生 物 数 学 学 报 2000 ,15(1) : 103-109

Journal of Biomathematics

基因型值多次聚类法构建作物种质资源核心库*

胡 晋 徐海明 朱 军

(浙江大学农学系 杭州 310029)

摘 要: 采用合适的遗传模型无偏预测基因型值，用基因型值进行聚类分析. 采用马氏距离计算遗传材料间的遗传距离，并用不加权类平均法 (UPGMA) 进行聚类. 根据树型图，从遗传变异相似的每组二个遗传材料中随机选取一个遗传材料，如组内只有一个遗传材料，则选取该遗传材料. 对所取的所有遗传材料再次聚类、取样，直至所取遗传材料的数量为总遗传材料的 20% ~ 30%，这些遗传材料作为核心资源. 用方差同质性测验、均值 t 测验评价核心资源与原资源群体遗传多样性的差异，并比较两者的变异系数和极差. 以棉花 168 个基因型 5 个纤维性状为例进行多次聚类，得到的 41 个基因型核心资源能够代表原棉花资源遗传材料的遗传多样性. 表明多次聚类构建资源核心库的方法可行.

关键词: 基因型值; 核心库; 种质资源; 多次聚类

文献标识码: A **文章编号** 1001-9626(2000)01-0103-07

1 引 言

种质资源的保存和利用，对人类选育高产、优质、抗逆新品种，具有重要意义. 各国都相继建立了不同作物的种质资源库，以保存地方品种及其野生近缘种. 育种工作者由于选育新品种的需要，往往也保存大量的种质资源. 随着种质资源的广泛征集和不断累积交换，种质资源库会越来越大. 但是单纯增加种质资源库的容量，并不一定能保证遗传变异也得到相应增加. 面对大量的材料，育种家在取得这些材料的详细信息之前是无法利用的，这就需要开展深入细致的评价鉴定工作. 由于资源种类、遗传材料总数的规模极大，以目前的人力、物力，育种单位也只能了解一些简单的表现性状，不可能对所有的遗传材料进行详细的评价、整理和有效地管理、利用. 鉴于以上状况，人们寻找既能包含种质的遗传多样性，而材料数量又较小的方法，这就是建立作物种质资源核心库 (core collection) 方法. 这是作物种质资源研究的一个新领域.

资源核心库在遗传资源的管理和利用方面发挥了重大作用 [1,2,3]. 资源核心库包含了应优先保存的材料，遗传材料具有代表性. 资源核心库的建立使基因库的繁重工作有了工作的重点. 在保存过程中，可优先对这些遗传材料进行生活力的监测和繁殖. 资源核心库在新遗传材料的增加、特性的描述、种质评价、种质利用、种质分发等方面也发挥了重大作用. 核心资源可以优先繁殖、包装、交换、准备好用于分送. 更重要的是核心资源具有代表性，可减少分发的规模. 对于复杂特性及花费昂贵的评价，核心资源可以优先进行. 核心资源也适宜于发展新的种质保存方法，例如超干种子保存 (ultra-dry seed storage) 、试管保存 (“ in vitro ” storage)

收稿日期: 1998-10-23 ; **收修改稿日期:** 1999-01-29

* **基金项目:** 国家自然科学基金资助项目

和超低温保存 (cryopreservation). 育种工作中可以利用核心资源的代表性，优先与当地品种进行配合力的测定，提高育种效率. 核心资源的选出有助于加速需要者的利用，从而提高利用率. 构建资源核心库的优点是有目的地减少一个种质资源库运行中所处理遗传材料的数量，大大节约人力和物力，提高种质库的管理和利用水平，为育种家提供更详细和有用的育种材料.

Franke 于 1984 年最早提出资源核心库这一概念 [4]，目前种质资源核心库的构建研究在国际上处于起步阶段 [1]，所见报导尚不多. 国内在这方面的研究也刚开始. 本研究探讨了多次聚类构建资源核心库的方法，并对构建的作物种质资源核心库进行了评价.

