Theor Appl Genet (2011) 122:49–61
DOI 10.1007/s00122-010-1421-7

ORIGINAL PAPER

A core collection and mini core collection of *Oryza sativa* L. in China

Hongliang Zhang · Dongling Zhang · Meixing Wang · Junli Sun · Yongwen Qi · Jinjie Li · Xinghua Wei · Longzhi Han · Zongen Qiu · Shengxiang Tang · Zichao Li

Received: 17 October 2009 / Accepted: 30 July 2010 / Published online: 18 August 2010

Abstract The extent of and accessibility to genetic variation in a large germplasm collection are of interest to biologists and breeders. Construction of core collections (CC) is a favored approach to efficient exploration and conservation of novel variation in genetic resources. Using 4,310 Chinese accessions of *Oryza sativa* L. and 36 SSR markers, we investigated the genetic variation in different sized sub-populations, the factors that affect CC size and different sampling strategies in establishing CC. Our results indicated that a mathematical model could reliably simulate the relationship between genetic variation and population size and thus predict the variation in large germplasm collections using randomly sampled populations of 700–1,500 accessions. We recommend two principles in determining the CC size: (1) compromising between genetic variation and genetic redundancy and (2) retaining the main types of alleles. Based on the most effective scheme selected from 229 sampling schemes, we finally developed a hierarchical CC system, in which different population scales and genetic diversities allow a flexible use of genetic resources. The CC, comprising 1.7% (932) of the accessions in the basic collection, retained more than 85% of both the SSR and phenotypic variations. A mini core collection, comprising 0.3% (189) of the accessions in the basic collection, retained 70.65% of the SSR variation and 76.97% of the phenotypic variation, thus providing a rational framework for intensive surveys of natural variation in complex traits in rice genetic resources and hence utilization of variation in rice breeding.

Communicated by T. Tai.

H. Zhang and D. Zhang contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-010-1421-7) contains supplementary material, which is available to authorized users.

H. Zhang · D. Zhang · M. Wang · J. Sun · Y. Qi · J. Li · Z. Li (✉)
Key Laboratory of Crop Genomics and Genetic Improvement of Ministry of Agriculture, Beijing Key Laboratory of Crop Genetic Improvement, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100193, China
e-mail: lizichao@cau.edu.cn

X. Wei · S. Tang
China National Rice Research Institute, Hangzhou 310006, Zhejiang, China

L. Han · Z. Qiu
Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing 100081, China

Introduction

Conservation and sustainable utilization of plant genetic resources are keys to improving agricultural productivity and sustainability, thereby contributing to national development, food security and poverty alleviation (FAO 1997). According to the World Information and Early Warning System (WIEWS) on Plant Genetic Resources for Food and Agriculture (PGRFA) (http://apps3.fao.org/wiews/wiews.jsp), more than 5,000,000 accessions of germplasm resources are held in worldwide collections. The extent of variation in these resources and their accessibility to biologists and breeders are essential factors affecting their utilization. Some collections have grown so large that they hinder the conservation, evaluation and accessibility of the genetic diversity they hold. Thus Frankel and Brown (1984)

proposed the concept of "core collection" (CC). Establishment of CC has proven to be a favored approach to facilitate efficient exploration of novel variation from genetic resources (Ellis et al. 1998; Holbrook et al. 2000; Malvar et al. 2004). However, the principles and methodologies for appropriate sampling proportions and the choice of individual accessions from basic collections are yet controversial. Various studies have suggested sampling proportions ranging from 5 to 30% (Brown 1989a; van Hintum et al. 2000; Yonezawa et al. 1995; Charmet and Balfourier 1995; Bisht et al. 1998; Noirot et al. 1996); however, the suggested sampling scales were either theoretically simulated or were not suitable for very large collections. Using a heuristic algorithm, Kim et al. (2007) developed the PowerCore program, which could allocate a subset of accessions to a CC that retained all characteristics for qualitative traits and all classes for quantitative ones. But with this program, different numbers and different entries are obtained with different data. MSTRAT, a program to implement the M strategy, provides a curve describing the relationship between redundancy and population size (Gouesnard et al. 2001). An appropriate size for the CC is determined near the inflection point of this curve (Gouesnard et al. 2001).

Among various methods for selecting individual accessions for CC, the stratified sampling and M (maximization) strategies are preferred by most researchers (Peeters and Martinelli 1989; Charmet and Balfourier 1995; Spagnoletti and Qualset 1993; Schoen and Brown 1993). Stratified sampling strategies include three steps: (1) grouping accessions by prior groups according to the available information such as origin and passport data, or by clustering according to the available phenotypic or molecular marker information; (2) allocating the number of accessions among groups according to the proportion or logarithmic proportion of the group size in the basic collection (such as the P strategy, L strategy, Brown 1989b), or according to the proportion of the diversity in each group of the basic collection (such as the H strategy in Schoen and Brown 1993, the D method in Franco et al. 2006, the G method in Li et al. 2002); and (3) choosing the individuals from each group randomly or according to genetic membership among accessions (such as the clustering method in Li et al. 2002). The M strategy aims at selecting the highest diversity among subsets (Schoen and Brown 1993) and is expected to perform particularly well (Bataillon et al. 1996). The MSTRAT program was able to implement the M strategy, providing the opportunity not only to determine an optimal CC size, but also to choose the representative individuals for a given sized CC (Gouesnard et al. 2001; Franco et al. 2006).

China is one of centers of origin of Asian cultivated rice (*Oryza sativa* L., hereafter called "cultivated rice") (Oka 1988). In the national germplasm collections, there are 61,479 accessions of cultivated rice and its ancestral wild relative, common wild rice (*Oryza rufipogon* Griff.). The Chinese germplasm resources of cultivated rice include 50,526 landrace or local varieties (LV, pureline varieties developed by farmers without artificial intercrossing) and 5,382 modern varieties. The latter includes 4,085 modern pureline varieties (MPV) and 1,297 parents of trilinear hybrids (PTH). Through a hierarchical sampling strategy, we constructed a primary CC of cultivated rice, comprising 4,310 varieties (including 3,632 LV, 604 MPV and 74 PTH accessions) and retained about 95% of the morphological variation in the basic collection (Li et al. 2003). Fifty traits of the primary CC were evaluated in 2000 and 2001, and all accessions of the primary CC were genotyped using 36 SSR markers. Here, we will develop a more efficient CC and a more practical mini core collection (MCC) to provide a rational framework for undertaking intensive surveys of natural variation, including the phenotyping of complex traits and genotyping of DNA polymorphisms, allowing more efficient utilization of natural variations for the study and breeding of complex traits.

Materials and methods

Plant materials

The research material used was the primary CC of cultivated rice comprising 4,310 accessions (Li et al. 2003), including landrace varieties, MPV and PTH (Table S1). According to the records of the Chinese Crop Germplasm Information Network (http://icgr.caas.net.cn/pztest.htm), all accessions could be allocated to six ecological zones (EZ) in China (Li et al. 2003) and to two subspecies (*indica* and *japonica*). In addition, by using the same 36 SSR markers as in Li et al. (2003) and the program STRUCTURE version 2.1 (Falush et al. 2003), we have inferred the population structures of LV (Zhang et al. 2009), MPV (Qi 2007) and PTH (unpublished). The distributions of varieties in different EZ and subspecies are listed in Supplementary Table S1 and the numbers in clusters inferred from STRUCTURE are listed in Supplementary Table S2.

