Shannon diversity index: a call to replace the original Shannon's formula with unbiased estimator in the population genetics studies

Maciej K. Konopiński

Institute of Nature Conservation, Polish Academy of Sciences, Kraków, Poland

ABSTRACT

Background. The Shannon diversity index has been widely used in population genetics studies. Recently, it was proposed as a unifying measure of diversity at different levels—from genes and populations to whole species and ecosystems. The index, however, was proven to be negatively biased at small sample sizes. Modifications to the original Shannon's formula have been proposed to obtain an unbiased estimator.

Methods. In this study, the performance of four different estimators of Shannon index—the original Shannon's formula and those of Zahl, Chao and Shen and Chao et al.—was tested on simulated microsatellite data. Both the simulation and analysis of the results were performed in the R language environment. A new R function was created for the calculation of all four indices from the genind data format.

Results. Sample size dependence was detected in all the estimators analysed; however, the deviation from parametric values was substantially smaller in the derived measures than in the original Shannon's formula. Error rate was negatively associated with population heterozygosity. Comparisons among loci showed that fast-mutating loci were less affected by the error, except for the original Shannon's estimator which, in the smallest sample, was more strongly affected by loci with a higher number of alleles. The Zahl and Chao et al. estimators performed notably better than the original Shannon's formula.

Conclusion. The results of this study show that the original Shannon index should no longer be used as a measure of genetic diversity and should be replaced by Zahl's unbiased estimator.

Subjects Biodiversity, Evolutionary Studies, Genetics, Population Biology

Keywords Genetic diversity, Shannon index, Coalescent simulations, Measures of genetic variation, Sample size effect, Statistical genetics

Submitted 17 December 2019
Accepted 29 May 2020
Published 29 June 2020

Corresponding author
Maciej K. Konopiński,
konopinski@iop.krakow.pl

Academic editor
Patricia Gandini

Additional Information and Declarations can be found on page 12

DOI **10.7717/peerj.9391**

INTRODUCTION

The Shannon diversity index (*Shannon, 1948*), also known as the Shannon-Wiener index, Shannon entropy or, incorrectly, the Shannon-Weaver index (*Spellerberg & Fedor, 2003*), has been used to estimate genetic diversity in numerous studies. It can be utilised to describe variation at multiple levels of genetic organisation from single nucleotide polymorphisms (SNP), through whole species or larger taxonomic units to ecosystems. Due to its additive properties (*Jost, 2007*), the Shannon index has recently been postulated

as a unifying measure for the partitioning of diversity at those levels (*Gaggiotti et al., 2018*; *Sherwin, 2018*). Additionally, *Sherwin et al. (2017)* showed its potential utility in genomic studies. Several population genetics programs and *R* packages calculate Shannon's *H,* e.g., GenAlEx (*Peakall & Smouse, 2012*), *DartR* (*Gruber et al., 2018*), *SpadeR* (*Chao, Ma & Chiu, 2016*), *vegan* (*Oksanen et al., 2019*), *poppr* (*Kamvar, Tabima & Grünwald, 2014*), GenoDive (*Meirmans & Tienderen, 2004*), *HierDpart* (*Qin, 2019*) and the index is still in use (e.g., *O'Reilly et al., 2018*; *Zhang et al., 2019*; *Chang, Huang & Liao, 2019*).

The measure was initially developed within information theory (*Shannon, 1948*) but it was soon adopted in studies on species diversity (e.g., *Good, 1953*; *Margalef, 1957*; *Crowell, 1961*) and in population genetics (*Jain et al., 1975*). In principle, Shannon's *H* takes into account the proportion of each species in an ecosystem studied; hence, it gives a better description of an ecosystem's diversity than a plain number of species. When the number of species is equal in two locations, the index is capable of distinguishing between sites dominated by a single or only a few predominant species and those where each species has comparable input to the whole biodiversity (*Margalef, 1957*). Similarly, in population genetics studies, Shannon's *H* allows distinguishing the level of variation between populations with the same number of alleles, when in some populations loci are dominated by only a few common alleles while in others variation is contributed more evenly by all alleles. The Shannon index is more sensitive to the loss of rare variants (e.g., due to genetic bottlenecks) than heterozygosity, and more informative than allelic richness or a plain number of alleles (*Sherwin et al., 2017*).

In the original formula of the Shannon index developed within information theory, it is assumed a researcher is capable of counting all words or letters in a text studied. In biological studies, however, researchers depend on a sample from the population and use it as a proxy for the population parameters. The index changes rapidly when the number of low-frequency occurrences grows, while their number depends on the sample size (*Basharin, 1959*; *Chao & Shen, 2003*). The probability that all the alleles are sampled falls dramatically when the sample size is small. At the same time, in small samples, the lack of some allele inflates the frequencies of the alleles that have been sampled. As a solution to that, a few unbiased estimators of *H* have been proposed. The methods use jack-knifing (*Zahl, 1977*), rarefaction (*Chao & Shen, 2003*) or the Good-Turing frequency formula (*Good, 1953*; *Chao, Wang & Jost, 2013*) to account for unsampled components of the system (i.e., species or alleles). Although the issue of sample size in population genetics has been addressed in several studies (e.g., *Marquez-Sanchez & Hallauer, 1970*; *Gorman & Renzi, 1979*; *Chakraborty, 1992*; *El Mousadik & Petit, 1996*; *Leberg, 2002*; *Pruett & Winker, 2008*), the dependence of Shannon's diversity estimation on sample size has never been thoroughly discussed with regard to genetic data. *Bashalkhanov, Pandey & Rajora (2009)* noticed an increase in the deviation of Shannon's *H* at small sample sizes; however, the only solution they suggested was increasing the sample size to 60–90 genotypes. However, *Sherwin et al. (2017)* pointed out that although methods for sampling correction of *H* exist, unbiased estimators remain rarely applied in population genetics studies.

