

关联图分析: 一种基于连锁图的 QTL 综合分析方法

匡峰磊, 汪霞, 周玲, 章元明*

南京农业大学作物遗传与种质创新国家重点实验室, 南京 210095

* 联系人, E-mail: soyzhang@njau.edu.cn; soyzhang@hotmail.com

2010-06-28 收稿, 2010-08-04 接受

国家自然科学基金(30971848)、江苏省自然科学基金(BK2008335)和教育部新世纪优秀人才支持计划(NCET-05-0489)资助项目

摘要 通过统合分析整合多个独立实验 QTL (quantitative trait locus)定位结果获得的一致性 QTL, 可为标记辅助育种和基因图位克隆提供坚实基础. 但是这种方法没有充足的生物学背景支撑, 而且遗传图谱间不同映射会严重影响分析结果. 为此, 本文提出一种基于连锁图的 QTL 综合分析法, 简称关联图分析. 首先, 利用连锁群构建图模型; 然后, 利用非监督分类法来寻找基因组间具有模式性共分离趋势的标记区间; 最后, 通过频繁集挖掘法对这些标记区间挖掘以获得与 QTL 紧密连锁的分子标记(区间). 新方法通过 Monte Carlo 模拟研究和 3 个棉花 QTL 定位结果的综合分析予以证实. 新方法的优点在于: 不需要构建参考图谱; 利用基因组间的分子标记模式性共分离趋势来查找特异性 QTL, 在一定程度上排除了初始 QTL 的干扰; 通过非特异性 QTL 获得的具有潜在育种价值的分子标记, 在作物育种上更有利用价值.

关键词

连锁图
数量性状基因座
统合分析
关联图分析

数量性状基因座(QTL)定位可为基因图位克隆^[1,2]和标记辅助育种^[3]提供重要的依据, 但是 QTL 定位结果受到亲本、作图群体类型、标记组合、实验误差和统计方法等因素的影响. 实践中, 在不同情形下的 QTL 定位结果是否具有一致性是非常重要的. 这就需要探索比较 QTL 结果间一致性的方法学.

统合分析是最早用于比较 QTL 结果间一致性的方法^[4]. 它通过关键标记(core marker)连结在不同独立研究中得到的 QTL 及其置信区间, 利用改良的 AIC (Akaike's information criterion)准则对不同的非生物学模型进行筛选获得最优模型, 得到一致性 QTL 的位置和置信区间. BioMercator 软件包可实现统合分析, 已在奶牛^[5]、玉米^[6]、小麦^[7]和棉花^[8]等动植物中得到应用. 统合分析整合遗传连锁图谱, 缩小了 QTL 置信区间, 提高了 QTL 定位精度^[9]. 但是, 统合分析也存在不足: 统合分析方法和 Homothetic 函数^[6]没有生物学背景支撑; 初始 QTL 位置、置信区间以及图谱之间的映射对统合分析结果干扰严重^[7];

AIC 准则在模型选择中功效较低^[10].

为了准确地进行 QTL 综合分析, 需要根据基因组之间的相似性来合并或分类 QTL. 虽然目前有利用蛋白质域^[11,12]、基因 Homology 矩阵^[13,14]和最大间距基因簇^[15,16]等多种识别方法来揭示基因组区域部分或全部保守性, 但是这些方法都是基于碱基序列的. 碱基序列虽然信息内容丰富, 但是不适用于基于连锁图谱的 QTL 分析. 减数分裂中的同源重组机制, 不仅能对同源染色体进行配对, 还可以有效地交换同源染色体间的基因组信息, 而且大部分重组发生在基因外部^[17], 不会对基因产生影响. 因此同源重组在真核生物减数分裂时期修复 DNA、保持生物保守性、引起染色体结构变异和促进生物进化上起着重要作用. 基因组锚点在同源重组中有共分离(co-segregation)趋势. 这种趋势表现了生物对基因组信息操作的自发过程, 较序列信息, 度量简单, 可以用重组率表示. 由此本文提出一种新的基于连锁图的 QTL 综合分析方法, 简称关联图分析. 该方法利用连

锁群构建图模型, 搜索连锁群上相似的标记区间, 以获得与 QTL 紧密连锁的分子标记.

1 算法

1.1 初始化连锁群

对于一个连锁群上的多个标记, 可用一个连锁 G 图表示. 在连锁 G 图中, 顶点为标记, 两两连锁标记用线连接, 这线称为边, 边的权重用标记间重组率表示. 若已知标记间遗传距离, 则可用 Kosambi 函数转换^[18]. 运算过程详见算法 1.