SSR genotyping

SSR markers show high polymorphism and are ubiquitous in occurrence (Tautz and Renz 1984). The 36 SSR markers (Table S3) were randomly distributed across all 12 rice chromosomes (3 markers per chromosome). Following the modified CTAB procedure (Saghai-Maroof et al. 1984), DNA of cultivated rice was extracted from fresh leaf tissues of 30-day seedlings planted in the greenhouse and that of wild rice from the silica gel-dried leaf tissues collected

from nurseries grown at the Guangdong Academy of Agricultural Sciences, Guangzhou, or Guangxi Academy of Agricultural Sciences, Nanning. PCR conditions were as described by Panaud et al. (1996). PCR products were run on 8% denaturing polyacrylamide gels at 2,000 V, and bands were visualized using silver staining as described by Zhang et al. (2007) and Wang et al. 2008.

Evaluation of morphological and agronomic traits

Fifty traits, including morphological traits diverse in rice and agronomic traits important in rice breeding, of the primary CC were assessed in 2001 and 2002 in Hangzhou (Zhejiang Province) and Sanya (Hainan Province). Four rows with six plants each were planted for each variety. Data for each measured quantitative trait were averaged from ten randomly sampled individuals. The traits were evaluated based on *Descriptors for Rice*, *O. sativa* L. and *Method for characterizing the morphological traits of rice*, *O. sativa* L. established by the International Rice Research Institute. The 34 discrete traits included basal sheath color, leaf color, the penultimate leaf angle, the penultimate leaf ear color, blade pubescence, ligule color, ligule shape, pedestal color, flag leaf curl, flag leaf angle, culm diameter, culm angle, lodging, culm node exposure, culm node color, internode color, culm angle, panicle type, panicle exsertion, stigma color, apiculus color, lemma and palea color, glume color, glume hair, sterile lemma length, awn length, awn color, awn distribution, seed coat color, stigma exposure, grain shape, panicle shattering, appearance accessibility and leaf senescence; and the 16 quantitative traits were length of penultimate ligule, length of penultimate leaf, width of penultimate leaf, length of flag leaf, width of flag leaf, plant height, days to flowering, length of panicles, number of panicles per plant, panicle weight per plant, number of primary branches, number of secondary branches, grain length, grain width, 1,000-grain weight and ratio of full seed.

Estimation of core collection size

The expected number of alleles in a sample of n accessions at a particular locus could be estimated according to the binomial distribution of the allele frequency (described by Crossa et al. 1993) or according to the binomial distribution of the genotype frequency (See supplementary material for detail). But could the sum of the expected number of alleles at each of l loci (according to Eq. 1) represent the allelic number at all l loci when the same n accessions are randomly sampled? The answer may be positive when all l loci are homozygous and independent of each other, otherwise not always. Thus, we tried to simulate the relationship between the allelic number at l loci and the population scale using a curve-fitting system (Hyams 2003). We carried out the simulation using the primary CC of cultivated rice. A series of subsets were sampled incrementally from 5 to 4,305 with a step of five accessions. Five independent samples were obtained for each subset. The relationship between allelic number (i.e., the numbers of allelic states in different subsets) and population size (i.e., the numbers of accessions in different subsets) was established using a curve-fitting system (Hyams 2003). Pre-fitting showed that the best mathematical model was the Morgan–Morgan–Finney (MMF, Eq. 1) (Morgan et al. 1975),

$$y = \frac{ab + cx^d}{b + x^d} \tag{1}$$

where a is the lowest asymptote, b the scale parameter, c the highest asymptote, and d is the parameter to determine the inflexion. In this study, the interval of the independent variable was $(1{-}\infty)$, and y was $L(1 + H_o)$ [here H_o and L are constants for certain populations, H_o is the real heterozygosity, and L is the number of loci (36 in this study)]. When only one accession is sampled, the parameter a can be taken as $[L(1 + H_o)(b + 1) - c]/b$, leading to Eq. 2. The relationship between allelic number (take y) and population size (take x) was fitted to this modified model.

$$y = \frac{L(1 + H_o)(b + 1) - c + cx^d}{b + x^d} \tag{2}$$

First, we determined the parameters of the MMF models using all 4,305 subsets. Using the fitted MMF models, we estimated the allelic numbers in the Chinese basic rice collection (50,526 accessions) and in a series of hypothetical basic collections with sizes from 5,000 to 60,000 accessions incrementally increasing by 5,000. To estimate the robustness of MMF in estimating the allelic number of collections of certain size, we obtained the parameters of 43 MMF models and sampled the subsets incrementally with a step of 5 accessions from each of 43 populations: 100, 200,…, till 4,300 with a step of 100 accessions. Using each of the 43 fitted MMF models, we estimated the allelic numbers in the same hypothetical basic collections as above. Thus, we obtained 44 estimations of allelic number in one hypothetical basic collection, one based on the MMF model from 4,300 accessions and the others based on those from 43 different sized populations. The robustness of the estimations using different sized populations was represented by the ratio between the latter and the former estimated number of alleles in the hypothetical basic collections.

In determining the optimal size of a CC, we considered three methods. One was based on the relationship between the allelic number and population size revealed by the

MMF model, as well as the retention of alleles with different frequencies. The second was provided by the program PowerCore (Kim et al. 2007). The third was inferred from the curve between redundancy and population size given by the program MSTRAT (Gouesnard et al. 2001).

Sampling strategy of core collection

We used several methods to construct the CC. In general, the nongroup-based strategy and the group-based (stratified) strategy were designed (Fig. 1). The nongroup-based strategy chose varieties from the primary CC in four ways: (1) random with three repeats, (2) clustering based, i.e., dividing all varieties into the given number of clusters in clustering analysis, then randomly selecting one from each cluster with two repeats, (3) by the PowerCore program and (4) by the MSTRAT program with three repeats. The core sets sampled by the four nongroup-based schemes were labeled as NG-R, NG-CR, NG-PC and NG-M, respectively. The group-based strategy included three steps: grouping the primary CC, then allocating the accessions among groups and finally selecting individuals from each group (Fig. 1).

We designed a hierarchical two-level grouping strategy. The first level related to only one grouping method, including three variety types: LV, MPV and PTH (Tables S1, S2). At the second level, each type of variety was grouped in three ways: (1) by EZ (Table S1), (2) by subspecies (Table S1) and (3) by population structure (Table S2). At both levels, we used five methods to allocate the varieties among groups: (1) *C*—constant number, (2) *G*—proportional to Nei's gene diversity index of the group in basic collection (Li et al. 2002, similar to the H strategy in Schoen and Brown (1993) and the D method in Franco et al. (2006), (3) *L*—proportional to the logarithmic group size in the basic collection, (4) *P*—proportional to the group size in the basic collection, (5) *S*—proportional to the square-root of the group size in the basic collection (Li et al. 2002). We chose the given number of varieties from each of the above groups in three ways: (1) random with two repeats (*R*), (2) dividing the varieties into given number of clusters equal to the given number of varieties based on clustering analysis, then randomly selecting one from each cluster with two repeats (CR), (3) dividing the varieties into given number of clusters equal to the given number of varieties based on clustering analysis, then selecting from each cluster the most representative individual—any one of the two being firstly clustered together during clustering analysis (CM). There were 225 sampling schemes for the above group-based strategy and these were designated using four combined symbols separated by short dashes: the first symbol representing the method of allocating varieties among variety types (5 methods), the second representing the method of grouping varieties within each variety type (3 methods), the third representing the methods of allocating varieties among the second level of groups (5 methods) and the fourth representing the methods choosing varieties from each second level group (3 methods). For example, VTC-EZ-C-R represents the scheme that allocates a constant number of varieties among different variety types (VTC), groups the varieties into different EZ, allocates a constant number of varieties among different EZ (*C*) and finally randomly chooses varieties from each EZ (*R*). We assessed the advantages of different sampling strategies and determined the optimal sampling scheme by analysis of variance and pairwise *t* tests using two indicators, retention of SSR alleles in each core set relative to the primary CC and Nei's gene diversity index H_e.

