Biodivers Conserv (2018) 27:1471–1486
https://doi.org/10.1007/s10531-018-1504-3

ORIGINAL PAPER

Measuring the representativeness of a germplasm collection

Carlos Hernandez-Suarez[1]

Received: 3 May 2017 / Revised: 7 January 2018 / Accepted: 15 January 2018 /
Published online: 20 January 2018

Abstract Many germplasm collections aim to preserve most of the genetic diversity present in a population so that the population could be regenerated, which provides genetic resources to ensure food security. This paper proposes a way to measure how well a germplasm collection achieve this goal. In the most common scenario, one has little information regarding the number and statistical distribution of alleles at every locus, and it is thus very difficult to measure the representativeness of the accession. Here, we show how to use samples of allelic diversity at a sample of loci to estimate the representativeness of an accession based on the coverage of a sample with point and interval estimates. Our approach avoids making unrealistic assumptions regarding the number of loci, the bounds for the number of alleles or their frequency distributions. Depending on the sampling scheme of a collection, we differentiate between absolute or relative coverage. Here, we demonstrate this methodology using data from the germplasm collection at the Leibniz Institute of Plant Genetics and Crop Plant Research.

Keywords Coverage · Allele conservation · Seed accession

Introduction

A gene bank is a collection of seeds and other plant reproductive material, primarily including cultivated plants and their wild relatives. These collections have become the gene pools for plants of agricultural importance. The main goals of these collections are to secure plant genetic resources and to provide access to them. Evaluating a gene bank requires to

Communicated by David Hawksworth.

This article belongs to the Topical Collection: Ex-situ conservation.

✉ Carlos Hernandez-Suarez
carlosmh@mac.com

[1] Facultad de Ciencias, Universidad de Colima, Bernal Diaz del Castillo 340, Colima, México

evaluate how well the methods for collecting, preserving and distributing the genetic material were designed. Although it is not difficult to establish measures of the performance of gene banks in preserving and distributing genetic materials, it is more challenging to evaluate the amount of genetic variation that the bank has captured. Trying to include most of the alleles in a whole genome would require prohibitive sample sizes and many biological and mathematical assumptions that are not always valid but are sometimes used due to a lack of alternatives. Intensive sampling to capture a high proportion of the total genetic diversity of a species is a vague goal. This is because the concept of high diversity can be relative, especially if there is little information regarding the genetic composition of a population. So, for the most common scenario in which the sampled fraction is small compared to the population size, the unsampled fraction could include many genotypes that are not included in the sample. In the setting of a gene bank, the answer to the question: what proportion of the alleles is included in the accession? is not possible to answer without making assumptions regarding the number of loci, and the number of alleles per locus or their frequency distribution.

Because a large number of individuals can be represented by a few genotypes and few individuals can represent a large number of genotypes, sampling to guarantee that a particular proportion of possible genotypes has been preserved can lead to prohibitively large sample sizes. Nevertheless, many alleles could be slight variants of other alleles, and thus, a set of alleles S could be used to reconstruct a larger set of alleles.

Another related question for the preservation of diversity is: what proportion of the alleles present in a population is represented in the sample? This is known as the coverage of a sample (Good 1953). To better explain the concept of coverage of a sample, consider for instance the preface to *The Origin of Species* (Darwin 1866). The preface uses 628 different words and is a total of 1633 words in length. Let's suppose we take a random sample of size 30 and this includes only the words *the*, *of*, *and*, *to* and *I*. Because 19% of the words in the preface belongs to this five words the coverage of the sample is 19%. In simpler terms, the coverage of a sample is the fraction of individuals in the population that is represented in the sample.

Put it in context of the present problem, if the worldwide coverage of an accession in terms of alleles is C, it means that if we sample one seed from the field at random and an allele at random from its genome, there is a probability C that the allele is included in the accession.

Clearly, the goal is to achieve high coverage at all loci and accurately estimate the proportion of existing alleles in a genome that is included in an accession. Here, we introduce a methodology that requires no previous knowledge of the number of loci, allele frequency or its distribution at each locus. This methodology is effective under mild conditions such as random sampling. Further, we assume only that we can sample seeds from several accessions at a gene bank and identify alleles at one or more loci (i.e., by using microsatellite markers). Here, we show that the key parameter for assessing the coverage of a sample is the number of *singletons*, that is, alleles that are detected only once in the sample. When the number of singletons for a locus is high in a sample, the variability of alleles at that locus is high and the coverage of that locus is low. We also show how to combine the estimate of coverage at different loci to construct point and interval estimates for coverage achieved at the genome level.

The coverage of a sample

As mentioned previously, the coverage of a sample is defined as the proportion of the individuals in the population that are represented in the sample. If p_i is the proportion of individuals of class i in the population, and x_i is the observed number of individuals of class i in the sample, then C, the coverage of the sample is

$$C = \sum_{i=1}^{M} p_i \, I(x_i > 0),$$

where M is the number of classes in the population. Good (1953) was the first to provide an estimate for the coverage of a sample, although he attributes the main result of his work to personal communications with A. M. Turing. If q_r is the population relative frequency of a species that is represented r times in the sample, and n_r is the number of species (alleles in our case) in the sample that are represented exactly by r individuals, then, according to Good (1953):

$$E\left[q_r\right] \approx \frac{(r+1)n_{r+1}}{n\, n_r}$$

An immediate consequence is that the expected total probability of all species (alleles) that are each represented r times ($r \geq 1$) in the sample is approximately

$$(r+1)\, n_{r+1}/n$$

Thus, the expected total chance of all species represented at all in the sample is approximately

$$\sum_{r=1} E\left[q_r\right] \approx (2n_2 + 3n_3 + 4n_4 + \ldots)/n = 1 - n_1/n \tag{1}$$

where $n_1 = s$, the number of singletons in the sample, which is the well-known estimate of coverage.

