作物学报 ACTA AGRONOMICA SINICA 2013, 39(1): 1–11 http://www.chinacrops.org/zwxb/
ISSN 0496-3490; CODEN TSHPA9 E-mail: xbzw@chinajournal.net.cn

DOI: 10.3724/SP.J.1006.2013.00001

QTL 作图中零假设检验统计量分布特征及 LOD 临界值估计方法

孙子淇 李慧慧 张鲁燕 王建康*

中国农业科学院作物科学研究所 / 农作物基因资源与基因改良国家重大科学工程 / CIMMYT 中国办事处, 北京 100081

摘 要: 研究 QTL 作图中零假设检验统计量分布特征, 可以帮助我们选取合适的 LOD 临界值, 以控制全基因组显著性概率水平下犯第一类错误的概率。本文利用模拟方法, 研究了 QTL 作图中单个扫描位点的似然比检验(LRT)统计量在零假设下的分布特征、影响最大 LOD 统计量累积分布的因素以及不同群体在不同标记密度下有效独立检验次数与染色体长度的关系。结果表明, 在定位加显性效应 QTL 的一维扫描和定位上位性互作 QTL 的二维扫描中, 单个扫描位置上的 LRT 统计量均服从卡方分布, 其自由度等于检测 QTL 遗传参数的个数; 染色体个数、群体大小和表型测量误差方差对零假设下检验统计量的分布没有影响, 即不影响 LOD 临界值的选取, 而群体类型、标记密度和染色体长度有明显影响, BC_1、RIL 和 F_2 三种类型的群体中, BC_1 群体的临界值最小, F_2 群体的临界值最大, 标记越密染色体越长, 对应的 LOD 临界值越大; QTL 一维扫描中有效独立检验次数与染色体长度呈正比, 二维扫描中有效独立检验次数与染色体长度呈二次幂关系。借助 Bonferroni 矫正, 给出了全基因组显著性水平与单个扫描位点显著性水平间的关系, 因此, 研究者可根据作图群体的群体类型、标记密度和基因组长度, 很方便地确定特定全局显著性概率水平下的 LOD 临界值。

关键词: QTL 作图; 似然比检验; LOD 统计量; 零假设; 显著性水平; 独立检验次数

Properties of the Test Statistic under Null Hypothesis and the Calculation of LOD Threshold in Quantitative Trait Loci (QTL) Mapping

SUN Zi-Qi, LI Hui-Hui, ZHANG Lu-Yan, and WANG Jian-Kang*

Institute of Crop Sciences / National Key Facility for Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences / CIMMYT China Office, Beijing 100081, China

Abstract: Selecting an appropriate LOD threshold is of great interest in QTL mapping studies. Many approaches can be considered to calculate the critical value throughout a genome, such as simulation-based method, analytical approximation, and empirical method based on permutation test. Many tests are conducted in QTL mapping, which are not mutually independent because the linkage relationship of adjacent markers on chromosomes. In order to declare a significant QTL at a genome-wide significance level, it is necessary to understand the behavior of test statistic under null hypothesis in QTL mapping and to deal with the dependent multiple-test problem arising in the genome-wide test. Our objectives in this study were (1) to investigate the properties of LRT (likelihood ratio test) statistic of one-point scanning under null hypothesis in QTL mapping, (2) to determine the factors affecting the cumulative distribution of maximum LOD score, and (3) to identify the relationship between the effective number of independent tests and the length of chromosome by simulation method. Results indicated that the LRT test statistic in one-dimensional scanning of additive-dominant QTL and two-dimensional scanning of epistatic QTL followed chi-square distributions, and the degree of freedom (*df*) was equal to the number of genetic parameters to be estimated. For example, degree of freedom in recombinant inbred lines (RIL) population was equal to 1 in one dimensional or two dimensional scanning. Degree of freedom in F_2 populations was equal to 2 in one-dimensional scanning and 4 in two-dimensional scanning. Number of chromosome, population size and phenotyping error variance did not have any effect on the distribution of LRT under null hypothesis, and therefore will not affect the selection of LOD threshold. On the contrary, population type, genome size and marker density had significant impacts. For BC_1, RIL, and F_2 populations, the threshold was the smallest in BC_1 population and the highest in F_2 population. Higher marker density and longer chromosome resulted in higher LOD threshold. It was identified that the effective

本研究由国家自然科学基金项目(31000540)资助。

* 通讯作者(Corresponding author): 王建康, E-mail: wangjk@caas.net.cn, jkwang@cgiar.org, Tel: 010-82105846

第一作者联系方式: E-mail: sunziqi777@163.com, Tel: 010-82108579

Received(收稿日期): 2012-05-13; Accepted(接受日期): 2012-09-05; Published online(网络出版日期): 2012-11-14.

URL: http://www.cnki.net/kcms/detail/11.1809.S.20121114.1642.007.html

number of independent tests (M_{eff}) was proportional to the length of chromosome in one-dimensional scanning of additive-dominant QTL. In two-dimensional scanning of epistatic QTL, it was identified that M_{eff} was in a squared relationship to the length of chromosome. With the help of Bonferroni correction, we could acquire the relationship between point-wise and genome-wide significance levels. Therefore, it is convenient to calculate the threshold LOD in QTL mapping, given the genome-wide significance level, the population type, marker density and genome size.

Keywords: QTL mapping; Likelihood ratio test; LOD score; Null hypothesis; Significance level; Number of independent tests

选取合适的 LOD 临界值是 QTL 作图中的一个基本问题，合适的 LOD 临界值是得到可靠 QTL 的保证，进而为精细定位、图位克隆及分子标记辅助育种奠定基础。基于区间作图的 QTL 定位方法通过计算各个扫描位点的 LOD 统计量判断 QTL 的存在，当某位点的 LOD 值超过预先设定的临界值时，就认为该位点可能存在控制数量性状的基因(quantitative trait locus, QTL)。选择不同的临界值，会得到不完全相同的定位结果。因此，究竟应选取多大的 LOD 临界值是研究学者多年来一直比较关心的问题。国内外学者对此问题开展了广泛研究并提出了许多不同的计算 LOD 临界值方法，这些研究方法可概括为模拟法[1-2]、公式法[1,3-11]、排列检验法[12]和求有效独立检验次数(M_{eff})法[13-16] 4 类。当标记密度适中时，Lander 和 Botstein [1]利用模拟法得到全基因组显著性水平为 0.05 时的 LOD 临界值应为 2~3。而当标记密度非常大或非常小时，Lander 和 Botstein [1]给出 2 个公式计算回交群体在 2 种极端情况下的 LOD 临界值。公式法[1,3-11]的适用条件比较复杂，如不同群体类型和标记密度适用不同的计算公式，故该法很少被采用。排列检验法[12]不受群体类型、表型分布和检测方法的影响，多年来被广泛使用，但该法具有专一性，每组表型数据都需要单独进行排列检验，特别是随着标记量和作图群体的增大，需要计算的时间越来越长。Cheverud [13]、Nyholt [14]、Li 和 Ji [15]、Gao 等[16]提出了通过求有效独立检验次数(M_{eff})的方法计算 LOD 临界值，但这些方法只考虑了加显性效应 QTL 的 LOD 临界值，随着 QTL 之间的互作效应[17-19]受到越来越多的重视，如何选取上位性 QTL 的 LOD 临界值是亟待解决的一个问题。