We also compared the advantages of our sampling strategies with the M strategy (Gouesnard et al. 2001). The M strategy-based sampling was performed by the program MSTRAT. The comparison comprised group-based methods and nongroup-based methods, both of which included three methods of choosing varieties, i.e., random (*R*), clustering-based random (CR) and M strategy based (M). In the comparison, we focused on the retention of SSR alleles and the Nei's gene diversity index for the unsupervised SSR loci, i.e., those not used in the process of clustering or simulation of MSTRAT.

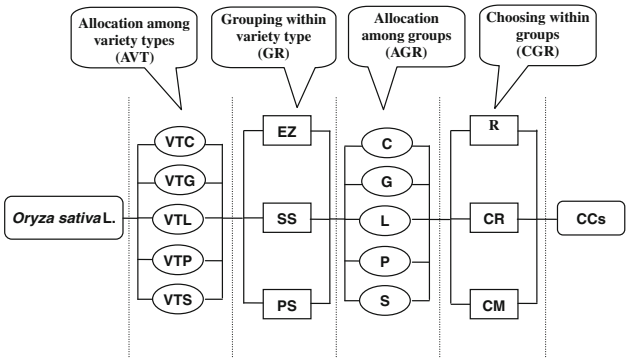

Fig. 1 Group-based (*hierarchical*) sampling schemes of core sets of cultivated rice in China. Here, VTC, VTG, VTL, VTP and VTS refer to the methods of allocating sampling scales among three variety types (VT): *C* constant number, *G* proportion to Nei's gene diversity index for the group in the basic collection, *L* proportion of the logarithm group size in the basic collection, *P* proportion of the group size in the basic collection, *S* proportion of the square-rooted group size in the basic collection. EZ, SS and PS refer to the grouping methods within each variety type according to ecological zones, two subspecies (*indica* and *japonica*) and population structure. C, G, L, P and S refer to the methods allocating sampling scales among groups within each variety type, same as those among variety types. *R* random sampling, *CR* clustering-based random sampling, *CM* clustering-based most representative sampling, *CCs* core collections

Statistics

Clustering analysis during sampling was conducted in SPSS version 15 program using simple match index and within group linkage. Number of allele (N_a), Nei's unbiased gene diversity index (H_e) and the observed heterozygosity (H_o) in populations or groups were calculated using the POWERMARKER version 3.25 software (Liu and Muse 2004; http://www.powermarker.net). The ANOVA with model III and pairwise t test were also conducted in the SPSS version 15 program.

Results

Genetic variation in the primary core collection

In total, 605 alleles at 36 SSR loci were detected in the primary CC of cultivated rice in China. The numbers of alleles per locus varied from 3 to 32, averaging 16.8. The numbers of SSR alleles of LV, MPV and PTH were, respectively, 588, 464 and 240, indicating that the genetic variation in LV was the most abundant, whereas that of PTH was comparatively limited.

The degree of genetic redundancy reflecting the proportions of alleles with higher frequency is the main factor that influences the retention of variation in a sample (Crossa et al. 1993). The normal distribution of the common logarithm of allele frequencies (Fig. S1) allowed us to classify alleles into four types according to their frequencies (P): we called them predominant alleles ($P > 0.1$), common alleles ($0.1 \geq P > 0.01$), rare alleles ($0.01 \geq P > 0.001$) and inferior alleles ($P \leq 0.001$). In the primary CC, 42.8% of the alleles belong to rare alleles, and 25.0, 17.2 and 15.0% of them belonged to common alleles, predominant alleles and inferior alleles, respectively.

Determination of core collection size

In determining the CC size and estimating the validity of a CC, it is important to know the magnitude of genetic variation in the basic collection and the relationship between the degree of variation and population size. We therefore estimated the allelic number in the Chinese basic collection by a modified MMF model (Morgan et al. 1975), using 4,310 accessions of the primary CC (Eq. 3): 'MMF-based estimation'. Accordingly, there were 644 alleles in the basic collection (55,908 accessions) (Table 1), i.e., 17.9 alleles per locus, only about 6% higher than that in the primary CC (16.8).

$$y = \frac{-465.0799 + 665.4138x^{0.4581}}{4.2190 + x^{0.4581}} \quad (3)$$

As expected, the robustness of estimating allelic numbers using different-sized populations (Fig. 2) showed that the smaller population gave the lower-biased estimation of the number of alleles in the basic collection. However, robustness rapidly increased when the population being used for the estimation increased from 100 to 300. A population with about 3,000 accessions gave a nearly unbiased estimation. However, 90% of the unbiased estimate could be achieved using populations of 200–600 accessions, and 95% of the unbiased estimation could be achieved using 700–1,500 accessions.

The curve line according to Eq. 3 indicated that increasing rates of allele number were not linear (Fig. 3). Along with the increase in population size, the rate of increase in allelic numbers (IRI = increased number of SSR alleles/unit of increased accession numbers) decreased and genetic redundancy increased. According to IRI, we divided the transition between genetic variation and genetic redundancy into three stages as shown in Fig. 3, viz. the stage of rapid increase in genetic variation (RiGV) (IRI $\geq$ 1) with a slight genetic redundancy, the stage of

Table 1 Entry numbers and marker and phenotypic parameters of various Chinese rice collections

Collection	Accessions			SSR				Phenotype			
	Num.	%[a]	%[b]	N_a	%[a]	%[b]	H_e	N_v	%[a]	%[b]	I
MCC	189	4.39	0.34	455	75.21	70.65	0.736	137	81.07	76.97	0.796
C500	500	11.6	0.89	516	85.29	80.12	0.748	152	89.94	85.39	0.824
CC	932	21.62	1.66	552	91.24	85.71	0.753	156	92.31	87.64	0.839
C1500	1,500	34.8	2.68	573	94.71	88.98	0.757	162	95.86	91.01	0.850
C2000	2,000	46.4	3.57	582	96.2	90.37	0.756	162	95.86	91.01	0.853
PCC	4,310	–	7.71	605	–	93.94	0.746	169	–	94.94	0.853
BC	55,908	–	–	644	–	–	–	178	–	–	–

CC, core collection; MCC, mini core collection; C500, the core set with 500 accessions (the same for C1500 and C2000); N_a, the number of SSR alleles; N_v, the number of phenotypic variations; H_e, Nei's unbiased gene diversity index being estimated using POWERMARKER version 3.25 (Liu and Muse 2004; http://www.powermarker.net)

[a] Percentage of estimators relative to those in the primary core collection (PCC)

[b] Percentage of the estimators relative to those in the basic collection (BC)

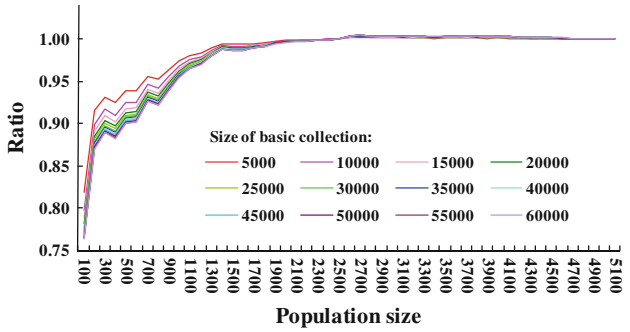

Fig. 2 Robustness in estimating allele numbers in different sized basic collections using different hypothetically sized simulated populations. The *vertical axis* indicates the ratio between the allelic number in the designated basic collection estimated by a certain sized simulation population relative to that estimated for the primary core collection (4,310 accessions)