Several authors have deal with issues related to the coverage of a sample: Good and Toulmin (1956) extended Good (1953) work and provided estimates of the number of species (alleles) in a population, and analyzed the effect of an additional sample on an increase in coverage. Esty (1985) was the first to use the concept of sample coverage to estimate the number of classes in a population. Esty (1986) analyzed the efficiency of the Good (1953) estimate against the best estimate under the assumption that all classes are equally likely, and found both that Good's estimate performed remarkably well, and that it is a lower bound for the true coverage, unless all classes have homogeneous frequencies. Chao and Lee (1993) generalized the definition of the coverage to the case in which there is dependence among classes. Chao and Lee (1992) and Lee and Chao (1994) provided estimates of the number of alleles in a population constructed from Good (1953) coverage estimate. Huang and Weir (2001) used Chao and Lee (1992) approach to estimate the number of alleles in a population. Zhang and Zhang (2009) showed the asymptotic normality of Good's estimate, thus allowing interval estimation of the coverage of a sample.

Other authors (see Harris 1959; Knott 1967; Robbins 1968; Starr 1979; Chao 1981; Esty 1982, 1983; Lo 1992) discussed Good (1953) estimate, or used it to obtain estimates on the number of classes in a population.

The most important drawback of the coverage is statistical in nature: coverage is not a parameter of the population, thus, cannot be estimated. This means that coverage is a property of the sample and not of the population. In the following section, we first introduce a new, more intuitive construction to estimate the coverage of a sample that is based on urn models. The goal of this new construction based on urn models is for the researcher to apprehend the assumptions involved in estimating coverage when the number of different types of individuals in a population (and hence their frequencies) is unknown.

Urn models

An urn problem is a thought experiment in which one or more hypothetical urns are filled with balls of different colors that represent actual items of interest. Urn models, which describe the probability of events arising from sampling balls from the urns, are among the most common models currently used in biology, engineering, medicine, social sciences and other fields. Estimating the composition of balls in an urn, i.e., the types of individuals it contains and their frequencies, and thereby predicting the outcome of sampling balls from the urn, is the most common problem for which these models are used. Traditionally, the question is posed as: in a sample of size n, there are x_i individuals of type $i, i = 1, 2, 3, \dots, k$. What is the maximum likelihood estimate (MLE) of the relative frequencies of the types? The likelihood function of the sample is written as:

$$P(\underline{\mathbf{X}} = \underline{\mathbf{x}} \,|\underline{\mathbf{p}}) = \binom{n}{x_1\, x_2\, \dots\, x_k} p_1^{x_1}\, p_2^{x_2}\, p_3^{x_3} \dots p_k^{x_k} \tag{2}$$

that is maximized at $p_i = x_i/n$.

The previous paragraph explains the usual rationale behind sampling a multinomial distribution, but there is an inherent misconception that turns out to be relevant when the number and frequencies of types in the urn is unknown. For instance, according to (2), if we take a sample of size $n = 3$ and this yields balls of three different colors (types), e.g., red, black and white, the urn that maximizes the likelihood of such a sample is an urn with 1/3 red, 1/3 black and 1/3 white balls. However, this only holds true if we concede that only those three colors exist in the urn.

However, if we do not know the colors (types) that are present in the urn in advance, and we obtain a sample of one red, one black and one white ball even though the urn contains at least one other color, then we have obtained a sample of only three of the different possible colors. In this situation, the urn composition that would maximize the probability of such a sample is one in which all balls are of different color. Take for example a sample of size 10 from an urn and suppose we obtain 4 black, 3 white, 1 red, 1 blue and 1 yellow ball. The classical estimate for these respective colors is:

$$\bar{\mathbf{p}} = \{4/10, 3/10, 1/10, 1/10, 1/10\}$$

The likelihood of the given sample is around 1.7×10^{-2}. Nevertheless, when we do not know anything about the composition, or numbers of types (colors) of balls in the urn, the correct question is not: what urn composition maximizes the probability of getting 4 black, 3 white, 1 red, 1 blue and 1 yellow? but rather: what urn composition maximizes the probability of getting balls at the observed frequencies? Answering this requires some care, because if there are singletons in the sample (colors represented by a single individual) there is always another urn composition that has higher likelihood than using the traditional estimate $p_i = x_i/n$.

