中国水稻科学 (Chinese J Rice Sci), 2007, 21 (1): 51～58
http://www.ricesci.cn; http://www.ricescience.org

建立在基因型值和分子标记信息上的水稻核心种质评价参数

王建成　胡　晋*　张彩芳　徐海明　张　胜
(浙江大学 农业与生物技术学院 农学系, 浙江 杭州 310029; * 通讯联系人, E-mail: jhu@dial.zju.edu.cn)

Evaluating Parameters of Rice Core Collections Based on Genotypic Values and Molecular Marker Information

WANG Jian-cheng, HU Jin*, ZHANG Cai-fang, XU Hai-ming, ZHANG Sheng
(Department of Agronomy, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310029, China; * *Corresponding author, E-mail: jhu@dial.zju.edu.cn*)

Abstract: Monte Carlo simulation combining with mixed linear model were used in the research of evaluating parameters for rice core collection based on genotypic values and molecular marker information, which eliminated the interference of environment and obtained more reliable results. The coincidence rate of range (*CR*) was the optimal evaluating parameter. Mean Simpson index (M_D), mean Shannon-Weaver index of genetic diversity (M_I) and mean polymorphism information content (M_{PIC}) were important evaluating parameters. The variable rate of coefficient of variation (*VR*) could act as an important referential parameter for evaluating the variation degree of core collection. Percentage of polymorphic loci (*p*) could act as a determination parameter for the size of core collection. Mean difference percentage (*MD*) was a determination parameter for the reliability judgement of core collection. The effective evaluating parameters for core collection selected by present research could be used in different plant germplasm population as criteria for sampling percentage.

Key words: core collection; genotypic value; molecular marker information; Monte Carlo simulation; mixed linear model; evaluating parameter; rice

摘　要: 采用蒙特卡洛模拟结合混合线性模型的方法, 直接从基因型值和分子标记水平上研究了水稻核心种质的 11 个评价参数, 排除了环境因素的干扰, 对各个评价参数做出了准确的评价。研究表明, 极差符合率 (*CR*) 可以作为评价核心种质代表性的首选参数。平均 Simpson 指数 (M_D)、平均 Shannon-Weaver 多样性指数 (M_I) 和平均多态信息含量 (M_{PIC}) 是评价核心种质代表性的重要参数。变异系数变化率 (*VR*) 可以作为评价核心种质变异程度的重要参考参数。多态位点百分率 (*p*) 可以作为判断核心种质取样规模的判定参数。均值差异百分率 (*MD*) 可作为判断核心种质是否具有代表性的判定参数。本研究筛选出的核心种质评价参数, 适用于不同的种质资源群体, 可以用作确定核心种质取样比例的判定依据, 进而解决了确定核心种质合理取样比例的问题。

关键词: 核心种质; 基因型值; 分子标记信息; 蒙特卡洛模拟; 混合线性模型; 评价参数; 水稻
中图分类号: Q943; S511.024　　文献标识码: A　　文章编号: 1001-7216 (2007) 01-0051-08

未来人类的生存在很大程度上将取决于人们保存种质资源的数量和多样性。许多国家和国际农业研究组织先后建设了大小不等的数百个基因库, 已有数以百万计的各类作物种质资源被保存起来[1]。然而, 随着种质资源的不断搜集, 种质资源库变得越来越大, 巨大的种质资源数量给保存、评价、研究和利用带来了许多困难。为此, Frankel[2] 提出构建核心种质的设想, 即用科学方法, 从整个种质资源中选择一部分样本, 以最小的遗传资源数量, 尽可能最大限度地代表整个遗传资源的多样性。建立核心种质的目的是为了对它进行优先评价和利用, 从而提高整个种质资源库的管理和利用水平。水稻是我国最主要的粮食作物之一, 中国水稻育种研究处于世界前列, 这与国内保存了丰富的水稻种质资源是分不开的。至 2000 年, 我国已编目稻种资源国内品种达 75 597 份[3]。庞大的水稻种质资源量给应用带来了不便, 因此, 水稻是国内开展核心种质研究比较早的作物[4]。

构建核心种质的主要方法是: 首先根据地理起源、生态等信息将种质资源材料分组, 然后在不同的组内抽取核心材料组成核心子集, 将所有的核心子集整合在一起, 就构成核心种质。通过取样生成的核心样品群体, 必须进行遗传多样性的分析, 评价其代表性, 符合要求才能成为核心种质[5]。代表性是核心种质最重要的性质, 核心种质的代表性评价是核心种质构建中极其重要的一环。评价核心种质的代表性, 必须借助于一系列评价参数。在核心种质构建方法的研究中, 确定一系列有效的代表性评价参数是极其重要的一项工作。只有确定了核心种质的评价参数, 对其取样方法、取样比例等的研究才

收稿日期: 2006-05-12; 修改稿收到日期: 2006-08-21。
基金项目: 国家自然科学基金资助项目 (30270759)。
第一作者简介: 王建成 (1979—), 男, 博士研究生。

有判定标准。然而,目前对众多核心种质评价参数尚缺乏系统性研究。

提高核心种质代表性的关键是正确的分组和组内科学的取样。只有组内核心子集的代表性得到提高,整个核心种质的有效性才能得到本质的提高。核心子集是核心种质的一个子样本,其代表性评价参数同核心种质是一致的。本研究以水稻为材料,从组内核心子集水平上研究核心种质的代表性评价参数。采用蒙特卡洛模拟的方法,直接从基因型值和分子标记水平上研究各个核心子集的评价参数,排除了环境或人为因素的干扰,以期对各个评价参数做出准确的评判,并从中筛选出有效的评价参数,为水稻种质资源研究提供可靠的判定依据。

1 材料与方法

1.1 种质资源群体的蒙特卡洛模拟

在 Bataillon 等[6]的方法基础上,采用蒙特卡洛模拟建立种质资源群体。由于本研究只探讨组内核心样本的代表性评价,因此将模拟进程简化如下:假设不存在突变、漂移、迁移和选择,初始化原始基因库,包含 600 个有关的 QTL (quantitative trait loci)。用正态分布的随机数给每个 QTL 赋值,然后分别随机抽取 200、300 和 400 个个体,将每个个体自交 100 代,直到基因型完全纯合。这样,每一个纯合个体都可以看作是一个地方品种,所有个体就组成了一个地理起源相同的异质纯合种质资源原始群体[7],用于核心子集的构建。

