Theor Appl Genet (2002) 104:1325–1334
DOI 10.1007/s00122-001-0854-4

S. Chandra · Z. Huaman · S. Hari Krishna · R. Ortiz

Optimal sampling strategy and core collection size of Andean tetraploid potato based on isozyme data – a simulation study

Received: 5 June 2001 / Accepted: 8 November 2001 / Published online: 26 April 2002

Abstract Selection of an appropriate sampling strategy is an important prerequisite to establish core collections of appropriate size in order to adequately represent the genetic spectrum and maximally capture the genetic diversity in available crop collections. We developed a simulation approach to identify an optimal sampling strategy and core-collection size, using isozyme data from a CIP germplasm collection on an Andean tetraploid potato. Five sampling strategies, constant (C), proportional (P), logarithmic (L), square-root (S) and random (R), were tested on isozyme data from 9,396 Andean tetraploid potato accessions characterized for nine isozyme loci having a total of 38 alleles. The 9,396 accessions, though comprising 2,379 morphologically distinct accessions, were found to represent 1,910 genetically distinct groups of accessions for the nine isozyme loci using a sort-and-duplicate-search algorithm. From each group, one accession was randomly selected to form a genetically refined entire collection (GREC) of size 1,910. The GREC was used to test the five sampling strategies. To assess the behavior of the results in repeated sampling, k = 1,500 and 5,000 independent random samples (without replacement) of admissible sizes n = 50(50)1,000 for each strategy were drawn from GREC. Allele frequencies (AF) for the 38 alleles and locus heterozygosity (LH) for the nine loci were estimated for each sample. The goodness of fit of samples AF and LH with those from GREC was tested using the χ^2 test. A core collection of size n = 600, selected using either the P or the R sampling strategy, was found adequately to represent the GREC for both AF and LH. As similar results were obtained at k = 1,500 and 5,000, it seems adequate to draw 1,500 independent random samples of different sizes to test the behavior of different sampling strategies in order to identify an appropriate sampling approach, as well as to determine an optimal core collection size.

Keywords Andean tetraploid potato · Core collection · Sampling strategy · Simulation

Communicated by H.C. Becker

S. Chandra · S. Hari Krishna
International Crops Research Institute
for the Semi-Arid Tropics (ICRISAT),
Patancheru, 502 324 AP, India

Z. Huaman
Centro Internacional de la Papa, Apartado 1558, Lima 100, Perú

R. Ortiz (✉)
International Institute of Tropical Agriculture (IITA),
c/o Lambourn & Co., Carolyn House, 26 Dingwall Road,
Croydon, CR9 3EE, UK
e-mail: r.ortiz@cgiar.org

Present address:
Z. Huaman, Pro Biodiversity of the Andes,
Av. Raul Ferrero # 1354, Lima 12, Perú

Introduction

Frankel and Brown (1984) proposed the concept of a core collection to enable efficient and cost-effective management and utilization of crop genetic resources. They defined a core collection as a limited subset of accessions from an existing germplasm collection that adequately represents the genetic spectrum of, and captures maximal genetic diversity in, a collection held in a genebank. An ideal core collection should include entries that are also ecologically and genetically distinct from one another (Brown 1989). A core collection that meets these requirements acts as a representative entry point to the whole collection in order to facilitate the processes of crop genetic improvement and research.

The gene bank at Centro Internacional de la Papa (CIP, Lima, Perú) maintains one of the largest collections of tetrasomic Andean potatoes (*Solanum tuberosum* subsp. *andigena*) (Huaman 1998). These accessions have been characterized both at morphological and genetic levels, the latter using isozyme markers. Isozymes markers in Andean potatoes have been employed for assessing genetic variation (Zimmerer and Douches 1991; Quiros et al. 1992), determining rates of out-crossing be-

tween primitive cultivated potatoes (Rabinowitz et al. 1990), characterizing North American tetraploid potato cultivars (Douches et al. 1991; Douches and Ludlam 1991) and for determining how human selection affects genetic diversity in tetraploid potatoes (Ortiz and Huaman 2001). Some of these isozymes are associated with important agronomic characters in potato-segregating populations (Ortiz et al. 1993; Freyre and Douches 1994; Freyre et al. 1994).

Recently, Huaman et al. (2000a) developed a core collection of 306 Andean tetraploid potatoes from a subset of morphologically distinct 2,379 accessions. The latter were selected from an existing whole collection of 10,722 accessions held in the CIP gene-bank after removing from it 8,343 duplicate accessions based on several morphological traits. A square-root sampling approach was used to select the core of 306 entries from each geographical division of Latin American countries, from which the 2,379 accessions were collected. Data on nine isozyme markers were subsequently used to investigate the genetic structure of the 2,379 accessions and to assess the genetic representativeness of the core of 306 entries in terms of allele frequencies and locus heterozygosity (Huaman et al. 2000b).

The objective of this research was to develop and apply a simulation approach to determine an optimal sampling strategy and core collection size for Andean tetraploid potato accessions using only the isozyme data.

Materials and methods

Genetic materials

The original Andean tetraploid collection at CIP consisted of 10,722 accessions from eight Latin American countries. Of these, only 9,396 accessions, characterized for morphology and nine isozymes, were included in this study. These 9,396 accessions represented 2,379 morphologically distinct genotypes (Huaman et al. 2000b).

Genetic markers

Allozyme diversity was determined using horizontal gel-electrophoresis and two buffer systems. The procedures for tissue processing, electrophoresis, gel staining and allozyme scoring were those of Douches and Quiros (1988) and Huaman et al. (2000b). These nine isozyme loci covering a total of 38 alleles were: isocitric acid dehydrogenase 1 (*Idh-1* in chromosome I), malate dehydrogenase 1 (*Mdh-1* in chromosome VII), malate dehydrogenase 2 (*Mdh-2*), and phosphoglucose isomerase 1 (*Pgi-1* in chromosome XII) for histidine-citrate at pH 5.7, and Diaphorase 1 (*Dia-1* in chromosome V), glutamate oxaloacetate transaminase 1 (*Got-1* in chromosome VIII), glutamate oxaloacetate transaminase 2 (*Got-2* in chromosome VII), phosphoglucomutase 1 (*Pgm-1* in chromosome III), and phosphoglucomutase 2 (*Pgm-2* in chromosome IV) for tris-borate at pH 8.3.

Creation of a genetically refined entire collection

The data consist of counts Yijk of allele $k = 1, \ldots, a_j \in [3,6]$ at locus j = 1, …, n_l (= 9) for accession i = 1, …, *N* (= 9,396) with the property $\sum_k Y_{ijk} = 4 \forall (i,j)$. The allele counts Y_{ijk} were transformed to allele frequencies $P_{ijk} = (1/4)Y_{ijk}$ with $\sum_k P_{ijk} = 1 \forall (i,j)$.