For instance, in our sample case above, we will show that there is an urn composition whose likelihood is 27 times higher for maximizing the probability of obtaining a particular frequency than that obtained using the usual MLE. To show this, we start by assuming that our sample comes from an urn containing N balls divided into M classes (types), and that both N and M are unknown. Without loss of generality, assume that K classes contain more than one individual each, whereas the remaining $S = M - K$ classes contain a single individual each. Clearly $S \geq 0$. Let θ be the fraction of the population occupied by these S classes. It is important to remark that every multinomial distribution is a subset of this general population. Now, suppose our sample of size n contains m different types or classes, of which s are singletons. A posteriori, we label the classes arbitrarily as $1, 2, 3, \dots, m$ and let x_i be the number of individuals in the sample belonging to class i. The likelihood of the sample is then:

$$P(\underline{\mathbf{X}} = \underline{\mathbf{x}} \;|\underline{\mathbf{p}}) = \begin{pmatrix} & n & \\ x_1\; x_2 & \dots & x_{m-s} \end{pmatrix} p_1^{x_1}\, p_2^{x_2}\, p_3^{x_3} \cdots p_{m-s}^{x_{m-s}}\, p_{m-s+1}^1\, p_{m-s+2}^1 \cdots p_m^1 \tag{3}$$

where p_i is the proportion in the population represented by the x_i individuals in the sample. Observe that in the traditional MLE for multinomial distributions, the dimension of the parameter space is restricted to the observed number of classes. Nevertheless, if the dimension of the parameter space is not restricted to a known number of classes, the product of the last s of the p_i's in (3), namely

$$p_{m-s+1}^1\, p_{m-s+2}^1 \cdots p_m^1$$

is in fact the probability of getting s different individuals with a sample of size s. This probability is maximized when a sample of this size is taken out of the fraction θ of the population whose individuals each belong to a different class. Then, the likelihood of the sample can be rewritten as:

$$P(\underline{\mathbf{X}} = \underline{\mathbf{x}} \;|\underline{\mathbf{p}}) = \begin{pmatrix} & n & \\ x_1\; x_2 & \dots & x_{m-s}\; s \end{pmatrix} p_1^{x_1}\, p_2^{x_2}\, p_3^{x_3} \cdots p_{m-s}^{x_{m-s}}\, \theta^s$$

Thus, the MLE's are:

$$\hat{p}_i = x_i/n, \quad \hat{\theta} = s/n$$

which yields a likelihood of 4.7×10^{-1} That is, the chance of obtaining a sample matching the one observed are larger if the urn contains 4/10 black, 3/10 white, and the remaining 3/10 is composed of balls that are each of different colors. Under the arguments leading to (3), the colors not detected in the sample must belong to the fraction θ whose individuals each belong to a different class. Thus, an estimate of the fraction of the population that has not been represented in the population is precisely $\hat{\theta}$ minus the fraction s/N corresponding to the population frequencies of the singletons detected in the sample, that is, $\hat{\theta} = s/N$. Thus, the MLE of the lower bound for of the coverage of a sample is

$$\widehat{C} = 1 - s/n$$

which is precisely Good (1953) estimate.

Properties of the coverage of several populations

We have established that the coverage of a sample is defined as the proportion of individuals in the population that is represented in the sample. This is the same as the probability that an individual selected at random from the population is represented in the sample. Some results would be needed to estimate the coverage of an accession at the genome level or to estimate the coverage of a group of accessions, i.e., a germplasm collection. We leave the proof of the following results to the Appendix.

Properties of the coverage

1. If every individual in a population has two possible attributes, X and Y and X takes values from a set S_X whereas Y takes values from a set S_Y, and these sets do not intersect, then, if a sample of size n has a coverage C_X from set S_X and C_Y from set S_Y, the overall coverage of attributes of the set $S_X \cup S_Y$ is then $(C_X + C_Y)/2$.
2. If two populations of sizes N_1 and N_2 are sampled with respective sample sizes of $n_1 = nf$ and $n_2 = n\,(1-f)$ where $f = N_1/(N_1 + N_2)$, then, the coverage of the mixture of both samples is the fraction of individuals in the mixture of populations that is represented in the sample. This will be defined as the *absolute coverage*.
3. For two populations of sizes N_1 and N_2, $N_1 \neq N_2$, each sampled with a sample of fixed size n, the coverage of a mixture of the two samples can be interpreted as the likelihood that an individual who is equally likely to be selected from either population is represented in the sample. This will be defined as the *relative coverage*.
4. If a random sample of s accessions is selected from a germplasm collection with S accessions, the coverage estimated from the sample of s accessions is an estimate of the coverage of the collection. Equivalent results apply if the collection is divided into groups and random samples are selected from each group.

Property 1 implies that if the coverages at loci 1 and 2 are respectively C_1 and C_2, then the coverage of alleles at both loci simultaneously is the average of C_1 and C_2, as long as no allele can exist at both loci. This result will prove useful when estimating coverage at the genome level.

Property 2 implies that we can estimate the coverage of a mix of two populations as long as the sample size denotes the weighted proportion of each population.

Property 3 implies that when sample sizes from each population are not weighted according to their relative presence in the population, there is a slight change in the interpretation of coverage. It is no longer *absolute coverage*, i.e., the likelihood that a random individual selected from the mixture of populations is represented in the sample. Rather it becomes *relative coverage*, or the likelihood that an individual selected at random from a population at random is represented in the sample. This result is particularly useful when accession sample sizes are equal and the accessions were not weighted according to their representation in the actual population in the field.