2 模型和方法

对于规模相对较大的种质资源遗传试验，可采用按田间行列编号顺序种植基因型，以一定间隔穿插对照基因型 (纯合的地方品种) 的试验设计方法. 用对照基因型控制田间不同位置的差异，其它基因型设置两次重复. 对于多年份的这类遗传试验，一般存在环境效应、环境内的行效应、环境内的列效应、基因型效应、基因型与环境互作效应，以及机误 (朱军， 1994；1996)[5,6]. 因此得到的观察值可作如下分解：

$$Y_{hg(ij)} = \mu + E_h + R_{i(h)} + C_{j(h)} + G_{g(ij)} + GE_{hg(ij)} + e_{hg(ij)},$$

其中 E_h 表示第 h 个环境的效应，固定效应；

$R_{i(h)}$ 表示环境 h 内第 i 行的效应，固定效应；

$C_{j(h)}$ 表示环境 h 内第 j 列的效应，固定效应；

$G_{g(ij)}$ 表示在环境 h 内第 i 行 j 列处的第 g 个基因型效应，随机效应，$G_{g(ij)} \sim (0, \sigma_G^2)$；

$GE_{hg(ij)}$ 表示环境 h 与基因型 g 的互作效应，随机效应，$GE_{hg(ij)} \sim (0, \sigma_{GE}^2)$；

$e_{hg(ij)}$ 是机误效应，随机效应，$e_{hg(ij)} \sim (0, \sigma_e^2)$.

采用朱军提出的混合线性模型统计分析方法进行统计分析，无偏预测基因型效应值 [7,8,9].

2.1 原始矩阵

主要有表现型协方差矩阵、基因型协方差矩阵及机误协方差矩阵. 由于表现型值的差异不能真正反映品种间的遗传差异，利用表现型协方差矩阵和表现型值进行聚类显然不能反映遗传差异. 合理的聚类分析应采用基因型值. 本试验利用合理的统计模型和统计分析方法，预测基因型值，用基因型值及其协方差矩阵计算马氏 (Mahalanobis) 距离，再进行品种聚类.

2.2 遗传距离的计算

遗传距离作为度量遗传群体差异的综合数量指标，直接影响聚类的结果，而性状间的相关又影响着遗传距离的计算. 而马氏距离可以不受量纲的影响，无须将原始数据作适当的变换 [10,11]. 因此，本试验采用马氏距离来计算遗传距离. 假设共有 n 个基因型，采用 m 个性状进行聚类. 第 i 个基因型与第 j 个基因型的基因型效应向量分别为 $g_i^T = (g_{i1}, g_{i2}, \cdots, g_{im}), g_j^T = (g_{j1}, g_{j2}, \cdots, g_{jm})$，则第 i 个基因型与第 j 个基因型间的马氏距离计算公式为 [11]

$$D_{ij}^2 = (g_i - g_j)^T V_G^{-1} (g_i - g_j),$$

其中 $V_G = (\sigma_{ij})$, $i = 1, 2, \cdots, n$; $j = 1, 2, \cdots, n$.

2.3 聚　类

采用系统聚类法中的不加权类平均法 (Unweighted pari-group average). 假设类 G_i 与 G_j 分别有 n_i 与 n_j 个基因型，其合并所得的新类为 G_r，基因型数为 $n_r(=n_i+n_j)$，它与其它各类 G_s 的类间距离计算公式为 [12]

$$D_{rs}^2=\frac{n_i}{n_r}D_{si}^2+\frac{n_j}{n_r}D_{sj}^2.$$

2.4 多次聚类构建资源核心库

从某一作物所有遗传材料中选取多少遗传材料构成资源核心库，一般认为按遗传材料总数一定比例 (p) 较合适. 根据遗传变异量和等位基因变异的保持情况，原遗传材料总数的 20% ~ 30% ($p=0.2\sim0.3$) 被认为具有较好的代表性 [13].

用遗传距离进行聚类后，得到树型图. 根据聚类的原理，最低分类水平上类 (组) 内的差异最小，因此从每组二个遗传材料中随机取一个遗传材料进入下一轮聚类分析，如组内只有一个遗传材料，则该遗传材料直接进入下一轮聚类分析；对所取的遗传材料再次聚类，按同样的方法取样，再进入下一轮聚类、取样，直至所取遗传材料量达到设定的要求，即为构建的资源核心库.