算法 1 初始化连锁图

输入: 收集各个独立实验的连锁群 $L_i, i \in \{1, 2, \dots, n\}$

输出: 各个连锁群的连锁 G 图 LG_i

1. foreach L_i 包含标记 $m_{i1}m_{i2}\dots m_{ik}$, 区间长度为 $l_j^{j+1}, j \in \{1, 2, \dots, k-1\}$
2. if l_j^{j+1} 为遗传距离时, 标记间重组率 r_j^{j+1} 用 Kosambi 函数转换
3. else $r_j^{j+1} = l_j^{j+1}$
4. return $LG_i = \left\{ \left\{ \{m_{i1}, m_{i2}\}, r_1^2 \right\}, \dots, \left\{ \{m_{i1}, m_{ik}\}, r_1^k \right\}, \dots, \left\{ \{m_{i(k-1)}, m_{ik}\}, r_{k-1}^k \right\} \right\}$

1.2 搜索相似区间

将上述连锁 G 图折分成一组数据, 每一数据对象为一个边的信息, 它包含该边的两端标记和标记间的重组率, 根据无监督聚类分析方法的“最大化组内相似性、最小化组间相似性”原则将边数据对象划分为若干组^[19,20]. 对于组内数据对象个数大于 1 的组, 说明组内标记区间在不同连锁群间的共分离现象有一定的模式性.

利用这些组中的边对原始连锁 G 图构建子图, 则子图中任何一边都有着模式性. 这个子图称为约简图. 约简图上每对标记的共分离趋势都能在其他连锁群上发现. 运算过程详见算法 2.

算法 2 图约简

输入: 各个连锁群的连锁 G 图 $LG_i, i \in \{1, 2, \dots, n\}$

输出: 各个连锁群的约简图 RLG_i

1. 定义标记集合 M , 这个集合中的标记至少存在于两个连锁群上
2. foreach LG_i
3. $LG_i = \left\{ \left\{ \{M_p, M_q\}, r_{M_q}^{M_p}(i) \right\}, \dots \right\}$

4. 定义边集合 E

5. 对 $\forall \left\{ \left\{ \{M_p, M_q\}, r_{M_q}^{M_p}(i) \right\}, i \in \{1, 2, \dots, n\} \right\}$ 聚类

得到 cluster

6. if $|\text{cluster}_b| \neq 1$, then $E := E \cup \text{cluster}_b$
7. foreach i
8. $RLG_i = \left\{ \left\{ \{M_p, M_q\}, r_{M_q}^{M_p}(i) \right\} \in E \right\}$
9. return RLG_i

一个连锁群上的标记彼此之间都有连锁关系, 因此在连锁 G 图中, 这些标记满足传递性关系, 能构成一个完全图. 对一个完全图中的 n 个标记构建子图, 这个子图仍为完全图. 因此连锁图中由任意 n 个标记组成的标记区间在连锁 G 图中为一个完全子图. 然而对于约简图, 由于图中仅保留了具有模式性共分离的边, 导致标记之间并不一定满足传递性. 因此约简图中任意 n 个标记不一定满足传递性并构成连锁图意义上的标记区间. 只有当这些标记构造子图为一个完全子图, 说明这些标记满足传递性关系, 并且这些标记构成的连锁图标记区间与其他连锁群上的具有结构相似性. 这种标记区间在这里称为相似区间. 运算过程详见算法 3.

算法 3 搜索标记所在的相似区间

输入: 一个约简图 RLG , 目标标记 $m_i, i \in \{1, 2, \dots, n\}$

输出: 相似区间 B_m

1. 定义 M 为相邻标记
2. foreach i
3. $M := M \cup$ 标记 m_i 在 RLG 中的的相邻顶点
4. 通过 M 生成子图
5. 对上述子图生成最大完全子图集合 C_i
6. foreach C_i
7. if $\forall m_i \in C_j$ 的顶点集合, then $B_m := B_m \cup C_i$ 的顶点集合
8. return B_m

1.3 QTL 综合分析

对于一个检测到的 QTL, 根据检测方法说明它与标记 $\{m_i\}$ 紧密连锁. 对 $\{m_i\}$ 及其相邻的标记在约简图上构建最大完全子图, 其中包含 $\{m_i\}$ 的最大完全子图 G 不仅说明 $\{m_i\}$ 在其他连锁群有相似的共分离趋势, 而且拓展了对 QTL 的认识. 运算过程详见算法 4.

算法 4 QTL 综合分析

输入: 目标 QTL $i, i \in \{1, 2, \dots, n\}$

输出: 与这些 QTL 有关联的标记

1. foreach QTL_i
2. 根据检测方法获得相关的标记集合 $\{m_i\}$
3. 根据所在连锁群的约简图, 搜索标记集合 $\{m_i\}$ 所在的相似区间集合 B_{m_i} , 其中 B_{m_i} 为 QTL_i 所有相关的相似区间
4. 对所有 B_{m_i} 利用频繁集模式方法挖掘最频繁的标记

2 Monte Carlo 模拟研究

随机模拟一条标记数为 21, 长度为 100 cM 的连锁群为初始连锁群, 并随机选择一段区域随机发生一次易位、倒位或缺失变异, 重复 9 次, 得到 1 次倒位、2 次易位和 6 次缺失的 9 条变异连锁群(表 S1). 将遗传率为 15% 的 QTL 设置在未发生染色体变异的标记 m12 上. 利用初始连锁群和 9 条变异连锁群, 分别产生样本容量为 250 的回交、 F_2 和重组自交系群体的标记基因型值和数量性状表型值. 利用 Windows QTL Cartographer v2.5 软件对这 30 个数据集进行 QTL 定位. QTL 定位结果见表 1, 其中缺失变异连锁群 8 的 RIL 群体未检测出 QTL. 利用新方法和统合分析分别对这些 QTL 定位结果进行分析.