The aim of this study was to assess the effect of sample size and locus properties (mutation rate and the maximum possible number of allelic states) on the estimation of Shannon's

Table 1 **Combination of mutation rates and a maximum number of allelic states in the 24 loci simulated in *fastsimcoal2*.**

		Max. number of alleles					
		3	6	9	12	15	20
Mutation rate	0.0001	L01	L02	L03	L04	L05	L06
	0.0002	L07	L08	L09	L10	L11	L12
	0.0005	L13	L14	L15	L16	L17	L18
	0.001	L19	L20	L21	L22	L23	L24

original index (H_{MLE}) and its three unbiased estimators proposed by (*Zahl, 1977*) (H_Z), (*Chao & Shen, 2003*) (H_{CS}), (*Chao, Wang & Jost, 2013*) (H_{Chao}). The performance of the four indices was tested extensively on data generated using coalescent simulations. The relative effects of sample size, locus properties and population diversity were analysed with a Generalised Linear Model. A wrapper R function was written to allow for estimation of the four indices directly on *adegenet*'s 'genind' objects.

MATERIALS & METHODS

Analyses were conducted in R 3.6.2 (*R Development Core Team, 2009*). Coalescent simulations as implemented in *fastsimcoal2* ver. 2.6 (*Excoffier et al., 2013*) were used to generate populations differing in levels of genetic variation due to their demographic histories. The program was called from within the R environment using the *strataG* package ver. 2.0.2 command *fastsimcoal* (*Archer, Adams & Schneiders, 2017*). Twenty-four microsatellite loci with four different mutation rates (0.0001, 0.0002, 0.0005 and 0.001 mutations per generation) and six different maximum numbers of alleles (3, 6, 9, 12, 15 and 20 alleles) were simulated (Table 1). The model assumed a large population of 10,000 diploid individuals divided into four populations containing 10,000 individuals each. Three of those underwent bottlenecks of different sizes (20, 50 and 500 individuals in populations P_{20}, P_{50} and P_{500} respectively) while the fourth, the control population (P_C remained at a stable size of 10,000 individuals. Each of the bottlenecked populations after 20 generations recovered to the original size of 10,000 and was simulated for another 20 generations until time T0 when samples equalling the whole populations were saved both from bottlenecked and control populations. Reference parametric values of four different Shannon's estimators were calculated for each total population sample:

- H_{MLE}—the maximum likelihood estimate of H based on the original Shannon's formula (*Shannon, 1948*, Theorem 2)—probably the most widely used in population genetics;
- H_{CS}—unbiased estimator proposed by *Chao & Shen* (*2003*, equation 8);
- H_{Chao}—unbiased estimator proposed by *Chao, Wang & Jost* (*2013*, equation 7);
- H_Z—jackknife estimate proposed by *Zahl* (*1977*, equation 7).

Additionally, expected heterozygosity (He) was calculated as a reference measure of the genetic variation of the total populations.

The four H estimators are calculated using the function *Diversity* in R package *SpadeR* (*Chao, Ma & Chiu, 2016*). Initial attempts to apply the function to the simulated set

showed that the function often fails due to a problem with another nested function that estimates species richness. Moreover, the function has to be run separately for each locus to obtain locus-specific *H*. Therefore, a generic function (*ShannonGen*) was written using the formulas from the *shannon_index* function nested in *Diversity*. The function takes *genind* objects as the input and transforms them into abundance data, which is required by functions taken from *shannon_index*. *ShannonGen* returns a list containing user-selected estimators of H for all loci and populations included in the input object. The function can be acquired from GitHub (https://github.com/konopinski/Shannon/).

The Shannon diversity index estimators were calculated for samples of $Ns = 5$, 20, 80 and 200 genotypes drawn randomly without replacement from the simulated populations, The numbers of samples used were selected to represent four sampling scenarios:

- limited availability of samples—a situation often faced in studies on rare or elusive animals—$Ns = 5$ samples;
- a minimum acceptable number of samples as suggested by *Pruett & Winker (2008)* commonly occurring in population genetics studies—$Ns = 20$ samples;
- optimal sampling, according to *Bashalkhanov, Pandey & Rajora (2009)*—$Ns = 80$ samples.
- a very large sample with presumably low sampling error—$Ns = 200$ samples.

Additionally, the parametric values were obtained from the simulated populations.

Sampling variance of each H estimator was assessed using 500 sets of samples randomly drawn from each of the simulated populations. The standard deviations (*SD*) of the results were calculated for each sample size and demographic scenario. The *SD* values distributions were compared pairwise between metrics. The level of significance was assessed based on 100,000 comparisons of the randomly drawn pairs of the *SD* values.

For each H estimator, a relative bias was calculated as $rB = ((\hat{H} - H)/H)$, where $\hat{H}$ is the estimate of a given metric in a sample, and H is the parametric value of a given estimator. The bias was estimated only once per each metric/population/iteration (i.e., 16,000 times). The error of the metric was estimated as a mean relative squared error (MRSE):

$MRSE = \frac{1}{500}\sum_{i=1}^{500}\frac{(\hat{H}_i - H)^2}{H}$, where $\hat{H}_i$ is an estimate of a metric in the i-th sample of the 500 resamplings that were carried out to estimate the mean.

To explore the factors influencing the errors of the analysed indices, the generalised linear model, *glm*, a function from R's *stats* package, was used. The model included *MRSE* as the response and four explanatory variables: H estimator, sample size, population gene diversity and locus. To avoid overparametrisation, the GLM analyses were performed in two steps. Firstly, *MRSE* of mean H values over the 24 simulated loci were provided to the model as an effect; secondly, *MRSE* calculated for each locus separately was used.

Because the relation between sample size and *MRSE* is asymptotic, and the effect size may depend on arbitrarily selected number of samples, the values were provided to the model as categorical values (factors). The best-fitting model was selected based on the Akaike Information Criterion (AICc) as implemented in the *model.sel* function from the R package *MuMIn* (*Bartoń, 2019*). The GLM results were analysed using the *Anova* function from the package *car* (*Fox & Weisberg, 2019*). The effects of the explanatory

Table 2 Summary of the simulations. Minimum, maximum and median values of the H estimators and the Nei's gene diversity (Hs) calculated for the four simulated demographic scenarios.