3 应用实例

以下以辽宁经济作物研究所棉花种质资源 168 个基因型，5 个纤维性状数据为例，构建资源核心库. 5 个纤维性状分别为 2.5% 跨长 (mm)、整齐度 (%)、强度 (gf/tex)、伸长度 (%)、麦克隆. 为方便起见，分别以 1 至 168 号代表各基因型. 首先采用朱军 (1993) 提出的调整无偏预测 (AUP) 法无偏预测遗传效应值 [7]. 用遗传效应值计算品种间的马氏距离. 聚类采用系统聚类法中的不加权类平均法. 根据聚类结果，得到树型图. 在最低分类水平上，选取遗传材料，组内只有一个遗传材料的，则此遗传材料直接进入下一轮聚类，组内有二个遗传材料的，则随机选取一个遗传材料进入下一轮聚类. 因此，进入第二轮聚类的遗传材料有 106 个. 对 106 个遗传材料的基因型值重新计算马氏距离，然后进行第二轮聚类，得到树型图 (图 1). 根据上述取样原则，进行取样，得到 68 个遗传材料. 对 68 个遗传材料的基因型值计算马氏距离，然后进行第三轮聚类，得到树型图 (图 2). 对 68 个遗传材料取样后，得到 41 个遗传材料. 根据 $p=0.2\sim0.3$，168 个遗传材料所构建的核心资源材料数为 33.6 ~ 50.4 个，现得到 41 个遗传材料，已符合要求，因此这 41 个遗传材料可作为原有棉花种质资源的核心资源 (遗传材料号为：167, 127, 93, 88, 19, 130, 17, 96, 35, 110, 43, 109, 23, 156, 145, 36, 15, 49, 147, 123, 112, 118, 108, 34, 114, 157, 159, 111, 83, 169, 31, 76, 121, 165, 122, 74, 37, 25, 162, 2, 1).

4 资源核心库的评价

核心资源材料应能代表原有种质资源遗传材料的遗传多样性. 对于构建的种质资源核心库是否能很好地代表原种质资源遗传材料的遗传多样性，本研究采用方差、极差、均值和变异系数来评价. 对方差的差异性进行同质性测验 (两个方差比较时为 F 测验)，对均值的差异性

2.3 聚　类

采用系统聚类法中的不加权类平均法 (Unweighted pari-group average). 假设类 G_i 与 G_j 分别有 n_i 与 n_j 个基因型，其合并所得的新类为 G_r，基因型数为 $n_r(=n_i+n_j)$，它与其它各类 G_s 的类间距离计算公式为 [12]

$$D_{rs}^2=\frac{n_i}{n_r}D_{si}^2+\frac{n_j}{n_r}D_{sj}^2.$$

2.4 多次聚类构建资源核心库

从某一作物所有遗传材料中选取多少遗传材料构成资源核心库，一般认为按遗传材料总数一定比例 (p) 较合适. 根据遗传变异量和等位基因变异的保持情况，原遗传材料总数的 20% ~ 30% ($p=0.2\sim0.3$) 被认为具有较好的代表性 [13].

用遗传距离进行聚类后，得到树型图. 根据聚类的原理，最低分类水平上类 (组) 内的差异最小，因此从每组二个遗传材料中随机取一个遗传材料进入下一轮聚类分析，如组内只有一个遗传材料，则该遗传材料直接进入下一轮聚类分析；对所取的遗传材料再次聚类，按同样的方法取样，再进入下一轮聚类、取样，直至所取遗传材料量达到设定的要求，即为构建的资源核心库.