假设该群体的基因存在“一因多效”和“多因一效”。从原始群体的每个个体中随机抽取一定数目(本研究采用的是 10～50 的随机数)的 QTL,个体间的 QTL 一一对应,组成一个数量性状,将每个个体该性状所有的 QTL 值相加,就得到了每个个体在该性状的基因型值[8]。重复抽取 30 次,得到了该群体 30 个数量性状的基因型值。假设分子标记与 QTL 紧密连锁,从原始群体的每个个体中随机抽取 100 个 QTL,个体间的 QTL 一一对应,组成该群体的分子标记信息。这样,这个群体中既有数量性状的基因型值,也有分子标记信息,并且两类数据是相互关联的,可对群体的遗传多样性进行全面评价。群体的所有数量性状在抽样构建核心种质前均经标准化处理,使所有性状都变成均值为 0,方差为 1。

1.2 核心种质的评价参数

本研究共选择 11 个评价参数,包括前人提出[9,10]和本文提出的。评价数量性状的参数为:均值差异百分率 (mean difference percentage, *MD*)、方差差异百分率 (variance difference percentage, *VD*)、最大值变化率 (changeable rate of maximum, CR_{MAX})、最小值变化率 (changeable rate of minimum, CR_{MIN})、平均值变化率 (changeable rate of mean, CR_{MEA})、极差符合率 (coincidence rate of range, *CR*) 和变异系数变化率 (variable rate of coefficient of variation, *VR*)。评价分子标记信息的参数为:多态位点百分率 (percentage of polymorphic loci, *p*)、平均 Simpson 指数 (mean Simpson index, M_D)、平均 Shannon-Weaver 多样性指数 (mean Shannon-Weaver index of genetic diversity, M_I) 和平均多态信息含量 (mean polymorphism information content, M_{PIC})。各参数计算公式如下:

$MD=\left(\frac{S_t}{n}\right)\times 100\%$,其中 S_t 是核心子集与原始群体进行 *t* 测验得到均值差异显著 ($\alpha = 0.05$) 的性状数,*n* 是数量性状总数;

$VD=\left(\frac{S_F}{n}\right)\times 100\%$,其中 S_F 是核心子集与原始群体进行 *F* 测验得到方差差异显著 ($\alpha = 0.05$) 的性状数,*n* 是数量性状总数;

$CR=\frac{1}{n}\sum_{i=1}^{n}\frac{R_{C(i)}}{R_{I(i)}}\times 100\%$,其中 $R_{C(i)}$ 是核心子集第 *i* 个性状的极差,$R_{I(i)}$ 是原始群体第 *i* 个性状的极差,*n* 是数量性状总数;

$VR=\frac{1}{n}\sum_{i=1}^{n}\frac{CV_{C(i)}}{CV_{I(i)}}\times 100\%$,其中 $CV_{C(i)}$ 是核心子集第 *i* 个性状的变异系数,$CV_{I(i)}$ 是原始群体第 *i* 个性状的变异系数,*n* 是数量性状总数;

$CR_{MAX}=\frac{1}{n}\sum_{i=1}^{n}\frac{Max_{C(i)}}{Max_{I(i)}}\times 100\%$,其中 $Max_{C(i)}$ 是核心子集第 *i* 个性状的最大值,$Max_{I(i)}$ 是原始群体第 *i* 个性状的最大值,*n* 是数量性状总数;

$CR_{MIN}=\frac{1}{n}\sum_{i=1}^{n}\frac{Min_{C(i)}}{Min_{I(i)}}\times 100\%$,其中 $Min_{C(i)}$ 是核心子集第 *i* 个性状的最小值,$Min_{I(i)}$ 是原始群体第 *i* 个性状的最小值,*n* 是数量性状总数;

$CR_{MEA}=\frac{1}{n}\sum_{i=1}^{n}\frac{Mea_{C(i)}}{Mea_{I(i)}}\times 100\%$,其中 $Mea_{C(i)}$ 是核心子集第 *i* 个性状的平均值,$Mea_{I(i)}$ 是原始群体第 *i* 个性状的平均值,*n* 是数量性状总数;

$p=\left(\frac{k}{n}\right)\times 100\%$,其中 *k* 是多态位点的数目,*n* 是分子标记位点总数;

$MD = \frac{1}{n}\sum_{i=1}^{n}(1-\sum_{j=1}^{m}p_{ij}^2)$，其中 p_{ij}是第 i 个位点第 j 种等位基因的频率，m是第 i 个位点等位基因的数目，n是分子标记位点总数；

$MI = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}p_{ij}\ln p_{ij}$，其中 p_{ij}是第 i 个位点第 j 种等位基因的频率，m是第 i 个位点等位基因的数目，n是分子标记位点总数；

$MPIC = \frac{1}{n}\sum_{k=1}^{n}(1-\sum_{i=1}^{m}p_{ki}^2-\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}2p_{ki}^2p_{kj}^2)$，其中 p_{ki}和 p_{kj}表示第 k 个分子标记位点上第 i 种和第 j 种等位基因的频率，m是第 k 个位点等位基因的数目，n是分子标记位点总数。

所有评价数量性状参数的计算均是基于未标准化的群体，即用标准化的群体构建核心种质，根据构建结果从未标准化的群体中找到对应的核心材料，计算各个评价参数的值。

1.3 核心子集的构建方法

根据 Hu 等[9]提出的逐步聚类构建核心种质的方法，进行了改进。具体构建方法为：首先给定取样比例。随后，计算原群体各个样品间的遗传距离，根据遗传距离进行聚类，根据聚类结果找出遗传距离最小的一个组，随机删除该组中两个样品之一，另一个样品被保留。然后对剩下的样品重新计算样品间的遗传距离并进行聚类分析，用同样的方法对群体进行缩减。如此循环，直到剩余样品数达到取样比例规定的规模，就构成了核心子集。

这种核心子集构建方法，能够最大限度地去除原始群体中冗余材料和遗传背景相近的材料，构建出的核心子集具有很强的代表性。并且，该方法经过改进之后，每次取样都是基于遗传距离最小的一个组，理论上只要计算遗传距离的方法不变，无论采用何种系统聚类方法，最后得到的结果都是相同的。本研究采用欧氏距离结合最短距离聚类法[11]构建核心子集。

