Received: 15 November 2021 | Revised: 3 June 2022 | Accepted: 27 July 2022

DOI: 10.1111/1755-0998.13707

MOLECULAR ECOLOGY RESOURCES WILEY

NEWS AND VIEWS

COMMENT

Average weighted nucleotide diversity is more precise than pixy in estimating the true value of π from sequence sets containing missing data

Maciej K. Konopiński

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

Correspondence
Maciej K. Konopiński, Institute of Nature Conservation Polish Academy of Sciences, al. Mickiewicza 33, 31-120 Kraków, Poland.
Email: konopinski@iop.krakow.pl

Funding information
Institute of Nature Conservation Polish Academy of Sciences, Grant/Award Number: Statutory funds

Handling Editor: Paul A. Hohenlohe

Abstract

Nucleotide diversity remains an important statistic in population genetic/genomic studies. Although recent advances in massive sequencing make generating sequence data sets cheaper and faster, currently used technologies often introduce substantial amounts of missing nucleotides in their output. A novel method of estimating π from data sets containing missing data – pixy - has also recently been proposed. In this study, the pixy estimator, π_{pixy}, was compared to average weighted nucleotide diversity, π_W. The estimators were tested both on sequences simulated in fastsimcoal and real sequence sets. Both sets were modified by random insertion of missing nucleotides. Weighted nucleotide diversity performed better in all pairwise comparisons. It was characterized by a smaller error and a narrower distribution of the results. π_{pixy} tends to overestimate the nucleotide diversity when both the proportion of missing data and the level of variation is low. Of the two estimators, only π_W estimated the true nucleotide diversity in a part of the simulations. A simple formula for estimating π_W allows for easy integration of the estimator in packages such as pixy, which would allow obtaining more precise estimates of nucleotide diversity either in a sliding window or for discrete genomic regions.

KEYWORDS
bioinfomatics/phyloinfomatics, genetic variation, missing data, next-generation sequencing, nucleotide diversity, statistics

1 | INTRODUCTION

Nucleotide diversity, π, was proposed by Nei and Li (1979), who defined it as "the average number of nucleotide differences per site between two randomly chosen DNA sequences". Since then, the statistic has been used routinely in genetic and genomic diversity studies and remains one of the central measures for describing sequence variation (e.g., Loeuille et al., 2021; Petit-Marty et al., 2021). The level of nucleotide diversity reflects both the functional importance of the DNA region and the demographic history of the species (Charlesworth, 2009). In times of rapid progress in massive DNA sequencing, nucleotide diversity becomes increasingly used to estimate the diversity of board range of genomic data types. Although the generation of genome-scale sequence data becomes cheaper and faster, the majority of currently available techniques often produce sequences containing a fraction of unresolved nucleotides. This problem is particularly visible in population-scale studies where missing sites from each individual accumulate in the final data set.

The original formula from Nei and Li (1979, equation 22) derives the nucleotide diversity from pairwise differences between whole sequences:

$$\pi = \sum_{ij} x_i x_j \pi_{ij} \quad (1)$$

where x_i and x_j are frequencies of ith and jth alleles and π_{ij} is the number of differences between them. The formula does not consider the possibility of missing nucleotides occurrence in one or more sequences in the data set. Several methods for handling missing nucleotides have been developed since. The most simplistic approach is a complete deletion of such sites across all individuals treating them equally to indels (available as an option in Mega software; Kumar et al., 2018). This method of handling missing data causes, however, substantial loss of information especially when a large number of individuals is sequenced. A more sophisticated method is a pairwise deletion of sites with missing nucleotides in each pairwise comparison (Paradis, 2010; Rozas et al., 2017), which allows retaining more information when calculating nucleotide diversity.