The difference between *relative* and *absolute* coverage will be provided with a simple example: suppose we have two populations of size N_1 and N_2 where $N_2 = 2N_1$. Lets suppose we take a sample of size n_1 and n_2 respectively from each population and mix both samples, and then we estimate the coverage of the sample mix using (3) yielding a coverage estimate $\hat{C}$. Now, If $n_2 = 2\,n_1$, then, the probability that an individual taken at random from the mix of populations will be represented in the sample mix is $\hat{C}$, which is *absolute* *coverage*. On the other hand, if $n_2 \neq 2\, n_1$ then, we can only conclude that if we select at random an individual equally likely from either population, there is a probability $\hat{C}$ that is represented in the sample mix. This is *relative coverage*.

Estimating the coverage of an accession at the genome level

Before we can estimate the coverage of an entire germplasm collection, we must first introduce how to estimate and interpret the coverage of a single accession at the genome level. Assume an accession has size N and that it is a random sample from a population. Let the true coverage at locus i be C_i. If there are M loci, the C_i values constitute a population with mean C_G. By Property 1, the coverage of the accession at the genome level is precisely:

$$C_G = (C_1 + C_2 + \ldots + C_M)/M,$$

which has some variance σ_G^2. Suppose a sample of size n is taken from this accession and every DNA strand in the sample is characterized at the same K randomly selected loci (Fig. 1), obtaining s_i singletons at locus $i, i = 1, 2, 3, \ldots K$. Let $\hat{c}_i = 1 - s_i/n$ be the estimate of coverage at locus i. Then,

$$\widehat{C}_g = \sum_{i=1}^{K} \hat{c}_i/K \tag{4}$$

follows a normal distribution with some expected value μ and variance σ_G^2/K. Since c_i is constructed by sampling n individuals from the accession, its expected value is at most equal to the coverage of the accession, that is $E[c_i] \leq C_i$ for $n \leq N$. Thus, (4) is a lower bound for the coverage at the genome level, C_G.

Let $\widehat{\sigma}$ be the standard deviation of the $\hat{c}_i$ values, then a $1 - \alpha$ confidence interval (CI) for the coverage of the accession would be:

$$\widehat{C}_g \ \pm \ Z_{1-\alpha} \frac{\widehat{\sigma}}{K} \tag{5}$$

where $Z_1 - \alpha$ can be substituted with Student's t distribution with the appropriate degrees of freedom when the number of sampled loci is small. For instance, using the example in Fig. 1, where $n = 8$ and $K = 9$, the estimate of coverage using (4) is $\widehat{C_g} = 0.833$. That is, at least 83 % of all alleles in the population are represented by the 32 different alleles in the sample. Using (5) with $(t_8, 0.95)$ we obtain a 95 % CI for the lower bound of the coverage as (0.797, 0.869)

Estimating the coverage of a group of accessions

We might sample each accession proportionally to the relative amount in the population that the accession represents. Then, from Property 2, after alleles are identified at K loci in the mixed sample, the estimated coverage of the germplasm collection is calculated using (4) and the estimates of the lower bound can be calculated as (5). This coverage would be *absolute coverage*.

However, such a sampling scheme is impractical mainly because the relative amount in the population is unknown. In practice, n seeds would be taken from each accession and

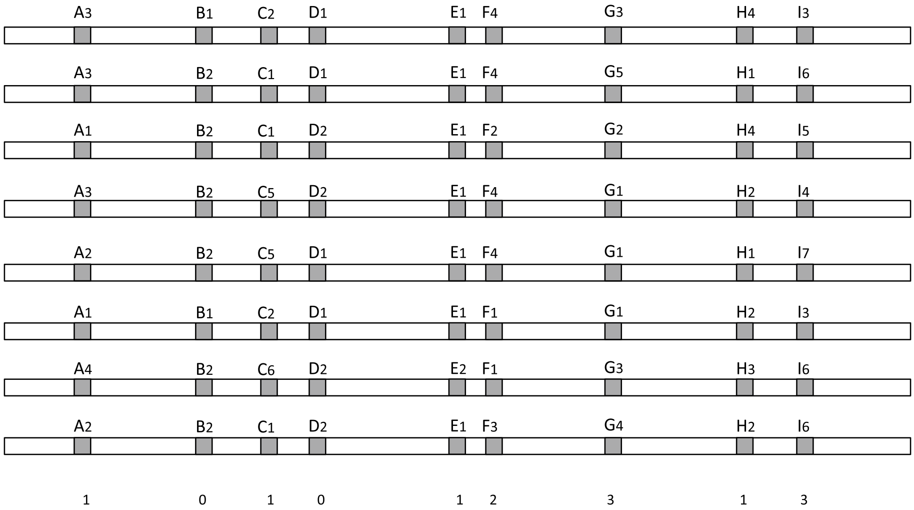
A3 B1 C2 D1 E1 F4 G3 H4 I3
A3 B2 C1 D1 E1 F4 G5 H1 I6
A1 B2 C1 D2 E1 F2 G2 H4 I5
A3 B2 C5 D2 E1 F4 G1 H2 I4
A2 B2 C5 D1 E1 F4 G1 H1 I7
A1 B1 C2 D1 E1 F1 G1 H2 I3
A4 B2 C6 D2 E2 F1 G3 H3 I6
A2 B2 C1 D2 E1 F3 G4 H2 I6
1 0 1 0 1 2 3 1 3

Fig. 1 A sample of 8 DNA strands characterized at nine different loci showing the identified allele. Bottom row indicates the number of singletons found in the sample in every locus. In this example, 32 different alleles were detected in the sample

each individual analyzed to identify alleles at the same loci across individuals. Although the coverage estimate at the genome level (4) and its CIs (5) remain the same, according to Property 3, the interpretation of coverage changes: instead of *absolute coverage*, it would now be *relative coverage*. That is, if we chose a strain at random (equally likely among the strains contained in the germplasm collection) and an allele is selected at random from a single individual of that strain in the field, the probability is $\widehat{C_g}$ that there is a copy of that allele in the collection. This is more stringent than *absolute coverage*, as it gives equal weight to each strain in the accession, instead of giving greater weight to strains that are more abundant in the field.