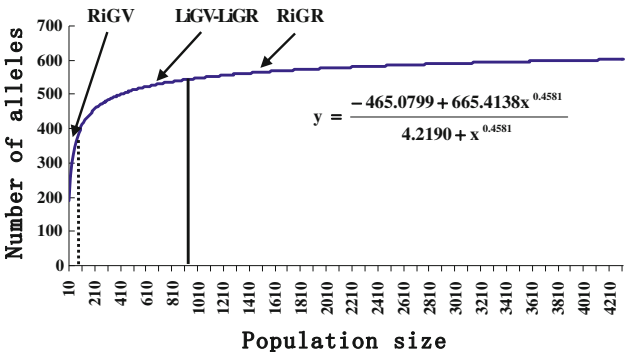

Fig. 3 Relationship between allelic number and population size, as well as the dynamic transition between genetic variation and genetic redundancy. *RiGV* stage of rapid increase in genetic variation with slight genetic redundancy, *LiGV-LiGR* stage of low increases in both genetic variation and genetic redundancy, *RiGR* stage of rapid increase in genetic redundancy with a slight increase in genetic variation

low increases in both genetic variation ($1 > \text{IRI} \geq 0.05$) and genetic redundancy (LiGV-LiGR), and the stage of rapid increase in genetic redundancy (RiGR) with slight increase in genetic variation ($\text{IRI} < 0.05$). To determine an appropriate size of a CC, it is necessary to compromise between the minimum number of accessions and the maximum allelic number, i.e., to find a point with high sampling efficiency through a compromise between genetic variation and genetic redundancy. We considered, therefore, that the population size, near the point of transition from LiGV-LiGR to RiGR (indicated by the vertical solid line) where there are 932 accessions and where, at most, five new alleles could be sampled per 100 accessions (Fig. 3), would be the most appropriate point of compromise for our expected CC. To achieve the highest sampling efficiency, we recommend a further option, which we call the MCC. The MCC should be sampled near the point of transition from RiGV to LiGV-LiGR, as marked by the vertical dotted line, where there are 82 accessions and at least 100 new alleles could be sampled per 100 accessions (Fig. 3).

We randomly sampled different sized collections from the primary CC, and the relative retentions of four types of alleles on average over five repeats showed that 100% of predominant alleles and 95% of common alleles in the primary CC were retained in samples of 189 accessions (Fig. 4). However, the inclusion of 95% of rare alleles required a sample size of 900 accessions. Retention of most inferior alleles required very large collections; for example, more than 3,000 accessions were needed if 90% of inferior alleles were retained. The distribution and retention of different types of alleles prompted us to suggest a second criterion of sampling CC, that is, the inclusion of most predominant, common and rare alleles in a CC, whereas most of the predominant and common alleles in an MCC would be sufficient and practical.

Finally, we determined the sizes of the CC and MCC by two principles. The first related to sampling efficiency and required a "point of compromise" between genetic variation and genetic redundancy contained in the CC, and the second related to the sampling validity and required a "point" at which the main variation types can be preserved in the CC (see "Discussion" for details). Because the two principles set different population sizes (82 vs. 189 for MCC, and 930 vs. 900 for CC), we took one of the larger populations to sample as much variation as possible. Finally, 189 and 932 accessions were recommended for MCC and CC, respectively, accounting for 0.3 and 2.2% of the accessions in the respective basic collections (Table 1). Simulation using PowerCore generated a core set of 174 accessions, the size and genetic diversity (0.78) of which were similar to our recommended MCC. Redundancy estimation given by the program MSTRAT showed that the

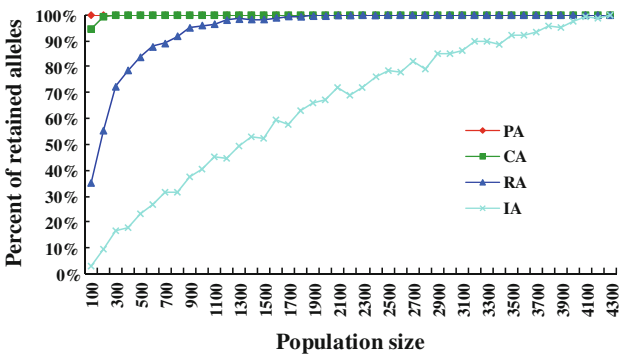

Fig. 4 Retention of four types of alleles in different sized randomly sampled collections relative to the primary core collection. *PA* predominant alleles ($P > 0.1$), *CA* common alleles ($0.1 \geq P > 0.01$), *RA* rare alleles ($0.01 \geq P > 0.001$) and *IA* inferior alleles ($P \leq 0.001$)

increase in allele number reached a platform stage and showed no distinct difference between random sampling and optimized sampling in a population of about 932 individuals (Fig. S2).

Effect of sampling strategy on the SSR allele retention and genetic diversity of core collections

Sampling strategies had significantly different effects on SSR allelic retentions (RT) and especially Nei's gene diversity indexes (H_e) of CC (Tables 2, S4; Fig. 5). Among five methods of allocation among variety types, the proportional method captured significantly more alleles and higher genetic diversity in most cases, except when varieties were grouped according to EZ and were randomly chosen within groups. These deviant cases may be attributed to interactions between methods of allocation among variety types, grouping within variety type and choice within groups. Among the three methods of grouping within variety type, grouping by EZ led to apparently less SSR alleles and lower gene diversity than that by population structure and subspecies. Because of the interaction of grouping method within variety type with allocation among variety types, allocation among groups and method of

Table 2 ANOVA of 225 group-based sampling schemes using SSR allele retentions (RT) and Nei's gene diversity indexes (H_e) as indicators

Source	Dependent variable	Sum of squares	*df*	Mean square	*F*	Sig.
Main effects						
Allocation methods among variety types (AVT)	RT**	0.0057	4	0.0014	28.44	1.4E−17
	H_e**	0.0014	4	0.0004	66.15	3.8E−32
Grouping methods within each variety type (GR)	RT**	0.0127	2	0.0063	125.70	8.7E−33
	H_e**	0.0011	2	0.0005	99.81	2.7E−28
Allocation methods among groups within each variety type (AGR)	RT**	0.0014	4	0.0004	7.02	3.3E−05
	H_e**	0.0003	4	0.0001	15.08	2.2E−10
Choice methods within groups (CGR)	RT**	0.0096	2	0.0048	94.95	2.3E−27
	H_e**	0.0088	2	0.0044	819.93	1.8E−81
Interaction effects						
AVT × GR	RT**	0.0015	8	0.0002	3.76	4.8E−04
	H_e**	0.0003	8	0.0000	5.91	1.4E−06
AVT × AGR	RT	0.0009	16	0.0001	1.13	3.4E−01
	H_e	0.0001	16	0.0000	0.99	4.7E−01
AVT × CGR	RT	0.0005	8	0.0001	1.33	2.3E−01
	H_e**	0.0006	8	0.0001	14.19	2.8E−15
GR × AGR	RT**	0.0012	8	0.0002	2.98	4.0E−03
	H_e**	0.0006	8	0.0001	14.40	1.8E−15
GR × CGR	RT**	0.0015	4	0.0004	7.23	2.4E−05
	H_e**	0.0003	4	0.0001	15.19	1.9E−10
AGR × CGR	RT	0.0003	8	0.0000	0.71	6.8E−01
	H_e*	0.0001	8	0.0000	2.31	2.3E−02
AVT × GR × AGR	RT	0.0012	32	0.0000	0.73	8.5E−01
	H_e	0.0002	32	0.0000	1.23	2.0E−01
AVT × GR × CGR	RT	0.0007	16	0.0000	0.88	6.0E−01
	H_e**	0.0004	16	0.0000	5.17	1.6E−08
AVT × AGR × CGR	RT	0.0010	32	0.0000	0.61	9.5E−01
	H_e	0.0001	32	0.0000	0.27	1.0E+00
GR × AGR × CGR	RT	0.0008	16	0.0001	1.03	4.3E−01
	H_e	0.0001	16	0.0000	1.08	3.8E−01
AVT × GR × AGR × CGR	RT	0.0025	64	0.0000	0.78	8.7E−01
	H_e	0.0001	64	0.0000	0.43	1.0E+00
Error	RT	0.0076	150	0.0001		
	H_e	0.0008	150	0.0000		
Total	RT	0.0490	374			
	H_e	0.0154	374			