Churchill 和 Doerge [12]指出 QTL 作图过程中选取合适的 LOD 临界值需要解决 2 个比较棘手的问题：①零假设下检验统计量的分布特征；②全基因组扫描过程中的有效独立检验次数 M_{eff}。Piepho [9]、Zou 等[10]、Chang 等[11]认为零假设下，定位加显性效应 QTL 的一维扫描中单个扫描位点的 LRT 统计量渐近服从卡方分布，自由度(df)等于 QTL 作图中遗传参数的个数。Cheverud [13]、Nyholt [14]、Li 和 Ji [15]、Gao 等[16]给出了一维扫描中计算有效独立检验次数(M_{eff})的方法。本研究利用模拟方法研究了不同群体类型在定位加显性效应 QTL 的一维扫描和定位上位性互作 QTL 的二维扫描中 LRT 统计量的分布特征；影响零假设下 LOD 统计量分布的因素；以及不同标记密度下一维和二维扫描中的有效独立检验次数(M_{eff})与染色体长度的关系，并利用 3 个真实群体比较不同方法得到的 LOD 临界值。

1 材料与方法

1.1 LOD 统计量分布的影响因素

影响零假设下 LOD 统计量分布即 LOD 临界值选取的可能因素有很多，如群体类型、基因组大小、标记密度、染色体条数、群体大小、标记奇异分离、标记缺失和表型鉴定的随机误差方差等[1,10,12,20]。本文研究了群体类型、基因组大小、标记密度、染色体条数、群体大小和表型测量误差方差 6个因素对 LOD 统计量分布的影响。由于基因组内各染色体之间是独立的，因此除了研究染色体条数对统计量分布的影响外，其他情况下本文以一条染色体为例进行研究。

研究的群体类型包括回交(BC_1)、重组自交系(RIL)和 F_2；染色体长度(length of chromosome, CL)包括 50、100 和 150 cM；标记密度(marker density, MD)包括 1、5 和 20 cM；染色体条数包括 1、2 和 4 条，基因组总长为 200 cM；群体大小(population size, PS)包括 100、300 和 500；表型测量误差方差(error variance, EV)包括 0.2、0.4 和 0.6。上述是在零假设下即没有 QTL 效应存在的情况下进行的，各个组合模拟 10 000 次。当研究其中一个因素时，其他因素保持不变。模拟研究中，特定全局显著性概率水平对应的 LOD 临界值是由零假设下每次模拟试验得到的最大 LOD 统计量的累积分布决定的。因此，本研究通过比较各种情况下最大 LOD 统计量的累积分布来分析各因素是否影响 LOD 临界值的选取。

1.2 作图群体的模拟

定位加显性效应 QTL 的一维扫描中，研究的群

体类型包括 BC_1、RIL 和 F_2 群体; 基因组内各染色体之间是相互独立的, 故只以一条染色体为例进行研究, 染色体长度包括 50、60、……、200 cM; 标记在染色体上平均分布, 标记密度包括 1、2、5、10 和 20 cM; 各组合的群体大小均为 200, 表型测量误差方差为 0.4。利用连锁图谱和 QTL 作图集成软件 QTL IciMapping (可从 http://www.isbreeding.net/下载)模拟每个组合 10 000 次即可得到 10 000 个作图群体。

定位上位性互作QTL的二维扫描中, 群体类型仍为 BC_1、RIL 和 F_2 群体, 染色体长度仍为 50~200 cM。不仅同一条染色体上的 2 个位点间可能存在互作效应, 不同染色体上的 2 个位点间也可能存在互作效应, 因此, 在研究如何选取上位性QTL的LOD临界值时, 假定一个基因组中包含 2 条长度相等的染色体。二维扫描中标记密度的情况少于一维扫描, 只研究了 MD = 5、10 和 20 cM 时的情况。为了有足够的自由度检测互作 QTL, 各组合的群体大小设为 500, 表型测量误差方差仍为 0.4, 模拟次数为 10 000 次。

1.3　真实群体

为了评价本研究提出的计算 LOD 临界值方法的可靠性, 基于 3 个真实群体(即小麦 DH 群体[21]、玉米 RIL 群体[22]和水稻 F_2 群体[23]), 比较了模拟法[1-2]、排列检验法[12], 以及 Cheverud [13]、Li 和 Ji [14]、Gao 等[15]和本研究提出的求有效独立检验次数 M_{eff} 方法得到的 LOD 临界值。小麦 DH 群体的基因组长度为 3112.98 cM, 405 个标记, 21 条染色体, 平均标记密度为 7.6 cM。玉米 RIL 群体的基因组长度为 2250.9 cM, 132 个标记, 10 条染色体, 平均标记密度为 17 cM。水稻 F_2 群体的基因组长度为 1532.7 cM, 117 个标记, 12 条染色体, 平均标记密度为 13 cM。

1.4　作图方法及 M_{eff} 的计算

本研究利用区间作图(IM)的加性和上位性检测方法(即实现在 QTL IciMapping 软件[24]中的 IM-ADD 和 IM-EPI 两种方法)分别对得到的模拟群体进行一维加显性 QTL 扫描和二维互作 QTL 扫描。M_{eff} 的计算步骤为: ①统计 10 000 次模拟得到的最大 LOD 统计量, 并求出最大统计量的频率分布及其累积分布。②根据 Bonferroni 矫正$\alpha_{\text{p}}=\alpha_{\text{g}}/M_{\text{eff}}$, 可以得到:

$$M_{\text{eff}}=\alpha_{\text{g}}/\alpha_{\text{p}} \quad (1)$$

其中, α_{g} 为最大 LOD 统计量累积分布对应的收尾概率(全局第一类错误概率), α_{p} 为卡方分布的单尾概率(每次检验犯第一类错误概率)。因为 QTL IciMapping 软件输出的是 LOD 统计量, 而服从卡方分布的是似然比(LRT)统计量, 故计算 α_{p} 时, 频率分布的区间数组需乘以 2ln10。③统计 α_{g} 为 0.05 和 0.01 时对应的 M_{eff}。

2　结果与分析

2.1　零假设下单个扫描位点的 LRT 统计量分布特征

由图 1 可知, 零假设下单个扫描位点的 LRT 统计量服从卡方分布。图 1-a 和 b 分别为定位加显性效应 QTL 的一维扫描中 RIL 和 F_2 群体的卡方分布图。图 1-c 和 d 分别为定位上位性互作 QTL 的二维扫描中 RIL 和 F_2 群体的卡方分布图。一维扫描和二维扫描中, RIL 和 F_2 群体的 10 000 个 LRT 统计量的密度分布与卡方分布的概率密度完全拟合。与 Piepho [9]、Zou 等[10]、Chang 等[11]的结论一致, 定位加显性效应QTL的一维扫描中单个扫描位点的LRT统计量服从卡方分布, RIL 群体的自由度为 1, F_2 群体的自由度为 2。定位上位性互作 QTL 的二维扫描中单个扫描位点的 LRT 统计量仍服从卡方分布, RIL 群体的自由度仍为 1, 而 F_2 群体的自由度为 4。一维扫描和二维扫描中, BC_1 群体单个扫描位点的 LRT 统计量分布特征与 RIL 群体的相同, 均服从自由度为 1 的卡方分布。