A sample of accessions could be taken to estimate the coverage of a germplasm collection. If a sample of s accessions is selected from S, Property 4 guarantees that the estimates are still valid. Although it might seem odd that the value of s does not play a role in calculating the expression $\widehat{C_g}$, nevertheless, it is still included in the sample size n. This tends to reduce the number of singletons found and hence reduces both the variance of coverage estimated at each locus and that between loci.

Simulations

Here, we evaluate the performance of (4) and (5) for estimating the coverage of single accession when the true coverage is known. First, we generated a fictitious genome with 1×10^5 loci, and from this we generated three accessions of size $N = 50, 100$ or 250 seeds. We then sampled each one of these accessions with several sample sizes n and identified alleles in each sample at $K = 1, 5, 10, 20$ or 30 loci. In each simulation, the n individuals and the K loci for allele identification where selected at random from the accession. We reiterated this latter step 10, 000 times for each sample combination of n and K to obtain point and interval estimates for the lower bound of the coverage of the accession and then

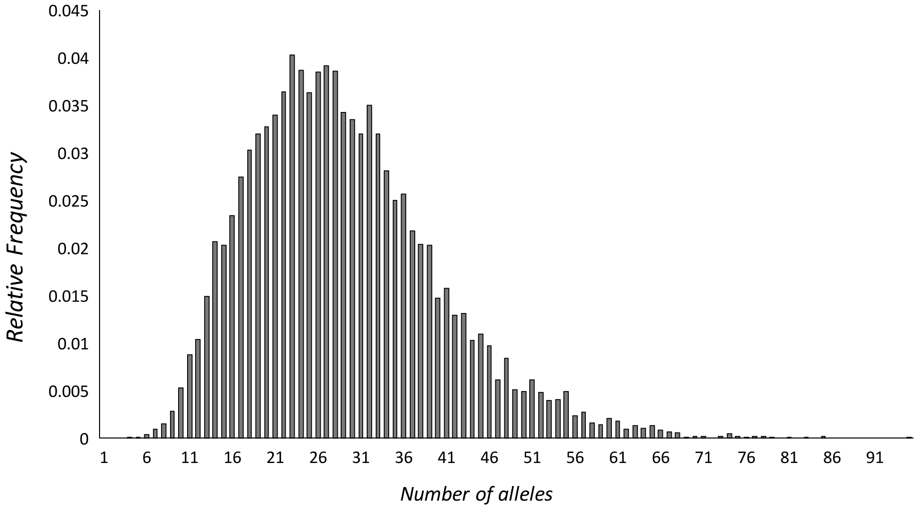

Fig. 2 The distribution of the number of alleles per locus in the simulated genome. For every locus, we used a Gamma (4, 4) to simulate the parameter of a Poisson distribution which in turn was used to simulate the alleles at that locus for every individual in the accession. The maximum number of simulated alleles in a locus was 94 with an average of 16

Table 1 Coverage estimated in accession 1

Seeds sampled (n)	Sampled loci (K)				
	1	5	10	20	30
25	0.788	0.783 (0.046)	0.784 (0.032)	0.784 (0.023)	0.785 (0.019)
50	0.914	0.914 (0.021)	0.913 (0.015)	0.913 (0.011)	0.913 (0.009)

Size of accession: 50 seeds. True Coverage of accession is: 0.916

Mean coverage over 10,000 simulations

Numbers in parenthesis are SD

compared these against the true coverage of the same. For the simulated genome, the number and frequency of alleles per locus was randomly generated using a Poisson random variable, with varying parameter, according to a Gamma distribution with parameters $\alpha = 4$ and $\beta = 4$ for an average of 16 alleles per locus (see Fig. 2).

Tables 1, 2 and 3 show the estimated coverage for each accession at different sample sizes, while varying the number of loci analyzed. The estimate of coverage improves with sample size, but increasing the number of loci has little effect on the point estimate, although clearly reducing its variance. It is important to recall that these are samples from the accessions, so these are lower bounds for their true coverage.