Estimating π using whole sequences is problematic when the amount of data is large, or the statistic is calculated not from discrete regions such as genes but, for example, in a sliding window along large stretches of genome. In such cases π can be calculated as an average over all sites of a region. For a single site, nucleotide diversity can be calculated as:

$$\pi = \frac{N_{\text{diff}}}{N_{\text{comp}}} \quad (2)$$

where N_{diff} is the number of differences and N_{comp} is the number of comparisons. This approach has already been used in several R packages, for example, in nucleotideDiversity function from strataG (Archer et al., 2017) or diversity. Stats function from PopGenome (Pfeifer et al., 2014). To obtain the average nucleotide diversity of the whole fragment π is divided by the number of sites. Because N_{diff} and N_{comp} are calculated separately for each site, the resulting π is weighted by the number of comparisons:

$$\pi_W = \frac{\sum_1^n \frac{N_{\text{diff}}}{N_{\text{comp}}}}{n} \quad (3)$$

where n is the sequence length. Because N_{diff} equals 0 at invariant sites, $N_{\text{diff}}/N_{\text{comp}}$ always equals 0 irrespective of the number of missing nucleotides at the site.

Nucleotide diversity can also be derived from site frequency spectra (Achaz, 2008, 2009). Mathematically the estimator is equal to Nei's π, but theta is derived directly from the SNP frequencies instead of a number of the pairwise differences. It is also equal to π_w in the way missing nucleotides are considered as site frequency spectrum used for estimation of θ_π is made up only of polymorphic sites. Ferretti et al. (2012) showed that frequency-based estimators can be adapted to handle sequences containing missing data.

Korunes and Samuk (2021) proposed another method to control for missing nucleotides - pixy. The nucleotide diversity is estimated using the number of differences and number of comparisons calculated for each site separately, omitting missing nucleotides:

$$\pi_{\text{pixy}} = \frac{\sum_1^n N_{\text{diff}}}{\sum_1^n N_{\text{comp}}} \quad (4)$$

The pixy package gained immediate attention and was cited several times within just a few months since its publication. This illustrates the need for fast and precise estimation of nucleotide diversity from genomic data. The main reason for pixy's popularity is that the method used in the package is fast. According to the results presented in the study, the pixy estimator gives more precise estimates of π than rough methods of handling missing data in other programs (Korunes & Samuk, 2021).

The aim of this paper is to compare the precision of π_{pixy} and the weighted nucleotide diversity π_w in the estimation of the true nucleotide diversity from sequence sets containing missing data. The efficiency of the methods of calculating π is discussed.

2 | MATERIALS AND METHODS

All simulations were performed in R 4.0.5 environment (R Development Core Team, 2009). The formulas were tested on:

- sequences simulated by fastsimcoal coalescent simulator
- two real DNA sequence sets obtained from Sanger sequencing, containing no missing data.

Fastsimcoal simulations were performed using functions from the strataG package (Archer et al., 2017). Evolution of a 10,000 bp sequence was simulated in three populations with different *thetas*. The mutation rate was set to 10^{-7} site/individual/generation, while the differences in *thetas* were obtained by setting three different effective population sizes (Deme.1 $N_e = 1000$, Deme.2 $N_e = 10{,}000$ and Deme.3 $N_e = 100{,}000$ of haploid individuals). Deme.1 and Deme.2 were derived from Deme.3 in a single event 10,000 generations before the sampling. Samples of 100 haplotypes were recorded from each generation. The simulations were repeated 10,000 times.

The two real sequence sets differ in their lengths and levels of variation: 301 bp fragment of a highly variable mitochondrial control region in 53 domestic cats and three wildcats (Branicki et al., 2006) and 79 sequences of concatenated three mitochondrial genes of *Barbus carpathicus* totalling 2579 bp (Konopiński et al., 2013). The latter species shows a very low level of genetic variation due to a deep genetic bottleneck it underwent at the edge of the last glaciation and Holocene.

The reference nucleotide diversity was calculated using nuc.div function from the pegas package version 1.0.1 (Paradis, 2010), which estimates π using the original Nei's and Li's formula. Weighted nucleotide diversity, π_w, and π_{pixy} were calculated according to the Equations (3) and (4). For calculation of π_w, the number of differences can be estimated only when the number of samples sequenced at a given site is larger than one. Thus only samples that contain at least two unambiguously sequenced nucleotides are included in n. In π_{pixy} formula, these loci return 0 both in numerator and denominator; therefore, they do not alter the result of both sums.