既然单个扫描位点的 LRT 统计量服从卡方分布, 那么单个扫描位点的显著性水平 α_{p} 对应的临界值可由卡方分布表求得。但是, 我们通常所要求的是全局显著性概率水平 α_{g} 对应的 LOD 临界值。若知 α_{g} 与 α_{p} 之间的关系, 即可根据卡方分布求出全局显著性概率水平 α_{g} 对应的 LOD 临界值。

2.2　最大 LOD 统计量分布的影响因素

若某因素在不同水平下, 最大 LOD 统计量的分布保持不变, 说明该因素不影响最大 LOD 统计量的分布; 反之, 则说明在选取 LOD 临界值时应考虑该因素。由图 2-d~f 可知, 在不同水平的染色体条数、群体大小和表型测量误差方差下, 最大 LOD 统计量分布曲线几乎完全重合, 因此, 这 3 个因素对最大 LOD 统计量的分布无影响, 即不影响 LOD 临界值的选取。而不同群体类型、染色体长度和标记密度对应的最大 LOD 统计量分布曲线明显不一致(图 2-a~c), 即这 3 个因素影响 LOD 临界值的选取。且一定的显著性概率水平下, BC_1 群体的临界值小于 RIL 群体, RIL 群体小于 F_2 群体; 染色体越长, 标记越密, 临界值越大。

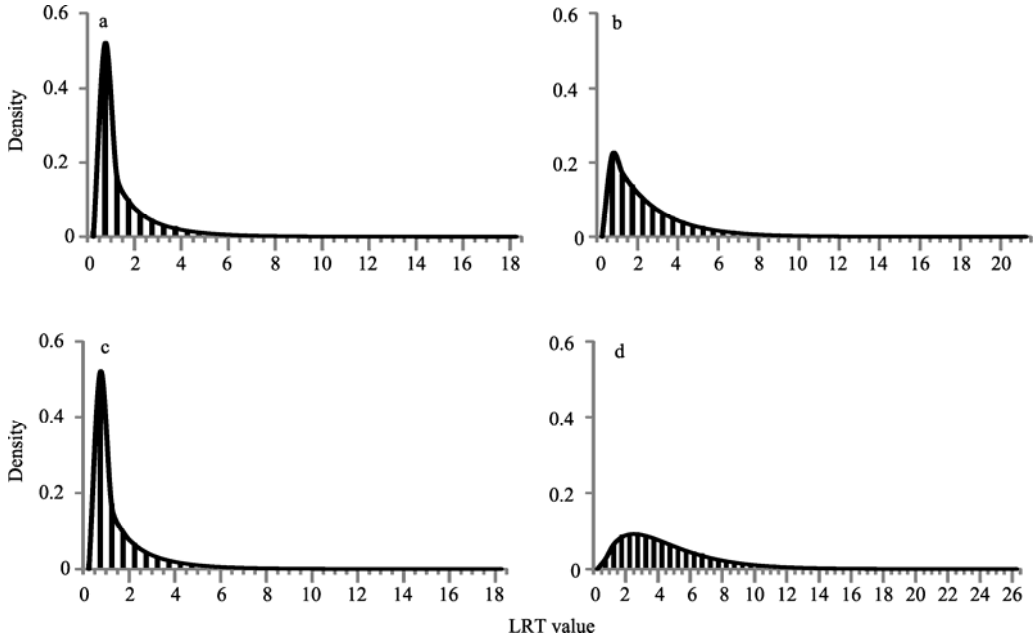

图 1 一维和二维 QTL 作图扫描过程中单个扫描位点 LRT 统计量与卡方分布的拟合

Fig. 1 Goodness of fit of single point LRTs in one-dimensional and two-dimensional scanning in QTL mapping

曲线表示卡方分布，竖直线表示观测分布, a、b 分别为定位加显性效应 QTL 的一维扫描中 RIL、F_2 群体的观测分布和卡方分布，卡方分布的自由度分别为 1、2; c、d 分别为定位上位性互作 QTL 的二维扫描中 RIL、F_2 群体的观测分布和卡方分布，卡方分布的自由度分别为 1、4。

The curve line stands for the chi-square distribution and the vertical line stands for the observed distribution, a and b correspond to RIL and F_2 population respectively, in one-dimensional scanning of additive-dominant QTL, and the degrees of freedom are 1 and 2, respectively; c and d correspond to RIL and F_2 population respectively, in two-dimensional scanning of epistatic QTL, and the degrees of freedom are 1, 4 respectively.

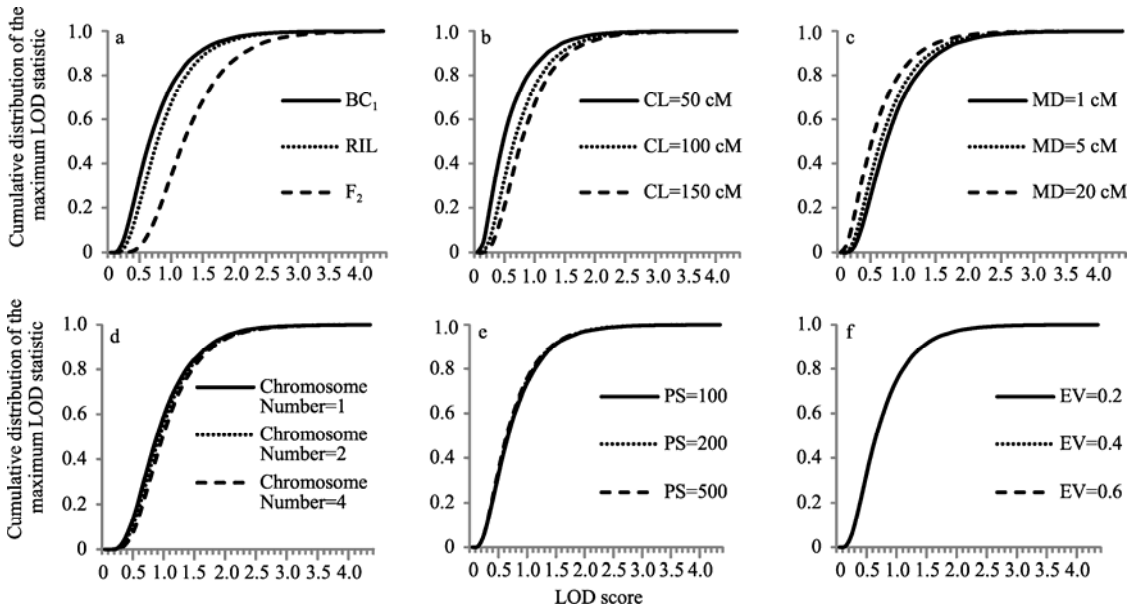

图 2 不同因素对最大 LOD 统计量累积分布的影响

Fig. 2 Factors affecting the cumulative distribution of the maximum LOD statistic

a 为群体类型, b 为染色体长度, c 为标记密度, d 为染色体个数, e 为群体大小, f 为表型测量误差方差。

a for population type, b for length of chromosome, c for marker density, d for chromosome number, e for population size, and f for error variance.

