频率相差较小时，连锁分析的功效较高。

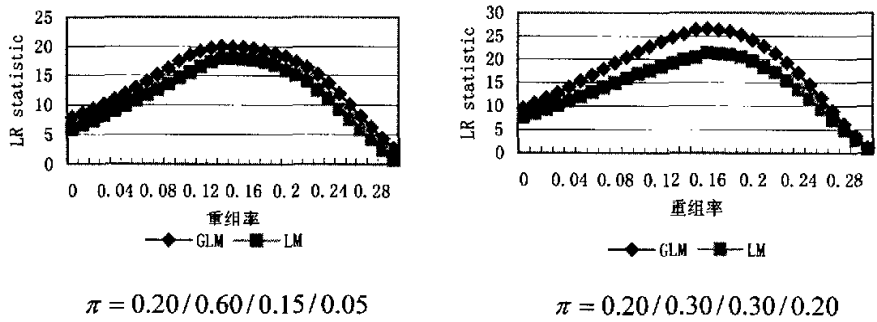


图 3-2 不同表型发生率下的平均似然率曲线（4 分类等级性状， $h^2 = 0.4, \Delta_q = 0.5, r = 0.3$ ）
Figure 3-2 The likelihood profile produced by averaging the analyses of 100 replicated simulations using two method under the different categories of incidence.

双侧翼标记间距对 QTL 的准确性和检验功效都有一定程度的影响。标记位点间距越小，QTL 与侧翼标记的连锁越紧密，QTL 定位的准确性和功效也就越高。从本研究设计的三种标记密度来看，随着标记间距的缩小（ $r = 0.3 \rightarrow 0.1$ ），检验统计量的值相应增大，从而 QTL 定位的功效也就相应增加（图 3-3）。因此，与常规数量性状一样，较高密度的标记图谱对阈值性状 QTL 连锁分析也是非常必要的。

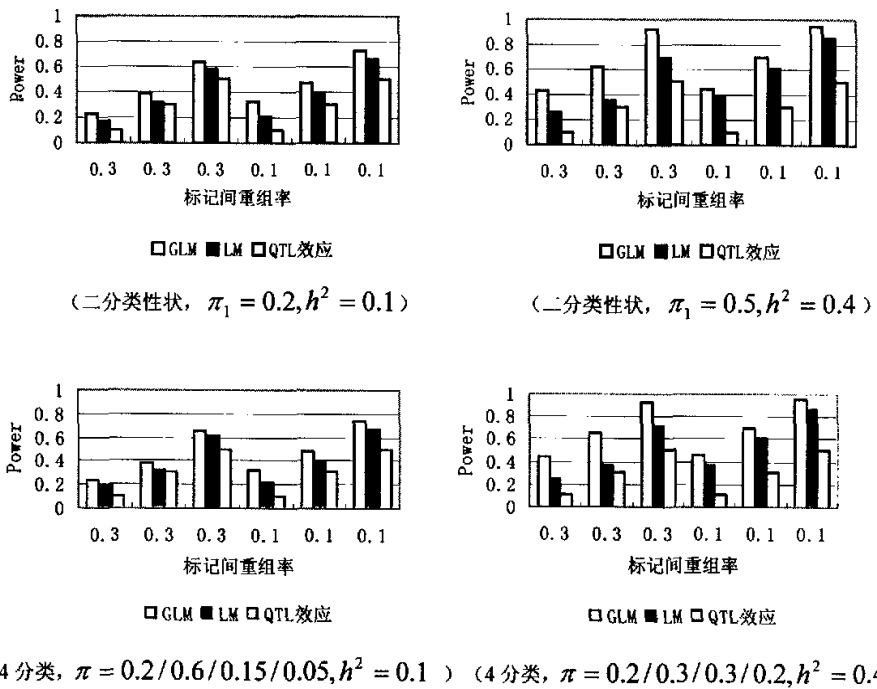


图 3-3 不同参数组合下 GLM 与 LM 方法的检验功效

Figure 3-3 Comparison of power with different parameter case

表 3-7 单阈值二分类性状在不同参数组合下 QTL 定位功效
Table 3-7 Comparison of statistical powers between GLM and LM method for QTL analysis of binary traits

表型发生率 (π)	重组率 (r)	遗传力	QTL 方差 贡献	功效	
				GLM	LM
$\pi_1 = 0.2$	$r = 0.3$	0.10	0.10	22.7	16.8
			0.30	38.6	31.5
			0.50	63.3	57.6
		0.40	0.10	39.4	25.2
			0.30	53.7	34.2
			0.50	89.4	65.8
	$r = 0.1$	0.10	0.10	32.6	20.8
			0.30	46.7	39.4
			0.50	72.5	65.7
		0.40	0.10	43.3	36.1
			0.30	68.6	57.5
			0.50	92.2	81.4
$\pi_1 = 0.5$	$r = 0.3$	0.10	0.10	25.6	17.3
			0.30	41.2	28.7
			0.50	65.8	56.6
		0.40	0.10	42.5	25.8
			0.30	61.7	35.2
			0.50	92.4	69.5
	$r = 0.1$	0.10	0.10	35.5	21.3
			0.30	49.2	39.7
			0.50	78.5	68.5
		0.40	0.10	44.3	37.6
			0.30	69.6	60.5
			0.50	94.8	85.2

注： r 为双侧翼标记间的重组率，QTL 与左侧标记间的重组率 (r_1) 为 r 的一半； π_1 为二分类性状中第一种状态的发生率。

表 3-8 多阈值等级性状在不同参数组合下 QTL 定位功效
Table 3-8 Comparison of statistical powers between GLM and LM method for QTL analysis of ordinal traits