1.4 核心子集的构建

对于 3 个模拟群体（分别为 200、300 和 400 个个体），取样比例从 1%到 30%，每个比例 3 次重复，各构建出 90 个核心子集，3 个模拟群体共得到 270 个核心子集。为了验证蒙特卡洛模拟的正确性和有效性，使用了一个水稻群体作为实例。该群体共 90 个基因型，2 次重复，种植 3 年，考查 8 个农艺性状，检测 60 个分子标记位点的多态性信息。对于 8 个农艺性状的表现型数据，使用混合线性模型的方法无偏预测出基因型值[12]，将基因型预测值和分子标记位点的多态性信息用于本研究。由于该水稻群体规模较小，因此对于该群体，按取样比例从 2%到 30%，每个比例 3 次重复，构建出 87 个核心子集，用于与模拟数据进行对比。

1.5 数据处理

对所有核心子集（共 357 个），均计算出 11 个评价参数的值，并绘出每个评价参数随取样比例变化的趋势图。对于各个评价参数，均采用方差分析的方法，比较同一个参数在各个取样比例间的差异显著性，用 Tukey 法进行多重比较（$\alpha = 0.05$），用标记字母法表示多重比较的最终结果。用 Tukey 测验的同质群数目（如：多重比较的标记字母按字母表顺序最大出现 c 时，则为 3 个同质群；最大出现 f 时，则为 6 个同质群）来评价各个参数的有效性。同质群数越多，表示该参数能够显著区分的核心子集越多，该参数对于区分不同核心子集的有效性越好。

本研究的蒙特卡洛模拟、核心子集构建和各个评价参数值的计算均在 MATLAB 环境中编程实现，对各个评价参数的分析均在 SAS 系统下进行。

2 结果与分析

2.1 各个评价参数的有效性分析

在取样比例为 1%～30%或 2%～30%时（表 1），无论是在各个模拟群体还是真实水稻群体中，CR的同质群数均大于其他参数，排在第 1 位。MD、VD和 CR_{MEA}这 3 个参数在各个群体中的同质群数基本上均排在最后几位，且最大不超过 3，说明这 3 个参数难以将不同的核心子集有效区分开。3 个基于分子标记的参数 MD、MI 和 $MPIC$在各个群体中均表现稳定，在单个群体中同质群数及其位次差别不大，3 个参数都能比较有效地将不同核心子集区分开。CR_{MAX}在 3 个模拟群体中，同质群数均排在前列，且至少为 8，因此可以有效地将不同核心子集区分开，但在 90 个水稻个体的群体中表现稍差，同质群数为 5。CR_{MIN}在 200 个模拟个体和 90 个水稻个体两个群体中，同质群数仅次于 CR，排在第 2 位，可以有效地将不同核心子集区分开，但是在 300 个模拟个体的群体中表现较差，同质群数位次仅排在第 7 位。VR在 300 个模拟个体和 400 个模拟个体两个群体中的同质群数分别为 6 和 14，可以比较有效地将不同的核心子集区分开，而在 200 个模拟个体和 90 个水稻个体两个群体中的同质群数分别为 1 和 2，难以有效地将不同核心子集区分开。p仅在 90 个水稻个体的群体中同质群数为 5，可以

将不同核心子集区分开, 而在 3 个模拟群体中同质群数均为 3, 难以将不同核心子集区分开。

在取样比例为 10%～30%时 (表 2) , *CR* 的同质群数在 200 个模拟个体和 90 个水稻个体两个群体中, 排在第 1 位, 在 300 个模拟个体和 400 个模拟个体两个群体中分别排在第 4 位和第 2 位。*MD*、*VD*、CR_{MEA} 和 *p* 这 4 个参数在各个群体中的同质群数均排在末位。与表 1 相同, M_D、M_I 和 M_{MAX} 这 3 个参数的同质群数在各群体中均较大, 且 M_D 的同质群数在 300 个模拟个体群体中排在第 1 位。CR_{MAX} 在各个群体中的同质群数均排在前列, 但综合比较起来, 其有效性差于 *CR*、M_D、M_I 和 M_{PIC} 这

表 1 不同群体中 11 个评价参数各自在取样比例为 1%～30%或 2%～30%时的 Tukey 测验 ($\alpha = 0.05$) 同质群数目

Table 1 Number of homogeneous populations by Tukey's test ($\alpha = 0.05$) of 11 evaluating parameters in different germplasm populations under the sampling percentage at 1%－30% or 2%－30%.

参数 Parameter	200 个模拟个体 200 simulated accessions		300 个模拟个体 300 simulated accessions		400 个模拟个体 400 simulated accessions		90 个水稻个体 90 rice accessions	
	同质群数 No. of homogeneous populations	秩[1)] Rank[1)]	同质群数 No. of homogeneous populations	秩[1)] Rank[1)]	同质群数 No. of homogeneous populations	秩[1)] Rank[1)]	同质群数 No. of homogeneous populations	秩[1)] Rank[1)]
均值差异百分率 *MD*	1	9	1	9	1	11	2	8
方差差异百分率 *VD*	1	9	1	9	3	8	1	11
最大值变化率 CR_{MAX}	11	3	8	2	10	4	5	5
最小值变化率 CR_{MIN}	17	2	4	7	12	3	7	2
平均值变化率 CR_{MEA}	3	7	1	9	2	10	2	8
极差符合率 *CR*	18	1	15	1	17	1	9	1
变异系数变化率 *VR*	1	9	6	6	14	2	2	8
多态位点百分率 *p*	3	7	3	8	3	8	5	5
平均 Simpson 指数 M_D	9	4	8	2	7	5	5	5
平均 Shannon-Weaver 多样性指数 M_I	8	5	8	2	6	6	6	3
平均多态信息含量 M_{PIC}	8	5	8	2	6	6	6	3

[1)] 各个参数对应的同质群数按大小排列的等级。

[1)] The grade of the number of homogeneous populations for each evaluating parameters.