2.3 定位加显性效应 QTL 的一维扫描中有效独立检验次数(M_{eff})与标记密度(MD)和染色体长度(CL)的关系

由图 2 可知，影响 LOD 临界值选取的主要因素有群体类型、基因组大小和标记密度。因此，本文主要统计了 BC_1、RIL 和 F_2 三个群体在标记密度为 1、2、5、10 和 20 cM，染色体长度为 50、60、……、200 cM 时的有效独立检验次数(M_{eff})。由图 3 可知，零假设下定位加显性效应 QTL 的一维扫描中，不同标记密度下有效独立检验次数(M_{eff})与染色体长度呈线性关系。图 3-a、c 和 e 分别为 α_g=0.05 时 BC_1、RIL 和 F_2 群体的 M_{eff} 与染色体长度的关系。图 3-b、d 和 f 分别为 α_g=0.01 时 BC_1、RIL 和 F_2 群体的 M_{eff} 与染色体长度的关系。横坐标代表不同的染色体长度(CL)，纵坐标表示有效独立检验次数(M_{eff})。图中的虚线表示 M_{eff} 分布的趋势线，对应的函数式表示一定标记密度下，M_{eff} 与染色体长度的比例关系。y 表示 M_{eff}, x 表示染色体长度。

不同的群体类型、标记密度和显著性水平 α_g, 趋势线斜率不同；标记越密，显著性水平越高，斜率越大。如图 3-a 所示, BC_1 群体标记密度为 1 cM 时的斜率最大，为 0.153；标记密度为 20 cM 时的斜率最小，仅为 0.054。相同标记密度下, α_g=0.01 时的斜率比 α_g=0.05 时的大。对于不同的群体类型, BC_1 群体的斜率比 RIL 群体的小，F_2 群体的斜率与 RIL 群体的很接近甚至在某些情况下二者相等，如 α_g=0.05，标记密度为 5 cM 和 10 cM 时(图 3-c 与 e)。由于染色体个数对 LOD 临界值的选取没有影响，故图中的染色体长度(CL)也可以是整个基因组的长度(genome length, GL)。

已知某一作图群体的群体类型、标记密度和基因组大小时，可以根据图 3 估算出该作图群体的 M_{eff}, 然后根据 Bonferroni 矫正 $\alpha_p=\alpha_g/M_{eff}$ 可得到 α_g=0.05 和 0.01 时的 α_p，查卡方分布表即可得到该作图群体在全基因组显著性水平下的 LOD 临界值。为了避免数学计算和查表带来的不便，可以借助 Microsoft Excel 进行数学运算并利用 Excel 中的函数 *Chiinv* 直接计算出对应的 LOD 临界值。需要注意的是，利用 *Chiinv* 函数时，BC_1 和 RIL 群体的自由度为 1，F_2 群体的自由度为 2。且此时得到的临界值为 LRT 值，除以 2ln 10 后才能得到 LOD 临界值，即

$$LOD=Chiinv(\alpha_p, df)/2\ln 10 \quad (2)$$

因此，估算 LOD 临界值的一般步骤为，首先利用图 3 中有效独立检验次数(M_{eff})与染色体长度的函数关系得到作图群体的有效独立检验次数(M_{eff})，然后利用 Bonferroni 矫正得到全基因显著性水平 α_g=0.05 或 0.01 对应的单个扫描位点的显著性水平 α_p，最后利用公式(2)求得该作图群体的 LOD 临界值。

表 1 和表 2 分别给出定位加显性效应 QTL 的一

表 1　定位加显性效应 QTL 的一维扫描中, MD = 1 cM 时 3 种群体在不同基因组长度下的 LOD 临界值

Table 1　LOD threshold of three populations with different genome lengths in one dimensional scanning of additive-dominant QTL when MD = 1 cM

基因组长度 Genome length (cM)	BC_1		RIL		F_2	
	α_g=0.05	α_g=0.01	α_g=0.05	α_g=0.01	α_g=0.05	α_g=0.01
250	2.24	3.02	2.49	3.22	3.10	3.88
500	2.52	3.31	2.77	3.50	3.40	4.18
750	2.69	3.47	2.93	3.67	3.57	4.36
1000	2.80	3.59	3.05	3.79	3.70	4.49
1250	2.89	3.68	3.14	3.88	3.79	4.58
1500	2.97	3.76	3.22	3.95	3.87	4.66
1750	3.03	3.82	3.28	4.02	3.94	4.73
2000	3.09	3.88	3.33	4.07	4.00	4.79
2250	3.13	3.93	3.38	4.12	4.05	4.84
2500	3.18	3.97	3.43	4.17	4.10	4.88
2750	3.22	4.01	3.47	4.21	4.14	4.93
3000	3.25	4.05	3.50	4.24	4.17	4.96
3250	3.28	4.08	3.53	4.27	4.21	5.00
3500	3.32	4.11	3.56	4.31	4.24	5.03
3750	3.34	4.14	3.59	4.33	4.27	5.06
4000	3.37	4.16	3.62	4.36	4.30	5.09

表 2 定位加显性效应 QTL 的一维扫描中, MD = 20 cM 时 3 种群体在不同基因组长度下的 LOD 临界值
Table 2 LOD threshold of three populations with different genome lengths in one dimensional scanning of additive-dominant QTL when MD = 20 cM

基因组长度 Genome length (cM)	BC_1		RIL		F_2	
	α_g=0.05	α_g=0.01	α_g=0.05	α_g=0.01	α_g=0.05	α_g=0.01
250	1.83	2.64	1.92	2.63	2.58	3.35
500	2.10	2.92	2.20	2.91	2.88	3.65
750	2.27	3.09	2.36	3.07	3.06	3.82
1000	2.38	3.21	2.48	3.19	3.18	3.95
1250	2.47	3.30	2.57	3.28	3.28	4.05
1500	2.55	3.37	2.64	3.36	3.36	4.13
1750	2.61	3.44	2.70	3.42	3.42	4.19
2000	2.66	3.49	2.76	3.47	3.48	4.25
2250	2.71	3.54	2.80	3.52	3.53	4.30
2500	2.75	3.58	2.85	3.57	3.58	4.35
2750	2.79	3.62	2.89	3.61	3.62	4.39
3000	2.83	3.66	2.92	3.64	3.66	4.43
3250	2.86	3.69	2.95	3.67	3.69	4.46
3500	2.89	3.72	2.98	3.71	3.73	4.49
3750	2.92	3.75	3.01	3.73	3.76	4.52
4000	2.94	3.78	3.04	3.76	3.78	4.55