统合分析使用 BioMercator v2.1 软件. 新方法使用 R 语言^[21]编译: 无监督分类由 cluster 包实现, 最大完全子图搜索由 igraph 包^[22]实现, 频繁集挖掘由 arules 包^[23]实现.

在统合分析中, 以原始图谱为参考图谱进行分析, 认为一致性 QTL 个数不少于 4, 即无法进一步分析(图 1). 但仍然发现在标记 m12 上存在 QTL. 若以易位的第 6 连锁图谱作参考图谱, 获得 4 个一致性 QTL

(图 1, 表 2). 虽然有一个一致性 QTL 位于标记 m12 上且明显缩小了置信区间, 但是其他的一致性 QTL 则是假的. 相比较于第 6 连锁图谱, 原始图谱信息更为准确, 但是却影响了统合分析的进一步实施. 因此, 参考图谱的设定影响统合分析结果. 此外, 两种映射均发现存在一些无法排除的假阳性一致性 QTL.

利用新方法对表 1 结果进行 QTL 综合分析时不需要设置参考图谱. 首先, 把 10 个连锁群转换为连锁 G 图(表 S2), 进行图约简后得到各个连锁群的约简图(表 S3). 结果表明, 第 10 连锁图在约简后与原图保持一致, 说明原始信息得到有效的保存和利用, 但是模式性识别在其他连锁图中存在一定偏差, 说明需要引入更好的聚类方法. 其次, 得到每个 QTL 相关的相似区域(表 S4). 最后, 对所有可辨 QTL 的相似区间进行频繁集挖掘, 分析结果见表 3. 结果表明, 这些 QTL 与标记区间 $\{m11, m12\}$ 具有模式性共分离趋势, 说明新方法更为准确.

可以发现, 第 2 连锁图中与标记 m6 关联的假阳性 QTL, 在第 2 次统合分析中被映射到 m9 标记上的同时认为属于一个一致性 QTL(图 2), 而该 QTL 被新方法视为具有特异性, 不会进入到下一步的分析; 图 2 说明了这个 QTL 所在标记区间 $\{m6, m10, m11\}$ 为变异区域. 在总共 29 个 QTL 中, 有 2 个被排除. 但是通过相似区间来寻找特异性 QTL 能力是有限的, 如图 3 中第 9 连锁图的一个假阳性 QTL 存在于一个未变异的区域内, 新方法不能识别. 尽管如此, 这些 QTL 最终对 QTL 综合分析影响不大. 表 3 表明, $\{m11, m12\}$ 是置信度 100% 和支持率 85.7% 条件下的 QTL 相关区域, 即这些标记是与 QTL 关联的标记.

表 1 QTL 定位结果

连锁群	回交群体				F_2 群体				重组自交系群体			
	位置	LOD	r^2	置信区间	位置	LOD	r^2	置信区间	位置	LOD	r^2	置信区间
1	38.0	6.69	0.1180	34.8~42.2	34.0	10.92	0.1679	31.3~40.4	37.0	3.61	0.0659	32.7~42.7
2	32.0	8.03	0.1309	29.4~33.8	39.0	6.20	0.0856	38.0~41.9	44.0	4.07	0.0924	43.3~47.6
3	88.0	5.58	0.0969	87.5~93.0	72.0	6.39	0.0501	71.0~73.8	50.0	10.12	0.1758	48.0~55.0
4	56.0	5.90	0.0998	53.3~59.9	43.0	8.0	0.1006	40.7~51.0	45.0	2.94	0.0586	40.8~48.0
5	55.0	5.40	0.1027	52.8~59.5	57.0	9.93	0.0275	55.7~60.0	56.0	3.32	0.0542	54.3~61.8
6	49.0	7.36	0.1266	46.5~53.0	50.0	7.50	0.1480	48.0~52.3	50.0	2.56	0.0431	45.1~54.1
7	56.0	4.57	0.0680	53.8~62.0	56.0	10.96	0.0543	53.8~57.9	60.0	7.97	0.1432	56.0~63.1
8	48.0	3.70	0.0634	41.8~48.5	51.0	6.71	0.0458	48.6~52.0				
9	27.0	2.75	0.0508	25.7~30.0	38.0	9.41	0.0810	37.0~41.2	38.0	3.75	0.0574	37.0~43.4
10	54.0	6.08	0.1089	52.0~60.9	56.0	6.59	0.0350	49.5~61.8	54.0	7.03	0.1173	48.7~58.2