There are several methods of calculating the number of differences. nucleotideDiversity function from strataG package (Archer et al., 2017) estimates the N_{diff} from allele frequencies

$$N_{\text{diff}} = \frac{N * \sum 1 - p_i^2}{2}$$

where p_i is the frequency of *i*th allele and N is the number of sequenced nucleotides at the site. This approach allows for including both biallelic SNPs and loci with a larger number of alternative nucleotides. On the other hand, estimating additional floating-point variables is computationally more demanding than operating solely on integers. At biallelic sites, N_{diff} can be calculated as:

$$N_{\text{diff}} = c_0 * c_1$$

where c_0 and c_1 are alleles counts. The method is used both in pixy and in *PopGenome*. For sites with more than two alleles, the most straightforward method is:

$$N_{\text{diff}} = \frac{N^2 - \sum c_i^2}{2}$$

This equation uses only integers (including the resulting N_{diff}); thus, it should be very efficient when a large number of operations are to be performed. The final equation used in the simulations is:

$$\pi_W = \frac{N^2 - \sum c_i^2}{N * (N - 1)} \quad (5)$$

To test the robustness of the two methods missing nucleotides (*N*) were randomly introduced to the original sequence sets. The missing nucleotides distribution created by, first, discretised gamma distribution and, second, uniform distribution from 0 to 30% of the sequence length. Overlapping these two distributions allowed to mimic the true distribution of N's observed in the data set obtained by Illumina sequencing of molecular inversion probes from the raccoon immunological genes that contained 10% of missing data on average (M. K. Konopiński, A. Fijarczyk, A. Biedrzycka, unpublished data). A separate simulation of missing data was performed for each simulated data set, while the real data sets were subjected to the simulation 10,000 times each. The different proportions of missing data were obtained by randomly changing the gamma distribution parameters and the proportion of sites subjected to gamma and uniformly distributed N's (from 0.7 to 0.9 and from 0.1 to 0.3, respectively).

The efficiency of the two estimators was compared using summary statistics: median, mean and standard deviation of the estimate, and absolute relative error (ARE = $|\pi_{\text{est}} - \pi|/\pi$, where π_{est} is an estimated value and π is a true value), its median and 95th percentile. Summary statistics were calculated within seven classes of missingness depending on the frequency of missing nucleotides. The results were visualized using ggplot 3.3.5 (Wickham et al., 2021) and ggpubr 0.4.0 (Kassambara, 2020).

The scripts are attached to this study as Data S1. A function to estimate weighted π from fasta files, pi.weighted, has been included in ShannonGen R package (Konopiński, 2020).