维扫描中, MD = 1 cM 和 20 cM, α_g=0.05 和 0.01 时, BC_1、RIL 和 F_2 三个群体在不同基因组大小下的 LOD 临界值。基因组越大, 标记越密, 显著性水平越高, LOD 临界值越大。基因组大小相等时, RIL 群体的 LOD 临界值大于 BC_1 群体, 小于 F_2 群体(表 1 和表 2)。虽然 RIL 群体的趋势线斜率与 F_2 群体的比较接近(图 3-c 和 e, 图 3-d 和 f), 但是二者的自由度不同, RIL 群体的 df = 1, F_2 群体的 df = 2, 故 F_2 群体的 LOD 临界值较 RIL 群体的大。由表 1 和表 2 可知, 标记密度为 1~20 cM, 基因组长度为 250~4000 cM 时, BC_1、RIL 和 F_2 三个群体在全局显著性概率水平 α_g=0.05 时的 LOD 临界值范围分别为 1.83~3.37、1.92~3.62 和 2.58~4.30; 在全局显著性概率水平 α_g=0.01 时的 LOD 临界值范围分别为 2.64~4.16、2.63~4.36 和 3.35~5.09。

2.4 定位上位性互作 QTL 的二维扫描中有效独立检验次数(M_{eff})与标记密度(MD)和染色体长度(CL)的关系

定位上位性互作 QTL 的二维扫描中只研究了标记密度为 5、10 和 20 cM 时的情况。不仅同一条染色体上的 2 个位点间可能存在互作效应, 不同染色体上的 2 个位点间也可能发生互作, 因此我们以 2 条染色体为例研究上位性互作 QTL 临界值的选取。群体大小对最大 LOD 统计量的分布无影响的前提是样本量足够大。通常检测上位性 QTL 需要较大的样本量, 为使研究结果尽可能准确, 我们将群体大小设为 500。

由图 4 可知, 零假设下定位上位性互作 QTL 的二维扫描中, 有效独立检验次数(M_{eff})与染色体长度呈二次幂关系。与定位加显性效应 QTL 的一维扫描一样, 染色体越长, 标记越密, M_{eff} 越大, 且 α_g=0.01 时的 M_{eff} 比 α_g=0.05 时的大。如图 4-a 和 b 所示, BC_1 群体在标记密度为 5、10 和 20 cM, α_g=0.05 时, 函数式的系数分别为 0.037、0.021 和 0.012; α_g=0.01 时, 系数分别为 0.044、0.027 和 0.015。与一维扫描一致, BC_1 群体的系数小于 RIL 群体, F_2 群体的系数与 RIL 群体的接近甚至相等(图 4-c, e)。

在研究如何选取上位性 QTL 的 LOD 临界值时作了 2 个假设, 即基因组内各条染色体的长度都相等, 和基因组内只包含 2 条染色体。但是, 在 QTL 作图中, 基因组内各染色体的长度并不完全相等, 且大部分物种的基因组都包含 2 条以上的染色体。

二维扫描中的最大 LOD 统计量取决于基因组的总长, 与各染色体的长度无关。因此, 可以把整个基因组看作 n 条长度相等的染色体, 每条染色体的长度为整个基因组长度的平均值。二维扫描是在 2 条染色体之间进行的, 基因组内有 n 条染色体时, 可以看作有 C_n^2 种情况, 即整个基因组扫描的有效

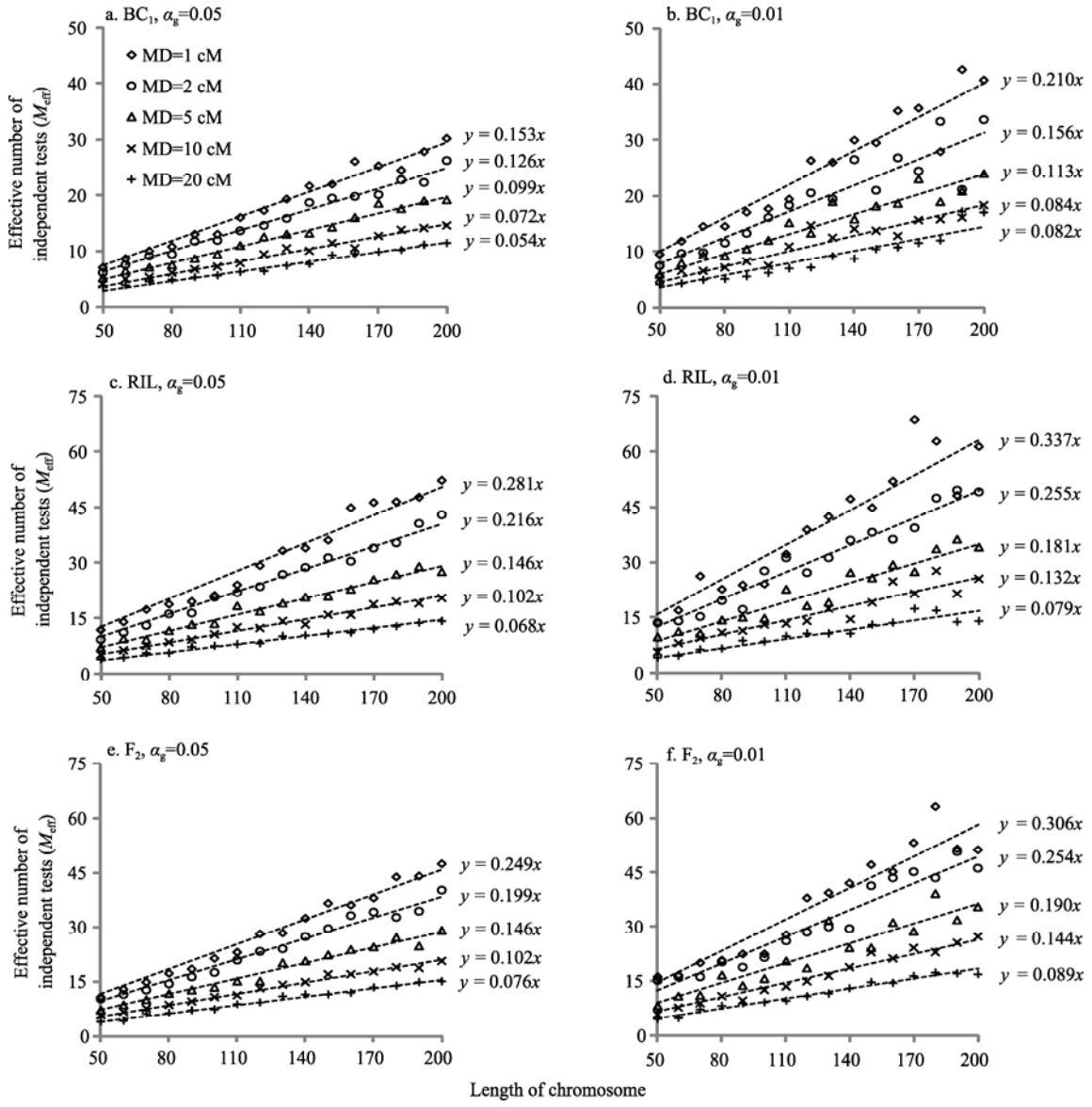

图 3　定位加显性效应 QTL 的一维扫描中不同标记密度下有效独立检验次数(M_{eff})与染色体长度(CL)呈线性关系

Fig. 3　Relationship between the effective number of independent test (M_{eff}) and the length of chromosome (CL) in one dimensional scanning of additive-dominant QTL

a, c, e 分别为 α_g=0.05 时 BC_1、RIL 和 F_2 群体的 M_{eff} 与 CL 的关系; b, d, f 为 α_g=0.01 时 BC_1、RIL 和 F_2 群体的 M_{eff} 与 CL 的关系。

a, c, e correspond to BC_1, RIL, and F_2 population respectively when α_g=0.05 and b, d, f correspond to BC_1, RIL, and F_2 population respectively when α_g=0.01.