Population		H_{MLE}	H_Z	H_{CS}	H_{Chao}	Hs
	min	1.5159	1.5161	1.5160	1.5161	0.6996
P_C	median	1.6706	1.6708	1.6706	1.6708	0.7568
	max	1.7834	1.7836	1.7834	1.7836	0.7889
	min	1.4220	1.4221	1.4221	1.4221	0.6764
P_{500}	median	1.6075	1.6077	1.6075	1.6077	0.7428
	max	1.7409	1.7411	1.7409	1.7411	0.7795
	min	0.9940	0.9941	0.9941	0.9941	0.5347
P_{50}	median	1.2007	1.2009	1.2008	1.2009	0.6266
	max	1.3441	1.3443	1.3442	1.3443	0.6843
	min	0.5927	0.5928	0.5929	0.5928	0.3179
P_{20}	median	0.8281	0.8282	0.8282	0.8282	0.4731
	max	1.0077	1.0079	1.0079	1.0079	0.5734

variables were visualised using the R package *effects* (*Fox & Weisberg, 2018*; *Fox & Weisberg, 2019*). Tukey's Honest Significant Difference (Tukey's HSD; *Tukey, 1977*) method was used to test whether the differences between the factors were significant. The function *glht* from the R package *multcomp* (*Hothorn, Bretz & Westfall, 2008*) was used to perform the analysis, while *cld* was used to summarise results and present them as compact letter displays (*Piepho, 2004*). The script used for simulations is deposited at Github (https://github.com/konopinski/Shannon/).

RESULTS

Each metric was estimated altogether 192 096 000 times in 24 loci, 4 populations (P_C, P_{500}, P_{50}, P_{20}), 5 sample sizes (i.e., 5, 20, 80 and 200 genotypes and for the whole simulated population to obtain parametric values), 500 randomizations and 1000 simulation repetitions. Mean expected heterozygosities calculated for the simulated total populations ranged from $He = 0.318$ to $He = 0.789$ with median $He = 0.683$. The parametric values of the four H indices were similar within each of the simulated demographic scenario both in terms of their median values and their ranges (Table 2).

Attempts to estimate the sampling variance of H_{CS} failed in 1,689 out of 16,000 resampling attempts, i.e., roughly 10%. The problem occurred only in the smallest simulated samples ($Ns = 5$) and only in the most variable populations: 937 in P_C and 752 in P_{500}. Similarly, mean H_{CS} could not be estimated in 10 cases in the 16 000 population sampling simulations extracted from the whole set to estimate bias. The problem occurred only in the most variable populations (7 failures in P_C and 3 in P_{500}) and in the smallest sample size. The loci that caused the problem were those simulated with a large number of possible allelic states (12, 15 and 20 alleles) and the problem was more frequent in the loci with higher mutation rates (Table S1). Due to the large proportion of missing data, the estimates of H_{CS} from the populations P_C and P_{500} simulated with the smallest sample size $Ns = 5$ were excluded from the sampling variance comparisons, and the 10 samples

that failed at estimation of H_{CS} in the simulations of population sampling were excluded from assessment of the performance of the H indices.

The standard deviation of the results distribution from the repeated sampling depended on sample size, demographic scenario and H estimator (Fig. 1). Standard deviations of H_{MLE} were significantly lower in all pairwise comparisons (Table 3). Among the remaining metrics, the estimates of H_Z had significantly narrower distribution then H_{Chao} in control populations (P_C) and the populations that underwent a bottleneck of 500 genotypes (P_{500}). In larger sample sizes, $Ns = 20$, 80 and 200, the distributions of standard deviations of H_Z, H_{Chao} and H_{CS} were not significantly different in any of the pairwise comparisons.

The median relative bias (rB) of the H estimates averaged over the 24 loci was inversely associated with the sample size in all cases (Fig. 2, Table 4). As compared to other H estimators at all sample sizes, the strongest negative departure from parametric values (i.e., calculated from the total population) was observed in H_{MLE} estimates. Among the remaining three H estimates, H_{Chao} and H_Z were the least biased (Table 4). Except for the H_{MLE} estimates at sample sizes below 80 genotypes, 95% confidence intervals of H estimators always spanned the parametric value of the simulated data. The observed relative bias ranges were markedly wider in the smallest samples.

The analyses of relative error ($MRSE$) provided similar findings. Based on AICc summarised by MuMIn function, the gamma distribution of $MRSE$ with a log link function was used in the GLM. According to ANOVA test of the GLM results, all four factors—locus, metric, sample size and expected heterozygosity of the total population (He)—were significantly associated with the error levels ($p = 10^{-15}$). The median $MRSE$ of H estimators was negatively associated with the sample size. When compared to $Ns = 200$ genotypes, the slope of the relation, β, increased on decreasing sample size, from $\beta = 0.97$ for $Ns = 80$ genotypes, to $\beta = 4.46$ for $Ns = 5$ individuals ($p = 10^{-15}$). Tukey's HSD analysis of GLM results confirmed the differences between error levels among all the different sample sizes were significant for all metrics. The strongest effect of sample size on $MRSE$ was observed for H_{MLE} (Fig. 3). The remaining H estimators were markedly less affected by a small sample size with H_{CS} performing slightly worse than H_Z and H_{Chao}. Analysis of GLM results using Tukey's HSD showed that among all the metrics, H_{Chao} and H_Z were significantly less affected by error than the other two estimators at the majority of sample sizes (Table 5). Only at the smallest sample size, the difference between H_{Chao}, H_Z and H_{CS} was not significant. The error levels were also strongly negatively associated with the He of the total population ($\beta = -0.71$, $p = 10^{-15}$, Fig. 4). In the case of H_{MLE} and H_Z, the effect depended on the sample size, being, positive at the smaller sample sizes: $Ns \leq 80$ in H_{MLE} and Ns = 5 in H_Z.