表 2 不同群体中 11 个评价参数各自在取样比例为 10%～30%时的 Tukey 测验 ($\alpha = 0.05$) 同质群数目

Table 2 Number of homogeneous populations by Tukey's test ($\alpha = 0.05$) of 11 evaluating parameters in different germplasm populations under the sampling percentage at 10%－30%.

参数 Parameter	200 个模拟个体 200 simulated accessions		300 个模拟个体 300 simulated accessions		400 个模拟个体 400 simulated accessions		90 个水稻个体 90 rice accessions	
	同质群数 No. of homogeneous populations	秩[1)] Rank[1)]	同质群数 No. of homogeneous populations	秩[1)] Rank[1)]	同质群数 No. of homogeneous populations	秩[1)] Rank[1)]	同质群数 No. of homogeneous populations	秩[1)] Rank[1)]
均值差异百分率 *MD*	1	7	1	9	1	10	1	7
方差差异百分率 *VD*	1	7	1	9	2	9	1	7
最大值变化率 CR_{MAX}	8	3	5	5	8	2	5	2
最小值变化率 CR_{MIN}	9	2	2	8	5	7	2	3
平均值变化率 CR_{MEA}	1	7	3	7	3	8	1	7
极差符合率 *CR*	11	1	8	4	8	2	6	1
变异系数变化率 *VR*	1	7	5	5	12	1	1	7
多态位点百分率 *p*	1	7	1	9	1	10	1	7
平均 Simpson 指数 M_D	6	4	10	1	8	2	2	3
平均 Shannon-Weaver 多样性指数 M_I	6	4	9	2	8	2	2	3
平均多态信息含量 M_{PIC}	6	4	9	2	8	2	2	3

[1)] 各个参数对应的同质群数按大小排列的等级。

[1)] The grade of the number of homogeneous populations for each evaluating parameters.

4 个参数。CR_{MIN}在 300 个模拟个体的群体中表现较差，同质群数仅为 2，在其他 3 个群体中的表现同 CR_{MAX}差别不大。同表 1 的规律类似，VR 在 300 个模拟个体和 400 个模拟个体两个群体中的同质群数均居前列，而在 200 个模拟个体和 90 个水稻个体两个群体中的同质群数仅为 1，难以有效地将不同核心子集区分开。

综合以上结果，对各个参数的有效性排序为：$CR > M_D$、M_I、$M_{PIC} > CR_{MAX}$、$CR_{MIN} > VR > p > MD$、VD、CR_{MEA}。

2.2　各个评价参数的稳定性分析

由于 MD、VD 和 CR_{MEA}参数在各个群体中均表现不理想，无法有效地将不同的核心子集区分开，因此对这 3 个参数不再进行稳定性分析。CR 在群体内基本上是稳步上升的，在不同群体内的变化速率基本一致(图 1)，M_I 也有相似的规律(图 2)。M_D 和 M_{PIC}两个参数在各个群体中随取样比例改变的变化趋势和 M_I 基本一致(图未列出)，因此用 M_I 代表 M_D、M_I 和 M_{PIC}这 3 个参数进行探讨。可以认为 CR、M_D、M_I 和 M_{PIC}这 4 个参数在群体内和群体间均表现稳定。CR_{MAX}和 CR_{MIN}两个参数在群体内基本上也是呈上升趋势，但是 CR_{MAX}在 400 模拟个体的群体中的变化速率明显大于另外 3 个群体，CR_{MIN}在各个群体中变化速率也差别较大(图 3，图 4)，所以 CR_{MAX}和 CR_{MIN}两个参数在群体内表现稳定，但在群体间表现不稳定。在所有群体中，p 的值基本上在前期就达到了 100%，随后保持不变，该参数的变化速率只在前期在不同群体中有差别(图 5)，因此，p 在群体内表现稳定，在群体间虽然表现不稳定，但是稳定性好于 CR_{MAX}和 CR_{MIN}这两个参数。在各个群体中，随取样比例逐渐增大，VR 的值波动很大，而且，该参数在不同群体内的变化趋势也差别很大(图 6)，说明 VR 在群体内和群体间表现均不稳定。因此，就稳定性而言，各个参数的排序为：CR、M_D、M_I、$M_{PIC} > p > CR_{MAX}$、$CR_{MIN} > VR$。

2.3　各个评价参数的敏感性分析

由于 VR 在群体内表现不稳定，其敏感性没有规律，因此对它在群体内的敏感性不进行探讨。对

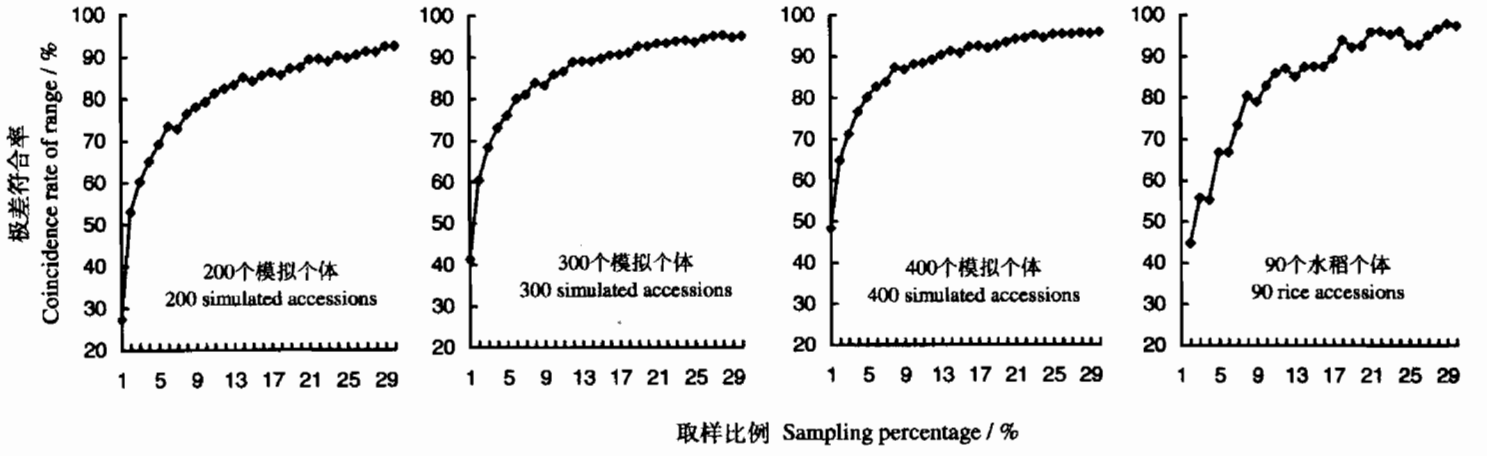

图 1　极差符合率 CR 在各个群体中的变化趋势

Fig. 1. Changing trend of coincidence rate of range in different populations.