表型发生率 (π)	重组率 (r)	遗传力	QTL 方差	功效	
			贡献	GLM	LM
$\pi_1 = 0.20$ $\pi_2 = 0.60$ $\pi_3 = 0.15$ $\pi_4 = 0.05$	$r = 0.3$	0.10	0.10	23.5	18.2
			0.30	38.4	31.7
			0.50	65.2	60.4
		0.40	0.10	39.6	25.5
			0.30	54.7	35.6
			0.50	90.5	66.6
	$r = 0.1$	0.10	0.10	31.8	21.2
			0.30	47.6	39.8
			0.50	73.8	67.2
		0.40	0.10	44.6	36.8
			0.30	70.2	59.3
			0.50	93.3	82.7
$\pi_1 = 0.20$ $\pi_2 = 0.30$ $\pi_3 = 0.30$ $\pi_4 = 0.20$	$r = 0.3$	0.10	0.10	26.7	17.6
			0.30	41.8	30.5
			0.50	67.2	56.5
		0.40	0.10	44.5	24.3
			0.30	65.1	35.7
			0.50	92.8	70.6
	$r = 0.1$	0.10	0.10	37.4	20.5
			0.30	50.8	38.6
			0.50	81.3	68.8
		0.40	0.10	44.7	35.7
			0.30	70.3	61.2
			0.50	96.2	85.9

注： r 为双侧翼标记间的重组率，QTL 与左侧标记间的重组率 (r_1) 为 r 的一半； $\pi_1 \sim \pi_4$ 分别为四等级性状中各状态的发生率。

3.1.3 显著性阈值

显著性阈值的高低直接影响着 QTL 定位的功效，在无 QTL 效应存在的零假设下，通过计算机模拟获得的经验显著性阈值见表 3-9 所示。

由表中的结果可以看出，对于不同分类性状或参数组合，计算机水平上的显著性阈值近似于自由度为 2 的卡方分布值（5.99），但多数参数组合下都要稍高于 $\chi^2_{df=2}$ 。相同参数设置下，GLM 方法的经验阈值要比 LM 方法的稍高，这主要是由于 GLM 方法能产生较高的似然率(LR)的原故，但二者之间的差异不明显。说明性状的遗传力和表型发生率对计算机水平上的经验阈值影响不大。

同一种方法，单阈值和多阈值离散性状的经验显著阈值也没有明显的差异，说明表型类型的差别不影响计算机水平上的经验阈值。

表 3-9 不同模拟条件下的显著性阈值

Table 3-9 Empirical critical values under different parameter case for testing the presence of a QTL using GLM and LM methods.

性状分类	发生率（ π ）	遗传力	经验阈值 (5%)	
			GLM	LM
二分类性状	$\pi_1 = 0.2$	0.10	6.43	6.12
		0.40	6.24	6.07
	$\pi_1 = 0.5$	0.10	6.15	5.95
		0.40	6.73	6.02
四等级性状	$\pi = 0.20/0.60/0.15/0.05$	0.10	6.24	6.20
		0.40	6.45	6.10
	$\pi = 0.20/0.30/0.30/0.20$	0.10	6.82	6.15
		0.40	6.23	5.97

3.4 讨论与结论

3.4.1 离散性状的 QTL 定位方法

随着现代动物育种理念的进一步发展,人们对离散性状的重视程度不断增大。多基因离散性状和数量性状有着相似的遗传基础,同样受多基因控制并受到环境的修饰,但表型上则为非连续分布,多呈现出二项分布、多项分布或者泊松分布,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低,在离散的表型中遗传信息的丢失较多。常规数量性状的表型服从正态分布及方差不依赖于均值(方差固定)的假设不能适用(Falconer, 1996; Gamal, 1999; Gianola, 1983)。因此,采用常规的线性方法对离散性状(或阈值性状)进行标记-QTL 连锁分析的功效不高。相对于常规的数量性状,多基因控制的离散性状或阈值性状的 QTL 定位及遗传分析就显得更具有挑战性。目前,对家畜离散性状 QTL 连锁分析的方法并不多,且 QTL 检测的效率较普通数量性状低,这主要是由于离散性状的表型值特殊性使很多亲代到子代的遗传信息丢失。

在动植离散性状的 QTL 定位研究方面主要还是采用一些常规的线性方法(Meijering & Gianola, 1985; Olesen, 1994; Weller, 2001; Knott, 1992; Haley, 1992; Gomez-Raya, 1997; Thaller, 1996)。Visscher 等(1996)以及 Rebai 等(1997)的研究表明,在回交及 F₂ 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有所损失。本研究的结果也证明了常规线性方法在分析离散性状方面的这一不足,不管是单阈值或是多阈值分类性,在 QTL 的检验功效上普遍低于 GLM 方法(表 3-7, 表 3-8),QTL 定位的效率较低。不论线性分析思想的理论假设有多么完美,但是离散性状的不连续性是不可否认的,性状表型服从于正态分布是线性分析思想的理论基石,因此对于呈现多项分布或者泊松分布的多基因离散性状需要寻求一种合理的 QTL 分析方法,以提高 QTL 定位的效率。

阈值模型(Falconer, 1996)的提出开辟了合理的离散性状遗传分析方法,阈值模型在分析思想上是基于潜在连续分布的假设基础之上,认为在离散表型的背后存在一个潜在的无法观察的连续分布,离散的表型值与潜在的连续值是通过一系列固定的阈相联系。阈值模型在分析方法上主要是基于广义线性模型(McCullagh, 1989),通过连接函数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。Hackett 和 Weller (1995)利用广义线性方法首次对近交系杂交设计下二分类性状的 QTL 参数进行了估计,之后出现了大量的相关研究(Xu & Atchley, 1996; Thaller, 1996; Heather, 2001; Yi, 1999; Visscher 等, 1996; Inge Riis Korsgaard, 2002),但大多方法主要是基于近交系杂交设计(如 BC、F₂ 群体),而在家畜群体 QTL 定位方面的应用相对较少。

Yi & Xu 等(1999)提出了适应于多个大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。这种固定效应模型主要适用于全同胞家系数数量较小时的 QTL 连锁分析,随着全同胞家系系数的上升和需要估计的参数增多,该方法的效率降低。在此基础上,相应发展了 QTL 随机模型(random model)和广义区间作图(generalised interval mapping method)(Kadarmideen 等, 2000, 2001),这些方法主要是基于广义的标记回归分析,

分析的群体可扩展到一般家系群体,但是在分析的性状上主要还是针对单阈值二分类性状。由于分析上的复杂性,对于多阈值等级性状的研究相对较少。

本研究在广义线性模型框架内,选用最大似然区间定位方法,对半同胞群体离散性状 QTL 参数的估计结果较为理想, QTL 的位置和效应参数估计值均在合理的范围内。在检验功效方面, GLM 方法在不同参数组合下的 QTL 检验功效也较常规线性方法高,说明阈模型方法对离散性状的 QTL 定位是适合的,并具有一定的优势,这与基于近交系设计内的多数研究报道 (Xu, 1996; Rao, 1998) 是一致的。