A sort-and-duplicate-search algorithm found the *N* accessions to fall into K = 1,910 distinct allelic-configuration/genotype classes, with $N_t \in [1, 198]$ duplicate genotypes present in class t = 1, …, K, $\sum_t N_t = N$. The original entire collection (OEC) of *N* accessions was therefore first reduced to a genetically refined entire collection (GREC) of K = 1,910 distinct tetraploid genotypes by randomly selecting one accession from each of the K genotype classes. The GREC, rather than the OEC, was used to investigate the suitability of different sampling strategies and to determine the optimal core collection size. Use of GREC ensures that the core contains genetically distinct entries.

Sampling strategies

Five sampling strategies were investigated, random (R), constant (C), proportional (P), logarithmic (L) and square root (S). For the R strategy, accessions were randomly selected from the GREC using simple random sampling without replacement (SRSWOR), in keeping with the fact that a core should include distinct entries. For C, P, L and S strategies, the 1,910 accessions in the GREC were first grouped into eight clusters according to the country of their collection as follows: Argentina 73, Bolivia 258, Colombia 105, Ecuador 131, Guatemala 24, Mexico 16, Peru 1,276 and Venezuela 27. From each of these eight clusters, the number of accessions n_u (u = 1, …, 8), to obtain a specified core sample of $n = \sum_u n_u$ accessions, was selected using intra-cluster SRSWOR as follows:

Strategy	Intra-cluster sample-size n_u	Admissible/tested core sample size *n*
C	$n_u = n/8$	*n* = 50(50)150*
L	$n_u = n[\log(K_u)/\sum_u \log(K_u)]$	*n* = 50(50)250
P	$n_u = K_u\,(n/K)$	*n* = 50(50)1,000
S	$n_u = n[\sqrt{K_u}/\sum_u \sqrt{K_u}]$	*n* = 50(50)400
R	–	*n* = 50(50)1,000

K_u = size of cluster u; *sample sizes varied between 50 and 150 with an increment of 50

Estimation of allele frequencies and locus heterozygosity

The allele frequencies (AF) P_{jk} for allele k at locus j, and locus heterozygosity (LH) H_j for locus j in the GREC were computed as follows:

$$P_{jk} = (a_{jk}/a_j) = \left(\sum_t a_{ijk}/\sum_t \sum_k a_{ijk}\right) = \left[\sum_t a_{ijk}/(4 \times K)\right] \forall j \quad (1)$$

$$H_j = \left\{1 - \left[\sum_k \#(P_{jk} = 1)/K\right]\right\}, \quad (2)$$

where a_{jk}: total count of the k-th-type allele at locus j across K genotypes, a_j: total count of all allele types at locus j across K genotypes, a_{ijk}: count of the k-th-type allele at locus j for genotype t = 1, …, K, $\#(P_{jk} = 1)$: number of genotypes homozygous for allele k at locus j.

Sample estimates p_{jk} and h_j of P_{jk} and H_j respectively for a sample of size *n* drawn from using any sampling strategy were obtained from equations (1) and (2) respectively, with K replaced by the sample size *n*.

Chi-square tests of goodness-of-fit

Goodness-of-fit of the sample estimates to the population values was tested using χ^2 tests as follows:

$$\chi^2(\mathrm{AF}) = \sum_j \chi_j^2(\mathrm{AF}) df = \sum_j (a_j - 1) = 28$$

at a level of signifi-cance (LOS) α for across-the-loci (genome-wide) fit of AF (H_0:p_{jk}=P_{jk}, j=1,…n_l, k=1,…, a_j),

$$\chi_j^2(\mathrm{AF}) = 4n\sum_k\left[(p_{jk}-P_{jk})^2P_{jk}\right] k = 1,\ldots,a_j \; df = a_j - 1$$

at LOS $\alpha_j = \alpha/8$ for an individual locus-wise fit of AF (H_0:$p_{(j)k}$ = $P_{(j)k}$, k = 1, …, a_j), $\chi^2(\mathrm{LH}) = \sum_j \chi_j^2(\mathrm{LH}) df = n_l = 9$ at LOS α for across-the-loci (genome-wide) fit of LH (H_0:h_j = H_j, j = 1, …, 9), and $\chi_j^2(\mathrm{LH}) = (h_j - H_j)^2[H_j(1-H_j)/n] df = 1$ at LOS $\alpha_j = \alpha/8$ for an individual locus-wise fit of LH (H_0:h_j = H_j).

Simulations

Inferences based on just one sample of a particular size *n* could be misleading as this does not give an idea of the likely variation in the results had we drawn more samples of that size. Repeated samples provide an objective assessment of the degree of consistency, stability and reproducibility of results. Therefore, k = 1,500 and 5,000 independent random samples of a particular size *n* = 50(50)1,000, as admissible for a given sampling strategy, were drawn according to the afore-stated five sampling strategies. Two values of k were chosen to determine the adequate number of random samples to be simulated.

A sample size and a strategy that consistently do not reject H_0 at a chosen level of significance α across all k repeated samples are the safest sample size and strategy to use. This is practically unlikely to happen as long as $n < N$. However, for a given sampling strategy, a sample size *n* for which, under H_0, the k-observed χ^2-values follow the corresponding theoretical χ^2 distribution, provides a lower bound, if that exists, on optimal sample size. We used the Kolmogorov-Smirnov (K-S) test (Sokal and Rohlf 1981) to identify this lower bound on the optimal sample size for each sampling strategy. Having identified the lower bound on an optimal *n*, the optimal *n* for a given sampling strategy can be determined from a suitably chosen characteristic of the frequency distribution of the k-observed χ^2-values. Some possible candidate-characteristics are the maximum, upper-0.05-quantile, and a median of the observed distribution of the k values of χ^2. The maximum is obviously the safest to use as it covers the maximum possible risk in terms of the largest possible discrepancy between GREC and sample values. However, since χ^2 can theoretically assume a maximum value of infinity, it is likely that, with increasing *n*, the observed maximum χ^2 values may show an erratic pattern, which they did, (see Tables 1 and 2). that This situation will make it difficult to clearly identify an optimal sample size and strategy. Use of the median, compared to using the observed upper-0.05-quantile χ^2, on the other hand, covers much-less risk. We therefore chose to use the upper-0.05-quantile of the observed distribution of k χ^2-values to judge the suitability of a sample size and strategy. Any upper-0.05-quantile χ^2-value that is non-significant at a chosen level of significance α implies that, for the corresponding sample size and strategy, all samples of that size will consistently deliver non-significant χ^2 values 95% of the time, and hence provide a good fit to the GREC. Also, the more the *P*-value of the observed upper-0.05-quantile χ^2-value exceeds the specified α, less is the discrepancy between GREC and sample values. From this perspective, one could choose an α-value more than the conven-