In the second analysis, locus properties were tested. ANOVA of the GLM results showed that locus predictor was significantly associated with the $MRSE$ ($p = 10^{-15}$) in all metrics. The size of the effect depended on mutation rates, the maximum number of alleles (Fig. 5) and expected heterozygosity at a given locus (Figs. S1–S4). Mutation rates had a more substantial effect on $MRSE$ (line colors in Fig. 5) than the maximum number of allelic states at a locus (line types in Fig. 5; Table S1), except for H_{MLE} at the smallest sample size in which case the error increased at fast mutating loci with the number of allelic states possible

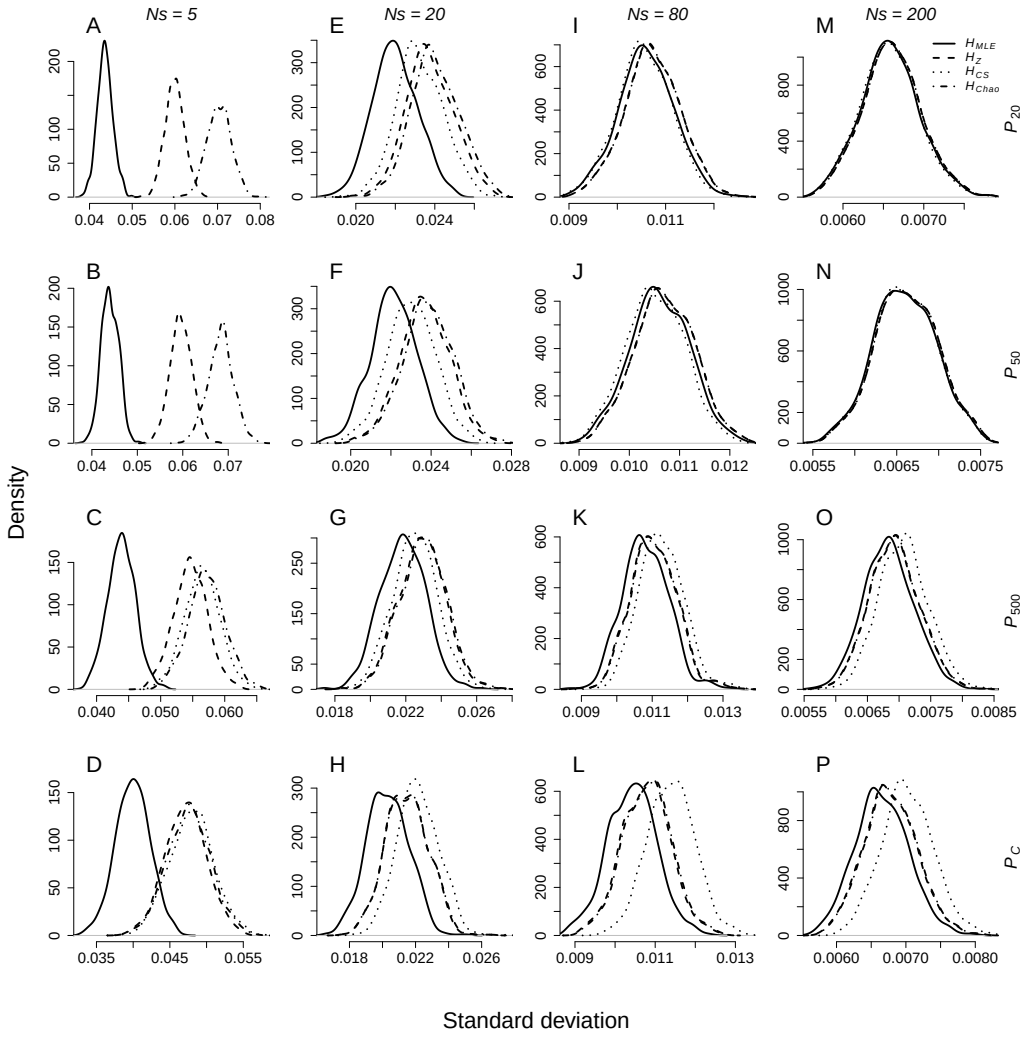

Figure 1 Density plots of SD distributions obtained through repeated sampling of *Ns* = 5, 20, 80 and 200 genotypes from each population representing different demographic scenario in each simulation. Sample sizes (columns): (A–D) *Ns* = 5, (E–H) *Ns* = 20, (I–L) *Ns* = 80, (M–P) *Ns* = 200. Demographic scenarios (rows): (A), (E), (I), (M) P_{20}, (B), (F), (J), (N) P_{50}, (C), (G) ,(K), (O) P_{500}, (D), (H), (L), (P) P_C.
Full-size DOI: 10.7717/peerj.9391/fig-1

(L13-L24). The error level of H_Z, H_{Chao} and H_{MLE} decreased with the total population's *He* of the locus (Figs. S1–S4). The slope of this relation was steeper in loci with fewer allelic states allowed (e.g., L01, L07, L13 and L19). In case of H_{CS} the MRSE at the smallest sample size, *Ns* = 5, was positively correlated with *He* at fast mutating loci with more allelic states allowed, while at the remaining sample sizes the relation was negative.

DISCUSSION

The problem of sample depencence of genetic diversity measures has been observed in estimation the number of alleles in populations; to tackle the issue, the rarefaction method

Table 3 Sampling variance of the H estimators calculated for $Ns = 5$ genotypes. Mean SD of each estimator in each demographic scenario, and the p-values of pairwise comparisons of the SD distributions.

Population	Metric	Mean SD of metric	p-values		
			H_Z	H_{CS}	H_{Chao}
P_C	H_{MLE}	0.0436	10^{-5}	–	10^{-5}
	H_Z	0.0599		–	0.0038
	H_{CS}	–			–
	H_{Chao}	0.0704			
P_{500}	H_{MLE}	0.0439	10^{-5}	–	10^{-5}
	H_Z	0.0596		–	0.0191
	H_{CS}	–			–
	H_{Chao}	0.0683			
P_{50}	H_{MLE}	0.0439	0.0001	0.0004	0.0016
	H_Z	0.0547		0.6397	0.5189
	H_{CS}	0.0565			0.8547
	H_{Chao}	0.0572			
P_{20}	H_{MLE}	0.0399	0.0418	0.0345	0.0498
	H_Z	0.0472		0.8478	0.9290
	H_{CS}	0.0480			0.9344
	H_{Chao}	0.0479			

was proposed for estimating allelic richness instead of the plain number of alleles (*El Mousadik & Petit, 1996*; *Kalinowski, 2004*). Allelic richness quickly gained attention and became a popular estimator of genetic variation. On the other hand, the advances in Shannon diversity index estimation proposed by *Zahl (1977)*, *Chao & Shen (2003)* and *Chao, Wang & Jost (2013)* remain unnoticed in population genetics studies.