Table 2 Coverage estimated in accession 2

Seeds sampled (n)	Sampled loci (K)				
	1	5	10	20	30
25	0.786	0.785 (0.047)	0.784 (0.033)	0.786 (0.023)	0.785 (0.019)
50	0.915	0.915 (0.021)	0.913 (0.015)	0.914 (0.010)	0.914 (0.009)
75	0.950	0.950 (0.012)	0.950 (0.009)	0.950 (0.006)	0.950 (0.005)
100	0.966	0.965 (0.009)	0.966 (0.006)	0.966 (0.004)	0.965 (0.004)

Size of accession: 100 seeds. True Coverage of accession is: 0.966

Mean coverage over 10,000 simulations

Numbers in parenthesis are SD

Table 3 Coverage estimated in accession 3

Seeds sampled (n)	Sampled loci (K)				
	1	5	10	20	30
25	0.784	0.784 (0.046)	0.784 (0.033)	0.784 (0.023)	0.785 (0.019)
50	0.913	0.914 (0.021)	0.914 (0.015)	0.914 (0.011)	0.914 (0.009)
75	0.951	0.950 (0.012)	0.951 (0.009)	0.950 (0.006)	0.951 (0.005)
100	0.966	0.965 (0.009)	0.966 (0.006)	0.965 (0.004)	0.966 (0.004)

Size of accession: 250 seeds. True Coverage of accession is: 0.988

Mean coverage over 10,000 simulations

Numbers in parenthesis are SD

Example data

This example is an analysis of data previously reported in Huang et al. (2002) from the germplasm collection at the Leibniz Institute of Plant Genetics and Crop Plant Research at Gatersleben, Germany. The collection contains more than 10,000 accessions of hexaploid bread wheat (*Triticum aestivum*). The study used a sample of 998 accessions that originated from 66 countries on five continents. Total genomic DNA was extracted from five grains of each accession and 24 microsatellite markers for 26 loci were used to evaluate the genetic diversity of the accessions, with at least one marker for each of the 21 chromosomes.

Table 4 shows some statistics previously reported in Huang et al. (2002) for several loci, together with the number of singletons and the coverage estimated at those loci. In Figs. 3 and 4 the number of alleles and Nei's diversity index (Nei 1973) are plotted against the coverage values. The lower bound for the estimated coverage of the collection at Gatersleben is 0.999

We grouped the samples by continent (Table 5) and region (Table 6) and obtained the estimates of the coverage for these groups. Some regions were underrepresented, so it was not possible to estimate CIs for their lower bounds. As for any sample, conclusions can be generalized to the population from which the sample came. Here, if 998 accessions were sampled at random from the 10,000 accessions at Gatersleben, the coverages here would be the estimated lower bounds of the coverage of the collection at Gatersleben. The fact that they were sampled to cover several continents and several regions within those continents does not affect the coverage estimates.

Table 4 Locus name, number of alleles, gene diversity, number of singletons and coverage estimated at each locus

Locus	Number of alleles[a]	Gene diversity[a,b]	Singletons	Coverage
Xgwm357	12	0.69	5	0.995
Xgwm95	13	0.78	1	0.999
Xgwm155	15	0.79	1	0.999
Xgwm160	12	0.70	1	0.999
Xgwm192c	4	0.43	0	1.000
Xgwm186	25	0.90	2	0.998
Xgwm459	46	0.94	2	0.998
Xgwm631	12	0.76	1	0.999
Xgwm18	13	0.77	3	0.997
Xtaglgap	26	0.82	5	0.996
Xgwm619	21	0.86	1	0.999
Xgwm389	19	0.88	1	0.999
Xgwm513	8	0.65	1	0.999
Xgwm192a	18	0.84	2	0.998
Xgwm408	20	0.84	1	0.999
Xgwm680	14	0.66	1	0.999
Xgwm46	24	0.88	0	1.000
Xgwm577	36	0.90	2	0.998
Xgwm337	24	0.83	1	0.999
Xgwm458	7	0.62	0	1.000
Xgwm261	22	0.56	4	0.997
Xgwm3	10	0.67	0	1.000
Xgwm192b	8	0.63	0	1.000
Xgwm190	23	0.82	5	0.995
Xgwm325	15	0.82	1	0.999
Xgwm437	23	0.90	0	1.000

[a] As reported in Huang et al. (2002)

[b] Nei's diversity index

Table 5 Relative coverages of the collection at the continent level

Continent	Coverage	Lower bound (0.95)
Africa	0.984	0.962
America	0.920	0.781
Asia	0.995	0.987
Europa	0.996	0.991
Oceania	0.038	[a]
Unknown	0.58	0.022
Total collection	0.999	0.996

[a] Sample size too small to estimate lower bound

Table 6 Relative coverages of the collection at the regional level

Region	Coverage	Lower bound (0.95)
Africa East	0.947	0.865
Africa Northwest	0.969	0.929
America Middle	0.633	0.083
America North	0.763	0.431
America South	0.896	0.702
Asia Middle East	0.902	0.733
Asia Far East	0.991	0.979
Asia Near East	0.984	0.964
Europe North	0.985	0.965
Europe Southeast	0.989	0.972
Europe Southwest	0.983	0.954
Oceania	0.038	[a]
Unknown	0.580	0.022