Table 1 Quantiles of 1,500 observed χ^2 values for allele frequencies for different sample sizes (*n*) under proportional strategy

n	Min	0.95-uq[a]	0.75-uq	0.50-uq	0.25-uq	0.05-uq	Max	D[b]
50	12.70	21.24	28.51	35.32	44.45	78.58	132.98	0.3603
	0.9941[c]	*0.8152*	*0.4377*	*0.1607*	*0.0251*	*0.0000*	*0.0000*	
100	12.24	21.94	28.90	35.40	43.60	62.66	129.76	0.3583
	0.9957	*0.7841*	*0.4176*	*0.1585*	*0.0304*	*0.0002*	*0.0000*	
150	12.30	20.94	28.20	34.86	42.48	56.49	108.00	0.3421
	0.9955	*0.8278*	*0.4539*	*0.1740*	*0.0390*	*0.0011*	*0.0000*	
200	10.16	20.40	28.00	34.48	42.68	55.60	92.96	0.3296
	0.9992	*0.8494*	*0.4644*	*0.1855*	*0.0374*	*0.0014*	*0.0000*	
250	11.60	20.20	27.60	34.00	40.90	52.45	75.50	0.3085
	0.9973	*0.8571*	*0.4858*	*0.2009*	*0.0548*	*0.0034*	*0.0000*	
300	9.60	20.04	27.00	32.52	39.06	51.12	80.52	0.2598
	0.9995	*0.8630*	*0.5182*	*0.2539*	*0.0800*	*0.0048*	*0.0000*	
350	10.22	19.88	26.32	32.06	38.85	49.84	76.02	0.2355
	0.9991	*0.8688*	*0.5555*	*0.2721*	*0.0834*	*0.0067*	*0.0000*	
400	8.64	19.04	25.60	30.88	36.96	47.52	74.08	0.1837
	0.9998	*0.8969*	*0.5950*	*0.3224*	*0.1197*	*0.0121*	*0.0000*	
450	10.44	18.36	24.48	29.88	35.82	46.26	75.60	0.1399
	0.9990	*0.9167*	*0.6560*	*0.3690*	*0.1472*	*0.0164*	*0.0000*	
500	9.60	17.80	23.80	29.00	35.20	44.00	69.40	0.0983
	0.9995	*0.9311*	*0.6920*	*0.4125*	*0.1641*	*0.0278*	*0.0000*	
550	10.34	17.16	22.88	27.72	33.66	42.68	81.40	0.0420
	0.9990	*0.9454*	*0.7390*	*0.4794*	*0.2123*	*0.0374*	*0.0000*	
600	8.88	16.56	22.08	26.88	31.92	41.04	60.24	0.0403
	0.9998	*0.9568*	*0.7776*	*0.5248*	*0.2778*	*0.0533*	*0.0004*	
650	9.36	16.12	21.58	26.00	30.68	39.26	62.66	0.0952
	0.9996	*0.9640*	*0.8004*	*0.5730*	*0.3315*	*0.0768*	*0.0002*	
700	8.68	15.68	20.72	24.78	29.96	38.08	61.88	0.1555
	0.9998	*0.9703*	*0.8368*	*0.6398*	*0.3651*	*0.0969*	*0.0002*	
750	7.50	15.30	20.10	24.00	28.80	36.90	48.90	0.2020
	1.0000	*0.9751*	*0.8608*	*0.6815*	*0.4227*	*0.1211*	*0.0086*	
800	9.60	14.24	18.88	23.04	27.52	35.04	52.16	0.2667
	0.9995	*0.9854*	*0.9018*	*0.7310*	*0.4901*	*0.1687*	*0.0037*	

[a] Upper quantile (uq)

[b] Kolmogorov-Smirnov test-statistic value ($D_{0.05}$ = 0.035, $D_{0.01}$ = 0.042 based on k = 1,500)

[c] *P*-value (*italics*) of the above observed χ^2 values

tional values of 0.05 and 0.01 to further minimize the risk of picking up an inappropriate sample size and strategy. The *P*-values corresponding to the observed upper-0.05-quantile χ^2-values, summarized in a tabular or graphical form, provide an objective probabilistic basis to compare the suitability of different sampling strategies to help determine the optimal sample size and strategy, with α chosen according to the risk one wants to cover.

Our strategy in determining an optimal sample size for a given sampling strategy was to adopt the approach of the preceding paragraph to first check the overall genome-wide fit. Having identified the genome-wide optimal sample size for a chosen α, the suitability of that sample size at individual loci was determined using a locus-wise level of significance $\alpha_j = \alpha/n_L$ based on the Bonferroni correction, where the denominator $n_L \leq n_l$ represents the number of independent linkage groups on which the n_l isozyme loci are located.

An optimal sampling strategy is defined as one that, for the observed upper-0.05-quantile χ^2, provides a smaller genome-wide optimal sample size with a *P*-value ≥ to the chosen level of significance, α. It is anticipated that, due to the difference in the way AF and LH are estimated, different sample sizes may turn out to be optimal for AF and LH. We took the larger of the two optimal sample sizes as the optimal sample size for both AF and LH.

Results

The results were similar for k = 1,500 and 5,000. Accordingly, we will subsequently report k = 1,500 in presenting and discussing the results. The K-S test showed that, for the admissible values of *n*, a lower bound on optimal *n* did not exist for the C, L and S strategies. At the same time, for all three strategies, the *P*-value corresponding to the upper-0.05-quantile χ^2 (AF) and the upper-0.05-quantile χ^2 (LH) never exceeded $\alpha = 0.05$ for any of the sample sizes. These three strategies, regarded as non-optimal because of the above reasons, are therefore not discussed further.

Allele frequencies

The genome-wide freuency distribution of the 1,500 observed χ^2(AF)-values for the P and R strategies for different sample sizes *n* are summarized in Tables 1 and 2 respectively. As expected from the law of large numbers, the χ^2 values show a generally decreasing trend as the sample size *n* increases. Figure 1 depicts the observed upper-0.05-quantile χ^2(AF)-values and their corresponding *P*-values for the P and R strategies. Figure 2 provides the values of the upper-0.05-quantile χ^2_j(AF)-values and their corresponding *P*-values for individual loci. The K-S test-statistic for the P-strategy (Table 1) is non-significant (at $\alpha = 0.01$) at *n* = 550,

Table 2 Quantiles of 1,500 observed χ^2 values for allele frequencies for different sample sizes (*n*) under random strategy