3.4.2 影响 QTL 定位的因素

影响 QTL 定位效率的因素很多,主要有资源群体的类型和规模、涉及的 QTL 数目、性状的遗传力、QTL 效应、标记密度和所采用的统计分析方法等 (Weller, 2001)。对于阈性状而言,除了上述因素以外还受到性状表型发生率的影响,而表型发生率又受到潜在阈值的控制,因此,潜在的阈值和阈值数对 QTL 定位的效率也有直接的影响。本研究设计的参数组合中考虑了性状的遗传力、QTL 效应、标记间距、性状表型发生率和阈值数等影响阈性状 QTL 定位的主要因素。研究结果表明,类似与常规数量性状 (Jansen, 1994; Knott, 1992; Meuwissen, 2000; Weller, 1990; Zeng, 1994), 性状遗传力和 QTL 方差贡献对离散性状的 QTL 定位准确度和功效都有较大的影响,当性状的遗传力较高, QTL 方差贡献较大时,参数估计的准确性高,同时检验统计量 LR 值也较大, QTL 的检验功效较高;在低遗传力和低的 QTL 效应下,检验功效较低 (见表 3-7, 表 3-8)。

双侧翼标记间距 (标记密度) 对 QTL 定位的准确性和功效也有较大程度的影响。标记位点间距越小, QTL 与侧翼标记间的连锁就越紧密, QTL 定位的准确性和精确性也就越高。相反,标记位点间距过大时, QTL 定位的偏差也随之增大,标准误增加,同时检验统计量 LR 值也有一定程度的下降。这与常规数量性状的 QTL 定位结果是一致的 (Liu, 2001; Zeng, 1994; Weller, 1990; Mackinnon, 1995), 因此较高密度的标记图谱对离散性状 QTL 连锁分析是非常重要的。

分析性状的表型发生率对离散性状的 QTL 定位影响也较大,但不如前面几个因素明显。潜在阈值的大小则决定着性状各表型的发生率。当其它条件不变时,不同状态的表型发生率相差越小 (二分类 50%-50%, 四分类 20%-30%-30%-20%), 则检验统计量 LR 值相应越大, QTL 定位的功效也就越高。本研究的结果与 Xu 等 (1996) 对回交群体中二项分布性状的研究结论是类似的,主要是由于表型发生率的大小直接影响着遗传信息的丢失程度。因此不同表型状态的表型发生率相互接近对离散性状的 QTL 连锁分析也是非常重要的。应该考虑到在自然群体中阈性状的潜在阈值是无法人为控制的,但可以通过一定的实验设计来对性状的表型发生率进行操作控制。

潜在的阈值数决定着性状的表型分类数,阈值数对 QTL 定位也存在一定程度的影响,随着阈值数的增加,性状的表型分类相应增加,其表型分布就愈接近正态分布。因此,潜在阈值数的增加对离散性状的 QTL 定位可能是有利的,理论上它可以提高 QTL 定位的准确性和检验功效,本研究的结果也证明了这一点。

3.4.3 GLM 方法的扩展

本研究虽然只模拟了半同胞群体和双侧翼标记的简单情况,但是这种广义线性模型框架内的离散性状 QTL 定位方法同样适用于一般群体,只是在连锁相及 QTL 基因型条件概率的推断方面更为复杂了。对于一般群体, QTL 基因型条件概率的推断主要依赖于历史重组的资料,利用上下代的亲缘关系来推断 QTL 位点上同源相同 (IBD) 的概率 (Kruglyak, 1995; Liu, 2001; Rao, 1998)。

同时,本研究中只考虑了双侧翼标记和一个 QTL 的情况,但对于多标记和多个 QTL 的连锁群来说 GLM 方法也是适用的 (Rao, 1998; Rodolphe, 1993)。在多标记的情况下,需要借助于多个标记的基因型来对 QTL 基因型进行联合推断;在考虑多个 QTL 时,可以在潜在值模型中加入多个 QTL 的效应以及 QTL 之间的互作效应,以实现影响性状的多个 QTL 的分析 (Zeng, 1994)。这方面的研究在常规的数量性状 QTL 连锁分析中已有了大量的报道,多 QTL 分析方法同样适用于离散性状的潜变量模型。

本研究中仅考虑了单阈值二项分布和三阈值多项分布的性状表型,阈模型方法同样适用于其它类型的表型分布性状 (如呈泊松分布或负二项分布的性状),只是在构建似然函数的方式上有所不同,比如对于呈泊松分布的离散性状可选用泊松回归模型来构建似然函数,其分析方法同样是基于广义线性混合模型 (Thomson, 2003; George, 2003; Tempelman, 1999)。Thomson 在广义线性混合模型的框架内,利用泊松回归模型对老鼠的产仔数进行拟合,并用 GEE 方法对影响老鼠产仔数的 QTL 参数进行了估计,研究结果显示泊松回归模型对产仔数进行遗传分析是非常有效的。

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,除了要不断改进试验设计外,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可靠性。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。连锁分析与连锁不平衡分析相结合应是今后家畜离散性状 QTL 精细定位的主要手段,根据连锁不平衡的原理,在 QTL 定位方面人们普遍考虑采用一些新的定位策略,如连锁不平衡 (LD) 定位、传递不平衡检验 (TDT) 定位和同源相同 (IBD) 定位等 (Carlier, 1996; Spielman, 1998; Xiong, 1997), 这些定位策略的共同前提是: (1) 通过连锁分析将 QTL 定位在一个较小区域内; (2) 在这个区域内配有高密度的分子标记。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,并且要利用群体历史上积累下来的重组事件。虽然在 QTL 精细定位方面连锁不平衡分析优势明显,但其定位效力受到标记与 QTL 间连锁不平衡程度的限制,因此有必要结合二者的优势进行多基因离散性状 QTL 的高效定位,即在连锁分析的基础上进行 LD、IBD 或 TDT 等精细定位。这也是多基因离散性状 QTL 定位的研究重点所在。

3.4.4 可获得的初步结论

从本次模拟研究的结果来看,可以获得以下两个初步结论:

(1) 相对于常规的线性方法, 广义线性模型框架内的阈模型方法能更为有效地进行离散性状的 QTL 连锁分析, 在 QTL 位置和效应参数的估计方面与 LM 方法无显著差异, 但是在检验功效上则明显优于常规的线性方法。

(2) GLM 方法在离散性状 QTL 定位方面所表现出的优势在很大程度上受到性状遗传力、QTL 效应、双侧翼标记间距、性状表型发生率和阈值数等多种因素的影响。其中, 性状遗传力、QTL 效应和标记密度对 QTL 定位的效果影响较明显。

这一结论是在本研究的特定试验设计中获得的, 对于其它群体、其它分布的性状仍需要进行深入的研究。

第四章 研究总结

4.1 多基因离散性状的遗传分析

在多基因离散性状的遗传分析方法方面,本研究主要对单阈值二分类性状进行了模拟研究,并分析了影响遗传参数估计的主要因素,研究结果可总结为如下几点:

(1) 在遗传力的估计准确度方面, GLMM 方法具有一定的优势。GLMM 方法可通过连结函数来实现响应变量的线性化,充分利用了遗传信息。相对而言, LM 方法估计遗传力的效果较差; TM 方法的遗传力估计效果介于 LM 与 GLMM 之间。因此,在多基因离散性状遗传分析方法的选择上, GLMM 方法应是首选。

(2) 从公畜个体育种值的排序结果来看,在不同的参数组合下, GLMM 方法的育种值估计效果也都不同程度地优于 LM 方法,能有效的给出公畜育种值的较为合理的排序,进一步证明了 GLMM 方法在进行多基因离散性状遗传分析上所具有的优势。这种优势在只有两种表型的二分类性状上体现的最为明显。

(3) 性状的遗传力和性状的表型发生率对遗传力估计和个体育种值预测的效果均有直接的影响,随着性状遗传力真值的提高, GLMM 方法出现一定程度的高估现象;随着性状发生率提高,三种方法的遗传分析效果都有上升的趋势,特别是对 LM 和 TM 的估计结果影响较大,另外, GLMM 方法的遗传力估计偏差及 MSE 有增大的趋势,这主要是由于 GLMM 方法对遗传力的高估引起。因此,在育种实践中应根据性状的性质(如遗传力和表型发生率)来合理的选择所要采用的遗传分析方法。对于具有较高遗传力和表型发生率的离散性状,可选用 LM 或阈模型方法进行分析,因为毕竟 GLMM 方法的计算成本较高。

(4) GLMM方法具有良好的统计性质,并可在多个方面进行扩展。本文仅研究了单阈值的二分类性状,但对于多于2个类型的多阈性状, GLMM方法的分析也同样可行有效,只是在连接函数及方差函数的选择上有所差异,这主要取决于离散性状的表型分布。但是,对于分类较多的多阈值性状,可以将其直接看作连续性状来处理,利用常规的线性方法可以获得与GLMM方法相当的参数估计效率和选择进展,特别是在阈值超过4个以上时。所以, GLMM方法主要适应于阈值数较少的性状遗传分析。

(5) 所有这些结果是在本研究的特定试验设计中获得的,虽然本研究考虑了不同半同胞群体大小、性状遗传力和表型发生率等因素的影响,但对于其它群体及其它分布的性状仍需要进行深入而全面的进一步研究。

4.2 多基因离散性状的 QTL 连锁分析

在离散性状的 QTL 连锁分析方法方面,本研究在半同胞设计内对单阈值和多阈值性状进行了模拟研究, QTL 参数估计采用最大似然方法,并分析了影响离散性状 QTL 连锁分析效果的主要因素,研究结果可总结为如下几点:

(1) GLM 方法能有效地估计出多基因离散性状的 QTL 参数,位置和效应参数估计值均在

合理的范围内。在检验功效方面, GLM 方法在不同参数组合下的 QTL 检验功效也较常规线性方法高, 说明阈模型方法对离散性状的 QTL 定位是适合的, 并具有一定的优势。

(2) 类似于常规的数量性状, 性状的遗传力、QTL 效应、标记间距等对离散性状的 QTL 定位都有显著的影响。当性状的遗传力较高, QTL 方差贡献较大时, 参数估计的准确性高, 同时检验统计量 LR 值也较大, QTL 的检验功效较高; 在低遗传力和低的 QTL 效应下, 检验功效也相应较低。

(3) 双侧翼标记间距对 QTL 定位的准确性和功效也有较大程度的影响。标记位点间距越小, QTL 与侧翼标记间的连锁就越紧密, QTL 定位的准确性和精确性也就越高。相反, 标记位点间距过大时, QTL 定位的偏差也随之增大, 标准误增加, 同时检验统计量 LR 值也有一定程度的下降。因此, 较高密度的标记图谱对离散性状 QTL 连锁分析是非常重要的。

(4) 对于阈性状而言, 除了上述因素以外, 性状表型发生率对 QTL 定位也有一定程度的影响, 而表型发生率又受到潜在阈值的控制, 因此, 潜在的阈值和阈值数对 QTL 定位的效率也有直接的影响。潜在阈值的大小则决定着性状各表型的发生率。当其它条件不变时, 不同状态的表型发生率相差越小 (二分类 50-50%, 四分类 20-30-30-20%), 则检验统计量 LR 值相应越大, QTL 定位的功效也就越高。因此, 不同表型状态的表型发生率相互接近对离散性状的 QTL 连锁分析也是非常重要的。潜在的阈值数决定着性状的表型分类数, 阈值数对 QTL 定位也存在一定程度的影响, 随着阈值数的增加, 性状的表型分类相应增加, 其表型分布就愈接近正态分布。因此, 潜在阈值数的增加对离散性状的 QTL 定位可能是有益的。

(5) 广义线性模型框架内的多基因离散性状 QTL 定位方法同样适用于一般群体、多标记和其它表型分布性状的情况。

(6) 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。连锁分析无法对 QTL 进行精细定位, 连锁分析与连锁不平衡分析相结合应是今后家畜多基因离散性状 QTL 精细定位的主要手段, 根据连锁不平衡的原理, 在 QTL 定位方面人们普遍考虑采用一些新的定位策略, 如连锁不平衡 (LD) 定位、传递不平衡检验 (TDT) 定位和同源相同 (IBD) 定位等, 在连锁分析的基础上进行 LD、IBD 或 TDT 等精细定位。