[a] Sample size too small to estimate lower bound

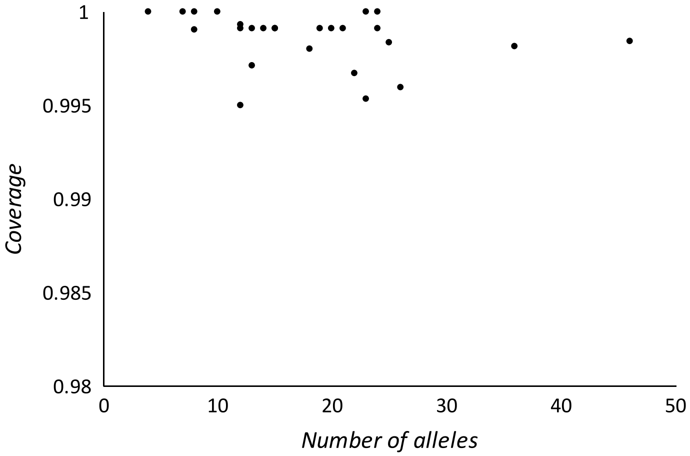

Fig. 3 Plot of the number of alleles found versus Coverage at each of 26 loci in Gatersleben collection

Discussion

It would be most desirable to measure representativeness of a collection in terms of the variability captured, so that one could estimate the number and distribution of the alleles not captured. This task is impossible unless we are willing to concede some assumptions about the frequency distribution of alleles. Such assumptions tend to make mathematicians lives simpler, but they can hide negative properties that might generally be unknown to biologists. Moreover, an awareness of the assumptions behind a method does not mean that the effects of deviations from those assumptions would be apprehended. In the worst-case

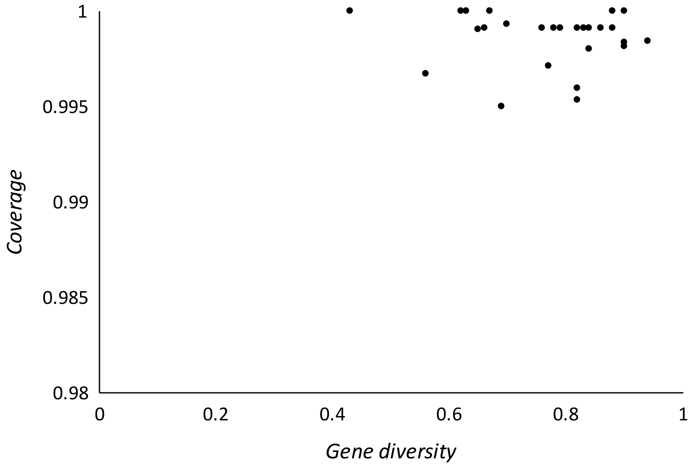

Fig. 4 Plot of Nei's gene diversity index versus coverage at each of 26 loci in Gatersleben collection

scenario, one might not distinguish whether some assumptions would result in overly conservative or optimistic estimates of the variability captured. The purpose of section "Urn models" was to provide a new construction of this problem that not only illustrates the importance of the singletons in estimating coverage, but also reveals the basic assumption that all singletons come from a fraction of the population not included in the sample in which all individuals are different and therefore not represented, which leads to a conservative estimate of coverage.

In practice, there is no way to guarantee that a collection has captured all of the genetic variability of a species even at a single-locus level. Nevertheless, if we assume that some alleles are variants of other alleles, then the alleles that occur at low frequency in nature might have been generated from other alleles that occur at higher frequency. In such a situation, the coverage of a sample would be a useful indicator of the potential for a collection not only to preserve the alleles it contains, but also to regenerate other alleles that might exist outside the collection.

For an individual genebank, a core collection consists of a limited number of the accessions in an existing collection, chosen to represent the genetic spectrum in the whole collection. It should include as much as possible of its genetic diversity but is not intended to replace a genebank (Brown 1995; van Hintum et al. 2000). By sampling the core collection and estimating the coverage we can estimate the proportion of alleles in the genebank that is represented in the core collection.

In Huang et al. (2002) a fixed sample of size five was taken from each of the 998 accessions sampled, thus, the coverage achieved must be interpreted as a *relative coverage*, that is, if we choose an accession at random from the collection and then we select (in the field) an allele at random belonging to the strain represented in the accession, there is a 0.99 chance that the allele is included in the collection.

Efforts to estimate the coverage of all germplasm collections of the most important crops worldwide should be combined, because some accessions might be duplicated between collections. This would provide a global overview of germplasm preservation efforts. It will be important to realize that results for two sets of strains could be combined even if they were not sampled at the same loci. For instance, if a study included M_1 strains identifying alleles at a set of loci L_1, and another study identified alleles in M_2 different strains at another set of loci L_2 (possibly intersecting with set L_1), the estimated coverages of both collections could be combined with appropriate weighting of the coverage estimates according to the relative representations of M_1 and M_2 in the combined collections and populations.

An important assumption in the present method is that the loci sampled are randomly selected from the genome. Nevertheless, this is not a stringent assumption, because the sampled region could be narrowed to, for instance, particular genomic regions of importance for crop performance. Sampling in genomic regions of high variability could lead to an even more conservative approach as any attempt to maintain high coverage of these regions would also maintain high coverage of other regions with less variability, In this instance, it would not be possible to sustain basic assumptions based on random sampling and the conclusion on the coverage are restricted to the sampled region.

It is important to point out that the degree of linkage between loci does not affect the estimates of the coverage because the average of coverages, as introduced in Property 1, is not affected by correlations between alleles detected at two loci in the same individual.