n	Min	0.95-uq[a]	0.75-uq	0.50-uq	0.25-uq	0.05-uq	Max	D[b]
50	12.96	21.81	29.39	36.52	45.61	66.23	169.92	0.3951
	0.9931[c]	*0.7900*	*0.3930*	*0.1298*	*0.0191*	*0.0001*	*0.0000*	
100	12.16	21.48	29.32	36.10	44.96	64.82	108.92	0.3842
	0.9959	*0.8048*	*0.3964*	*0.1401*	*0.0223*	*0.0001*	*0.0000*	
150	10.32	20.91	29.25	36.30	44.58	59.79	100.32	0.3799
	0.9991	*0.8291*	*0.3999*	*0.1351*	*0.0243*	*0.0004*	*0.0000*	
200	12.88	21.00	29.16	35.76	44.00	57.84	103.92	0.3811
	0.9934	*0.8253*	*0.4044*	*0.1488*	*0.0278*	*0.0008*	*0.0000*	
250	13.30	20.30	28.65	35.20	42.05	54.95	83.90	0.3535
	0.9915	*0.8533*	*0.4304*	*0.1641*	*0.0429*	*0.0017*	*0.0000*	
300	12.24	20.28	27.36	33.72	41.16	51.96	88.44	0.2918
	0.9957	*0.8540*	*0.4987*	*0.2102*	*0.0519*	*0.0039*	*0.0000*	
350	13.02	20.09	26.88	32.62	39.06	50.82	82.88	0.2607
	0.9928	*0.8612*	*0.5248*	*0.2500*	*0.0800*	*0.0052*	*0.0000*	
400	11.20	19.84	26.16	32.00	38.88	49.60	78.56	0.2339
	0.9980	*0.8702*	*0.5643*	*0.2745*	*0.0829*	*0.0072*	*0.0000*	
450	11.52	18.72	24.84	29.88	36.27	47.52	70.56	0.1419
	0.9975	*0.9066*	*0.6365*	*0.3690*	*0.1359*	*0.0121*	*0.0000*	
500	10.60	18.20	24.20	29.40	35.60	45.50	65.80	0.1182
	0.9988	*0.9210*	*0.6709*	*0.3925*	*0.1531*	*0.0196*	*0.0001*	
550	8.36	17.38	23.10	27.94	33.88	43.34	64.46	0.0616
	0.9999	*0.9407*	*0.7280*	*0.4676*	*0.2049*	*0.0323*	*0.0001*	
600	5.52	16.56	22.80	27.60	33.24	42.48	68.64	0.0415
	1.0000	*0.9568*	*0.7430*	*0.4858*	*0.2270*	*0.0390*	*0.0000*	
650	10.14	15.86	21.58	26.52	31.72	40.43	67.60	0.0681
	0.9992	*0.9678*	*0.8004*	*0.5445*	*0.2861*	*0.0605*	*0.0000*	
700	7.28	15.96	20.72	25.20	30.24	38.64	55.44	0.1494
	1.0000	*0.9664*	*0.8368*	*0.6169*	*0.3518*	*0.0869*	*0.0015*	
750	7.20	15.60	20.10	24.60	29.10	37.50	58.80	0.1743
	1.0000	*0.9714*	*0.8608*	*0.6495*	*0.4075*	*0.1082*	*0.0006*	
800	9.60	14.40	19.52	23.68	27.84	36.48	57.92	0.2452
	0.9995	*0.9841*	*0.8813*	*0.6983*	*0.4730*	*0.1308*	*0.0007*	

[a] Upper quantile (uq)

[b] Kolmogorov-Smirnov test-statistic value ($D_{0.05} = 0.035$, $D_{0.01} = 0.042$ based on k = 1,500)

[c] *P*-value (*italics*) of the above observed χ^2 values

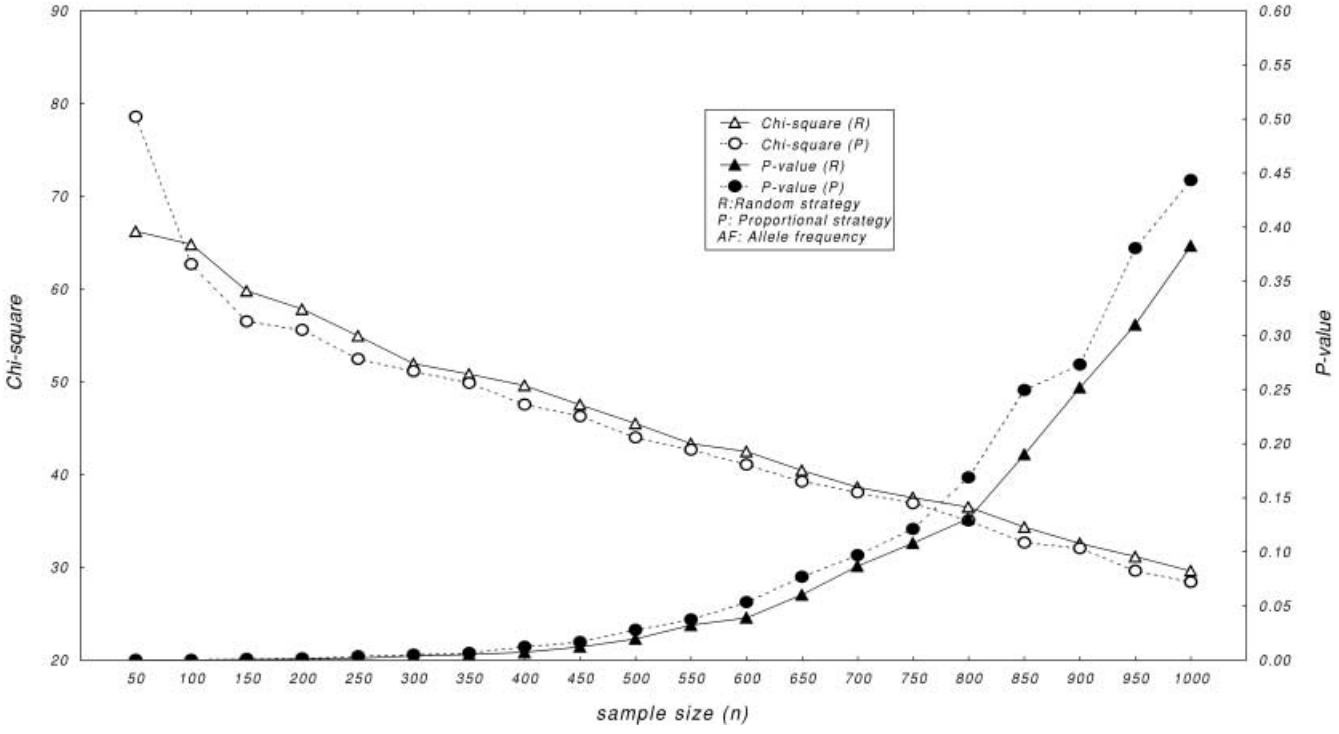

Fig. 1 Observed upper-0.05-quantile χ^2 and *P*-values for allele frequency under proportional and random strategies

which serves as a lower bound on an optimal *n* under the P strategy. However, for $\alpha = 0.05$, the corresponding upper-0.05-quantile χ^2 is significant at $n = 550$ having a *P*-value of 0.0374. At $n = 600$, the *P*-value (=0.0533) of the upper-0.05-quantile χ^2 exceeds $\alpha = 0.05$. At $\alpha = 0.05$ and $n = 600$, for each individual locus the *P*-value always exceeds $\alpha_j = 0.05/8 = 0.00625$ (Fig. 2). Therefore, $n = 600$ is the optimal *n* under the P strategy at $\alpha = 0.05$. For more risk to be covered by choosing say, e.g. $\alpha = 0.10$, the optimal *n* needs to be about 750 (Table 1, Fig. 1). Results for the R-strategy (Table 2, Figs. 1, 2) were similar to that of the P strategy with the difference that the K-S test-statistic was non-significant (at $\alpha = 0.01$) at $n = 600$, with $n = 650$ being the optimal *n*, which relative to the P strategy exceeds it by 50.