(7) 同样也应该考虑到本研究的这些结果是在特定试验设计中获得的, 对于不同群体、不同性状性质都需要进一步的深入研究。

参考文献

1. 丁向东. 利用系谱传递不平衡检验精细定位 QTL. 北京: 中国农业大学博士论文, 2004
2. 盛志廉, 陈瑶生编. 数量遗传学. 北京: 科学出版社, 1999
3. 宋九洲, 张沅, 张勤. 女儿设计和孙女设计连锁分析效率的比较研究. 畜牧兽医学报, 1998, 29:289~296
4. 吴常信. 分子数量遗传学与动物育种. 遗传, 1997, 19 (增刊): 1~3
5. 张勤. 基因组学时代的数量遗传学. 第 11 次全国动物遗传育种学术会议论文集, 2001: 151~157
6. 张沅, 张勤编. 畜禽育种中的线性模型. 北京: 北京农业大学出版社, 1993
7. 张沅主编. 家畜育种学. 北京: 中国农业出版社, 2001
8. Abecasis G.R., William O.C. and Lon R.C. Pedigree tests of transmission disequilibrium. *Eur. J. Hum. Genet.*, 2000,8:545~551
9. Alfonso L. & Haley C.S. Power of different F2 schemes for QTL detection in livestock. *Animal Science*, 1998,66:1~8
10. Allison D.B., Heo M, Kaplan N, Martin ER. Sibling-based tests of linkage and association for quantitative traits. *Am.J.Hum.Genet* 1999,64:1754~63
11. Andersson, L., and M. Georges. Domestic-animal genomics: deciphering the genetics of complex traits. *Nature Rev. Genet.*, 2004, 5:202~212.
12. Ashwell M.S., Van Tassell C.P., Sonstegard T.S. A genome scan to identify quantitative trait loci affecting economically important traits in a US Holstein population. *J Dairy Sci.*,2001,84(11):2543~2552
13. Beek S.vander, J.A.M.van Arendonk, and A.F.Groen. Power of two- and three-generation QTL mapping experiments in an outbred population containing full-sib or half-sib families. *Theor. Appl. Genet.*, 1995, 91:1115~1124
14. Beh K J, Hulme D J, Callaghan M J, Leish Z, Lenane I, Windon R G, Maddox J F. A genome scan for quantitative trait loci affecting resistance to *Trichostrongylus colubriformis* in sheep. *Animal Genetics*, 2002, 33: 97~106
15. Bishop S C, Chesnais J, Stear M J. Breeding for disease resistance: issues and opportunities. *Proceedings of the 7th world congress on genetics applied to livestock production. (Montpellier, France)*, 2002, 31:597~604
16. Blott S. Molecular dissection of a QTL: a phenylalanine to tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. *Genetics*, 2002, 163:253~266
17. Breslow N. E. , & Clayton D.G. Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.*, 1993, 88:9~25
18. Cardon L.R., and Abecasis G.R., Using haplotype blocks to map human complex trait loci. *Trends in genetics*. 2003,19:135~140

19. 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
20. Casas E, Bennett G.L., Bottema C.D. et al. Comprehensive linkage map of bovine chromosome 11. *Anim. Genet.* 2001. 32(2): 92~94
21. Chen W M, Broman K W, Liang K Y. Quantitative trait linkage analysis by generalized estimating equations: Unification of variance components and Haseman-Elston Regression. *Genetic Epidemiology*, 2004, 26:265~272
22. Chen W. X., Zhikang L. and Shizong X. Joint mapping of quantitative trait loci for multiple binary characters. *Genetics*, 2005, 169:1045~1059
23. Christoph L. Mapping quantitative trait loci using generalized estimating equations. *Genetics*, 2001, 159:1325~1337
24. Churchill G.A. and Doerge R.W. Empirical threshold values for quantitative trait mapping. *Genetics*, 1994, 138:963~971
25. Collins F.S., Green E.D., Guttmacher A.E. and Guyer M.S. A vision for the future of genomics research. *Nature*, 2003, 422:835~847
26. Darvasi A. and Soller M. A simple method to calculate resolving power and confidence interval of QTL map locations. *Behavior. Genetics*, 1997, 27:125~132
27. Darvasi A., A.Weinreb, V.Minke, et al. Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. *Genetics*, 1993, 134:943~951
28. Darvasi A., M. Soller. Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics*, 1995, 141:1199~1207
29. Darvasi A., M.Soller. Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. *Genetics*, 1994, 138:1365~1373
30. Dekkers, Jack C.M., and Frederic hospital. The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews*, 2002, 3:22~32
31. Devlin B, Roeder K. Genomic control for association studies. *Biometrics*, 1999, 55:997~1004
32. Devlin B. and Risch, N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 1995, 29:311~322
33. 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
34. Donoghue K. A., R. Rekaya, J. K. Bertrand, and I. Misztal. Threshold-linear analysis of measures of fertility in artificial insemination data and days to calving in beef cattle. *J. Anim. Sci.*, 2004, 82: 987
35. Engel B & Keen K. A simple approach for the analysis of generalized linear mixed models. *Statist. Neerlandica*, 1994, 48(1):1~22
36. Falconer D S, Mackay T F C. *Introduction to Quantitative Genetics*. 4th ed. Essex: Longman, 1996.
37. Fernando R.L. Grossman M. Marker-assisted selection using best linear unbiased prediction. *Genet*