The results of the present study confirm what has been known from species diversity studies (*Pielou, 1966*), that the original Shannon index, H_{MLE} is strongly dependent on sample size. This phenomenon is stronger both in more genetically variable populations and in more variable loci, particularly at small sample sizes. It is not possible to estimate the true value of Shannon index using H_{MLE} when the sample size is small. The most likely explanation for it is that in small samples, the probability that all alleles have been captured is lower than when the sample is large. The so-called *nearly unbiased* estimators also showed some level of bias, even in samples as big as 200 genotypes; however, the difference was negligible (less than 1‰), and the parametric values were well within the 95% confidence intervals of results from the simulated samples. Those measures performed better at more diverse loci and populations, where both *rB* and *MRSE* were, on average, smaller. On the other hand, the error levels of unbiased metrics were more dependent on mutation rates rather than the maximum possible number of allelic states at the locus, which may suggest that the occurrence of low-frequency alleles stemming from numerous mutation events has a stabilising effect on those estimators. Using H_{CS} bears a high risk of encountering

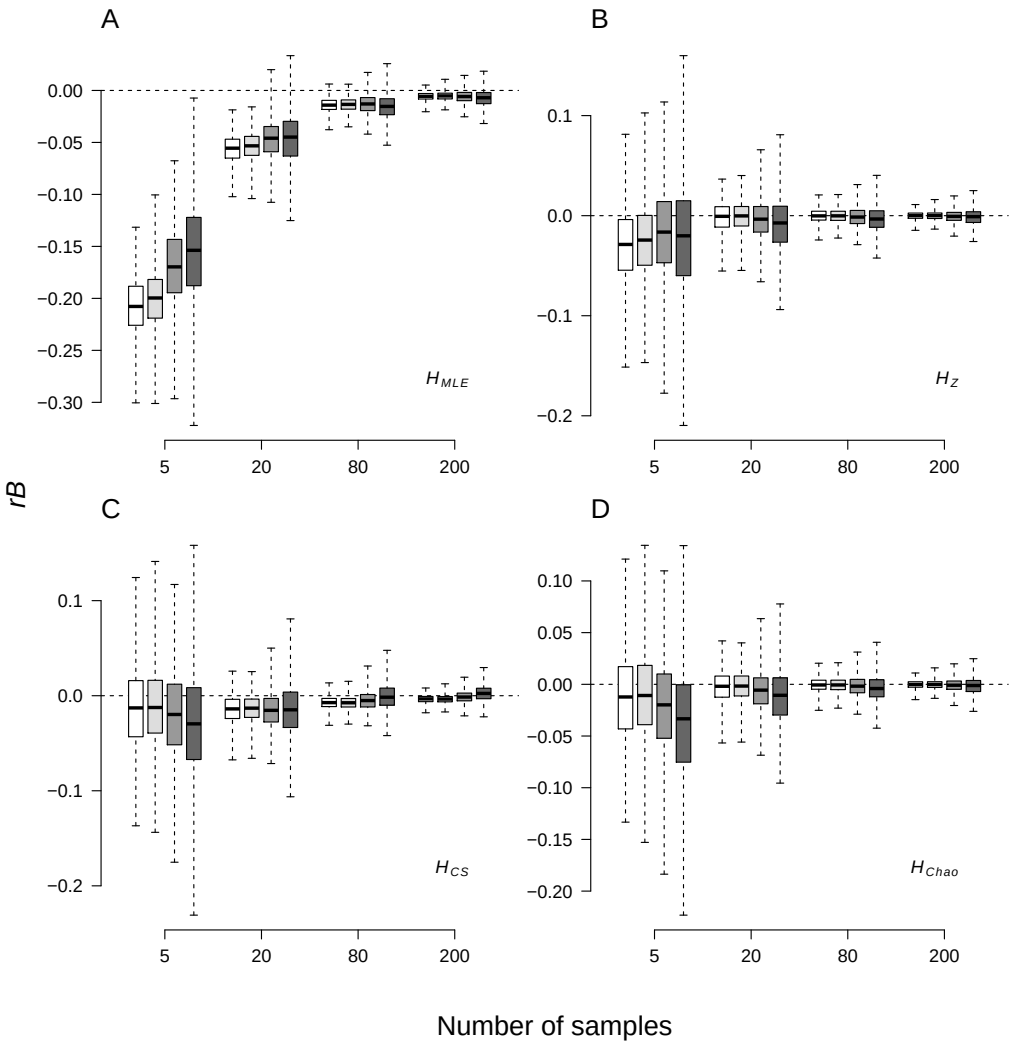

Figure 2 Box-whisker plot of relative bias (*rB*) of the four Shannon *H* estimators in all the simulated demographic scenarios and sample sizes. **(A)** H_{MLE}, **(B)** H_{MLE}, **(C)** H_{CS}, **(D)** H_{Chao}. The whiskers represent the range of maximum and minimum values, the top and bottom of the boxes represent the 75% and 25% quartiles. White—non-bottlenecked, control population (P_C), light-grey—bottleneck of 500 individuals (P_{500}), dark-grey—bottleneck of 50 individuals (P_{50}), anthracite—bottleneck of 20 individuals (P_{20}).
Full-size DOI: 10.7717/peerj.9391/fig-2

problems when the sample sizes are small, and the level of genetic variation is high, which makes this metric hardly useful in population genetics studies. The performance of this metric is similar to that of the other two unbiased *H* estimators; however, it proved to be less precise than H_{Chao} and H_Z. The analysis of the simulations results suggest the Zahl jackknife estimator H_Z as the most suitable estimator of Shannon diversity index to describe variation at multiallelic loci such as microsatellites. Among the three unbiased estimators, H_Z had the lowest sampling variance and the smallest bias, which also results in the lowest error as compared to the other metrics. For this reason, H_Z should replace traditionally used H_{MLE} in population genetics studies using microsatellites.