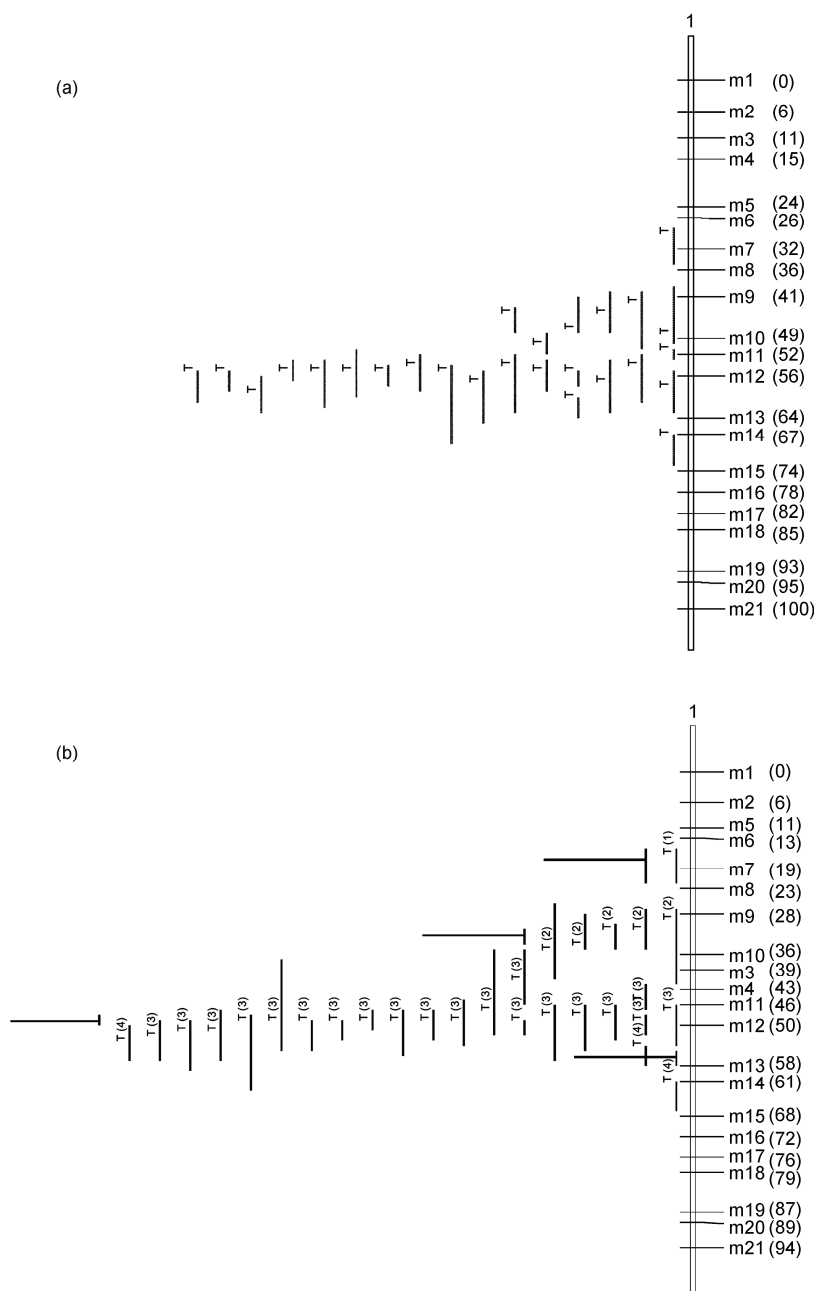


图 1 以第 6 (a)和 10 (b)连锁群为参考图谱时检测 QTL 的统合分析
QTL 位置及其置信区间用直线表示, 一致性 QTL 用 T 型图表示

表 2 以第 6 连锁群为参考图谱获得的一致性 QTL

一致性 QTL	相关标记	位置/cM	置信区间/cM
1	m6, m7	17.01	15.00~22.00
2	m9, m10	32.72	31.17~34.27
3	m12	49.65	48.86~50.32
4	m13	56.99	55.46~58.52

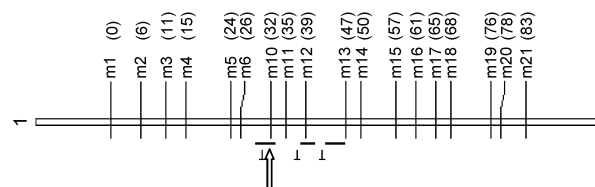


图 2 新方法获得的第 2 连锁图上的特异性 QTL
QTL 位置及置信区间在图中用直线表示, 特异性 QTL 用箭头标出

表 3 基于连锁图的 QTL 综合分析结果

相关标记	支持率	置信度
m11, m12	0.857	1
m10, m11, m12	0.75	1
m12	0.96	0.96

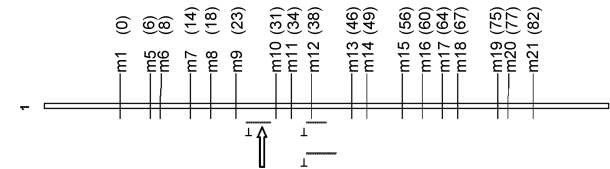


图 3 新方法未能检测出的第 9 连锁图上的假阳性 QTL
QTL 位置及置信区间在图中用直线表示, 假阳性 QTL 用箭头标出

3 实例分析

从棉花数据库 CottonDB (<http://www.cottontdb.org/cgi-bin/cmap>)下载了在 3 个独立研究中获得的 3 条高密度遗传连锁图谱. 群体 I 和 II 实施了 QTL 定位, 而群体 III 则未进行, 详细情况见表 4. 群体 I, II 和 III 遗传连锁群间相似性采用 CMap 平台 (<http://gmod.org/wiki/Cmap>)进行分析, 相似标记识别采用 CMap 的 Auto-name based 方法. 统合分析和新方法见 Monte Carlo 模拟研究.