3 应用实例

以下以辽宁经济作物研究所棉花种质资源 168 个基因型，5 个纤维性状数据为例，构建资源核心库. 5 个纤维性状分别为 2.5% 跨长 (mm)、整齐度 (%)、强度 (gf/tex)、伸长度 (%)、麦克隆. 为方便起见，分别以 1 至 168 号代表各基因型. 首先采用朱军 (1993) 提出的调整无偏预测 (AUP) 法无偏预测遗传效应值 [7]. 用遗传效应值计算品种间的马氏距离. 聚类采用系统聚类法中的不加权类平均法. 根据聚类结果，得到树型图. 在最低分类水平上，选取遗传材料，组内只有一个遗传材料的，则此遗传材料直接进入下一轮聚类，组内有二个遗传材料的，则随机选取一个遗传材料进入下一轮聚类. 因此，进入第二轮聚类的遗传材料有 106 个. 对 106 个遗传材料的基因型值重新计算马氏距离，然后进行第二轮聚类，得到树型图 (图 1). 根据上述取样原则，进行取样，得到 68 个遗传材料. 对 68 个遗传材料的基因型值计算马氏距离，然后进行第三轮聚类，得到树型图 (图 2). 对 68 个遗传材料取样后，得到 41 个遗传材料. 根据 $p=0.2\sim0.3$，168 个遗传材料所构建的核心资源材料数为 33.6 ~ 50.4 个，现得到 41 个遗传材料，已符合要求，因此这 41 个遗传材料可作为原有棉花种质资源的核心资源 (遗传材料号为：167, 127, 93, 88, 19, 130, 17, 96, 35, 110, 43, 109, 23, 156, 145, 36, 15, 49, 147, 123, 112, 118, 108, 34, 114, 157, 159, 111, 83, 169, 31, 76, 121, 165, 122, 74, 37, 25, 162, 2, 1).

4 资源核心库的评价

核心资源材料应能代表原有种质资源遗传材料的遗传多样性. 对于构建的种质资源核心库是否能很好地代表原种质资源遗传材料的遗传多样性，本研究采用方差、极差、均值和变异系数来评价. 对方差的差异性进行同质性测验 (两个方差比较时为 F 测验)，对均值的差异性

进行 t 测验.