* and **, significant at $P < 0.05$ and 0.01, respectively

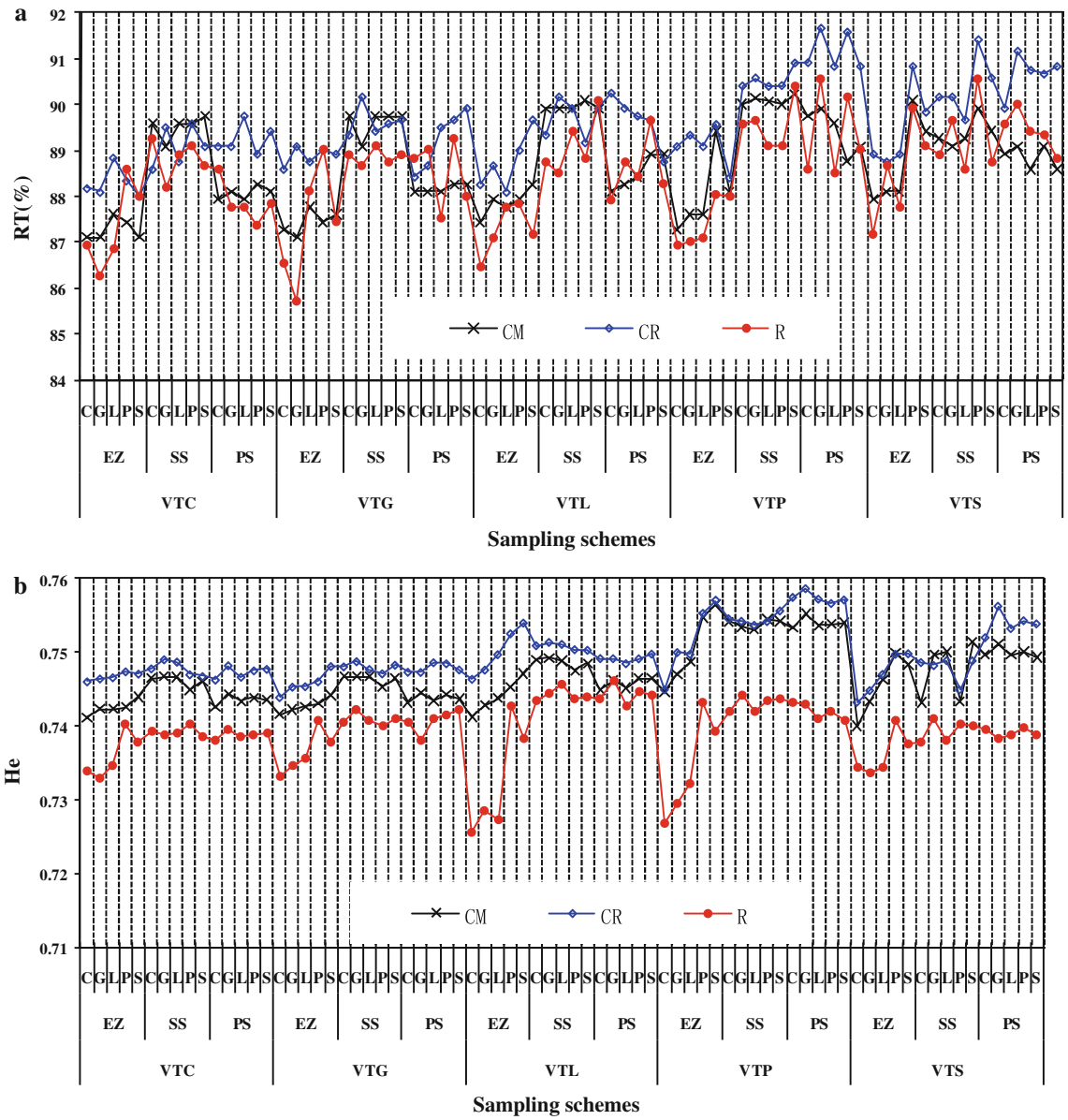

Fig. 5 Retention of SSR alleles (RT) (**a**) and Nei's gene diversity index (H_e) (**b**) in each core set developed by 225 group-based sampling schemes (Fig. 1)

choosing varieties within groups, the main effects of the two grouping methods within variety type (population structure and subspecies) on allelic retention and genetic diversity were not significantly different from each other. These were dependent on the methods of allocation among variety types, grouping within variety type and choice within group, but the highest values of two indicators both resulted from the population structure-based grouping method within variety type (Fig. 5). Among five methods of allocation among groups, the main effects of proportional method and square root-based proportional method were much greater in increasing allelic retention and gene diversity (Tables S4, S5), but were also dependent on other factors including the methods of allocation among variety types and grouping methods within variety groups because of their interactions (Fig. 5). Both the highest allele retention and the highest gene diversity were achieved by the genetic diversity-based proportional method (Fig. 5). Among three methods of choosing varieties within groups, randomly choosing varieties after clustering led to higher allelic retention and gene diversity than choosing the most representative varieties after clustering. However, choosing the most representative varieties after clustering was significantly better than randomly choosing varieties without

clustering (R) (Fig. 5; Table S4). Thus, the best sampling scheme among the 225 was VTP-PS-G-CR, i.e., allocating varieties among three variety types proportionally to their population sizes, grouping them by population structure within each variety type, allocating varieties among groups proportionally to the genetic diversities within the groups and, finally, randomly choosing varieties after clustering.

Evidently, scientists are often more interested in alleles with lower frequencies. Our results on four types of alleles showed that all sampling schemes retained all predominant and common alleles, but there were differences among sampling schemes for rare and inferior alleles. Both random and representative choice within group after clustering retained significantly more rare and inferior alleles than randomly choosing without clustering. Randomly choosing after clustering kept significantly higher retention than choosing "the most representative" variety after clustering (Table S5; Fig. S3). Grouping according to subspecies and population structure was more efficient in retaining rare and inferior alleles than grouping according to EZ (Table S5; Fig. S3).

Capture of unsupervised genetic diversity

Ideal sampling methods for CC should capture more genetic variation and have less redundancy, especially of 'unsupervised' variation. We therefore investigated retention of alleles (RT, Table 3) and gene diversity (H_e, Table 4) in core sets developed by different sampling methods. G-CR (VTP-PS-G-CR), the best sampling scheme among the 225, retained more alleles (higher RT) and removed more redundancy (higher H_e) of unsupervised loci than other methods except for two M strategy-based methods. There were no distinct differences between supervised loci (24), unsupervised loci (12) and all 36 loci for the core and MCC sampled by the optimal sampling scheme (VTP-PS-G-CR) (Fig. S4).