独立检验次数为 $C_n^2 M_{eff}$，故根据 Bonferroni 矫正，

$$\alpha_p=\alpha_g/(C_n^2 M_{eff}) \quad (3)$$

表 3 和表 4 分别给出了定位上位性互作 QTL 的二维扫描中，MD = 5 cM 和 20 cM，α_g=0.05 和 0.01 时 BC_1、RIL 和 F_2 三个群体在不同基因组大小下的 LOD 临界值。假定各个基因组内包含 10 条染色体，则每条染色体的长度为 GL/10。根据图 4 求出 M_{eff} 后，乘以 $C_{10}^2=45$ 即可得到整个基因组的独立检验次数。然后利用 Bonferroni 矫正得到全基因组显著性水平 α_g=0.05 或 0.01 对应的单个扫描位点的显著性水平 α_p，再利用公式(3)求得对应的 LOD 临界值。其中，BC_1 和 RIL 群体的 df=1，F_2 群体的 df=4。

由表 3 和表 4 可以看出，上位性 QTL 的 LOD 临界值显著大于加显性 QTL 的临界值。标记密度为 5~20 cM，基因组大小为 250~4000 cM，BC_1、RIL 和 F_2 三个群体在全局显著性概率水平 α_g=0.05 时的 LOD 临界值范围分别为 3.13~5.90、3.19~6.06 和 5.11~8.19；在全局显著性概率水平 α_g=0.01 时的 LOD 临界值范围分别为 3.88~6.65、3.98~7.79 和 5.98~9.90。

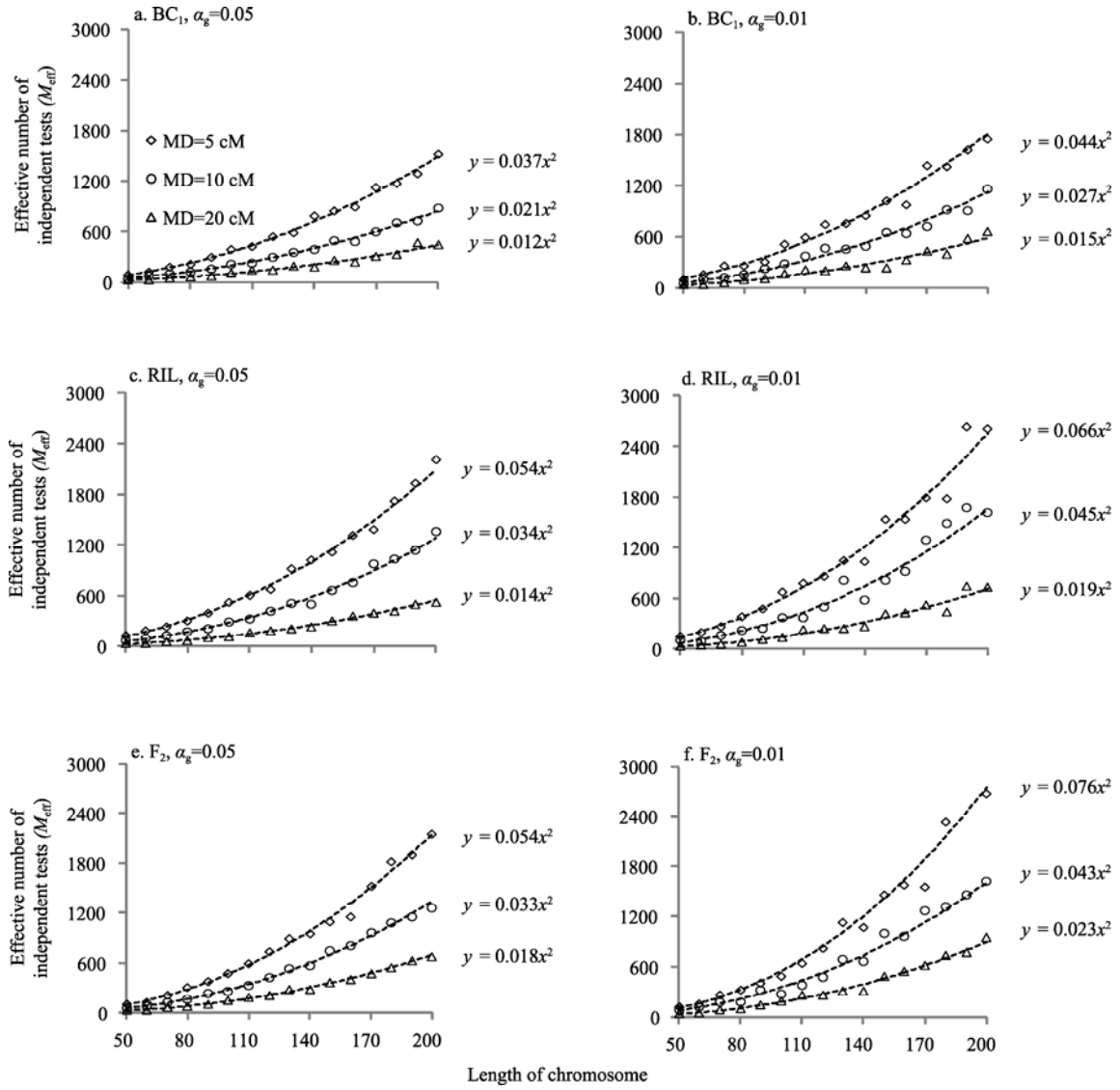

图 4 定位上位性互作 QTL 的二维扫描中不同标记密度下有效独立检验次数(M_{eff})与染色体长度(CL)呈二次幂关系

Fig. 4 Relationship between the effective number of independent test (M_{eff}) and the length of chromosome (CL) in two dimensional scanning of epistatic QTL

a、c、e 分别为 α_g=0.05 时 BC_1、RIL 和 F_2 群体的 M_{eff} 与 CL 的关系, b、d、f 分别为 α_g=0.01 时 BC_1、RIL 和 F_2 群体的 M_{eff} 与 CL 的关系。

a, c, e correspond to BC_1, RIL, and F_2 population respectively when α_g=0.05 and b, d, f correspond to BC_1, RIL, and F_2 population respectively when α_g=0.01.