Locus heterozygosity

Tables 3 and 4 list the genome-wide frequency distributions of the 1,500 observed χ^2(LH)-values for the P and R strategies respectively. Figure 3 depicts the observed upper-0.05-quantile χ^2(LH)-values and their corresponding *P*-values for the P and R strategies. Figure 4 shows the values of the upper-0.05-quantile χ^2_j(LH)-values and their corresponding *P*-values for individual loci. For the P strategy, the K-S test-statistic was always significant (at $\alpha = 0.05$) for all sample sizes *n*. However, for all *n*, the *P*-value corresponding to the upper-0.05-quantile χ^2 was always greater than $\alpha = 0.05$. Thus $n = 50$ could be taken as the minimum sample size for a genome-wide fit at $\alpha = 0.05$. Also, for $\alpha = 0.05$ at $n = 50$, Fig. 4 shows that for each individual locus the *P*-value always much exceeded $\alpha_j = 0.05/8 = 0.00625$. Therefore, $n = 50$ is the optimal *n* under the P strategy at $\alpha = 0.05$. For the R strategy (Table 4), the K-S test identified $n = 50$ as the lower bound on an optimal *n*, this *n* also being the optimal *n* as the *P*-value (= 0.0582) corresponding to the upper-0.05-quantile χ^2 exceeded $\alpha = 0.05$. The locus-wise results for the R-strategy (Figs. 3, 4) were similar to that of the P strategy. Accordingly, $n = 50$ is also the optimal *n* for the R strategy at $\alpha = 0.05$.

Optimal sampling strategy and core collection size

Results from the preceding two paragraphs indicate that, for AF and LH considered simultaneously, there is little difference in performance of the P and R strategy, with P performing slightly better than R. A core collection size of about 600 entries selected using either the P or the R strategy is optimal to adequately represent the genetic spectrum of, and to maximally capture the genetic diversity (in terms of LH) in, the GREC.

As evident from the results reported above, LH requires a much-smaller optimal sample size than AF. An optimal sample size chosen solely on the basis of LH is, therefore, not likely to adequately represent the genetic spectrum of the population. A safer approach in arriving at an optimal sample size therefore seems to be to consider the (larger) optimal sample size for AF as the optimal sample size.

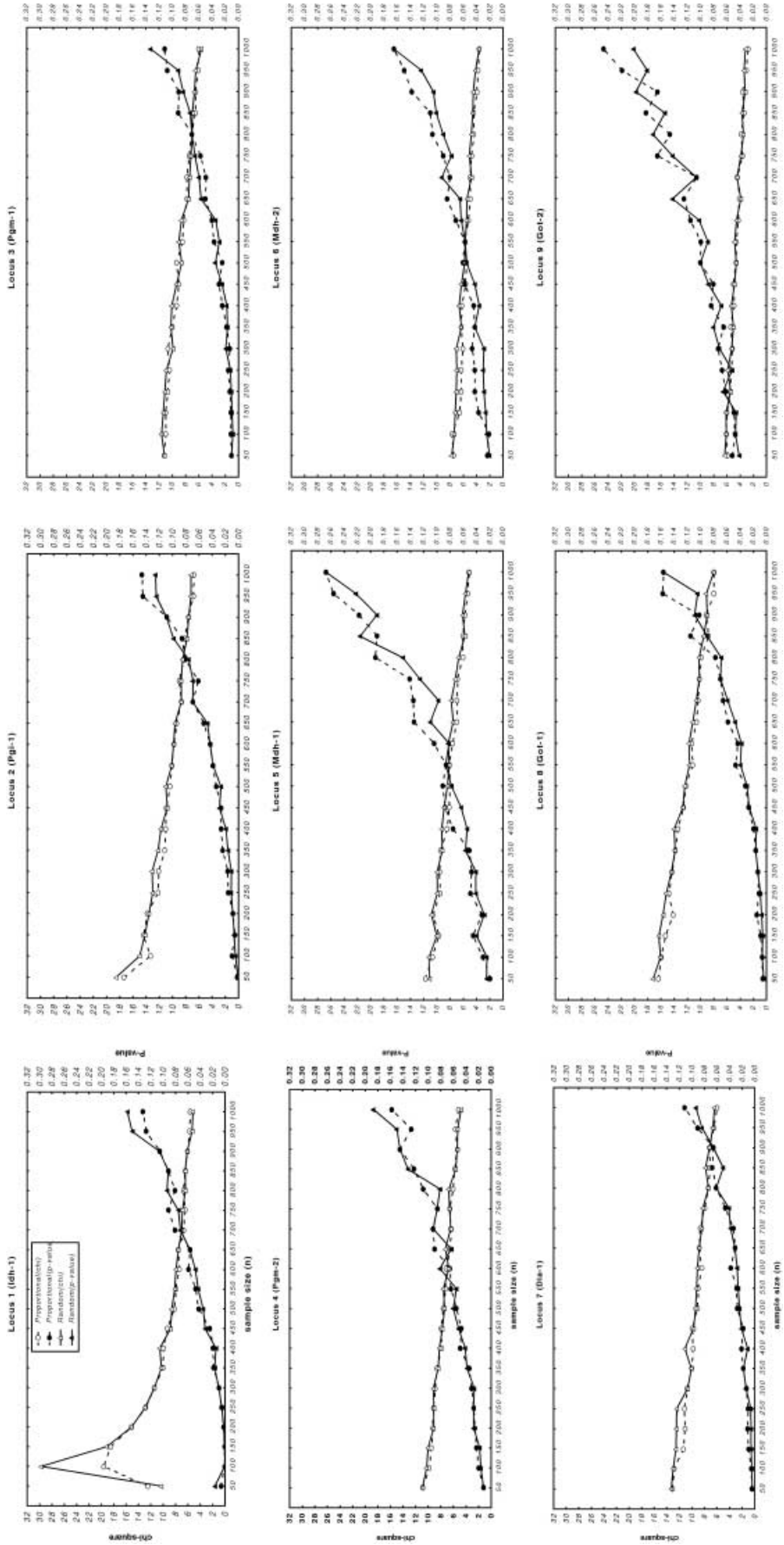

Fig. 2 Locus-wise observed upper-0.05-quantile χ^2 and *P*-values for AF under proportional (P) and random (R) strategies