Table 3 Pairwise t tests of SSR allele retentions (RT) of unsupervised loci between sampling methods

Sampling methods	RT	NG-R	G-CR	NG-CR	G-M	G-R	NG-M
NG-R	0.906	–	NS	NS	NS	NS	NS
G-CR	0.906	0.001	–	NS	NS	NS	*
NG-CR	0.903	0.176	0.355	–	NS	NS	NS
G-M	0.900	0.336	0.552	0.322	–	NS	NS
G-R	0.894	1.043	0.620	0.525	0.328	–	NS
NG-M	0.886	1.122	2.294	1.865	1.435	0.439	–

The t values are shown below the diagonal, and the significances are indicated above the diagonal. For NG-R, NG-CR, NG-M, G-R, G-CR and G-M, NG refers to nongroup-based method, G to the optimal grouping and allocation method revealed in 225 group-based sampling schemes, R to random sampling, CR to clustering-based random sampling and M to M strategy-based sampling

NS nonsignificant

* Significant at $P < 0.05$

Table 4 Pairwise t tests of Nei's gene diversity indices (H_e) of unsupervised loci between sampling methods

Sampling methods	H_e	G-CR	NG-CR	G-M	NG-R	G-R	NG-M
G-CR	0.752	–	NS	**	**	**	**
NG-CR	0.748	0.490	–	NS	NS	NS	NS
G-M	0.748	2.896	0.002	–	NS	*	*
NG-R	0.746	4.773	0.209	1.415	–	NS	NS
G-R	0.745	3.929	0.363	2.323	1.405	–	NS
NG-M	0.743	3.453	0.511	2.351	1.736	1.616	–

The t values are shown below the diagonal; significances are given above the diagonal. For NG-R, NG-CR, NG-M, G-R, G-CR and G-M, NG refers to nongroup-based method, G to the optimal grouping and allocation method revealed in 225 group-based sampling schemes, R to random sampling, CR to clustering-based random sampling and M to M strategy-based sampling

NS nonsignificant

* and **, significant at $P < 0.05$ and 0.01, respectively

Generation of the core collections

The above results showed that VTP-PS-G-CR was the optimal sampling scheme to establish the CC of Chinese cultivated rice. We established the CC and MCC using this scheme. The CC (932 accessions), comprising 785, 112 and 35 accessions of LV, MPV and PTH, respectively, retained 91.24% of the alleles in the primary CC and 85.71% of the alleles in the basic collection (Table 1). The MCC (189 accessions), comprising 159, 23 and 7 accessions of LV, MPV and PTH, respectively, retained 75.21% of the alleles in the primary CC and 70.65% of the alleles in the basic collection (Table 1). Both the CC and MCC retained almost all predominant alleles and common alleles, and the former also retained more than 95% of rare alleles and near 50% of inferior alleles, whereas the MCC retained more than 55% of rare alleles (Fig. 6).

In addition to the genetic diversity of SSR, we also checked the morphological diversity in the established CC and MCC. The CC retained 92.31% of the morphological variation, and diversity index (0.7957) higher than that of the primary CC (0.7911). This implied that the CC not only retained most alleles, but also removed most redundancy. Among 16 quantitative traits (Table S6), the variances and coefficients of variation of all traits, except the penultimate leaf, length of panicle and plant height for the CC, were higher than those for the primary CC. The phenotype

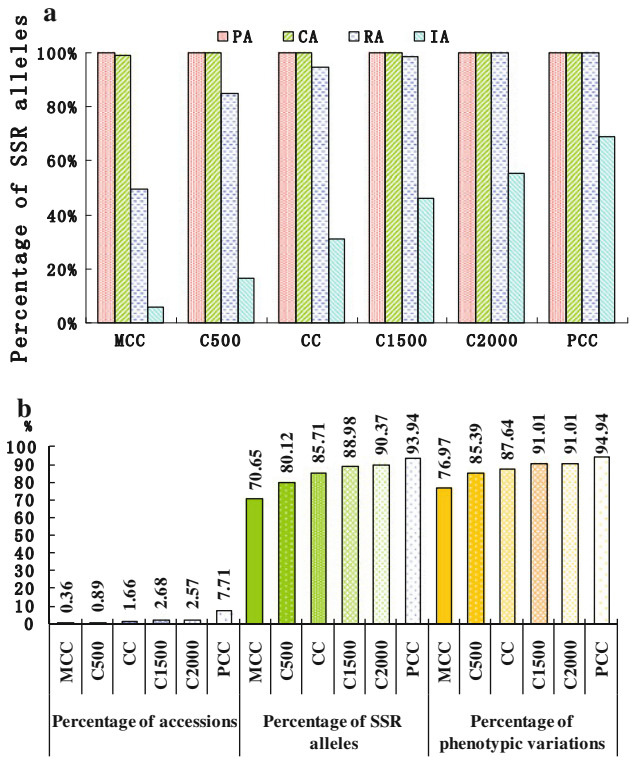

Fig. 6 Retention of different alleles (**a**) and percentage of accessions, SSR alleles and phenotypic variations (**b**) in the hierarchical core collections. *PA* predominant alleles ($P > 0.1$), *CA* common alleles ($0.1 \geq P > 0.01$), *RA* rare alleles ($0.01 \geq P > 0.001$), *IA* inferior alleles ($P \leq 0.001$), *MCC* mini core collection, *CC* core collection, *C500* core set with 500 accessions and similar for C1000, C1500 and C2000

ranges of eight traits in the CC covered more than 95% of those in the primary CC, those of four traits between 90 and 95%, and those of the four remaining traits between 85 and 90%. The average phenotype range in the CC for all 16 traits covered 93.94% of those in the primary CC. Considering retention of both morphological and quantitative trait variation, the CC retained 93.12% of the variation in the primary CC, whereas the primary CC retained more than 95% of the corresponding variation in the entire Chinese rice germplasm resources (Li et al. 2003), that is, the CC retained about 88.47% of the morphological variation found in the entire rice germplasm resources of China.

Thus, the CC comprised a representative collection retaining most of the total variation, and the MCC provided a practical small collection enabling efficient utilization in most breeding activities and intensive research investigations. However, different sized and representative collections are often required by scientists and breeders with different purposes. We therefore established a hierarchical CC system, with different numbers of accessions (Fig. 6, see "Discussion" for details).

Discussion

Principles in determining core collection size

CC size is a key factor affecting the availability of the genetic variation in large germplasm collections. Several general guidelines were proposed for predicting the appropriate population size or sampling proportion. For example, Brown (1989a) suggested that 10% of the accessions of an entire major crop collection, or no more than 3,000 accessions, would be adequate. Yonezawa et al. (1995) thought that 20–30% of the accessions of a parental collection should be sampled under most circumstances. Different studies have suggested a wide range of CC sizes, from 0.3% to even 50% of parental collections, and from tens to 2,500 accessions in terms of absolute numbers (van Hintum et al. 2000). Clearly, a single size or a fixed proportion of parental collections are too empirical and not appropriate for all circumstances. We propose two principles that should be considered in determining the size of a CC. The first relates to sampling efficiency and aims to find a "point of compromise" between gain of genetic variation and elimination of genetic redundancy in the CC. The second relates to the sampling validity and aims to find a "point" at which the prevalent variation types can be preserved in the CC. In practice, these two principles should be balanced, but often only one or the other is emphasized by different users or for different purposes.