利用 CMap 平台分析了群体 I, II 和 III 连锁图标记间的相似性, 结果见图 4. 由图 4 可知, 3 群体在 D8 连锁群上有较多相似标记, 但是标记间长度和顺序在不同群体间存在较大差异. 因而, 在选择参考图谱进行统合分析时, 任何群体的遗传图谱都不占绝对优势.

在统合分析中, 采用两种不同的映射方式. 方式 1: 首先把群体 I 的 D8 连锁群映射到群体 II 的 D8 连锁群, 再映射到群体 III 的 D8 连锁群, 即 (I → II) →

III; 方式 2: (I → III) → II. 结果表明, QTL 的信息会随着映射的不同而改变(图 5), 最终导致本例中的统合分析无法识别出一致性 QTL.

通过图 5 中的两种映射关系来说明新方法能克服统合分析受不同映射影响的缺点. 针对标记区间 {CIR202, BNL285}, 首先, 方式 1 的群体 I 和 II 的 D8 连锁群整合获得 3 个最大完全子图, 方式 2 的群体 III 和 I 的 D8 连锁群整合获得 1 个最大完全子图; 然后, 两种映射都获得了一致的 2 个最大完全子图(表 5). 结果表明, 不同组合方式对最终得到的相似区间没有影响.

应用新方法综合分析 D8 两连锁群上的 QTL. 首先, 把 3 群体的 D8 连锁群转换为连锁 G 图, 进行图约简, 得到各个连锁群的约简图; 其次, 在群体 I 和 II 中共定位的 11 个 QTL 中有 7 个 QTL 跟各自约简图上标记相关(表 6). 找到与这些 QTL 相关的相似区域, 结果表明第 4 个 QTL 为一个特异性 QTL, 不进入下一步分析. 最后, 对这 6 个 QTL 进行频繁集合挖掘, 第 1, 2, 5 和 7 个 QTL 与标记区间 {BNL852, BNL2448, CIR062, CIR224, BNL390} 具有模式性共分离趋势(表 7).

4 讨论

Hanocq 等人^[7]指出, 统合分析方法适合于同一作图群体的不同定位实验, 如果用不同群体构建参考图谱, 应当谨慎. 因为使用的经典投影法或齐序函数, 会误解标记的次序和标记之间的重组率. 在本文的统合分析中, 两种不同的映射方法获得了不同的结果也说明了这一问题. 在模拟研究中, 不同参考图谱一致性 QTL 数目不一致; 实例分析却无法识别出一致性 QTL. 应当指出, 新方法不需要建立图谱间映射关系, 减少了信息丢失, 获得了在不同群体中具有模

表 4 不同作图群体及相关信息

群体编号	作图群体	群体大小	标记个数	数量性状 ^{a)} (QTL 数)	参考文献
I	F ₂ [Handan208 × Pima90]	69	59	LI(1)	He 等人 ^[24]
II	BC ₂ [Guazuncho-2 × VH8-4602]	75	63	HS(2), MIC(2) H(1), CRd(2), FUI(1), FEL(2)	Lacape 等人 ^[25] Nguyen 等人 ^[26]
III	F ₂ (2) [CCRI-36 × Hai-7124]	186	89		Yu 等人 ^[27]

a) LI, 铃指标; HS, 纤维标准优异度; MIC, 玛科隆值; H, 纤维线性优异度; CRd, 颜色组分; FUI, 纤维长度一致性指数; FEL, 纤维延展度