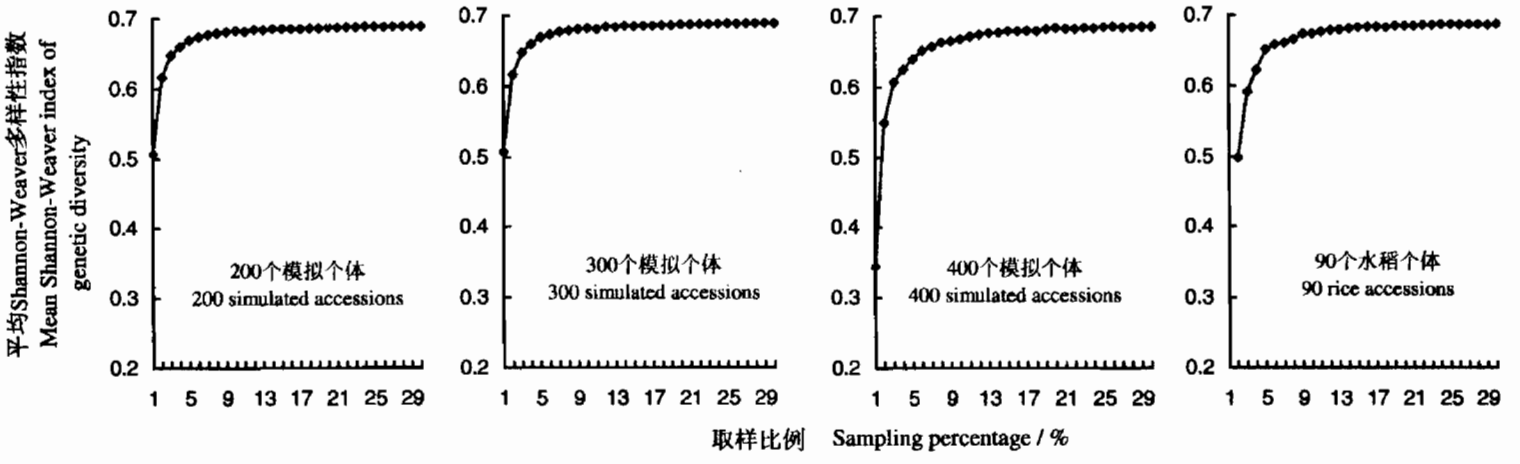

图 2　平均 Shannon-Weaver 多样性指数 M_I 在各个群体中的变化趋势

Fig. 2. Changing trend of mean Shannon-Weaver index of genetic diversity in different populations.

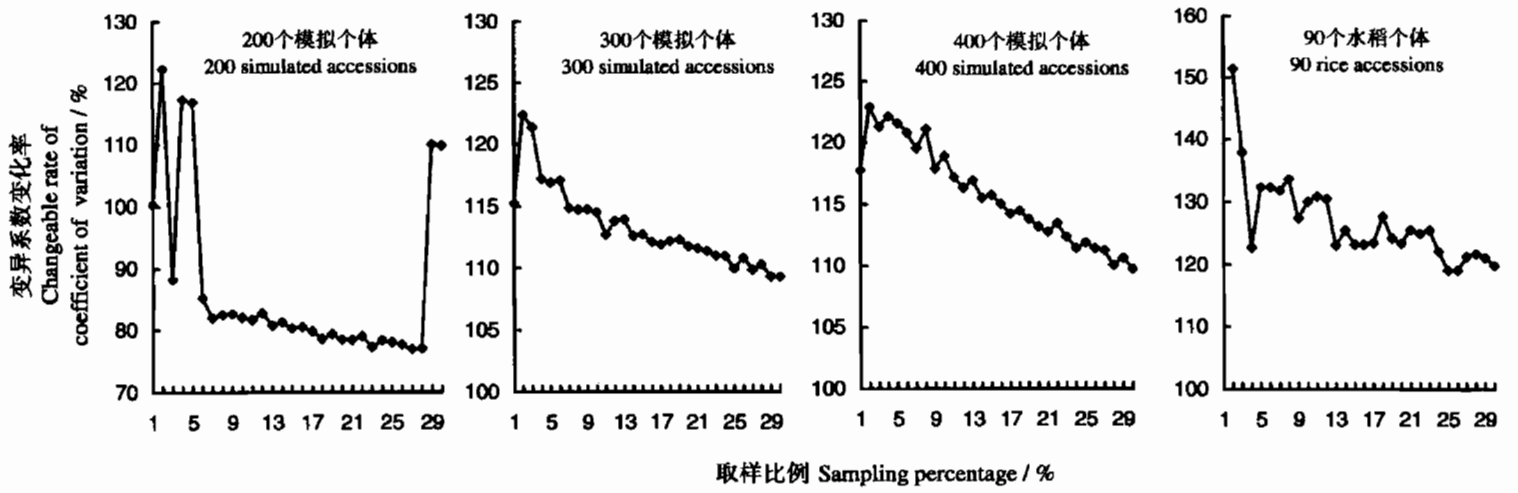

图 3 最大值变化率 CR_{MAX} 在各个群体中的变化趋势

Fig. 3. Changing trend of changeable rate of maximum in different populations.