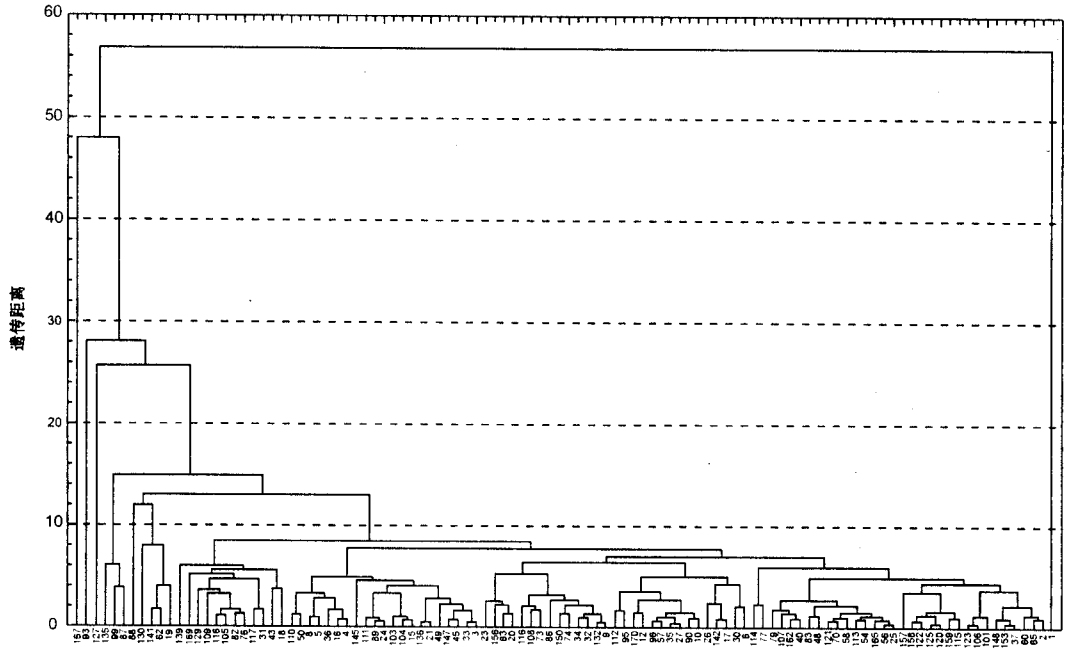

图 1 棉花 5 个纤维性状 106 个基因型的聚类树型图

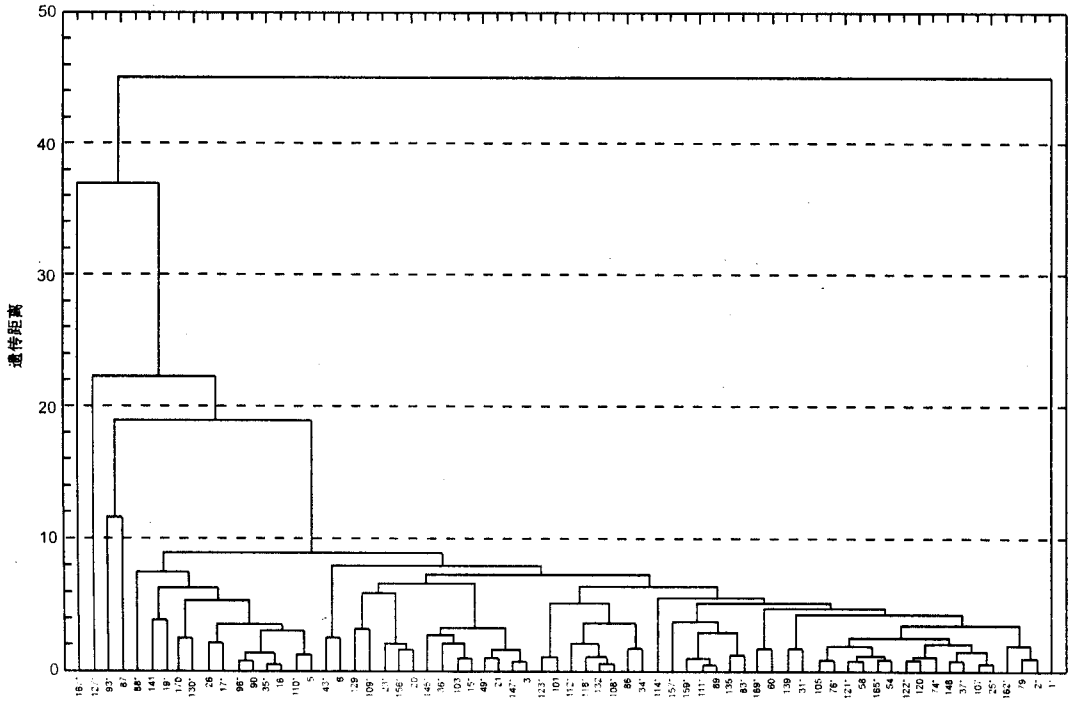

图 2 棉花 5 个纤维性状 68 个基因型的聚类树型图 (* 为选中的核心资源)

对棉花 5 个纤维性状的基因型值构建的资源核心库进行评价. 结果表明核心资源与原资

源群体之间跨长的均值无显著差异 (表 1). 核心资源的方差大于原资源群体的方差, 但未达显著差异. 极差小于原资源群体. 从变异系数看, 核心资源的变异系数稍大于原资源群体. 可以认为核心资源基本保持了原资源遗传材料跨长的遗传变异.

核心资源与原资源群体之间强度的均值无显著差异 (表 1), 核心资源的方差大于原资源群体的方差, 但未达显著差异. 极差稍小于原资源群体, 核心资源的的变异系数接近于原资源群体的变异系数. 可认为核心资源伸长度的遗传变异与原资源群体基本相同.

表 1　棉花资源核心库与原资源群体 5 个纤维性状的比较

Table 1　Comparison between Core Collection and Initial Population on Five Fiber Traits in Cotton

性　状		方　差	均　值	极　差	变异系数
2.5% 跨长	所有基因型	2.075	26.504	10.198	0.054
	41 个基因型	2.805	26.321	8.467	0.064
	P 值 *	0.196	0.482		
整齐度	所有基因型	1.963	50.720	9.924	0.028
	41 个基因型	2.345	50.541	8.560	0.030
	P 值	0.438	0.474		
强度	所有基因型	0.926	20.011	7.878	0.048
	41 个基因型	1.336	19.841	6.575	0.058
	P 值	0.116	0.332		
伸长度	所有基因型	0.176	6.064	2.794	0.069
	41 个基因型	0.246	6.004	2.525	0.083
	P 值	0.148	0.427		
麦克隆	所有基因型	0.087	4.053	2.045	0.073
	41 个基因型	0.130	4.036	1.905	0.089
	P 值	0.083	0.755		

* 所有基因型与 41 个基因型间差异显著性测验的概率值

从强度看, 核心资源与原资源群体之间的均值无显著差异 (表 1), 核心资源的方差大于原资源群体的方差, 但未达极显著差异. 极差稍小于原资源群体, 核心资源的变异系数稍大于原资源群体. 可认为核心资源强度的遗传变异与原资源群体基本相同.