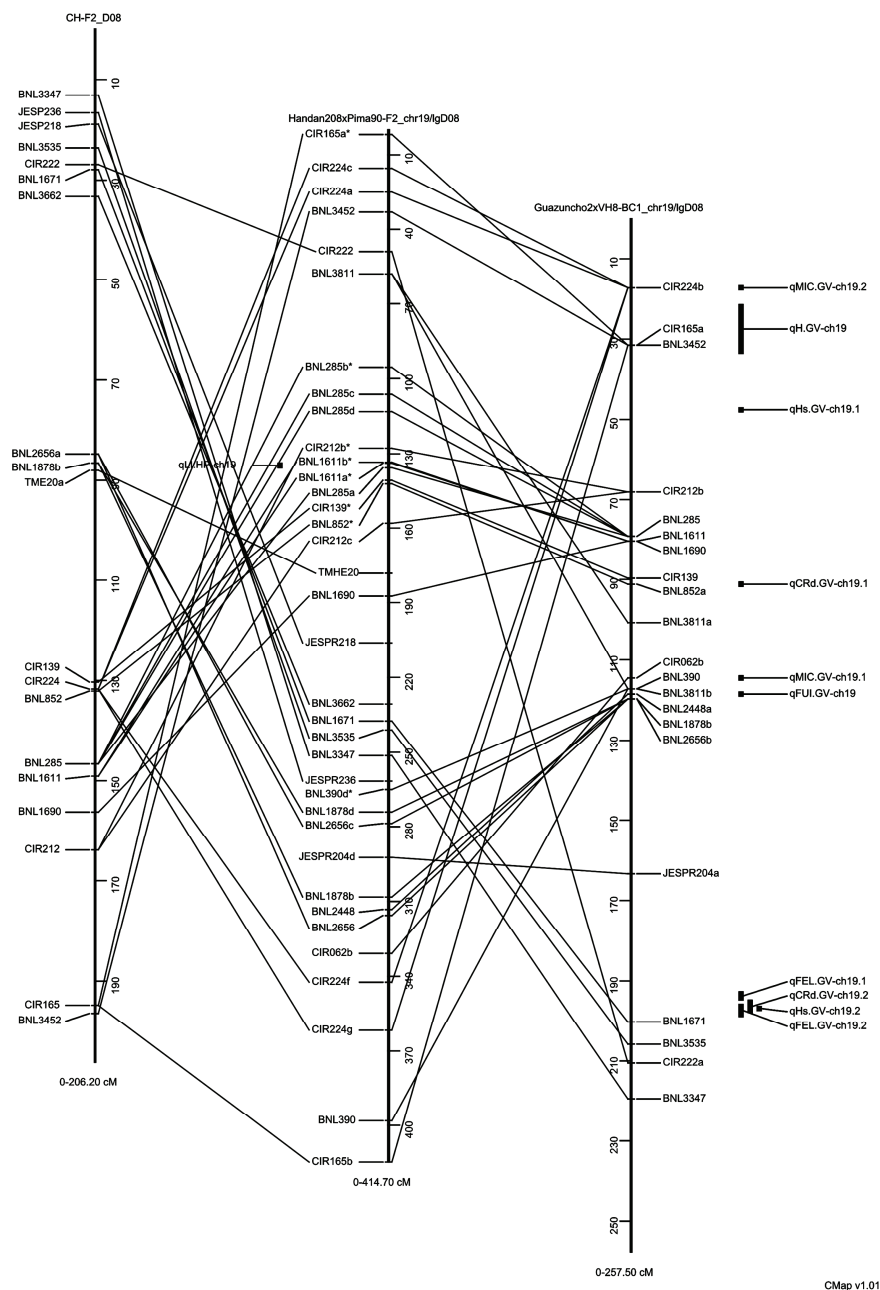


图 4 群体 I、II 和 III 的 D8 连锁群标记相似性分析
连线标记为相似标记, QTL 位置和置信区间列在连锁群旁

式性共分离趋势的标记区间。这些标记区间可以用来识别特异性 QTL, 一定程度排除初始 QTL 干扰, 表现在模拟分析中新方法比统合分析更好地找到了关键标记, 因而对于作物标记辅助育种更具指导意义。

当连锁图谱密度较低时, 所含有基因组信息有限, 会影响新方法的效果。这一缺点会随着分子生

物学的发展而得到解决。Kulathinal 等人^[28]指出, 连锁图中的图距(即一个重组单位含有的碱基对个数: Scale)对研究重组现象有着重要的影响。因为重组的发生是特异的, 重组率并不是随着基因组上两位点物理距离的增加而线性增加, 如人类 60%~70%的重组发生在 10%的基因组区域^[29]。目前通过群体家系

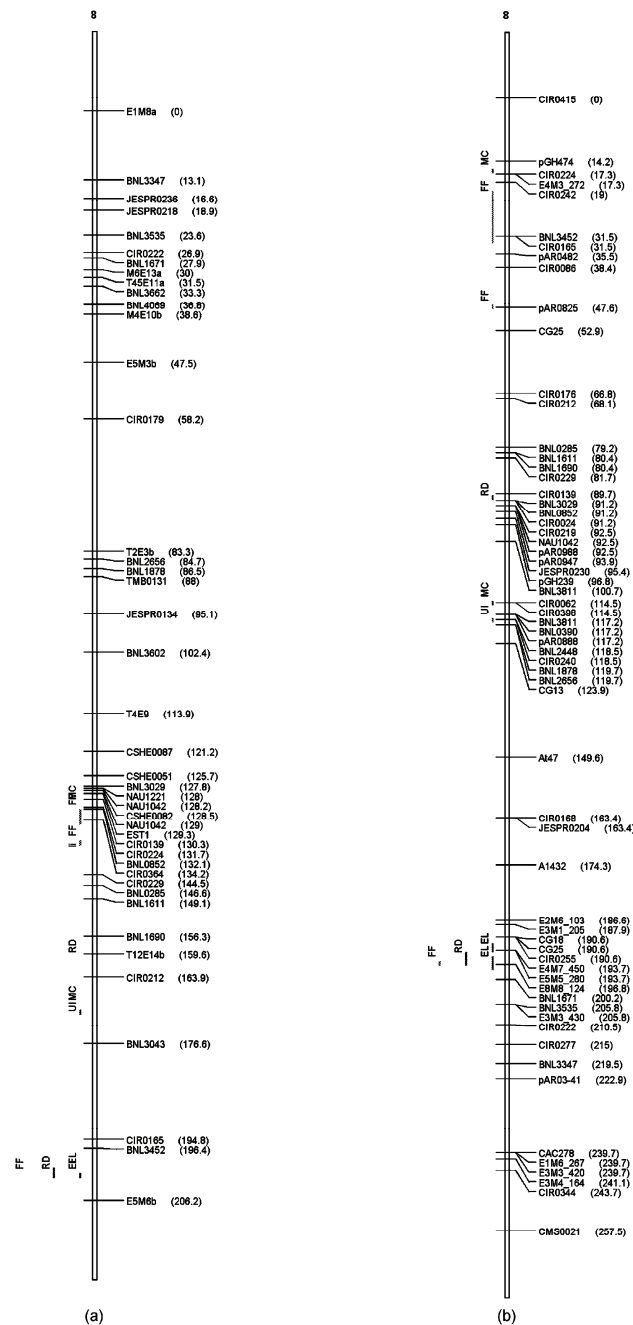


图 5 群体(I → II) → III 的 D8 连锁群间不同映射对统合分析的影响
(a) (I → II) → III; (b) (I → III) → II