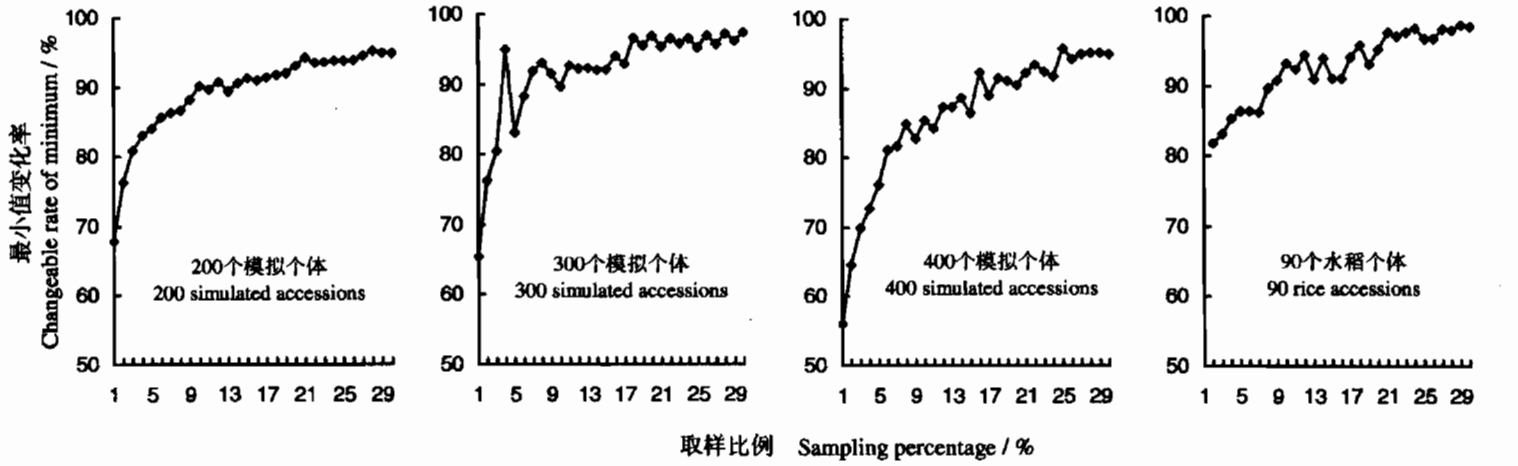

图 4 最小值变化率 CR_{MIN} 在各个群体中的变化趋势

Fig. 4. Changing trend of changeable rate of minimum in different populations.

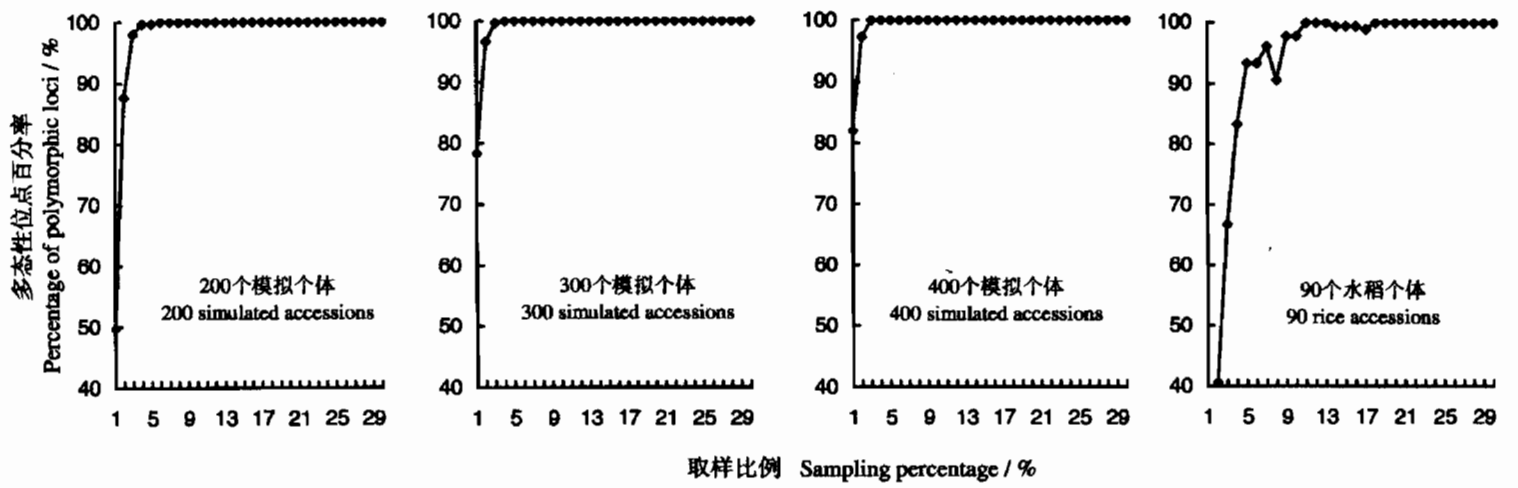

图 5 多态位点百分率 p 在各个群体中的变化趋势

Fig. 5. Changing trend of percentage of polymorphic loci in different populations.

比剩下的 7 个参数在各个群体中随取样比例改变的变化趋势图(图 1～图 5),发现 CR 在同样群体中的变化速率均大于其他参数,这表明 CR 比其他参数更容易在不同核心子集中表现出差异。CR_{MAX} 和 CR_{MIN} 两个参数在各个群体中的变化速率基本上是稳定保持在较高水平上,而 M_D、M_I 和 M_{PIC} 只在前期保持在较高水平上,后期变化缓慢。在 3 个模拟群体中,p 的值在取样比例大于 5% 之后,均达到 100%,在 90 个水稻个体的群体中,p 的值在取样比例超过 9% 之后,也趋于 100%(图 5)。因此,在较大取样比例的情况下,p 在所有核心子集中都是相同值,无法将不同核心子集区分开。综上所述,各个