核心资源与原资源群体伸长度之间的均值无显著差异 (表 1), 核心资源的方差大于原资源群体, 但未达显著差异. 极差则相接近. 核心资源的变异系数稍大于原资源群体. 可认为核心资源伸长度的遗传变异与原资源群体基本相同.

核心资源与原资源群体麦克隆之间的均值无显著差异 (表 1), 核心资源的方差大于原资源群体, 但未达显著差异. 两者的极差则接近. 核心资源的变异系数大于原资源群体. 表明核心资源麦克隆的遗传变异比原资源群体有所增加. 核心资源能够代表原有群体的遗传多样性.

5　讨　论

资源核心库由来自一个已存在的种质库中的一批有限遗传材料所组成, 用于代表整个种质库的遗传范围 (Brown,1989)[14]. IBPGR(国际植物遗传资源委员会) 认为资源核心库是指以最少的重复来代表一个种及其野生近缘种的遗传多样性. 核心库中入选的遗传材料都是有代表

性的，它们互相之间都存在着生态上或遗传上的距离．覆盖了一个种质库内的遗传变异和生态范围，目的是使遗传多样性尽可能最多，这就意味着资源核心库不应含有复份．

聚类分析作为一种重要的研究工具已被广泛地应用于种质资源的分类、品种遗传差异的评价等研究中[15]．构建资源核心库，首先要对现有的资源进行遗传分类，确定不同材料之间的遗传关系．从国内外研究报导看，在目前的聚类分析中，大多直接利用表现型值进行分类，构建核心资源所用的数据大多为表现型值．由于作物品种的性状多数为数量性状，与环境存在互作．因此，基于这些数据的遗传分类不能反映种质资源的遗传结构．且基因型多与环境存在一定的互作，同时遗传试验又存在试验误差，所以用表现型值计算的遗传距离不能正确度量基因型间的遗传差异．因此，种质资源的遗传聚类首先必须用合理的统计模型及统计分析方法，用预测的基因型值进行品种聚类，排除农业试验中存在的试验误差、环境效应、基因型与环境的互作效应．我们采用朱军 (1993) 提出的调整无偏预测 (AUP) 法无偏预测基因型值[7]，并用于聚类分析，使聚类的结果更具可靠性．

聚类后，如何确定分类 (组) 的最合适数目 (即确定分类标准的遗传距离阈值的划分)，在理论上尚未完全解决．在确定不同遗传材料之间的遗传关系后，即完成分组 (类) 后，要对每组进行抽样，但至今国际上仍无合适的抽样方法，常采用每组取同样数量的遗传材料，或以组内遗传材料数量按比例取样来构成核心资源．这些方法，前者未考虑到分组有大小，后者虽考虑到组的大小，却未能顾及到分组之间的遗传关系，因此取样得到的核心资源不能很好地代表原有的种质资源材料的遗传变异．本方法直接从最低分类水平开始，亦即从遗传变异最相似的基因型间开始取样，不需考虑分组数目和避免组有大小和组内取样不匀的问题．经多次聚类后，可以得到核心资源，并保留原资源遗传材料的遗传变异分布．对于遗传材料量大的种质资源，可先按它们的分布、地理起源、生态起源、遗传标记等进行划分大类，如棉花可先分为陆地棉、海岛棉、亚洲棉和非洲棉，再对每类进行多次聚类构建资源核心库．

对于构建的核心资源是否能很好地代表原有种质资源遗传材料的遗传多样性．国内外尚无有效的判断标准，有人曾用主成分分析来判断[16]，但应用主成分值丢失的信息较多，结果不够准确．我们建议采用统计方法判断核心资源与原资源群体遗传变异的差异，综合棉花纤维5个性状的结果，41个基因型与原有群体之间的方差和均值均无显著差异，极差接近，变异系数略变大．可认为核心资源能够代表原棉花资源遗传材料的遗传多样性．表明多次聚类方法构建资源核心库的方法可行．建立资源核心库的最后一步是入选遗传材料的处理，入选的核心资源仍然保存在原来的基因库内，只是在数据库中进行记载，表明某些遗传材料是核心资源．

致 谢 辽宁经济作物研究所李瑞祥研究员为本文提供实例研究所需的数据，特致谢意．

[参 考 文 献]

[1] Brown A H D. The core collection at the crossroads[A]. In: Hodgkin T, eds. Core Collection of Plant Genetic Resources[C]. Chichester: John Wiley & Sons, 1995. 3–19.