Table 4 Minimum, maximum and mean values of the relative bias (*rB*) of all Shannon *H* estimators and sample sizes tested.

Sample size		H_{MLE}	H_Z	H_{CS}	H_{Chao}
	min	−0.3222	−0.2097	NA	−0.2232
5	median	−0.1872	−0.0229	NA	−0.0186
	max	−0.0073	0.1599	NA	0.1344
	min	−0.1252	−0.0939	−0.1063	−0.0956
20	median	−0.0508	−0.0024	−0.0142	−0.0041
	max	0.0335	0.0809	0.0808	0.0777
	min	−0.0527	−0.0423	−0.0420	−0.0424
80	median	−0.0139	−0.0009	−0.0059	−0.0013
	max	0.0258	0.0403	0.0478	0,0407
	min	−0.0318	−0.0259	−0.0222	−0.0261
200	median	−0.0058	−0.0004	−0.0021	−0.0005
	max	0.0185	0.0250	0.0296	0.0248

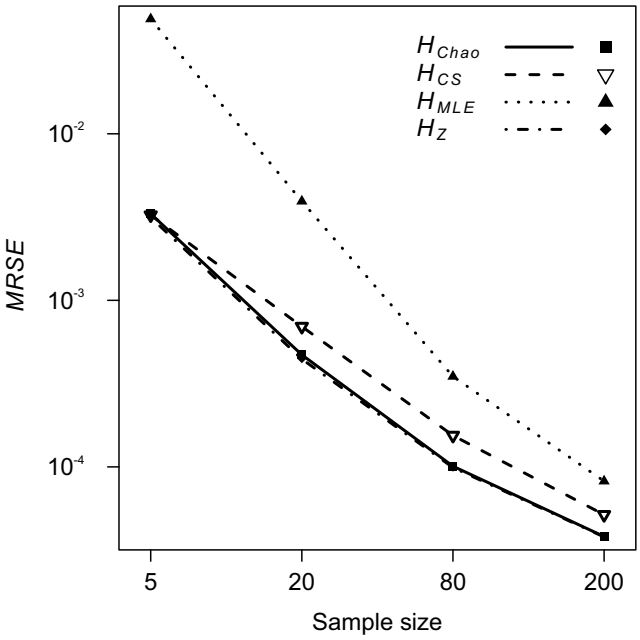

Figure 3 GLM results: the effect of the sample size predictor on the mean relative squared error (*MRSE*) of the four Shannon *H* estimators.

Full-size DOI: 10.7717/peerj.9391/fig-3

Although all the results presented here were derived from simulating microsatellite loci, the pattern of differences among them shows that the estimates of the Shannon index for less variable loci are more error-prone than for multi-allelic markers. However, as SNP markers are mostly biallelic and the *H* estimates might be more affected by error at individual loci, the effect of the errors averaged over a large number of loci usually used in SNP-based studies may become negligible. Further simulation and empirical tests are

Table 5 Compact letter display of Tukey HSD *post hoc* test of all pair-wise comparisons between the effects of the H estimators on *MRSE*. Four independent tests were performed on data with fixed sample sizes.

Sample size	Estimator			
	H_{MLE}	H_Z	H_{CS}	H_{Chao}
5	d	a	b	c
20	d	a	c	b
80	d	a	c	b
200	c	a	b	a

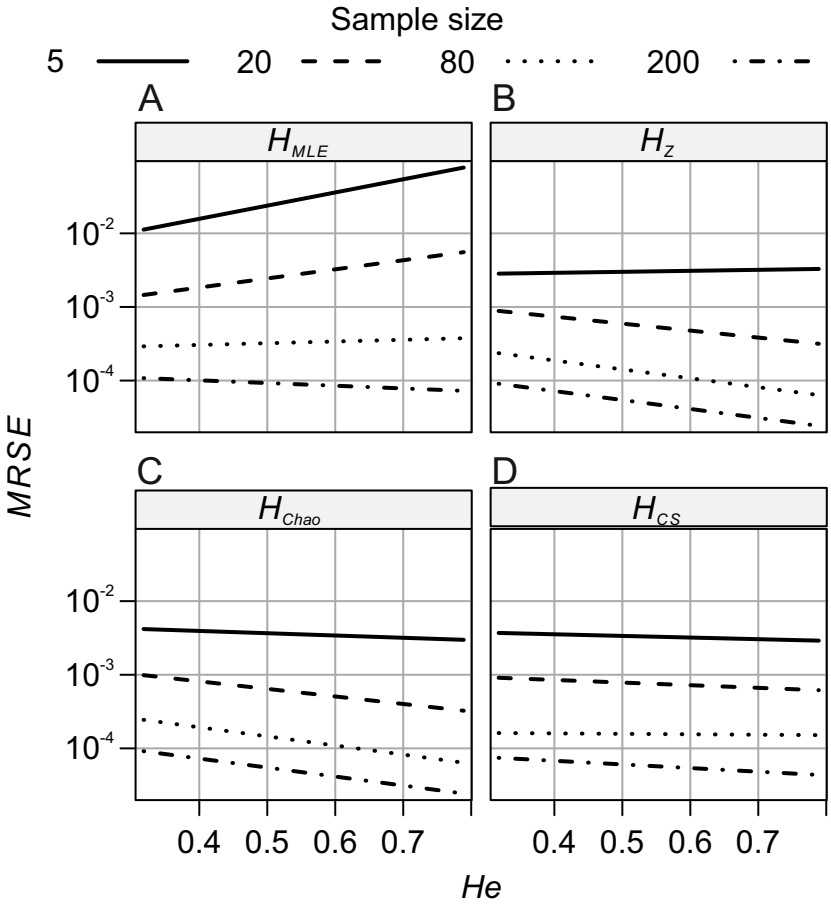

Figure 4 GLM results: the effects of total population's expected heterozygosity (*He*) and the sample size (*Ns*) predictors on the mean relative squared error (*MRSE*) of (A) H_{MLE}, (B) H_Z, (C) H_{CS} and (D) H_{Ch}.