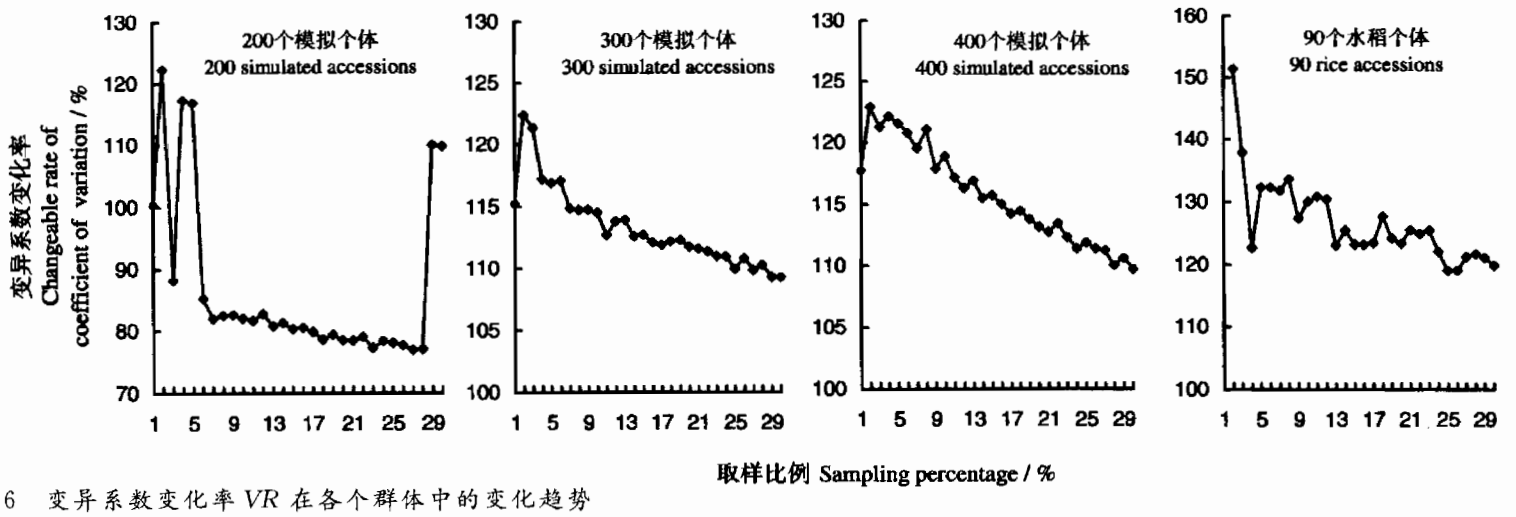

图 6　变异系数变化率 *VR* 在各个群体中的变化趋势

Fig. 6. Changing trend of changeable rate of coefficient of variation in different populations.

参数的敏感性排序为：$CR > CR_{MAX}$、$CR_{MIN} > M_D$、M_I、$M_{PIC} > p$。

3　讨论

在当前核心种质构建研究中，主要采用表现型值构建核心种质，对核心种质的代表性评价也是基于表现型值的[13-14]。基于基因型值的核心种质研究仅在棉花[9,11]和水稻[15]中有所报道。种质材料的表现型性状多为数量性状，容易受到环境因素和试验误差的影响，并且表现型的观察值中存在基因型与环境互作效应[9]。所以，基于表现型值的遗传分类与遗传多样性评价都可能是不准确的[1]。在本研究中，对各个核心子集的构建和评价都是在基因型值水平上的，排除了环境或人为因素的干扰，因此比用表现型值更能准确反映不同材料间的本质差异，对核心子集评价参数的研究结论也更加可靠。对于 3 个模拟群体，本研究采用蒙特卡洛模拟的方法得到种质材料的基因型值；对于真实的水稻群体，本研究采用混合线性模型的方法无偏预测出各个种质材料的基因型值。本研究使用的 11 个评价参数，各自在模拟群体中和真实群体中，其表现和变化趋势都是极为相似的。因此，可以认为本研究所使用的模拟方法是可靠的，模拟出的群体是适用于本研究的。本研究中核心子集的构建方法本质上是基于随机取样，由于随机取样的偶然性，虽然群体间数目差别不大(1%间隔取样)，但各群体间核心材料差别很大，不会对评价参数的筛选造成影响。

一个好的核心种质评价参数，应当能够有效地区分出不同的核心种质，而且在不同群体中都应有稳定的表现。所以，有效性和稳定性是首要的要求。敏感性表示评价参数在群体内随核心样本的变化而改变的剧烈程度，可作为核心种质评价参数的重要参考指标。综合所有参数的有效性、稳定性和敏感性的排序结果，可以认为 *CR* 在各个群体中，其有效性均排在第 1 位，而且其稳定性和敏感性都高于其他参数，因此可以作为评价核心种质代表性的首选参数。M_D、M_I 和 M_{PIC} 这 3 个参数，其有效性仅次于 *CR*，虽然敏感性稍差于 CR_{MAX} 和 CR_{MIN}，但是稳定性高于两者，可以认为这 3 个参数是评价核心种质代表性的重要参数。CR_{MAX} 和 CR_{MIN} 这两个参数在各个群体中的有效性与 M_D、M_I 及 M_{PIC} 这 3 个参数是相近的，而且具有较高的敏感性，因此也可以作为核心种质代表性的评价参数。从本质上说，*CR* 综合了这两类参数的优点，因而比这两类参数更加适用于核心种质的评价。*VR* 的有效性较 CR_{MAX} 和 CR_{MIN} 这两个参数差，在不同的群体中表现不稳定，但其他参数在单个群体中随取样比例改变而波动较大，因此可以作为评价核心种质的变异程度的重要参考参数。*p* 虽然有较高的稳定性，但是有效性和敏感性都较低，并且在分子标记数目较多的情况下，对于稍大规模的核心种质，该参数很容易达到 100%，无法区分出不同的核心种质，因此其使用受群体规模和标记数目的限制。但是，*p* 在某个取样比例前后有较高的变化率，因此可以作为一个判断核心种质取样规模的判定参数。本研究采用改进的逐步聚类方法构建核心子集，能够很好地保持原始群体的均值和方差。该方法构建的大部分核心子集，其均值和方差与原始群体接近，因此，*MD*、*VD* 和 CR_{MEA} 这 3 个参数没有表现出很好的有效性。然而，众多研究表明，*MD* 是判定核心种质是否可以接受的一个极其重要的参数[13,16]。通常情况下，只有一个核心种质的均值差异百分率小于 20%，才能认为该核心种质可以很好的代表原始群体的遗传多样性[9,15]。因此，*MD* 可作为一个判断