Table 3 Quantiles of 1,500 observed χ^2 values for locus heterozygosity for different sample sizes (n) under proportional strategy

n	Min	0.95-uq[a]	0.75-uq	0.50-uq	0.25-uq	0.05-uq	Max	D[b]
50	1.36	3.29	5.60	7.98	10.92	15.69	30.78	0.0485
	0.9980[c]	*0.9516*	*0.7794*	*0.5360*	*0.2810*	*0.0737*	*0.0003*	
100	1.22	2.92	5.36	7.77	10.64	15.54	30.61	0.0685
	0.9988	*0.9674*	*0.8018*	*0.5579*	*0.3011*	*0.0771*	*0.0003*	
150	1.08	3.05	5.43	7.54	10.31	16.11	24.95	0.0915
	0.9992	*0.9621*	*0.7958*	*0.5811*	*0.3261*	*0.0646*	*0.0030*	
200	0.72	2.83	5.21	7.50	10.20	14.89	26.58	0.1033
	0.9999	*0.9705*	*0.8155*	*0.5848*	*0.3348*	*0.0941*	*0.0016*	
250	0.95	2.98	5.14	7.13	9.85	14.41	32.20	0.1311
	0.9995	*0.965*	*0.8216*	*0.6231*	*0.3631*	*0.1083*	*0.0002*	
300	0.38	2.74	4.99	7.05	9.59	13.80	25.57	0.1429
	1.0000	*0.9736*	*0.8355*	*0.6318*	*0.385*	*0.1294*	*0.0024*	
350	0.87	2.72	4.77	6.78	9.39	14.16	26.36	0.1673
	0.9997	*0.9744*	*0.8542*	*0.6598*	*0.4025*	*0.1168*	*0.0018*	
400	0.95	2.47	4.72	6.93	9.51	13.81	21.33	0.1565
	0.9995	*0.9817*	*0.8577*	*0.6443*	*0.3915*	*0.1292*	*0.0113*	
450	0.93	2.55	4.69	6.51	8.94	13.25	21.72	0.2058
	0.9996	*0.9794*	*0.8607*	*0.6877*	*0.4424*	*0.1518*	*0.0098*	
500	0.61	2.36	4.56	6.28	8.65	13.00	25.70	0.2297
	0.9999	*0.9843*	*0.8705*	*0.7119*	*0.4706*	*0.1628*	*0.0023*	
550	0.61	2.45	4.34	6.20	8.35	12.76	20.08	0.2528
	0.9999	*0.9822*	*0.8878*	*0.7193*	*0.499*	*0.1738*	*0.0174*	
600	0.64	2.24	4.00	5.71	8.07	12.12	24.11	0.2946
	0.9999	*0.9871*	*0.9115*	*0.7686*	*0.5274*	*0.2069*	*0.0041*	
650	0.63	2.27	4.01	5.79	8.11	11.26	21.17	0.2848
	0.9999	*0.9865*	*0.9106*	*0.761*	*0.5229*	*0.2581*	*0.0119*	
700	0.35	2.05	3.83	5.54	7.67	11.79	19.85	0.3250
	1.0000	*0.9907*	*0.9219*	*0.7848*	*0.5679*	*0.2254*	*0.0189*	
750	0.51	2.22	3.77	5.26	7.29	10.61	17.35	0.3616
	1.0000	*0.9874*	*0.9259*	*0.811*	*0.6071*	*0.3035*	*0.0435*	
800	0.88	2.16	3.75	5.30	7.13	10.82	19.21	0.3799
	0.9997	*0.9886*	*0.9273*	*0.8077*	*0.6236*	*0.2882*	*0.0235*	

[a] Upper quantile (uq)

[b] Kolmogorov-Smirnov test-statistic value ($D_{0.05}$ = 0.035, $D_{0.01}$ = 0.042 based on k = 1,500)

[c] *P*-value (*italics*) of the above observed χ^2 values

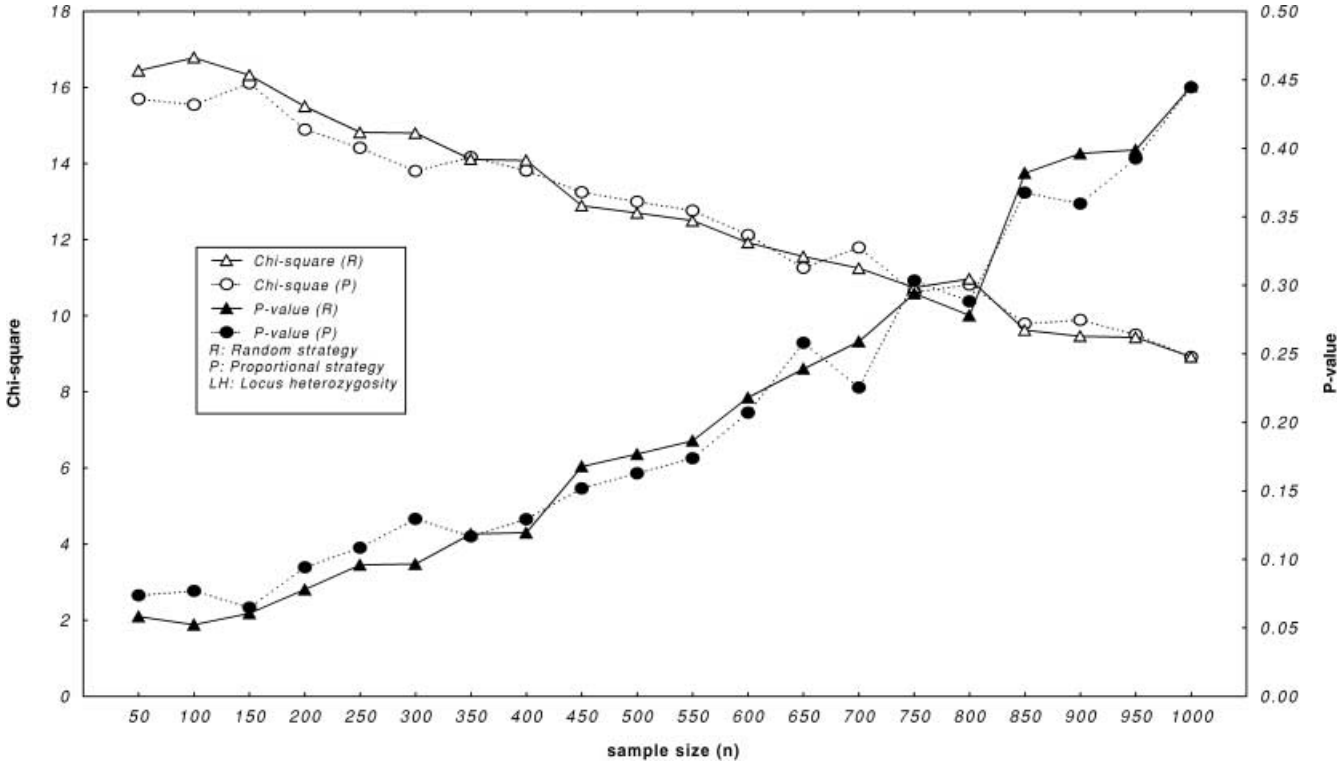

Fig. 3 Observed upper-0.05-quantile χ^2 and *P*-values for locus heterozygosity under proportional and random strategies