- Sel Evol, 1989, 21: 467~477
38. Foulley J.L. Prediction of selection response for threshold dichotomous traits. *Genetics*, 1992, 132:1187~1194
39. Gamal A. and Berger P.J. Properties of threshold model predictions. *J. Anim. Sci.*, 1999, 77: 582~590
40. George A. W., Thompson E. A. Discovering disease genes: Multipoint linkage analysis via a new Markov Chain Monte Carlo approach. *Statistical Science*, 2003, 18(4): 515~531
41. Georges M. Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics*. 1995, 139:907~920
42. Gianola D. Sire evaluation for ordered categorical data with a threshold model. *Genet. Sel. Evol.*, 1983, 15:210~224
43. Grignola F. and Hoeschele I. Mapping linked quantitative trait loci via Residual Maximum Likelihood. *Genet. Sel. Evol.* 1997 29:539~544
44. Gianola D. Sire evaluation for ordered categorical data with a threshold model. *Genet. Sel. Evol.*, 1983, 15:210~224
45. Gianola D. Theory and analysis of threshold characters. *J. Anim. Sci.*, 1982, 54:1079~1096.
46. Grignola F.E., I Hoeschele and B Tier. Residual maximum likelihood to map quantitative trait loci: Methodology. *Genet. Sel. Evol.* 1996. 28: 479~490
47. Gomez-Raya L, Klemetsdal G. Two-stage selection strategies utilizing marker-quantitative trait locus information and individual performance. *J Anim Sci*, 1997, 77: 2008~2018
48. Grisart B. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine *DGATI* gene with major effect on milk yield and composition. *Genome Res.*, 2002, 12:222~231
49. Goddard M.E. A mixed model for analysis of data on multiple genetic markers. *Theor. Appl. Genet.*, 1992, 83:878~886
50. Hackett C A, Weller J I. Genetic mapping of quantitative trait loci for traits with ordinal distributions. *Biometrics*, 1995, 51:1252~1263
51. Haley C.S. and S.A. Knott. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, 1992,69:315~324
52. Heather J., Cordell J. A., Todd N. J. Statistical modeling of interlocus interactions in a complex disease. *Genetics*, 2001, 158:357~367
53. Henderson C.R. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics*, 1976,32:69~83
54. Heringstad B. Analysis of clinical mastitis in Norwegian cattle with a longitudinal threshold model. 7th world congress on genetics applied to livestock production. Montpellier, France, 2002
55. 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

56. Hoeschele I., P. Vanranden. Bayesian analysis of linkage between genetic markers and quantitative trait loci. II. Combining prior knowledge with experimental evidence. *Theor. Appl. Genet.* 1993b, 85:946~952
57. Hyne V. QTL analysis-unreliability and bias in estimation procedures. *Mol. Breed.*, 1995, 1:273~282
58. Inge Riis Korsgaard. Prediction error variance and expected response to selection, when selection is based on the best predictor for Gaussian and threshold characters, traits following a poisson mixed model and survival traits. *Genet. Sel. Evol.* 2002, 34:307~333
59. Jansen R.C. and P. Stam. High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, 1994, 136:1447~1455
60. Jansen R.C. Controlling the type I and type II errors in mapping quantitative trait loci. *Genetics*, 1994, 138:871~881
61. Jansen R.C. Interval mapping of multiple quantitative trait loci. *Genetics*, 1993, 135:205~211
62. Jensen J. Estimation of recombination parameters between a quantitative trait locus (QTL) and two marker gene loci. *Theor. Appl. Genet.*, 1989, 78:613~618
63. Johan A.M., Arendonk V., Tier, B. and Kinghorn B.P. Use of multiple genetic markers in prediction of breeding values. *Genetics*, 1994, 137: 319~329
64. John M. Henshall and Michael E. Goddard. Multiple-trait mapping of quantitative trait loci after selective genotyping using logistic regression. *Genetics*, 1999, 151:885~894
65. Johnson G.C. Haplotype tagging for the identification of common disease genes. *Nature Genet.* 2001, 29:233~237
66. Jorde L.B. Linkage disequilibrium and the search for complex disease genes. *Genome Research*, 2000, 10:1435~1444
67. Kadarmideen H.N., Janss L.L.G. & Dekkers J C M. Power of quantitative trait locus mapping for polygenic binary traits using generalized and regression interval mapping in multi-family half-sib designs. *Genetical Research*, 2000, 76:305~317
68. 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
69. Knott S. A., Haley C. S. Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. *Genet Res*, 1992, 60:139~151
70. Kruglyak L., E.S. Lander. A nonparameteric approach for mapping quantitative trait loci. *Genetics*. 1995. 139: 1421~1428
71. Kruglyak L. What is significant in whole-genome linkage disequilibrium studies? *Am J Hum Genet*, 1997, 61:810~812
72. Lande R., Thompson R. Efficiency of marker assisted selection in the improvement of quantitative traits. *Genetics*, 1990, 124: 743~756
73. Lander E.S. and D. Botstein. Mapping mendelian factors underlying quantitative traits using RFLP

- linkage maps. *Genetics*, 1989, 121:185~199
74. Lee E.J., Yoshizawa K., Mannen H., et al. Localization of the muscular dystrophy AM locus using chicken linkage map constructed with the Kobe University resource family. *Anim. Genet.* 2002. 33(1): 42~48
75. Lindstrom M.J. Nonlinear mixed effects models for repeated measures data. *Biometrics*, 1990, 46:673~687
76. Lipkin E, Mosig M.O., Darvasi A. et al. Quantitative trait locus mapping in dairycattle by means of selective milk DNA pooling using dinucleotide microsatellite markers: analysis of milk protein percentage. *Genetics*, 1998, 149:1557~1567
77. Liu H.Y., Zhang Q., Zhang Y. Relative efficiency of marker assisted selection when marker and QTL are incompletely linked. *Chinese Sci. Bull.* 2001, 46(24):2058~2063
78. Liu Y., Jansen G. B., Lin C. Y. Quantitative trait loci mapping for dairy cattle production traits using a maximum likelihood method. *J Dairy Sci*, 2004, 87: 491~500
79. 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
80. Luo Z.W. and M.J. Kearsey. Maximum likelihood estimation of linkage between a marker gene and a quantitative locus. II. Application to backcross and double haploid populations. *Heredity*, 1991, 66:117~124
81. Mackinnon M. J., Weller J. I. Methodology and accuracy of estimation of quantitative trait loci parameters in a half-sib design using maximum likelihood. *Genetics*, 1995, 141:755~770
82. Maddox J.F., Davies K.P., Crawford A.M., et al. An enhanced linkage map of the sheep genome comprising more than 1000 loci. *Genome Res.* 2001. 11(7): 1275~89
83. Martinez Q. and Crunow R.N. Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theor. Appl. Genet.* 1992, 85:480~488
84. McCullagh P., Nelder J. A. *Generalized linear models*. 2nd edn Chapman-Hall, London, 1989
85. McGuirk B.J. The genetic evaluation of beef sires used for crossing with dairy cows in the UK. 2. Genetic parameters and sire merit predictions for calving survey traits. *Anim. Sci.* 1998, 66:47~54
86. McGuirk B.J. The genetic evaluation of UK Holstein Friesian sires for calving ease and related traits. *Anim. Sci.* 1999, 68:413~422
87. Meijering A., and D. Gianola. Linear versus non-linear methods of sire evaluation for categorical traits: a simulation study. *Genet. Sel. Evol.*, 1985, 17:115
88. Meuwissen T. H. E., B. Engel, and J. H. J. van der Werf. Maximizing selection efficiency for categorical traits. *J. Anim. Sci.*, 1995, 73:1933
89. Meuwissen T.H.E. and Luo Z. Computing inbreeding coefficients in large populations. *Genet. Sel. Evol.*, 1992, 24:305~313