[2] 胡晋. 植物遗传资源核心库及其建立 [J]. 种子, 1996, (5):22–24.

[3] Boukema I W, Hintum van T J L, Astley D. Creation and composition of the *Brassica oleracea* core collection[J]. *Plant Genetic Resources Newsletter*, 1997, (111):29–32.

[4] Frankel O H. Genetic perspectives of germplasm conservation[A]. In: Arber W, eds. Genetic manipulation: Impact on man and society[C]. Cambridge: Cambridge University Press, 1984. 161–170.

[5] 朱军. 广义遗传模型与数量遗传分析新方法 [J]. 浙江农业大学学报, 1994, **20**(6):551–559.

[6] 朱军. 包括基因型环境互作效应的种子模型及其分析方法 [J]. 遗传学报, 1996, **23**(1):53–58.

[7] 朱军. 作物杂种后代基因型值和杂种优势的预测方法 [J]. 生物数学学报, 1993, **8**(1):32–44.

[8] Zhu J, Weir B S. Diallel analysis for sex-linked and maternal effects[J]. *Theor Appl Genet*, 1996, **92**(1):1–9.

[9] Zhu J. Mixed model approaches for estimation genetic variances and covariances[J]. 生物数学学报, 1992, **7**(1):1–11.

[10] 裴鑫德. 多元统计分析及其应用 [M]. 北京: 北京农业大学出版社, 1991. 89–195.

[11] Mahalanobis P C. On the generalized distance in statistics[J]. *Proc Natl Inst Sci India*, 1936, **2**(1):49–55.

[12] Sokal R R, Michener C D. A statistical method for evaluating systematic relationships[J]. *Univ Kansas Sci Bull*, 1958, **38**(6):1409–1438.

[13] Yonezawa K, Nomura T, Morishima H. Sampling strategies for use in stratified germplasm collections[A]. In: Hodgkin T, eds. Core Collection of Plant Genetic Resources[C]. Chichester: John Wiley & Sons, 1995. 35–53.

[14] Brown A H D. Core collections: A practical approach to genetic resources managenment[J]. *Genome*, 1989, **31**(5):818–824.

[15] Hintum van T J L. Hierarchical aproaches to the analysis of genetic diversity in crop plants[A]. In: Hodgkin T eds. Core Collections of Plant Genetic Resources[C]. Chichester: John Wiley & Sons, 1995. 23–34.

[16] Crossa J, Delacy I H, Taba S. The use of multivariate methods in developing a core collection[A]. In: Hodgkin T eds. Core Collections of Plant Genetic Resources[C]. Chichester: Jonh Wiley & Sons, 1995. 77–92.

Constructing Core Collection of Crop Germplasm by Multiple Clusters Based on Genotypic Values

Hu Jin　Xu Haiming　Zhu Jun

(*Department of Agronomy, Zhejiang University, Hangzhou* 310029)

Abstract: A genetic model for controlling systematical errors in fields and a robust statistical method were used to analyze the data. Predicted genotype values were used to cluster crop germplasm resources. Mahalanobis distance was used to calculate genetic distance among accessions. Unweighted pair-group average method of hierarchical cluster was used for grouping. According to a dendrogram, one accession of each group with two accessions of similar genetic variation was randomly chosen for next cluster, the accession went into next cluster if there was only one accession in a group. The sample from the first cluster was clustered and chosen again in the same way. When the number of chosen accessions was 20% ~ 30% of the initial accessions, the cluster was stopped and core collection could be constructed by these accessions. The difference in genetic diversity between core collection and initial accessions was measured by F-test for variance, t-test for means, coefficient of variation and range. The example of multiple clusters on 168 accessions of cotton with five fiber traits was made. The results showed that the 41 genotypes of core collection could represent the genetic diversity of initial resources. It indicated that the method of multiple clusters to construct core collection was applicable.

Key words: Genotypic value ; Core collection ; Germplasm resources ; Multiple clusters