Table 4 Quantiles of 1,500 observed χ^2 values for locus heterozygosity for different sample sizes (n) under random strategy

n	Min	.95-uq[a]	0.75-uq	0.50-uq	0.25-uq	0.05-uq	Max	D[b]
50	0.85	3.17	5.88	8.36	11.16	16.44	30.53	0.0226
	0.9997[c]	*0.9572*	*0.7515*	*0.4988*	*0.2651*	*0.0582*	*0.0004*	
100	0.74	3.00	5.57	7.95	10.86	16.78	29.62	0.0609
	0.9998	*0.9643*	*0.7821*	*0.5392*	*0.2856*	*0.0523*	*0.0005*	
150	0.86	3.20	5.53	7.86	10.42	16.32	29.44	0.0726
	0.9997	*0.9559*	*0.7863*	*0.5485*	*0.3177*	*0.0605*	*0.0005*	
200	0.97	2.79	5.29	7.50	10.21	15.50	28.81	0.1050
	0.9995	*0.9722*	*0.8085*	*0.5854*	*0.3335*	*0.0780*	*0.0007*	
250	0.83	2.97	5.10	7.40	10.12	14.82	24.11	0.1120
	0.9997	*0.9653*	*0.8259*	*0.5953*	*0.3407*	*0.0959*	*0.0041*	
300	0.98	2.82	5.03	7.08	9.99	14.8	25.21	0.1384
	0.9995	*0.9711*	*0.8319*	*0.6286*	*0.3516*	*0.0966*	*0.0027*	
350	0.98	2.71	4.92	6.92	9.53	14.11	27.61	0.1618
	0.9995	*0.9745*	*0.8413*	*0.6457*	*0.3895*	*0.1185*	*0.0011*	
400	0.87	2.74	4.90	6.91	9.58	14.08	31.87	0.1568
	0.9997	*0.9737*	*0.8430*	*0.6464*	*0.3860*	*0.1194*	*0.0002*	
450	1.00	2.60	4.51	6.37	8.66	12.89	27.19	0.2333
	0.9994	*0.9780*	*0.8744*	*0.7028*	*0.4689*	*0.1677*	*0.0013*	
500	0.87	2.50	4.49	6.31	8.60	12.70	23.47	0.2318
	0.9997	*0.9808*	*0.8767*	*0.7085*	*0.4748*	*0.1768*	*0.0052*	
550	0.99	2.39	4.35	6.29	8.45	12.50	27.9	0.2560
	0.9995	*0.9837*	*0.8872*	*0.7107*	*0.4895*	*0.1865*	*0.001*	
600	0.81	2.26	4.16	5.86	8.14	11.92	24.09	0.2933
	0.9998	*0.9866*	*0.9007*	*0.754*	*0.5202*	*0.2181*	*0.0042*	
650	0.91	2.42	4.07	5.74	7.93	11.56	21.03	0.3071
	0.9996	*0.983*	*0.9067*	*0.7653*	*0.5413*	*0.2390*	*0.0125*	
700	0.83	2.26	3.91	5.48	7.54	11.25	20.67	0.3338
	0.9997	*0.9866*	*0.9170*	*0.7906*	*0.5812*	*0.2589*	*0.0142*	
750	0.65	2.23	3.78	5.48	7.52	10.74	18.24	0.3394
	0.9999	*0.9872*	*0.9251*	*0.7907*	*0.5834*	*0.2941*	*0.0325*	
800	0.27	2.18	3.71	5.22	7.19	10.97	18.98	0.3739
	1.0000	*0.9883*	*0.9295*	*0.8149*	*0.6176*	*0.2780*	*0.0254*	

[a] Upper quantile (uq)

[b] Kolmogorov-Smirnov test-statistic value ($D_{0.05}$ = 0.035, $D_{0.01}$ = 0.042 based on k = 1,500)

[c] *P*-value (*italics*) of the above observed χ^2 values

Discussion

Brown (1989) used the sampling theory of Ewens (1972) to propose a fraction *n*textsubscriptr/N_e = 0.10 as an optimal sampling fraction for randomly sampling n_r core entries from a germplasm collection of effective population size N_e. By doing so, Brown expected that at least 70% of existent alleles could be retained with 95% certainty. Ewens' sampling theory assumed that the finite germplasm collection contained selectively neutral alleles whose frequencies were in Hardy-Weinberg equilibrium. However, randomly sampling n_r core entries from a finite population of effective size N_e is not genetically equivalent to sampling n_r core entries from N accessions unless genetic duplicates are first removed. We achieved this by choosing to work with GREC, rather than with the original entire collection. However, the assumption of selectively neutral alleles may not hold for many genes that control adaptive traits since these are products of long-term natural and artificial selection. In fact, as pointed out by Yonezawa et al. (1995), the neutrality principle may not hold for some isozymes. The assumption of Hardy-Weinberg equilibrium may also not be valid since accessions in the collection do not interbreed with one another.

The major theoretical argument for core collections in seed crops is that a small number of samples may be efficient in retaining alleles at single loci (Brown 1989). This leads one to presume that the breeders would assemble alleles into genotypes at will in crossing programs. The relative efficiency of a few samples (approximately 10% of N) is attributable to the expectation that the number of alleles increases in proportion to the logarithm of the number N of available samples in the entire collection. However, in clonal crops like potato, much more interest surrounds the whole genotype; specific combinations of genes in highly heterozygous combinations could be worth preserving, and the number of genotypes (genets) preserved increases in direct proportion to the number of samples, assuming duplicates are removed. Realizing these specific features in clonal crops, Brown (1995) suggested that the proportion of entries in the core, rather than fixing at 10%, might have to be higher or lower than 10%. The findings of this research, giving an optimal sampling fraction of 600/1,910 = 0.31, agrees with Brown's views.

Huaman et al. (2000b) found that a core collection of 306 entries adequately represented their morphologically duplicate-free collection of 2,379 accessions. However, an examination of their Table 2 shows that, with n = 306,