An MCC aims to define a collection that is easy to use in research and breeding. Here, the first principle, high sampling efficiency, was emphasized, and the sample size should be based on the transition point from the stage of rapid increase of genetic variation to the stage of both low increase of genetic variation and genetic redundancy (Fig. 3). The population size for cultivated rice at this point was 82 (Fig. 3). It is impossible to include all the variations, especially the rare alleles, in such small a collection, but the main alleles [here predominant alleles ($P > 0.1$) and common alleles ($0.1 \geq P > 0.01$)] at least should be retained. The population size meeting the second principle, to retain at least 95% of common alleles, was 189. We adopted the larger of the two estimated population sizes on the premise of high sampling efficiency (IRI $\sim$ 0.38) and manageable numbers. Thus, 189 accessions were recommended for the MCC of cultivated rice, accounting for 0.34% of the accessions in the basic collection (Table 1). The randomly sampled MCC retained about 70% of the alleles in the basic collection. For the CC, we emphasized the second principle and set a higher retention level, i.e., 95% of rare alleles ($0.01 \geq P > 0.001$), where the population size was 900. For the first principle, we recommended the point near transition from the stage of both low increase of genetic variation and genetic redundancy to the

stage of rapid increase of genetic redundancy (Fig. 3), where the population size was 932. As for the CC, we took the larger population size, i.e., 932 accessions, which accounted for 1.7 and 11.3% of the accessions in the basic collection (Table 2).

Simulation of the relationship between genetic variation and population size can be used to investigate the above two principles. Of course, it is impossible to do such a simulation for a very large collection with tens of thousands of accessions. In the case of the Chinese rice genetic resources, we adopted a two-stage sampling strategy. First, we established a primary CC, which was large enough to retain 90–95% of the variation in the basic collection (Li et al. 2003). In the present study, we conducted a simulation using the primary CC according to a mathematic model. If a primary CC was not established, our results showed that a population with 700–1,500 individuals (accessions in the case of a self-pollinated species) should be sufficient to do this simulation. Although the relationship between genetic variation and population size could also be simulated using the Crossa et al. (1993) allele frequency-based method, or our genotype frequency-based estimation when the allelic frequencies or genotype frequencies in the parental collections were known, both estimations were lower biased relative to real random sampling, especially when there was a higher degree of linkage disequilibrium between loci. Allele frequency-based estimation proved to be lower biased than genotype frequency-based estimation, especially when the heterozygosity was higher than 10%. More importantly, the frequency of alleles or genotypes in a large basic collection are usually not available, thus MMF-based estimation, i.e., simulation according to mathematical models (such as MMF in the present study), is reliable and feasible for investigating the relationship between allelic number and population size. This could be used for similar research on other genetic resources using 700–1,500 individuals.

Hierarchical core collections of Chinese cultivated rice

The concept of CC has been a controversial issue. Curators feared that basic collections would be overlooked when scientists focused on CC and were concerned about retention of all variations. On the other hand, breeders and researchers preferred rationally manageable numbers of accessions to achieve their germplasm objectives. Even for breeders, different sized CC may be required to satisfy their different requirements and manipulation capabilities. We therefore developed additional core sets consisting of different numbers of accessions, such as core sets C500, C1500 and C2000. In this system, all accessions in the smaller core set are included in the next bigger core set. To develop a representative and practical MCC, we integrated the optimal sampling scheme and advice from professional breeders, germplasm curators and scientists. First, we sampled the MCC by the optimal sampling scheme, VTP-PS-G-CR. Second, a five-person advisory group including professional breeders, germplasm curators and scientists checked each of the 189 clusters in the primary CC to decide if there was a distinctly elite cultivar, which had special phenotype(s) or genotype(s) and had been a main parent in breeding. If there was, we chose that cultivar; otherwise one cultivar of this cluster was randomly chosen. In developing the CC and other core sets (such as C500, C1500 and C2000), we chose cultivars in lower levels of core sets where possible, otherwise the choice was random. The hierarchical CC are shown in Fig. 6.

The hierarchical CC system, with different population scales and genetic diversities, will allow the use of genetic resources to be more flexible. The primary CC contains 95% of the variation in the cultivated rice resources in China, but is too large in scale (4,300 accessions) to be intensively used and studied. With its small size (189 accessions) and high genetic diversity (70% of the variation in the Chinese cultivated rice resources), the MCC provides us a rational framework to systematically survey the natural variation and to study complex traits. The MCC has been distributed to more than 60 research or breeding organizations and used in their research and breeding program. We have completed the genotyping of 300 genome-wide SSR markers and phenotyping of more than 50 traits at three locations (Beijing, Hangzhou, and Hainan) for 2 years. We are developing MCC-based introgression lines and isogenic lines, and the resequencing of MCC is also in progress. These efforts will help us to determine the extent of natural variation in cultivated rice and the molecular mechanisms underlying complex traits.

How to capture more rare alleles?

Allele numbers increase from MCC to primary CC, but the gain of new alleles rapidly decreases. For example, 15% additional alleles were retained in CC (932 accessions) relative to MCC (189 accessions), and more than 70% of them were rare alleles. However, only about 5% additional alleles were captured in core set with 2,000 accessions relative to CC (932 accessions), and 90% of those were inferior alleles (Fig. S3). If we further sample larger core sets, almost all the additional alleles will be inferior alleles. The question is whether it is necessary to sample a larger core set to gain a few inferior alleles. The answer is rather dubious. It is difficult to intensively investigate a large collection (larger than 2,000 accessions) in research and breeding activities. Furthermore, the majority of inferior alleles will not contribute to the genetic diversity needed to develop elite cultivars. Various studies (Allard 1992;

Frankel et al. 1995) have argued that less frequent alleles only occasionally affect quality or other traits and are generally unlikely to be of future value. Of course, the specific roles of SSRs are mostly unknown, and we could not distinguish whether the alleles examined here were subjected to direct selection per se, or to indirect selection through linkage with genes under selection. However, many studies have shown that SSR alleles are not always neutral or adaptively neutral, and that they are functionally significant in some instances (e.g., Zhang et al. 2009; and for review, Kashi and King 2006). Thus, most inferior alleles were inferior not only in frequency, but also in economic usefulness. Thus, it is not appropriate to establish CC of more than 2,000 accessions solely to increase the likelihood of retaining rare alleles.

Although it is impossible to get most of the rare and inferior alleles in a small collection, our results showed that certain factors can impact the efficiency of capturing rare and inferior alleles. If there are associations among alleles or genotypes, the sampling efficiency even for random sampling can be increased (Fig. 3), but the individuals in a cluster should be randomly sampled (Figs. 5, S2; Tables S4, S5). Brown (1978), Schoen and Brown (1993) and our results show that the localization and correlation of alleles are important factors in influencing sampling efficiency. Population structure and stratification are the main forms of localization and resources of allele correlation. In our case, the population structure-based grouping methods are distinctly better in capturing more alleles, especially the rare and inferior ones than other grouping methods (Figs. 5, S2; Tables S4, S5). If the accession numbers differ dramatically, other strategies such as the group-based L strategy (Spagnoletti and Qualset 1993; Brown 1989b) or S strategy (Li et al. 2002) should be used.