90. Meuwissen T.H., Karlsten A., Lien S., Olsaker I. and Goddard M.E. Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics*, 2002, 161:373~379
91. Meuwissen, T. H. E and Goddard M.E. Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics*, 2000, 155:421~430
92. Mikawa S., Akita T., Hisamatsu N. A linkage map of 243 DNA markers in an intercross of Gottingen miniature and Meishan pigs. *Anim Genet*. 2000. 31(1): 80~81
93. Nelder J. A., Mead R. A simplex method for function minimization. *Computer J*, 1965, 7:308~313
94. Nengjun Y. Bayesian mapping of Quantitative Trait Loci for complex binary traits. *Genetics*, 2000, 155:1391~1403
95. Olesen I., Perez-Enciso M., Gianola D. and Thomas D.L. A comparison of normal and nonnormal mixed models for number of lambs born in Norwegian sheep. *J. Anim.Sci.*, 1994, 72:1166~1178
96. Pritchard J.K., Stephens M., Rosenberg N.A., et al. Association mapping in structured populations. *Am. J. Hum. Genet.*, 2000, 67:170~181
97. Pritchard J.K., Rosenberg N.A. Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.*, 1999, 65:220~228
98. Ramon C.L., George A.M., Walter W.S., Russell D.W. SAS system for mixed models. SAS Institute Inc., Cary, NC, 1996
99. Rebai A. Comparison of methods for regression interval mapping in QTL analysis with non-normal traits. *Genetical Research*, 1997, 69:69~74
100. Rao S Q, Xu S Z. Mapping quantitative trait loci for ordered categorical traits in four-way crosses. *Heredity*, 1998, 81:214~224
101. Rekaya R., Gianola D. Application of a structural model for genetic covariance in international dairy sire evaluations. *J. Dairy Sci*. 2001, 84:1525~1530
102. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science*, 1996, 273: 1516~1517
103. Rodolphe F. and M. Lefort. A multi-marker model for detecting chromosome segments displaying QTL activity. *Genetics*, 1993, 134:1277~1288
104. Rothschild M F, C Jacobson, D Vaske, et al. The estrogen receptor locus is associated with a major gene influencing litter size in pigs. *Proceedings of the National Academy of Science*, 1996, 93:201~205
105. Ruane J., Colleau J.J. Marker-assisted selection for a sex-limited character in a nucleus breeding population. *J. of Dairy Science*, 1996, 79:1666-1678
106. San Cristobal-Gaudy. Genetic components of litter size variability in sheep, *Genet.Sel.Evol*. 2001, 33:249~271
107. Schaeffer L.R. and B.W. Kennedy. Computing strategies for solving mixed model equations.

- Journal of Dairy Science, 1985, 69:575~579
108. Schaid D.J. General score tests for associations of genetic markers with disease using cases and their parents. *Genetic Epidemiology*, 1996; 13: 423~449
109. Schulze T.G. and Francis J.M. Genetic association mapping at the crossroads: which test and why? Overview and practical guidelines. *American Journal of Medical Genetics*, 2002, 114:1~11
110. Shi zhong Xu. Mapping quantitative trait loci for complex binary diseases using line crosses. *Genetics*, 1996, 143:1417~1424
111. Shorepy S.A. and D. R. Notter. Genetic variation and covariation for ewereproduction, lamb growth and lamb scrotal circumference in a fall-lambing sheep flock. *J. Anim. Sci.* 1996, 74:1490
112. Soller M. Marker assisted selection-an overview. *Animal Biotechnology*, 1994,5(2):193
113. Spielman R.S., Ewens W J. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *American Journal of Human Genetics*, 1998, 62: 450~458
114. Spyrides M. H., Demetrio C. G. B., Camargo L. E. A. Proportional odds model applied to mapping of disease resistance genes in plants. *Genetics and Molecular Biology*, 2000, 23:223~227
115. Tempelman R.J. A mixed effects model for overdispersed count data in animal breeding. *Biometrics*. 1996, 52:265~279
116. Tempelman R.J., Gianola D. Genetic analysis of fertility in dairy cattle using negative binomial mixed models. *J. Dairy Sci.*, 1999, 82:1834~1847
117. Teng J., Risch N. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases: II, individual genotyping. *Genome Res.*, 1999, 9:234~241
118. Thaller G. and I. Hoeschele. A Monte Carlo method for Bayesian analysis of linkage between single markers and quantitative trait loci: I. Methodology. *Theor. Appl. Genet.* 1996a, 93:1161~1166
119. Thaller G. and I. Hoeschele. A Monte Carlo method for Bayesian analysis of linkage between single markers and quantitative trait loci: II. A simulation study. *Theor. Appl. Genet.* 1996b, 93:1167~1174
120. Thomsen H., Reinsch N., Xu N. Comparison of estimated breeding values, daughter yield deviations and de-regressed proofs within a whole genome scan for QTL. *J. Anim. Breed. Genet.*, 2001,118:357~370
121. Thomson P. C. A generalized estimating equations approach to quantitative trait locus detection of non-normal traits. *Genetics Selection Evolution*, 2003, 35:257~280
122. Tu I.P, Whittemore A.S. Power of association and linkage tests when the disease alleles are unobserved. *American Journal of Human Genetics*, 1999; 64: 641~649
123. Van Arendonk J. A. M., Tier B., Kinghorn B. P. Use of multiple genetic markers in prediction of breeding values. *Genetics*, 1994, (137): 319~329