2.5 不同方法估算的 LOD 临界值

为评价本研究提出的方法的可靠性, 我们利用 3 个真实群体小麦 DH 群体、玉米 RIL 群体、水稻 F_2 群体比较了 Cheverud [13]、Li 和 Ji [15]、Gao 等[16]和本研究提出的估算有效独立检验次数(M_{eff})的方法, 并将它们与模拟法[2]和排列检验法[12]得到的临界值比较。由于 Cheverud [13]、Li 和 Ji [15]、Gao 等[16]法仅能估算加显性效应 QTL 的 LOD 临界值, 因此, 本研究只比较了定位加显性效应 QTL 的一维扫描中全局显著性水平为 0.05 时的 LOD 临界值。

由表 5 可以看出, Cheverud [13]、Li 和 Ji [15]、Gao 等[16] 3 种方法中, Li 和 Ji [15]法得到的临界值均最小, Gao 等[16]法次之, Cheverud [13]法得到的临界值最大。与模拟法和排列检验法相比, 除了 Cheverud [13]估算小麦 DH 群体的 LOD 临界值较高之外, 在其他情况下, 这 3 种方法得到的临界值均较小。特别是水稻 F_2 群体, 模拟法和排列检验法得到的 LOD 临界值显著高于前 3 种方法。

表 3　定位上位性互作 QTL 的二维扫描中, MD = 5 cM 时 3 种群体在不同基因组长度下的 LOD 临界值
Table 3　LOD threshold of three populations with different genome lengths in two dimensional scanning of epistatic QTL when MD = 5 cM

基因组长度 Genome length (cM)	BC_1		RIL		F_2	
	α_g=0.05	α_g=0.01	α_g=0.05	α_g=0.01	α_g=0.05	α_g=0.01
250	3.59	4.33	3.74	5.45	5.63	6.98
500	4.16	4.90	4.32	6.04	6.27	7.71
750	4.50	5.24	4.65	6.38	6.65	8.14
1000	4.74	5.48	4.89	6.62	6.92	8.44
1250	4.92	5.67	5.08	6.81	7.12	8.68
1500	5.08	5.82	5.23	6.96	7.29	8.87
1750	5.20	5.95	5.36	7.09	7.43	9.03
2000	5.32	6.06	5.47	7.20	7.55	9.17
2250	5.41	6.16	5.57	7.30	7.66	9.29
2500	5.50	6.25	5.66	7.39	7.76	9.40
2750	5.58	6.33	5.74	7.47	7.85	9.50
3000	5.66	6.40	5.81	7.55	7.93	9.59
3250	5.72	6.47	5.88	7.61	8.00	9.68
3500	5.78	6.53	5.94	7.68	8.07	9.76
3750	5.84	6.59	6.00	7.74	8.13	9.83
4000	5.90	6.65	6.06	7.79	8.19	9.90

表 4　定位上位性互作 QTL 的二维扫描中, MD = 20 cM 时 3 种群体在不同基因组长度下的 LOD 临界值
Table 4　LOD threshold of three populations with different genome lengths in two dimensional scanning of epistatic QTL when MD = 20 cM

基因组长 Genome length (cM)	BC_1		RIL		F_2	
	α_g=0.05	α_g=0.01	α_g=0.05	α_g=0.01	α_g=0.05	α_g=0.01
250	3.13	3.88	3.19	3.98	5.11	5.98
500	3.70	4.45	3.76	4.55	5.76	6.62
750	4.03	4.79	4.09	4.89	6.14	7.00
1000	4.27	5.03	4.33	5.13	6.41	7.26
1250	4.45	5.22	4.52	5.32	6.61	7.47
1500	4.61	5.37	4.67	5.47	6.78	7.64
1750	4.73	5.50	4.80	5.60	6.92	7.78
2000	4.85	5.61	4.91	5.71	7.05	7.90
2250	4.94	5.71	5.01	5.81	7.16	8.01
2500	5.03	5.80	5.10	5.90	7.25	8.10
2750	5.11	5.88	5.18	5.98	7.34	8.19
3000	5.18	5.95	5.25	6.05	7.42	8.27
3250	5.25	6.02	5.32	6.12	7.49	8.34
3500	5.31	6.08	5.38	6.18	7.56	8.41
3750	5.37	6.14	5.44	6.24	7.63	8.48
4000	5.42	6.19	5.49	6.29	7.69	8.53

表 5　不同方法估算的 LOD 临界值
Table 5　LOD threshold calculated by different methods

群体 Population	Cheverud [13]	Li & Ji [15]	Gao [16]	M_{eff} [a]	模拟法 Simulation	排列检验法 Permutation test
小麦 DH 群体 Wheat DH population	3.18	2.78	2.81	3.08	3.00	3.12
玉米 RIL 群体 Maize RIL population	2.74	2.60	2.70	2.80	2.88	2.86
水稻 F_2 群体 Rice F_2 population	2.69	2.52	2.63	3.50	4.10	4.28

[a] 为本研究提出的方法。[a] Method proposed by this study.

对于 3 个真实群体来说，本研究提出的方法估算的临界值较前 3 种方法更接近排列检验和模拟法得到的临界值。小麦 DH 群体和玉米 RIL 群体的临界值与排列检验和模拟法相差不超过 0.1。虽然本研究提出的方法估算水稻 F_2 群体的临界值与排列检验和模拟方法相差较大，但与前 3 种方法相比，是最接近的方法。因此，可以说明本研究提出的方法是准确可靠的。

3 讨论

本研究通过模拟方法估算全局显著性水平为 0.05 和 0.01 时定位加显性效应 QTL 的一维扫描和定位上位性互作 QTL 的二维扫描有效独立检验次数(M_{eff})与染色体长度和标记密度的关系，并将其推广至整个基因组，再利用 Bonferroni 矫正和卡方分布，即可推算出作图群体的 LOD 临界值。需要注意的是，图 3 和图 4 中，α_g=0.01 时的有效独立检验次数的波动较 α_g=0.05 时的大，原因可能是随机误差的影响和模拟次数的限制。为保证较高显著性水平下得到更准确的结论，群体大小和模拟次数应尽可能大。故对于 10 000 次模拟来说，α_g=0.05 时得到的有效独立检验次数更稳定，即得到的结论更准确可靠。一般而言，检测 BC_1 和 RIL 群体的加显性效应 QTL 显著性水平为 0.05 (即 α_g=0.05)时 LOD 临界值应为 2~3 [1]。对于 F_2 群体，应提高至 2.6~4.3。上位性互作 QTL 的 LOD 临界值显著大于加显性效应 QTL，BC_1 和 RIL 群体上位性 QTL 的临界值范围为 3.1~6.0, F_2 群体的为 5.1~8.2。

本文研究的 3 种群体类型 BC_1、RIL 和 F_2 群体，具有广泛的代表性。其中, BC_1 群体含有 2 种基因型，经过一次重组产生; RIL 群体含有 2 种基因型，经过多次重组产生; F_2 群体含有 3 种基因型，经过一次重组产生。此外, DH 群体也是常用的 QTL 作图群体。DH 群体的临界值计算方法与 BC_1 群体的相同，因为它也含有 2 种基因型，经历一次重组产生。表 5 中小麦 DH 群体的 LOD 临界值是根据 BC_1 群体推算出来的。