Acknowledgments This study was supported by the National Basic Research Program of China ("973" Program, 2010CB125904, 2004CB117201) and the National Natural Science Foundation of China (30600388, 30871506). We thank Professor Robert A McIntosh, University of Sydney, for suggested revisions to the manuscript, and Professor Qifa Zhang (Huazhong Agricultural University), Professor Jizeng Jia (Institute of Crop Science, Chinese Academy of Agricultural Sciences) and Professor Zhensheng Li (Institute of Genetics and Developmental Biology, Chinese Academy of Sciences) for their constructive advice on the development of the core collection and on the writing of this paper, and Prof Xiangkun Wang (China Agricultural University) and Prof Lifang Hong (Beijing Academy of Agricultural and Forestry Sciences) for their advice on sampling the mini core collection.

References

Allard RW (1992) Predictive methods for germplasm identification. In: Stalker HT, Murphy JP (eds) Plant breeding in the 1990's. CAB International, Wallingford, pp 119–146

Bataillon TM, David JL, Schoen DJ (1996) Neutral genetic markers and conservation genetics: simulated germplasm collections. Genetics 144:409–417

Bisht IS, Mahajan RK, Lokknathan TR, Agrawal RC (1998) Diversity in Indian sesame collection and stratification of germplasm accessions in different diversity groups. Genet Resour Crop Evol 45:325–335

Brown AHD (1978) Isozymes, plant population genetic structure and genetic conservation. Theor Appl Genet 52:145–157

Brown AHD (1989a) The case for core collections. In: Brown AHD, Frankel OH, Marshall DR, Williams JT (eds) The use of plant genetic resources. Cambridge University Press, Cambridge, pp 136–156

Brown AHD (1989b) Core collections: a practical approach to genetic resources management. Genome 31:818–824

Charmet G, Balfourier F (1995) The use of geostatistics for sampling a core collection of perennial ryegrass populations. Genet Resour Crop Evol 42:303–309

Crossa J, Hernandez M, Bretting P, Eberhart SA, Taba S (1993) Statistical genetic considerations for maintaining germplasm collections. Theor Appl Genet 86:637–678

Ellis PR, Pink DAC, Phelps K, Jukes PL, Breeds SE, Pinnegar AE (1998) Evaluation of a core collection of *Brassica oleracea* accessions for resistance to *Brevicoryne brassicae*, the cabbage aphid. Euphytica 103:149–160

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

FAO (1997) Report on the State of the World's Plant Genetic Resources for Food and Agriculture. Food and Agriculture Organization of the United Nations, Rome

Franco J, Crossa J, Warburton ML, Taba S (2006) Sampling strategies for conserving maize diversity when forming core subsets using genetic markers. Crop Sci 46:854–864

Frankel OH, Brown AHD (1984) Plant genetic resources today: a critical appraisal. In: Holden JHW, Williams JT (eds) Crop genetic resources: conservation and evaluation. Allen and Unwin, London, pp 249–257

Frankel OH, Brown AHD, Burdon JJ (1995) The conservation of plant biodiversity. Cambridge University Press, UK

Gouesnard B, Bataillon TM, Decoux G, Rozale C, Schoen DJ, David JL (2001) MSTRAT: an algorithm for building germplasm core collections by maximizing allelic or phenotypic richness. J Hered 92:93–94

Holbrook CC, Timper P, Xue HQ (2000) Evaluation of the core collection approach for identifying resistance to *Meloidogyne arenaria* in peanut. Crop Sci 40:1172–1175

Hyams D (2003) CurveExpert 1.3. A comprehensive curve fitting system for windows. http://curveexpert.webhop.net/

Kashi Y, King DG (2006) Simple sequence repeats as advantageous mutators in evolution. Trends Genet 22:253–259

Kim KW, Chung HK, Cho GT, Ma KH, Chandrabalan D, Gwag JG, Kim TS, Cho EG, Park YJ (2007) PowerCore: a program applying the advanced M strategy with a heuristic search for establishing core sets. Bioinformatics 23:2155–2162

Li ZC, Zhang HL, Zeng YW, Yang ZY, Shen SQ, Sun CQ, Wang XK (2002) Studies on sampling strategies for establishment of core collection of rice landrace in Yunnan, China. Genet Resour Crop Evol 49:67–74

Li ZC, Zhang HL, Cao YS, Qiu ZE, Wei XH, Tang SX, Yu P, Wang XK (2003) Studies on the sampling strategy for primary core collection of Chinese indigence. Acta Agron Sin 29:20–24

Liu K, Muse S (2004) PowerMarker: new genetic data analysis software, version 2.7. http://www.powermarker.net/

Malvar RA, Butron A, Alvarez A, Ordas B, Soengas P, Revilla P, Ordas A (2004) Evaluation of the European Union maize

landrace core collection for resistance to *Sesamia nonagrioides* (Lepidoptera: Noctuidae) and *Ostrinia nubilalis* (Lepidoptera: Crambidae). J Econ Entomol 97:628–634

Morgan PH, Mercer LP, Flodin NW (1975) A general model for nutritional responses of higher organisms. Proc Natl Acad Sci USA 72:4327–4331

Noirot M, Hamon S, Anthony F (1996) The principal component scoring: a new method of constituting a core collection using quantitative data. Genet Resour Crop Evol 43:1–6

Oka HI (1988) Origin of cultivated rice. Jnp. Sci. Soc. Press, Tokyo

Panaud O, Chen XL, McCouch SR (1996) Development of microsatellite markers and characterization of simple sequence length polymorphism (SSLP) in rice (*Oryza sativa* L.). Mol Gen Genet 252:597–607

Peeters JP, Martinelli JA (1989) Hierarchical cluster analysis as a tool to manage variation in germplasm collections. Theor Appl Genet 78:1142–1148

Qi YW (2007) Analysis of the genetic diversity of the rice cultivars (*Oryza sativa* L.) using SSR and HD1, HD3a gene. PhD Dissertation, China Agricultural University, Beijing

Saghai-Maroof MA, Soliman KM, Jorgensen RA, Allard RW (1984) Ribosomal DNA spacer length polymorphism in barley: mendelian inheritance, chromosomal location and population dynamics. Proc Natl Acad Sci USA 81:8014–8018

Schoen DJ, Brown AHD (1993) Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. Proc Natl Acad Sci USA 90:10623–10627

Spagnoletti PLZ, Qualset CO (1993) Evaluation of five strategies for obtaining a core subset from a large genetic resource collection of durum wheat. Theor Appl Genet 87:295–304

Tautz D, Renz M (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res 12:4127–4138

van Hintum ThJL, Brown AHD, Spillane C, Hodgkin T (2000) Core collections of plant genetic resources. IPGRI Technical Bulletin No. 3. International Plant Genetic Resources Institute, Rome, Italy

Wang MX, Zhang HL, Zhang DL, Qi YW, Fan ZL, Li DY, Pan DJ, Cao YS, Qiu ZE, Yu P, Yang QW, Wang XK, Li ZC (2008) Genetic structure of *Oryza rufipogon* Griff. in China. Heredity 101:527–535

Yonezawa K, Nomura T, Morishima H (1995) Sampling strategies for use in stratified germplasm collections. In: Hodgkin T, Brown AHD, van Hintum ThJL, Morales EAV (eds) Core collections of plant genetic resources. IPGRI, Wiley, Baffins Lane, pp 35–54

Zhang HL, Sun JL, Wang MX, Liao DQ, Zeng YW, Shen SQ, Yu P, Mu P, Wang XK, Li ZC (2007) Genetic structure and phylogeography of rice landraces in Yunnan, China revealed by SSR. Genome 51:72–83

Zhang DL, Zhang HL, Wang MX, Sun JL, Qi YW, Wang FM, Wei XH, Han LZ, Wang XK, Li ZC (2009) Genetic structure and differentiation of *Oryza sativa* L. in China revealed by microsatellites. Theor Appl Genet 119:1105–1117