核心种质有效性的判定参数。

构建核心种质的目的在于以最小的样本规模最大限度地代表整个遗传资源群体的遗传多样性[2,17]。Basigalup 等[18]把极差和方差作为评价核心种质代表性的有效参数,认为性状具有较大极差和方差的核心种质有较强的代表性。Hu 等[9]把极差符合率作为一个重要的评价参数应用于棉花核心种质的构建。本研究在水稻核心种质中进一步证明了极差符合率的优越性,还发现基于群体平均数的参数有效性不佳,而 Shannon-Weaver 多样性指数具较好的有效性,这与张洪亮等[10]的研究结果是一致的。目前,在核心种质的构建研究中,所使用的信息是非常有限的,必须对种质资源的形态、农艺、生物化学、分子标记等不同类型性状综合评价,才能较为全面地反映遗传资源的多样性[19]。然而,众多的性状既有数量型指标,也有质量型指标,对不同类型的指标数据进行评价需要借助不同的评价参数。对于数量型指标,可以采用平均数和标准差把数量型指标的数据进行分级,从而将数量型数据转换为质量型数据[20-21]。

确定合理的取样比例是构建植物种质资源核心种质的重要环节。Brown[17]提出在样品数不少于 3 000 时,以 10%的取样比例就可以代表原始群体 70%的遗传多样性。Upadhyaya 和 Ortiz[13]在构建鹰嘴豆核心种质过程中,原始群体为 16 991 份样品,得到的核心种质为 1 956 份样品,取样比例约为 10%。Zewdie 等[22]构建的高粱核心种质也采用了 10%的取样比例。王丽侠等[23]构建的长江春大豆核心种质则只占原始群体(2 148 份样品)的 8.58%。徐海明等[11]构建的棉花核心种质,其原始群体只有 168 个样品,取样比例高达 30%。一般来说,对于大容量的原始群体,可以采用较小的取样比例;而对于小规模的原始群体,则应适当增大取样比例。然而,对于合理取样比例的确定,应当根据不同种质资源本身的特点进行。保存的种质资源的物种虽然各不相同,但其核心种质代表性的评价参数却是可以一致的。本研究筛选出的核心种质评价参数,适用于不同的种质资源群体,可以用作确定核心种质取样比例的判据。因此,解决了核心种质评价参数的筛选问题,很大程度上就解决了确定核心种质合理取样比例的问题。例如,可以把 $MD = 20\%$ 且 $CR = 80\%$时对应的取样比例定为合理取样比例。若要构建更为稳妥的核心种质,需要把上述标准调高并结合其他参数作为标准。

参考文献:

[1] Tanksley S D, McCouch S R. Seed bank and molecular maps: Unlocking genetic potential from the wild. *Science*, 1997, 277: 1063-1066.

[2] Frankel O H. Genetic perspectives of germplasm conservation//Arber W, Llimensee K, Peacock W J. Genetic Manipulation: Impact on Man and Society. Cambridge, UK: Cambridge University Press, 1984: 161-170.

[3] 徐匡迪, 沈国舫. 依靠稻作科技创新, 推动中国水稻产业发展. 中国稻米, 2002 (6): 8-11.

[4] 李自超, 张洪亮, 曾亚文, 等. 云南地方稻种资源核心种质取样方案研究. 中国农业科学, 2000, 33 (5): 1-7.

[5] 崔艳华, 邱丽娟, 常汝镇, 等. 黄淮夏大豆 (*G. max*) 初选核心种质代表性检测. 作物学报, 2004, 30 (3): 284-288.

[6] Bataillon T M, David J L, Schoen D J. Neutral genetic markers and conservation genetics: simulated germplasm collections. *Genetics*, 1996, 144: 409-417.

[7] 徐云碧, 朱立煌. 分子数量遗传学. 北京: 中国农业出版社, 1994: 108-110.

[8] 季道藩. 遗传学. 2 版. 北京: 中国农业出版社, 2000: 87-91.

[9] Hu J, Zhu J, Xu H M. Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. *Theor Appl Genet*, 2000, 101: 264-268.

[10] 张洪亮, 李自超, 曹永生, 等. 表型水平上检验水稻核心种质的参数比较. 作物学报, 2003, 29 (2): 252-257.

[11] 徐海明, 邱英雄, 胡 晋, 等. 不同遗传距离聚类和抽样方法构建作物核心种质的比较. 作物学报, 2004, 30 (9): 932-936.

[12] 朱 军. 作物杂种后代基因型值和杂种优势的预测方法. 生物数学学报, 1993, 8 (1): 32-44.

[13] Upadhyaya H D, Ortiz R. A mini core subset for capturing diversity and promoting utilization of chickpea genetic resources in crop improvement. *Theor Appl Genet*, 2001, 102: 1292-1298.

[14] Okpul T, Singh D, Guarua T, et al. Assessment of diversity using agro-morphological traits for selecting a core sample of Papua New Guinea taro (*Colocasia esculenta* (L.) Schott) collection. *Genet Res Crop Evol*, 2004, 51: 671-678.

[15] 李长涛, 石春海, 吴建国, 等. 利用基因型值构建水稻核心种质的方法研究. 中国水稻科学, 2004, 18 (3): 218-222.

[16] Upadhyaya H, Gowda C, Pundir R, et al. Development of core subset of finger millet germplasm using geographical origin and data on 14 quantitative traits. *Genet Res Crop Evol*, 2006, 53 (4): 679-685.

[17] Brown A H D. Core collection: A practical approach to genetic resources management. *Genome*, 1989, 31: 818-824.

[18] Basigalup D H, Barnes D K, Stucker R E. Development of a core collection for perennial Medicago plant introductions. *Crop Sci*, 1995, 35: 1163-1168.

[19] Singh S P, Gutierrez J A, Molina A, et al. Genetic diversity in cultivated common bean: II. Marker-based analysis of morphological and agronomic traits. *Crop Sci*, 1991, 31: 23-29.

[20] Balakrishnan R, Nair N V, Screenivasan T V. A method for establishing a core collection of *Saccharum officinarum* L. germplasm based on quantitative morphological data. *Genet Res Crop Evol*, 2000, 47: 1-9.

[21] Li Y, Shi Y S, Cao Y S, et al. Establishment of a core collection for maize germplasm preserved in Chinese National Genebank using geographic distribution and characterization data. *Genet Res Crop Evol*, 2004, 51: 845-852.

[22] Zewdie Y, Tong N K, Bosland P. Establishing a core collection of *Capsicum* using a cluster analysis with enlightened selection of accessions. *Genet Res Crop Evol*, 2004, 51: 147-151.

[23] 王丽侠, 李英慧, 李 伟, 等. 长江春大豆核心种质构建及分析. 生物多样性, 2004, 12 (6): 578-585.