构建的连锁图还远远没有达到精细定位连锁图的精度. 配子型连锁作图方式^[29]的发展, 能解决连锁群信息含量低的问题, 适合于快速揭示种内差异. 因此, 利用新方法可以扩展连锁图的线性模型适用于复杂问题的研究, 尤其是目前研究表明重组热点和连锁不平衡有着很重要的联系^[30-33].

Spencer 等人^[34]认为重组率和核苷酸多样性关系并不是由自然选择引起, 而是在减数分裂时期修复 DNA 双链而引起的, 是一个种私有现象. 其他研究也发现重组热点(Hotspots)在物种之间很少共享^[30-33]. 因此, 这种基于重组率方法在种间的比较需要进一步的探讨和研究.

表 5 不同映射中含有标记对{CIR202, BNL285}的相似区间

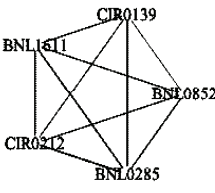
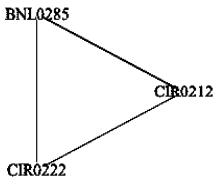
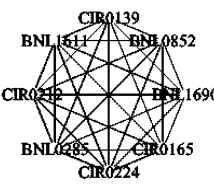
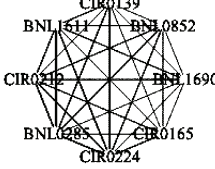
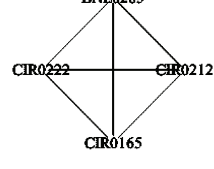
映射方式	第 1 步		第 2 步
(I → II) → III			
(I → III) → II			

表 6 QTL 信息

QTL	作图群体	连锁群	连锁的标记	数量性状
1	I	D8	BNL1611	LI
2	II	D8	CIR224	MIC
3	II	D8	CIR224, BNL3452, CIR165	H
4	II	D8	CIR139	CRd
5	II	D8	CIR062	MIC
6	II	D8	BNL0390	FUI
7	II	D8	BNL1671	FEL CRd Hs FEL

表 7 QTL 紧密关联的相似标记区间

相似区间	置信度	支持率	重组率	相关的 QTL
{BNL852, CIR224}	1.00	0.86	0.1210±0.1398	1, 2, 5, 7
{BNL390, BNL2448}	1.00	0.81	0.0011±0.0004	1, 2, 5, 7
{BNL390, CIR062}	1.00	0.81	0.0017±0.0011	1, 2, 5, 7
{BNL390, BNL852}	1.00	0.81	0.0294±0.0031	1, 2, 5, 7
{BNL2448, BNL852}	1.00	0.81	0.0310±0.0041	1, 2, 5, 7
{BNL2448, CIR224}	1.00	0.81	0.3047±0.0033	1, 2, 5, 7
{CIR062, BNL852}	1.00	0.81	0.0292±0.0011	1, 2, 5, 7
{CIR062, CIR224}	1.00	0.81	0.3029±0.0004	1, 2, 5, 7
{BNL2448, CIR062}	1.00	0.81	0.0021±0.0026	1, 2, 5, 7
{BNL390, CIR224}	1.00	0.81	0.3032±0.0025	1, 2, 5, 7

参考文献

- Wan X, Weng J, Zhai H, et al. QTL analysis for rice grain width and fine mapping of an identified QTL allele gw-5 in a recombination hotspot region on chromosome 5. *Genetics*, 2008, 179: 2239—2252
- Xue W, Xing Y, Weng X, et al. Natural variation in *Ghd7* is an important regulator of heading date and yield potential in rice. *Nat Genet*, 2008, 40: 761—767
- Wang J, Wan X, Li H, et al. Application of identified QTL-marker associations in rice quality improvement through a design-breeding approach. *Theor Appl Genet*, 2007, 115: 87—100
- Goffinet B, Gerber S. Quantitative trait loci: A meta-analysis. *Genetics*, 2000, 155: 463—473
- Khatkar M S, Thomson P C, Tammen I, et al. Quantitative trait loci mapping in dairy cattle: Review and meta-analysis. *Genet Sel Evol*, 2004, 36: 163—190
- Chardon F, Virlon B, Moreau L, et al. Genetic architecture of flowering time in maize as inferred from quantitative trait loci meta-analysis and synteny conservation with the rice genome. *Genetics*, 2004, 168: 2169—2185