124. Van Laere, A. S. Positional identification of a regulatory mutation in *IGF2* causing a major QTL effect on muscle growth in the pig. *Nature*, 2003, 425: 832~836
125. Visscher P.M., Haley C.S., Knott S.A. Mapping QTLs for binary traits in backcross and F2 populations. *Genet. Res.*, 1996, 68:55~63
126. Visscher P.M., Thompson R., and Haley C.S. Confidence intervals in QTL mapping by bootstrapping. *Genetics*. 1996. 143: 1013~1020
127. Waldman I.D, Robinson B.F., Rowe D.C. A logistic regression based extension of the TDT for continuous and categorical traits. *Ann. Hum. Genet.*, 1999, 63:329~340
128. Weller J. I., Kashi Y., Soller M. Power of "daughter" and "granddaughter" designs for genetic mapping of quantitative traits in dairy cattle using genetic markers. *J Dairy Sci*, 1990, 73:2525~2537
129. Weller J. I. Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. *Biometrics*, 1986, 42:627~641
130. Weller Joel Ira. Quantitative trait loci analysis in animals. printed and bound in the UK by Cromwell Press, Trowbridge. 2001
131. Whittaker J.C, Lewis C.M. The effect of family structure on linkage tests using allelic association. *American Journal of Human Genetics* 1998; 63: 889~97
132. 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. *Proc. Natl Acad. Sci. USA*, 2002, 99: 9300~9305
133. Wright S. An analysis of variability in number of digits in an inbred strain of guinea pigs. *Genetics*, 1934, 19:506~536
134. Xiong M., Guo S.W. Fine-scale mapping of quantitative trait loci using historical recombinations. *Genetics*, 1997, 145:1201~1218
135. Xu S, Atchley W R. Mapping quantitative trait loci for complex binary diseases using line crosses. *Genetics*, 1996, 143:1417~1424
136. Xu S. Further investigation of the regression method for mapping quantitative trait loci. *Heredity*, 1998, 80:364~373
137. Y.Da, P.M.VanRaden, N.Li, et al, Population structures for QTL detection in poul-try by DNA markers. *Journal of Agricultural Biotechnology*, 1999, 7(4):49~62
138. Yang Q., Rabinowitz D., Isasi C, Shea S. Adjusting for confounding due to population admixture when estimating the effect of candidate genes on quantitative traits. *Human Heredi*, 2000, 50:227~33
139. 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
140. Yi N.J., Xu S.Z. Mapping quantitative trait loci for complex binary traits in outbred population. *Heredity*, 1999, 82: 668-676
141. Zeng Z. B. Precision mapping of quantitative trait loci. *Genetics*, 1994, 136:1457~1468

142. Zeng Z. B. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc. Natl. Acad. Sci. USA*. 1993,90:10972~10976
143. Zhang Hao, Zhang Yuan, Zhang Qin. Pre-selecting young bulls before progeny-testing with marker-assisted BLUP. *Chinese Science Bulletin*. 2003,48(3):259~264
144. Zhang Qin, Didier Boidhard, Ina Hoeschele, et al. Mapping Quantitative Trait loci for milk production and health of dairy cattle in a large outbred pedigree. *Genetics*, 1998, 149: 1959~1973
145. Zhao H. Family-based association studies. *Statistical Methods in Medical Research*, 2000,9:563~587

致 谢

本论文是在导师张勤教授的悉心指导下完成的。从论文的选题、研究设计、实施到论文写作的整个过程,都浸透了导师大量的智慧和心血。导师严谨的治学态度、渊博的学识、高尚的人格以及对科学研究的高度热爱精神使我受益匪浅,终身难忘,导师的言传身教,将永远激励我奋进。在日常的学习和生活中,张勤教授和师母周少之老师也给予了无微不至的关怀,谨向他们全家致以最崇高的敬意和最衷心的感谢!

博士生学习期间,得到了张沅教授的悉心关怀,不管是在学习和工作上,给了我很多中肯的意见,热心为我解决遇到的困难。从他身上我领悟到了很多东西,与张沅教授相处的日子令我终身难忘。

感谢动物科学技术学院吴常信教授、张芳教授、杨宁教授、王爱国教授、王楚端教授、连正兴教授、孙东晓副教授、俞英副教授、王雅春博士、丁向东博士、白丽华老师以及动物遗传育种专业的其他各位老师对本人的学习和生活都给予了极大的关心和帮助。从各位老师身上我深深体会到遗传系浓厚的学术氛围和深厚的人文底蕴,为自己能在这个美好的环境里学习感到由衷的自豪。各位老师的敬业精神和对科学执着的追求是我学习的榜样,在此对各位老师表示衷心的感谢。

三年的学习生活转眼即逝,同学间真诚友好的相处和坦诚的交流令人难忘。感谢刘会英博士、亓小红博士、张俊博士、王栋博士、王慧博士、徐如海博士的帮助,也感谢张纪刚、姜润深、张浩、王春考、公维嘉、白俊艳、陈慧勇、颜炳学、刘剑锋、王希龙、张毅、徐青、何峰、邱小田、阴层层、李何君、李玉华、刘杨、刘锐、张剑、罗维真、王志鹏、舒娟、徐晶晶、胡芳、李玉红、陈留红、初芹、王重龙、关龙、梅瑰、孙传禹、郭宏伟、董沙沙、滕晓红、陈军、刘林、孙艺、田菲等同学的支持和帮助,感谢大家共同营造的这个令人难忘的大家庭。

感谢安徽农业大学的领导以及畜牧水产学院的章孝荣院长、陶传开书记、李力书记、魏建忠副院长、张伟力教授、陈宏权教授、耿照玉教授、刘旭光副教授、王力生副教授对我工作方面的支持与帮助。特别要感谢张伟力教授在我求学期间承担的繁重教学任务以及给予的各种关怀。

我的妻子郑慧女士、我的父母多年来始终如一地关爱着我,是她们的理解与支持才使我能够顺利完成学业;也感谢我的女儿殷嘉敏健康地成长,三年来她几乎每天都给予我特别的鼓励。

最后,向所有关心、支持和理解我的老师、同学、家人和朋友们表示最诚挚的谢意!

殷宗俊

2005年5月