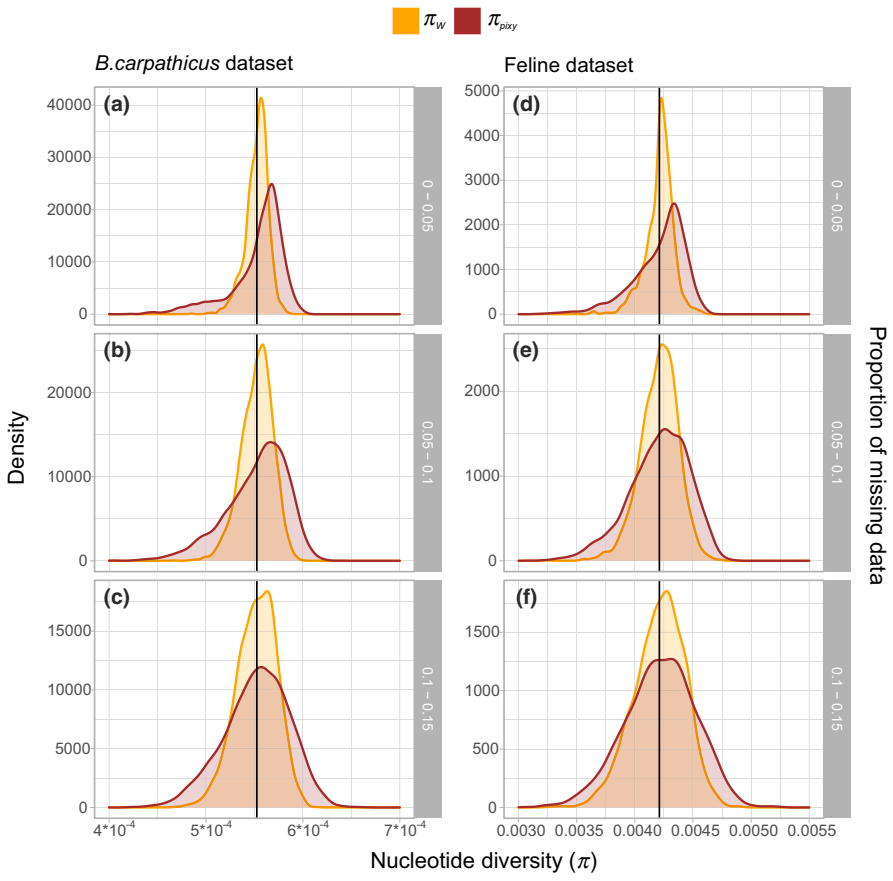

FIGURE 1 The distributions of nucleotide diversity estimates (π) from data sets containing simulated missing information in *B. carpathicus* mitochondrial genes (a–c) and feline hypervariable control region (d–f). Proportion of missing data: a,d > 0%–5%; b,e > 5%–10%; c,f > 10%–15% of missing information