本文用 IM 法[1]分析得到的结果也适用于完备区间作图(ICIM)法[17-18,24]。因为零假设下一维扫描和二维扫描中是不包含任何变量的。此外，在用 IM 和 ICIM 等方法进行 QTL 作图时，常需要设置扫描步长这一作图参数，即 QTL 作图时在全基因组上每隔一定的遗传距离进行一次假设检验。通常，扫描步长应小于平均标记密度。步长大于平均标记密度导致标记信息不能被充分利用，相当于减少了标记量，从而也会对 LOD 临界值的选取造成影响。当步长小于平均标记密度时，步长的大小不会影响 LOD 临界值的选取。检测上位性 QTL 时，常将扫描步长设为 5 cM，因此，在研究上位性 QTL 的 LOD 临界值选取时只研究了标记密度为 5、10 和 20 cM 时的情况。

选取 LOD 临界值时还要结合不同的研究目标[25]。如果 QTL 作图只是初步确定基因在染色体上的位置, 然后根据作图结果构建其他次级群体对检测到的 QTL 进行精细定位、图位克隆和转基因工作, 这类研究几乎不容许假 QTL 的发生。此时要适当提高检验 QTL 时的 LOD 临界值，保证后续研究中 QTL 的可靠性。另一方面, 如果研究目标是把 QTL 作图结果用于标记辅助选择聚合育种, 这时只有尽可能多地检测出控制育种目标性状的 QTL，才能保证对所有控制育种性状的基因进行选择，因此有必要适当降低检验 QTL 时的 LOD 临界值,使得遗传效应较小的 QTL 也有机会被检测出来。即使有一些假 QTL 的存在，也不会造成很大的损失。

4 结论

定位加显性效应 QTL 的一维扫描和上位性互作 QTL 的二维扫描中，单个扫描位置上的 LRT 统计量在零假设下均服从卡方分布，其自由度等于检测 QTL 遗传参数的个数; 影响 LOD 临界值选取的因素有群体类型、标记密度和基因组长度; 在 BC_1、RIL 和 F_2 三种群体中, BC_1 群体的临界值最小, F_2 群体的最大; 标记越密，基因组越大，对应的 LOD 临界值越高。定位加显性效应 QTL 的一维扫描中，有效独立检验次数(M_{eff})与基因组长度呈正比; 定位上位性互作 QTL 的二维扫描中，有效独立检验次数(M_{eff})与基因组长度呈二次幂关系。借助 Bonferroni 矫正，给出了全基因组显著性水平与单个扫描位点显著性水平间的关系，研究者可根据作图群体的群体类型、标记密度和基因组长度，选定特定全局显著性概率水平下的 LOD 临界值。通常情况下，全基因组显著性概率为 0.05 时，BC_1、RIL 和 F_2 群体加显性效应 QTL 的 LOD 临界值范围分别为 1.83~3.37、1.92~3.62 和 2.58~4.30; 上位性互作 QTL 的 LOD 临界值范围分别为 3.13~5.90、3.19~6.06 和 5.11~8.19。

References

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

[2] van Ooijen J W. LOD significance thresholds for QTL analysis in experimental populations of diploid species. *Heredity*, 1999, 83: 613–624

[3] Dupuis J, Siegmund D. Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics*, 1999, 151: 373–386

[4] Feingold E, Brown P O, Siegmund D. Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *Am J Hum Genet*, 1993, 53: 234–251

[5] Rebaï A, Goffinet B, Mangin B. Approximate thresh olds of interval mapping tests for QTL detection. *Genetics*, 1994, 138: 235–240

[6] Dupuis J. Statistical Problems Associated with Mapping Complex and Quantitative Traits from Genomic Mismatch Scanning Data. PhD Dissertation of Stanford University, 1994

[7] Kruglyak L, Lander E S. A nonparametric approach for mapping quantitative trait loci. *Genetics*, 1995, 139: 1421–1428

[8] Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet*, 1995, 11: 241–247

[9] Piepho H P. A quick method for computing approximate thresholds for quantitative trait loci detection. *Genetics*, 2001, 157: 425–432

[10] Zou F, Fine J P, Hu J H, Lin D Y. An efficient resampling method for assessing genome-wide statistical significance in mapping quantitative trait loci. *Genetics*, 2004, 168: 2307–2316

[11] Chang M N, Wu R, Wu S S, Casella G. Score statistics for mapping quantitative trait loci. *Stat Appl Genet Mol Biol*, 2009, 8: 16

[12] Churchill G A, Doerge R W. Empirical threshold values for quantitative trait mapping. *Genetics*, 1994, 138: 963–971

[13] Cheverud J M. A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, 2001, 87: 52–58

[14] Nyholt D R. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet*, 2004, 74: 765–769

[15] Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 2005, 95: 221–227

[16] Gao X, Starmer J, Martin E R. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol*, 2008, 32: 361–369

[17] Li H, Ribaut J M, Li Z, Wang J. Inclusive composite interval mapping (ICIM) for digenic epistasis of quantitative traits in biparental populations. *Theor Appl Genet*, 2008, 116: 243–260

[18] Zhang L, Li H, Li Z, Wang J. Interactions between markers can be caused by the dominance effect of QTL. *Genetics*, 2008, 180: 1177–1190

[19] Zhang L, Li H, Wang J. Statistical power of inclusive composite interval mapping in detecting digenic epistasis showing common F_2 segregation ratios. *J Integr Plant Biol*, 2012, 54: 270–279

[20] Doerge R W, Rebaï A. Significance thresholds for QTL interval mapping tests. *Heredity*, 1996, 76: 459–464

[21] Wang J, Chapman S C, Bonnett D G, Rebetzke G J. Simultaneous selection of major and minor genes: use of QTL to increase selection efficiency of coleoptiles length of wheat (*Triticum aestivum* L.). *Theor Appl Genet*, 2009, 119: 65–74

[22] Ribaut J M, Hoisington D A, Deutsch J A, Jiang C, González-de-León D. Identification of quantitative trait loci under drought conditions in tropical maize: 1. Flowering parameters and the anthesis-silking interval. *Theor Appl Genet*, 1996, 92: 905–914

[23] Zhang L, Wang S, Li H, Deng Q, Zheng A, Li S, Li P, Li Z, Wang J. Effects of missing marker and segregation distortion on QTL mapping in F_2 populations. *Theor Appl Genet,* 2010, 121: 1071–1082

[24] Li H, Ye G, Wang J. A modified algorithm for the improvement of composite interval mapping. *Genetics*, 2007, 175: 361–374

[25] Li H-H(李慧慧), Zhang L-Y(张鲁燕), Wang J-K(王建康). Analytical answers to frequently asked questions in quantitative trait locus mapping. *Acta Agron Sin* (作物学报), 2010, 36(6): 918–931 (in Chinese with English abstract)