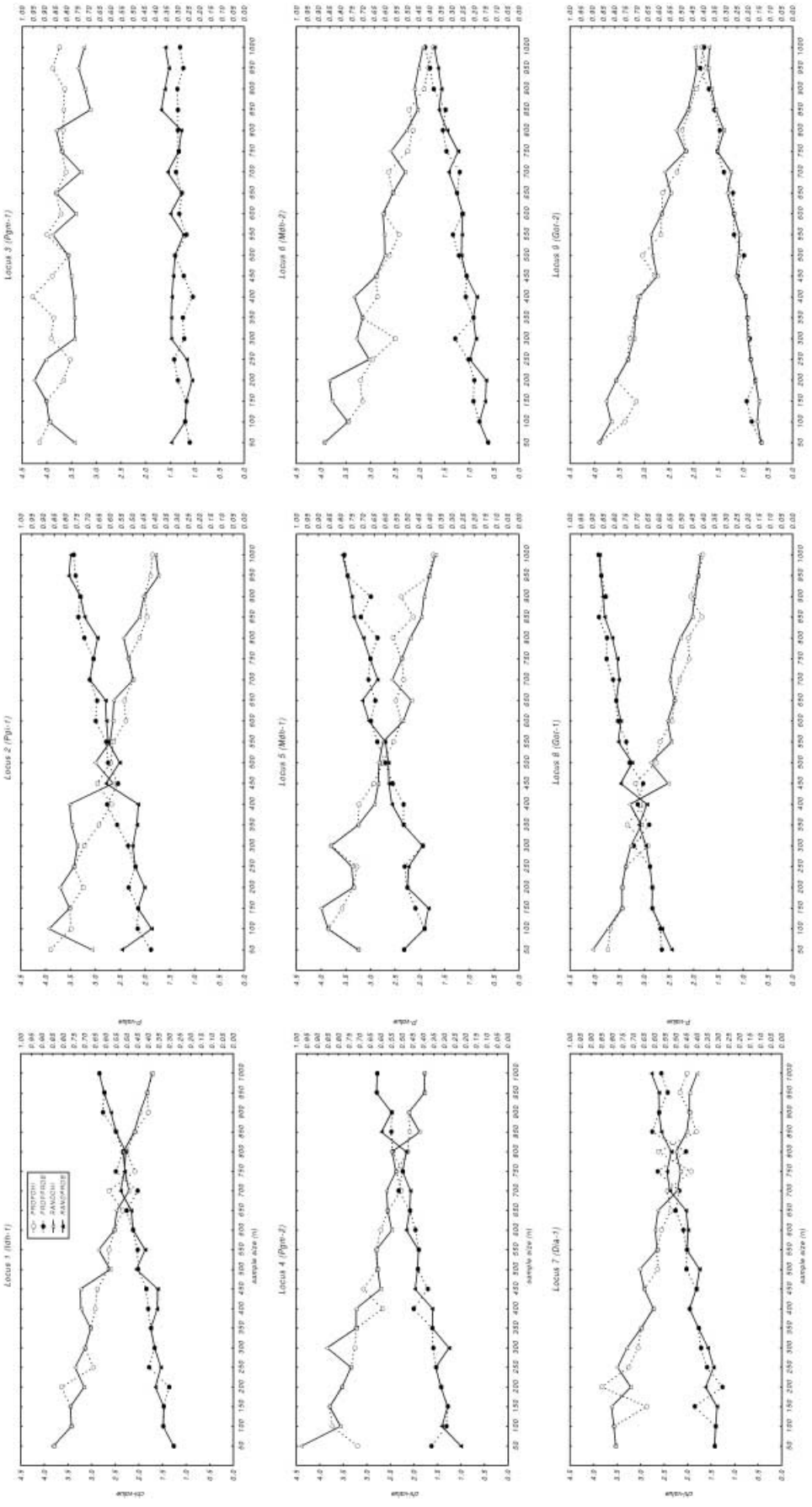

Fig. 4 Locus-wise observed upper-0.05-quantile χ^2 and *P*-values for locus heterozygosity under proportional and random strategies

two loci (Got-1 and Pgi-1) fail to be adequately represented in the population. The sum of individual-locus χ_j^2(AF) values in their Table 2 comes to χ^2(AF) = 55.385 (*df* = 28) with a *P*-value of 0.0015. This value of χ^2(AF), corresponding to *n* = 306, is included in the range of 1,500 χ^2(AF) values for *n* = 300 for both P and R strategies (Tables 1 and 2). This result provides validity to, and confidence in, the simulation approach employed in this study. Table 3 in Huaman *et al*. (2000b) also needs correction in the value of χ^2(LH), which should have been computed according to the χ^2(LH) formula given in Materials and methods and should have been 15.647 (*df* = 9; *P* = 0.075) in place of 5.729 (*df* = 8; *P* = 0.678) as reported. Simulation results clearly establish that Huaman et al. (2000b) need to revise their optimal core collection size from 306 to about 600 using either the P or the R strategy.

The conclusions regarding optimal core sample size and strategy arrived at for the potato collection obviously hold for the nine isozyme loci for which the accessions in the available collection were characterized. These may change when additional loci are used to characterize the collection.

The simulation approach, developed here using potato isozyme data, could be generally applied on genetic or molecular data of any crop species for identifying the optimal sampling strategy and core collection size, with suitable minor modifications as necessary.

References

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

Brown AHD (1995) The core collection at the crossroads. In: Hodgkin T, Brown AHD, van Hintum ThJL, Morales EAV (eds) Core collections of plant genetic resources. John Wiley and Sons, NewYork, pp 3–19

Douches DS, Ludlam K (1991) Electrophoretic characterization of North American potato cultivars. Am Potato J 68:767–780

Douches DS, Quiros CF (1988) Additional loci in tuber-bearing solanums: inheritance and linkage relationships. J Hered 79: 377–384

Douches DS, Ludlam K, Freyre R (1991) Isozyme and plastid DNA assessment of pedigrees of nineteenth-century potato cultivars. Theor Appl Genet 82:192–200

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Pop Biol 3:87–112

Frankel OH, Brown AHD (1984) Current plant genetic resources – a critical appraisal. In: Genetics: new frontiers (vol IV). Oxford and IBH Publishing, New Delhi, India, pp 1–13

Freyre R, Douches DS (1994) Development of a model for marker-assisted selection of specific gravity in diploid potato across environments. Crop Sci 34:1361–1368

Freyre R, Warnke S, Sosinski B, Douches DS (1994) Quantitative trait locus analysis of tuber dormancy in diploid potato (*Solanum* spp.). Theor Appl Genet 89:474–480

Huaman Z (1998) Collection, maintenance and evaluation of potato genetic resources. Plant Var Seeds 11:29–38

Huaman Z, Ortiz R, Gomez R (2000a) Selecting a *Solanum tuberosum* ssp. *andigena* core collection using morphological, geographical, disease and pest descriptors. Am J Potato Res 77:183–90

Huaman Z, Ortiz R, Zhang D, Rodriguez F (2000b) Isozyme analysis of entire and core collections of *Solanum tuberosum* subsp. *Andigena* potato cultivars. Crop Sci 40:273–276

Ortiz R, Huaman Z (2001) Allozyme polymorphism in tetraploid potato gene pools and the effect of human selection. Theor Appl Genet (in press)

Ortiz R, Douches DS, Kotch GP, Peloquin SJ (1993) Use of haploids and isozyme markers for genetic analysis in the polysomic polyploid potato. J Genet Breed 47:283–288

Quiros CF, Ortega R, van Raamsdock L, Herrera-Montoya M, Cisneros P, Schmidt E, Brush S (1992) Increase of potato genetic resources in their center of diversity: the role of natural outcrossing and selection by the Andean farmers. Genet Res Crop Evol 39:107–112

Rabinowitz D, Linder CR, Ortega R, Begazo D, Murguia H, Douches DS, Quiros CF (1990) High levels of interspecific hybridization between *Solanum sparsipilum* and *S. stenotomum* in experimental plots in the Andes. Am Potato J 67:73–81

Sokal RR, Rohlf FJ (1981) Biometry. W. H. Freeman and Co, New York

Yonezawa K, Nomura T, Morisima 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. John Wiley and Sons, New York, pp 35–53

Zimmerer KS, Douches DS (1991) Geographical approaches to crop conservation: the partitioning of genetic diversity in Andean potatoes. Econ Bot 45:176–189