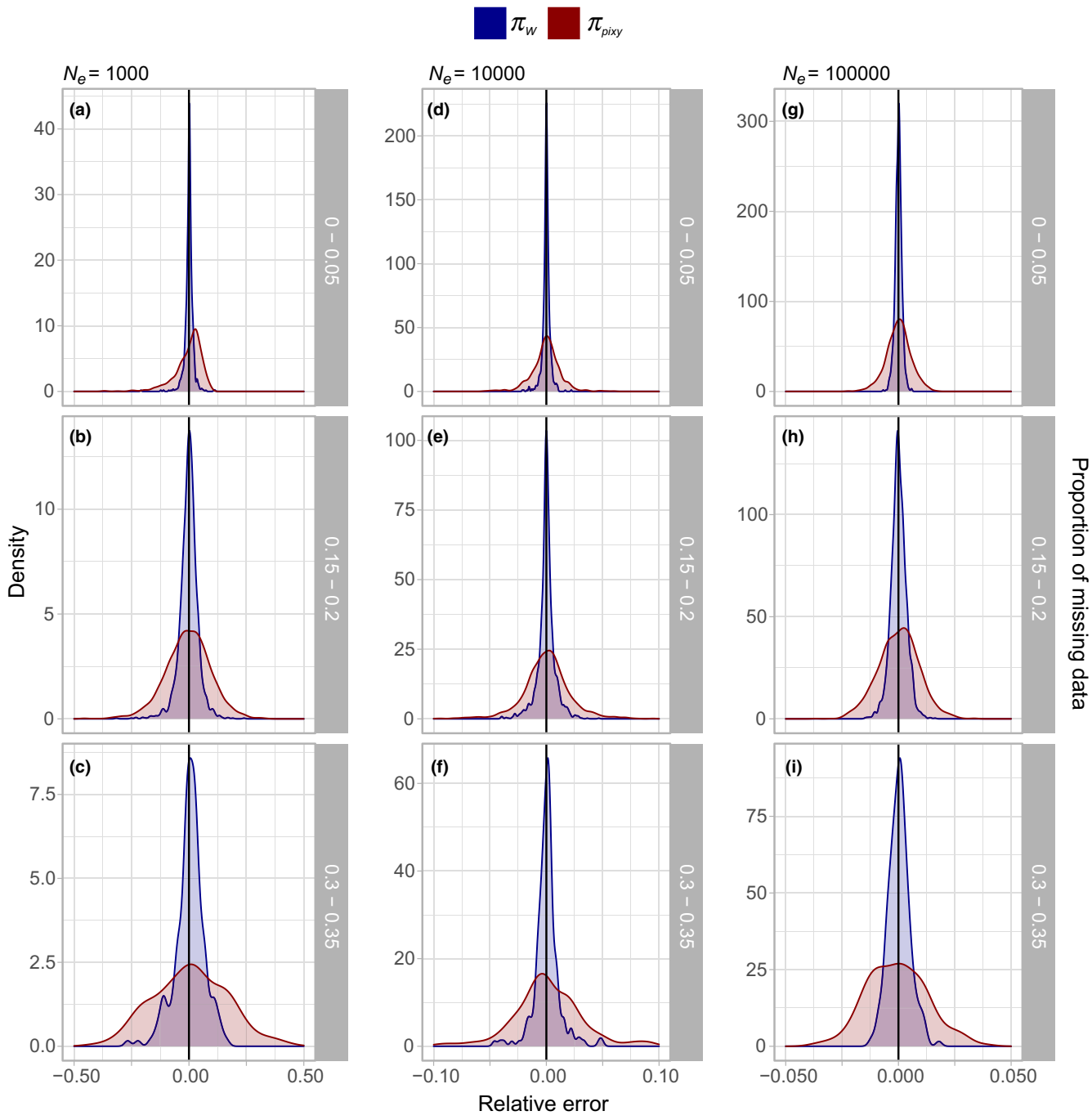

FIGURE 2 The distributions of relative errors (RE) of nucleotide diversity estimates (π) from data sets containing simulated missing information in the three simulated populations - Deme.1 (a–c), Deme.2 (d–f) and Deme.3 (g–i). Proportion of missing data: a,d,g > 0%–5%; b,e,h > 15%–20%; c,f,i > 30%–35% of missing information

3 | RESULTS

The mean nucleotide diversities in simulated data sets before introduction of missing nucleotides were $\pi = 0.0002$ (SD = 0.00015) in Deme.1, $\pi = 0.0083$ (SD = 0.008174) in Deme.2 and $\pi = 0.0194$ (SD = 0.0090) in Deme.3. When applied to complete data sets, both estimators returned identical values - $\pi = 0.00421$ for the feline control region and $\pi = 0.00055$ for the *Barbus carpathicus* mitochondrial genes, matching estimates from the original Nei and Li formula. The final proportion of the simulated N's ranged from 1.4% to 31.8% in the fastsimcoal data set, 1.4% to 13.6% in the *B. carpathicus* sequences, and 1.2% to 18.7% in the feline sequences.

Both π_w and π_{pixy} returned π estimates in all the simulations. In the real data sets, the distributions of π estimates were bell-shaped, with median values close to the true π (Figure 1). The relative error and the absolute relative error, ARE, depended on the proportion of missing nucleotides (Figures 1–3). Except for the estimates of π_{pixy} in the data sets with the smallest proportion of *N*'s no systematic bias