Full-size DOI: 10.7717/peerj.9391/fig-4

necessary to investigate the performance of the Shannon index in SNP loci. While the cost NGS analyses has dropped significantly in recent years, and the present computational power enables analyses of a large amount of data, the problem of small sample sizes in genomic studies remains critical in studies of vulnerable or elusive species, i.e., the cases where the Shannon index is still widely used.

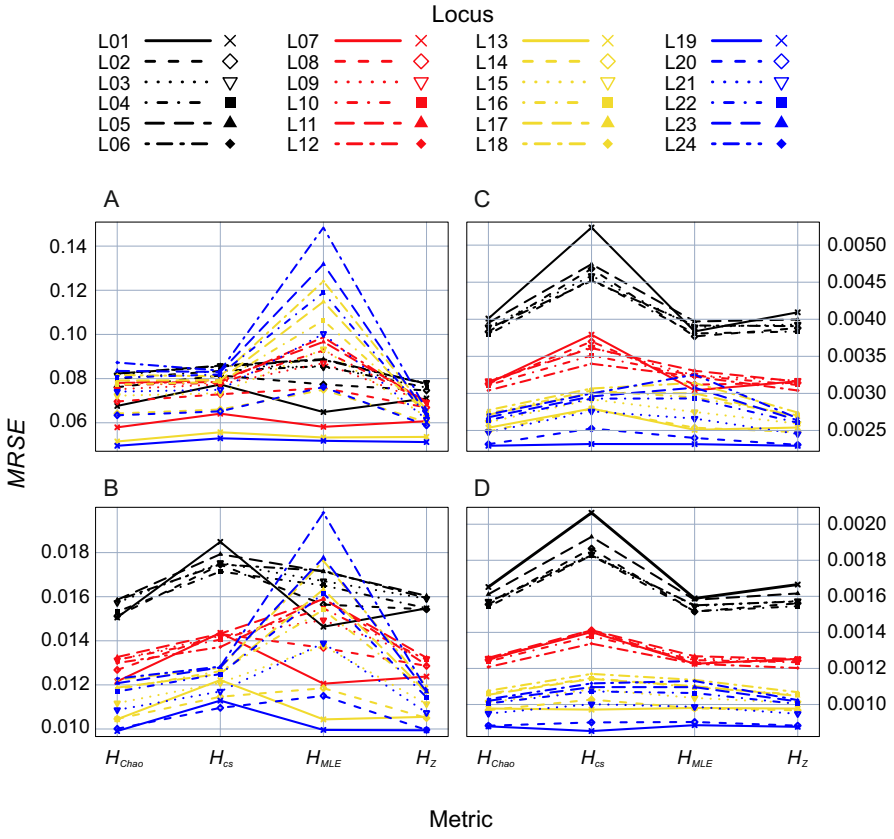

Figure 5 **GLM results: locus (L01–L24) effects on mean relative squared error (*MRSE*) of the four Shannon *H* estimators at four sample sizes (A) *Ns* = 5, (B) *Ns* = 20, (C) *Ns* = 80 and (D) *Ns* = 200.**
Full-size DOI: 10.7717/peerj.9391/fig-5

ACKNOWLEDGEMENTS

I would like to thank Kamil Bartoń, Aleksandra Biedrzycka, Adam Ćmiel and Piotr Skórka for fruitful discussions and valuable comments they made on the manuscript. I would like to thank the editor (Patricia Gandini), Oscar Gaggiotti and the two anonymous referees for help on improving this manuscript.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The work was entirely funded by the Institute of Nature Conservation, Polish Academy of Sciences. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the author:
Institute of Nature Conservation, Polish Academy of Sciences.

Competing Interests

The author declares he has no competing interests.

Author Contributions

- Maciej K. Konopiński conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The code used for simulations is available as a Supplemental File.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.9391#supplemental-information.

REFERENCES

Archer FI, Adams PE, Schneiders BB. 2017. stratag: an r package for manipulating, summarizing and analysing population genetic data. *Molecular Ecology Resources* **17**:5–11 DOI 10.1111/1755-0998.12559.

Bartoń K. 2019. MuMIn: multi-model inference. R package version 1.43.15. *Available at https://cran.r-project.org/package=MuMIn*.

Bashalkhanov S, Pandey M, Rajora OP. 2009. A simple method for estimating genetic diversity in large populations from finite sample sizes. *BMC Genetics* **10**:84 DOI 10.1186/1471-2156-10-84.

Basharin GP. 1959. On a statistical estimate for the entropy of a sequence of independent random variables. *Theory of Probability & Its Applications* **4**:333–336 DOI 10.1137/1104033.

Chakraborty R. 1992. Sample size requirements for addressing the population genetic issues of forensic use of DNA typing. *Human Biology* **64**:141–159.

Chang J-T, Huang B-H, Liao P-C. 2019. Genetic evidence of the southward founder speciation of Cycas taitungensis from ancestral C. revoluta along the Ryukyu Archipelagos. *Conservation Genetics* **20**:1045–1056 DOI 10.1007/s10592-019-01193-1.

Chao A, Ma KH, Chiu TCH. 2016. SpadeR: species-richness prediction and diversity estimation with R. R package version 0.1.1. *Available at https://cran.r-project.org/package=SpadeR*.

Chao A, Shen T-J. 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics* **10**:429–443 DOI 10.1023/A:1026096204727.

Chao A, Wang YT, Jost L. 2013. Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species. *Methods in Ecology and Evolution* **4**:1091–1100 DOI 10.1111/2041-210X.12108.

Crowell K. 1961. The effects of reduced competition in birds. *Proceedings of the National Academy of Sciences of the United States of America* **47**:240–243 DOI 10.1073/pnas.47.2.240.