In the particular case of the Gatersleben collection, these results indicate that there is no correlation between estimated coverage and other measures of diversity such as the number of alleles or Nei's index. The coverage attained at the locus level was high, yet there was still high variability in both the number of alleles and Nei's diversity index. This suggests that coverage express a distinct measure of diversity. The number of alleles and Nei's index measure attributes of the sample, whereas coverage attempts to provide information regarding what is left out of the sample.

Coverage, either absolute or relative, is a lower bound estimate for true coverage. Even with this limitation, it could be used for comparison purposes or as a measure of the progression or performance of a germplasm collection towards including maximum variability without duplicating accessions.

Acknowledgements The author would like to thank Dr. Marion Roder, who kindly shared the data set used in Huang et al. (2002) paper.

Author contributions Carlos Hernandez-Suarez developed the methodology, performed the simulations, wrote the manuscript.

Compliance with ethical standards

Conflict of interest The author declares no conflict of interest.

Appendix: Proof of properties of the coverage of several populations

1. If we select an individual at random from the population and then select one of its attributes X or Y, this attribute will be included in the sample with respective probabilities C_X and C_Y. Because the selected attribute is equally likely to be X or Y, the probability that the attribute selected is in the sample is $(C_X + C_Y)/2$.
2. If two populations of sizes N_1 and N_2 are mixed and a sample of size n is taken from the mix, the probability that an individual selected at random from the mixed population is represented in the sample is defined as the coverage of the sample, this follows from the fact that a randomly selected individual from the mix of populations

belongs to each initial population with respective probabilities $f = N_1/(N_1 + N_2)$ and $1 - f = N_2/(N_1 + N_2)$. It follows that there is no need to mix both populations as long as each population is sampled with sample sizes $n_1 = N_1/(N_1 + N_2)$ and n_2, respectively.

3. Suppose we have two populations 1 and 2 of sizes N_1 and N_2, respectively, where $N_1 = N_2$. Suppose we take a sample of size n from each population and let C represent the coverage of the mixed sample of size $2n$. By property 2, C can be interpreted as the probability that a random individual selected from the mix of both populations is represented in the sample, i.e., the *absolute coverage*. Now suppose that the size of population 2 is increased by a factor of k, where $k > 1$, keeping the relative frequency of alleles fixed. Clearly, the previous interpretation of the coverage (absolute coverage) no longer holds because an individual selected randomly from the mixture of populations 1 and 2 is k times more likely to come from population 2. But if we can guarantee that the individual selected is equally likely to come from either population, then the probability that this individual is already represented in the sample is still C. The restriction imposed by requiring that it must be equally likely that the individual comes from either population defines the *relative coverage*. It follows that if we have two populations of general sizes N_1 and N_2, $N_1 \neq N_2$, and take a sample of the same size n from each population, the coverage of the sample mix follows the definition of *relative coverage*.
4. This property follows from properties of random sampling.

References

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

Chao A (1981) On estimating the probability of discovering a new species. Ann Stat 9(6):1339–1342

Chao A, Lee SM (1992) Estimating the number of classes via sample coverage. J Am Stat Assoc 87(417):210–217

Chao A, Lee SM (1993) Estimating population size for continuous-time capture-recapture models via sample coverage. Biom J 35(1):29–45

Darwin C (1866) On the origin of species by means of natural selection: or the preservation of favoured races in the struggle for life. John Murray, London

Esty WW (1982) Confidence intervals for the coverage of low coverage samples. Ann Stat 10(1):190–196

Esty WW (1983) A normal limit law for a nonparametric estimator of the coverage of a random sample. Ann Stat 11(3):905–912

Esty W (1985) Estimation of the number of classes in a population and the coverage of a sample. Math Sci 10:41–50

Esty WW (1986) The efficiency of good's nonparametric coverage estimator. Ann Stat 14(3):1257–1260

Good IJ (1953) The population frequencies of species and the estimation of population parameters. Biometrika 40(3–4):237–264

Good I, Toulmin G (1956) The number of new species, and the increase in population coverage, when a sample is increased. Biometrika 43(1–2):45–63

Harris B (1959) Determining bounds on integrals with applications to cataloging problems. Ann Math Stat 30(2):521–548

Huang SP, Weir B (2001) Estimating the total number of alleles using a sample coverage method. Genetics 159(3):1365–1373

Huang X, Börner A, Röder M, Ganal M (2002) Assessing genetic diversity of wheat (*triticum aestivum* l.) germplasm using microsatellite markers. Theor Appl Genet 105(5):699–707

Knott M (1967) Models for cataloguing problems. Ann Math Stat 38(4):1255–1260

Lee SM, Chao A (1994) Estimating population size via sample coverage for closed capture-recapture models. Biometrics 50(1):88–97

Lo SH (1992) From the species problem to a general coverage problem via a new interpretation. Ann Stat 20(2):1094–1109

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci 70(12):3321–3323

Robbins HE (1968) Estimating the total probability of the unobserved outcomes of an experiment. Ann Math Stat 39(1):256–257

Starr N (1979) Linear estimation of the probability of discovering a new species. Ann Stat 7(3):644–652

van Hintum TJ, Brown AHD, Spillane C, Hodkin T (2000) Core collections of plant genetic resources (IPGRI Technical Bulletin No. 3., Rome, Italy, 2000)

Zhang C-H, Zhang Z (2009) Asymptotic normality of a nonparametric estimator of sample coverage. Ann Stat 37:2582–2595