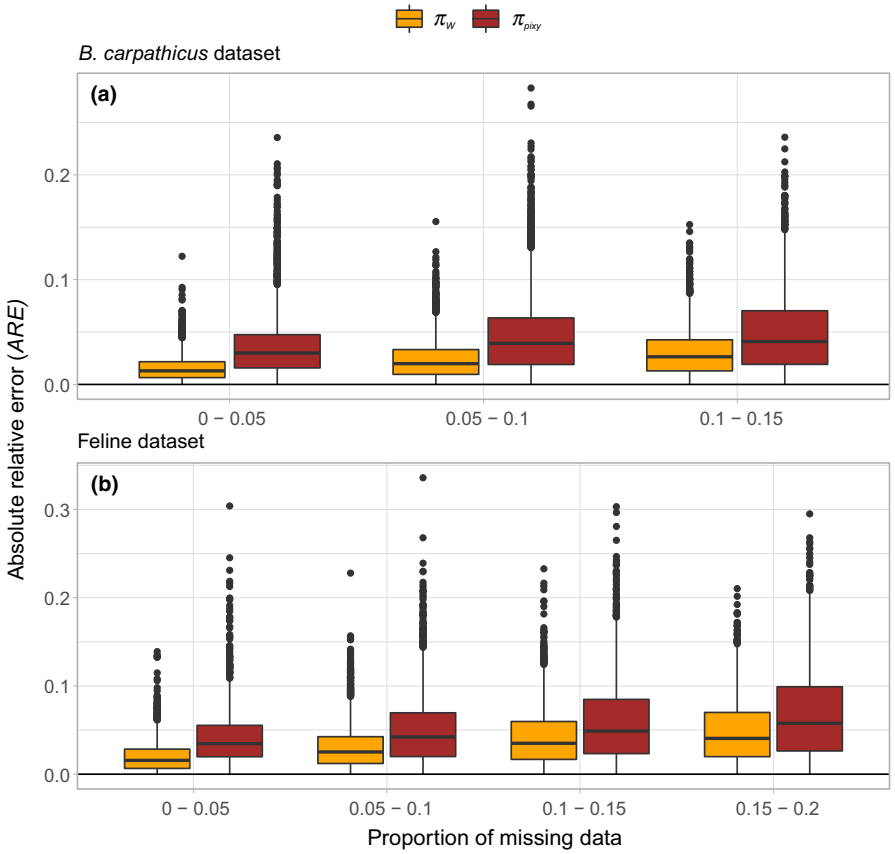

FIGURE 3 Box and whisker plots of absolute relative errors (ARE) of nucleotide diversity estimates from data sets containing simulated missing information in (a) *B. carpathicus* mitochondrial genes and (b) feline hypervariable control region, in three and four classes of missingness respectively. Median values of ARE are represented by horizontal line within box. Boxes show distribution of central 50% of observations

causing under- or overestimation was not observed in any estimator. The positive bias of π_{pixy} was strongest in the data sets characterized by the smallest actual nucleotide diversity. The median values were slightly smaller or bigger than the true π depending on the simulated data set (data not shown). The ARE strongly depended on the proportion of missing nucleotides for both estimators reaching the highest values in the group of simulations with the highest proportion of missing data (Figure 3).

In all pairwise comparisons, π_w performed better than π_{pixy}. The higher precision of π_w was observed in simulations using both sequences differing in length and the level of variation. The estimates obtained with π_w were less dispersed around the mean (lower standard deviation). Mean and median values of π_w estimates were closer to the actual π in the real data sets. The RE was less dispersed around 0 in the simulated data set (Figure 2), while the ARE of π_w had lower median, 75th percentile and maximum values than π_{pixy} in all classes of missingness in the real data sets (Figure 3). In seven simulations on *Barbus carpathicus* genes and 43 simulations on the feline control region π_w provided correct values of nucleotide diversity, while π_{pixy} failed to do so in any of the 10,000 simulations.

4 | DISCUSSION

The results presented in this study show that average weighted nucleotide diversity, π_w, performs better than the pixy method in estimating nucleotide diversity. Although both methods precisely calculated nucleotide diversity for the data sets containing no missing data, when applied to data sets containing missing nucleotides, π_w is more precise than π_{pixy} and, although rarely, only π_w in some simulations estimated the correct value of π. The reason for the latter is that missing values at monomorphic loci do not alter the estimation of π_W because those sites return zero, regardless of how many missing nucleotides are at the given locus. On the other hand, pixy sums the number of differences for monomorphic loci in the denominator; thus, even a single missing nucleotide alters the result. Although Korunes and Samuk called their estimate "π_{avg}", it is equal to average π only when the number of nucleotides sequenced at each site is equal across the whole fragment.

More precise results obtained with π_w would increase the power of statistical tests performed in downstream analyses, such as estimation of effective population size (Charlesworth, 2009; Subramanian, 2019; Wilson et al., 1985), scaled diversity (Booker & Keightley, 2018) or functional importance of genomic regions (Tatarinova et al., 2016). The difference may be of particular importance in the studies on natural selection, as it has been hypothesised that the difference in fitness may include many genes with small effects (Rockman, 2012).