El Mousadik A, Petit RJ. 1996. High level of genetic differentiation for allelic richness among populations of the argan tree [Argania spinosa (L.) Skeels] endemic to Morocco. *Theoretical and Applied Genetics* **92**:832–839 DOI 10.1007/BF00221895.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust demographic inference from genomic and SNP data. *PLOS Genetics* **9**:e1003905 DOI 10.1371/journal.pgen.1003905.

Fox J, Weisberg S. 2018. Visualizing fit and lack of fit in complex regression models with predictor effect plots and partial residuals. *Journal of Statistical Software* **87**:1–27 DOI 10.18637/jss.v087.i09.

Fox J, Weisberg S. 2019. *An R companion to applied regression*. Thousand Oaks: Sage.

Gaggiotti OE, Chao A, Peres-Neto P, Chiu C-H, Edwards C, Fortin M-J, Jost L, Richards CM, Selkoe KA. 2018. Diversity from genes to ecosystems: a unifying framework to study variation across biological metrics and scales. *Evolutionary Applications* **11**:1176–1193 DOI 10.1111/eva.12593.

Good IJ. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika* **40**:237–264 DOI 10.1093/biomet/40.3-4.237.

Gorman GC, Renzi J. 1979. Genetic distance and heterozygosity estimates in electrophoretic studies: effects of sample size. *Copeia* **1979**:242–249 DOI 10.2307/1443409.

Gruber B, Unmack PJ, Berry OF, Georges A. 2018. dartr: an r package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Molecular Ecology Resources* **18**:691–699 DOI 10.1111/1755-0998.12745.

Hothorn T, Bretz F, Westfall P. 2008. Simultaneous inference in general parametric models. *Biometrical Journal. Biometrische Zeitschrift* **50**:346–363 DOI 10.1002/bimj.200810425.

Jain SK, Qualset CO, Bhatt GM, Wu KK. 1975. Geographical patterns of phenotypic diversity in a world collection of durum wheats 1. *Crop Science* **15**:700–704 DOI 10.2135/cropsci1975.0011183X001500050026x.

Jost L. 2007. Partitioning diversity into independent alpha and beta components. *Ecology* **88**:2427–2439 DOI 10.1890/06-1736.1.

Kalinowski ST. 2004. Counting alleles with rarefaction: private alleles and hierarchical sampling designs. *Conservation Genetics* **5**:539–543 DOI 10.1023/B:COGE.0000041021.91777.1a.

Kamvar ZN, Tabima JF, Grünwald NJ. 2014. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2**:e281 DOI 10.7717/peerj.281.

Leberg PL. 2002. Estimating allelic richness: effects of sample size and bottlenecks. *Molecular Ecology* **11**:2445–2449 DOI 10.1046/j.1365-294X.2002.01612.x.

Margalef R. 1957. La teoría de la información en Ecología. *Memorias de la Real Academia de Ciencias y Artes de Barcelona* **32**:373–436.

Marquez-Sanchez F, Hallauer AR. 1970. Influence of sample size on the estimation of genetic variances in a synthetic variety of maize. I. Grain yield 1. *Crop Science* **10**:357–361 DOI 10.2135/cropsci1970.0011183X001000040012x.

Meirmans PG, Tienderen PHV. 2004. genotype and genodive: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes* **4**:792–794 DOI 10.1111/j.1471-8286.2004.00770.x.

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. 2019. vegan: community ecology package. R package version 2.5-6. *Available at https://cran.r-project.org/package=vegan*.

O'Reilly GD, Jabot F, Gunn MR, Sherwin WB. 2018. Predicting Shannon's information for genes in finite populations: new uses for old equations. *Conservation Genetics Resources* **12**:245–255 DOI 10.1007/s12686-018-1079-z.

Peakall R, Smouse PE. 2012. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics* **28**:2537–2539 DOI 10.1093/bioinformatics/bts460.

Pielou EC. 1966. Shannon's formula as a measure of specific diversity: its use and misuse. *The American Naturalist* **100**:463–465 DOI 10.1086/282439.

Piepho H-P. 2004. An algorithm for a letter-based representation of all-pairwise comparisons. *Journal of Computational and Graphical Statistics* **13**:456–466 DOI 10.1198/1061860043515.

Pruett CL, Winker K. 2008. The effects of sample size on population genetic diversity estimates in song sparrows Melospiza melodia. *Journal of Avian Biology* **39**:252–256 DOI 10.1111/j.0908-8857.2008.04094.x.

Qin X. 2019. HierDpart: partitioning hierarchical diversity and differentiation across metrics and scales, from genes to ecosystems. R package version 0.5.0. *Available at https://cran.r-project.org/package=HierDpart*.

R Development Core Team. 2009. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Shannon CE. 1948. A mathematical theory of communication. *Bell System Technical Journal* **27**:379–423 DOI 10.1002/j.1538-7305.1948.tb01338.x.

Sherwin WB. 2018. Entropy, or information, unifies ecology and evolution and beyond. *Entropy* **20**:727 DOI 10.3390/e20100727.

Sherwin WB, Chao A, Jost L, Smouse PE. 2017. Information theory broadens the spectrum of molecular ecology and evolution. *Trends in Ecology & Evolution* **32**:948–963 DOI 10.1016/j.tree.2017.09.012.

Spellerberg IF, Fedor PJ. 2003. A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon–Wiener' index. *Global Ecology and Biogeography* **12**:177–179 DOI 10.1046/j.1466-822X.2003.00015.x.

Tukey JW. 1977. *Exploratory data analysis*. Reading: Addison-Wesley.

Zahl S. 1977. Jackknifing an index of diversity. *Ecology* **58**:907–913 DOI 10.2307/1936227.

Zhang X, Su H, Yang J, Feng L, Li Z, Zhao G. 2019. Population genetic structure, migration, and polyploidy origin of a medicinal species Gynostemma pentaphyllum (Cucurbitaceae). *Ecology and Evolution* **9**:11145–11170 DOI 10.1002/ece3.5618.