- 7 Hanocq E, Laperche A, Jaminon O, et al. Most significant genome regions involved in the control of earliness traits in bread wheat, as revealed by QTL meta-analysis. *Theor Appl Genet*, 2007, 114: 569—584
- 8 Rong J, Feltus F A, Waghmare V N, et al. Meta-analysis of polyploid cotton QTL shows unequal contributions of subgenomes to a complex network of genes and gene clusters implicated in lint fiber development. *Genetics*, 2007, 176: 2577—2588
- 9 Arcade A, Labourdette A, Falque M, et al. BioMercator: Integrating genetic maps and QTL towards discovery of candidate genes. *Bioinformatics*, 2004, 20: 2324—2326
- 10 Bartlett P L, Boucheron S, Lugosi G. Model selection and error estimation. *Mach Learn*, 2002, 48: 85—113
- 11 Pasek S, Bergeron A, Risler J L, et al. Identification of genomics features using microsynteny of domain: Domain teams. *Genome Res*, 2005, 15: 867—874
- 12 Peer D, Van Y. Computational approaches to unveiling ancient genome duplications. *Nat Rev Genet*, 2004, 5: 752—763
- 13 Vandepoele K, Saeys Y, Simillion C, et al. The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. *Genome Res*, 2002, 12: 1792—1801
- 14 Wang X, Shi X, Li Z, et al. Statistical inference of chromosomal homology based on gene colinearity and applications to *Arabidopsis* and rice. *BMC Bioinformatics*, 2006, 7: 447
- 15 Ling X, He X, Xin D, et al. Efficiently identifying max-gap clusters in pairwise genome comparison. *J Comput Biol*, 2008, 15: 593—609
- 16 Luc N, Risler J L, Bergeron A, et al. Gene teams: A new formalization of gene clusters for comparative genomics. *Comput Biol Chem*, 2003, 27: 59—67
- 17 Myers S, Bottolo L, Freeman C, et al. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 2005, 310: 321—324
- 18 Kosambi D D. The estimation of map distance from recombination values. *Ann Eugenics*, 1944, 12: 172—175
- 19 Bradley P, Fayyad U, Reina C. Scaling clustering algorithms to large databases. *Knowledge Discovery and Data Mining*. Boston: Kluwer Academic Publishers, 1998. 9—15
- 20 Jain A K, Murty M N, Flynn P J. Data clustering: A review. *ACM Computing Surveys*, 1999, 31: 264—323
- 21 R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2009. Vienna, Austria. <http://www.R-project.org>
- 22 Csardi G, Nepusz T. The igraph software package for complex network research. *Inter J Complex Syst*, 2006. 1695
- 23 Hahsler M, Gruen B, Hornik K. A rules—A computational environment for mining association rules and frequent item sets. *J Statist Software*, 2005, (14): 15
- 24 He D H, Lin Z X, Zhang X L, et al. QTL mapping for economic traits based on a dense genetic map of cotton with PCR-based markers using the interspecific cross of *Gossypium hirsutum* × *Gossypium barbadense*. *Euphytica*, 2007, 153: 181—197
- 25 Lacape J M, Nguyen T B, Courtois B, et al. QTL analysis of cotton fiber quality using multiple *Gossypium hirsutum* × *Gossypium barbadense* backcross generations. *Crop Sci*, 2005, 45: 123—140
- 26 Nguyen T B, Giband M, Brottier P, et al. Wide coverage of the tetraploid cotton genome using newly developed microsatellite markers. *Theor Appl Genet*, 2004, 109: 167—175
- 27 Yu J, Yu S, Lu C, et al. High-density linkage map of cultivated allotetraploid cotton based on SSR, TRAP, SRAP and AFLP markers. *J Integr Plant Biol*, 2007, 49: 716—724
- 28 Kulathinal R J, Bennett S M, Fitzpatrick C L, et al. Fine-scale mapping of recombination rate in *Drosophila* refines its correlation to diversity and divergence. *Proc Natl Acad Sci USA*, 2008, 105: 10051—10056
- 29 Coop G, Wen X, Ober C, et al. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science*, 2008, 319: 1395—1398
- 30 Ptak S E, Roeder A D, Stephens M, et al. Absence of the *TAP2* human recombination hotspot in chimpanzees. *PLoS Biol*, 2004, 2: e155
- 31 Ptak S E, Hinds D A, Koehler K, et al. Fine-scale recombination patterns differ between chimpanzees and humans. *Nat Genet*, 2005, 37: 429—434
- 32 Wall J D, Frisse L A, Hudson R R, et al. Comparative linkage-disequilibrium analysis of the β -Globin hotspot in primates. *Am J Hum Genet*, 2003, 73: 1330—1340
- 33 Winckler W, Myers S R, Richter D J, et al. Comparison of fine-scale recombination rates in humans and chimpanzees. *Science*, 2005, 308: 107—111
- 34 Spencer C C A, Deloukas P, Hunt S, et al. The influence of recombination on human genetic diversity. *PLoS Genet*, 2006, 2: e148

补充材料

表 S1 Monte Carlo 模拟研究中的连锁图谱

表 S2 各个连锁群中的连锁 G 图

表 S3 各个连锁群中的约简图

表 S4 各个连锁群中与 QTL 相关的相似区间

本文的以上补充材料见网络版 csb.scichina.com。补充材料为作者提供的原始数据，作者对其学术质量和内容负责