Average weighted nucleotide diversity is suitable for fast genomic analyses due to the simplicity of its calculation, as compared to classical Equation (1), which requires handling of the whole sequences for estimating the fragment's nucleotide diversity. It can be easily integrated in statistical packages such as pixy or popgenome (Pfeifer et al., 2014) for precise estimation of nucleotide

diversity both in a sliding window and for discrete genomic regions. Because π_W does not require invariant sites to be included in the input file, it is less demanding in terms of computer memory resources than with π_{pixy}. It is, however, important to properly include information on fragment length in vcf regions. Some programs (e.g., DNAsp in module "Multi-MSA data file analysis", Rozas et al., 2017) use positions of the first and the last polymorphic nucleotide of the genomic regions in vcf file as their limits. This leads to an overestimation of nucleotide diversities due to the exclusion of monomorphic regions lying outside of the range. In extreme cases, if the whole region contains only a single polymorphic nucleotide π would be estimated only for this single site. The information on fragment lengths should thus be provided in gtf or gtt file format. This problem does not apply to files in fasta or fastq formats containing information on the whole sequence of the fragment.

ACKNOWLEDGEMENTS

The study has been funded from statutory resources of the Institute of Nature Conservation. I would like to thank Professor Aleksandra Biedrzycka for her comments on the draft version of the manuscript. Also I would like to thank the two reviewers and the editor for their valuable comments on the manuscript that have improved it.

CONFLICT OF INTEREST

The author declares that he has no conflicts of interests.

DATA AVAILABILITY AND BENEFIT-SHARING STATEMENT

The R code used for generating the data and the data that supports the findings of this study are available in the Data S1 of this article.

ORCID

Maciej K. Konopiński https://orcid.org/0000-0003-0893-0846

REFERENCES

Achaz, G. (2008). Testing for neutrality in samples with sequencing errors. *Genetics*, *179*(3), 1409–1424. https://doi.org/10.1534/genetics.107.082198

Achaz, G. (2009). Frequency Spectrum neutrality tests: One for all and all for one. *Genetics*, *183*(1), 249–258. https://doi.org/10.1534/genetics.109.104042

Archer, F. I., Adams, P. E., & Schneiders, B. B. (2017). Stratag: An R package for manipulating, summarizing and analysing population genetic data. *Molecular Ecology Resources*, *17*(1), 5–11. https://doi.org/10.1111/1755-0998.12559

Booker, T. R., & Keightley, P. D. (2018). Understanding the factors that shape patterns of nucleotide diversity in the house mouse genome. *Molecular Biology and Evolution*, *35*(12), 2971–2988. https://doi.org/10.1093/molbev/msy188

Branicki, W., Olszańska, O., & Konopiński, M. K. (2006). Sequence variation in the control region of mitochondrial DNA within a population sample of domestic cats *Felis catus* Linnaeus - Implications for domestic and wild cats differentiation. *Z Zagadnien Nauk Sadowych*, *67*, 279–288.

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, *10*(3), 195–205. https://doi.org/10.1038/nrg2526

Ferretti, L., Raineri, E., & Ramos-Onsins, S. (2012). Neutrality tests for sequences with missing data. *Genetics*, *191*(4), 1397–1401. https://doi.org/10.1534/genetics.112.139949

Kassambara, A. (2020). Ggpubr: "ggplot2" based publication ready plots (Version 0.4.0). https://CRAN.R-project.org/package=ggpubr

Konopiński, M. K. (2020). Shannon diversity index: A call to replace the original Shannon's formula with unbiased estimator in the population genetics studies. *PeerJ*, *8*, e9391. https://doi.org/10.7717/peerj.9391

Konopiński, M. K., Amirowicz, A., Kotlík, P., Kukuła, K., Bylak, A., Pekarik, L., & Šediva, A. (2013). Back from the brink: The holocene history of the carpathian barbel *barbus carpathicus*. *PLoS ONE*, *8*(12), e82464. https://doi.org/10.1371/journal.pone.0082464

Korunes, K. L., & Samuk, K. (2021). Pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources*, *21*(4), 1359–1368. https://doi.org/10.1111/1755-0998.13326

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, *35*(6), 1547–1549. https://doi.org/10.1093/molbev/msy096

Loeuille, B., Thode, V., Siniscalchi, C., Andrade, S., Rossi, M., & Pirani, J. R. (2021). Extremely low nucleotide diversity among thirty-six new chloroplast genome sequences from Aldama (Heliantheae, Asteraceae) and comparative chloroplast genomics analyses with closely related genera. *PeerJ*, *9*, e10886. https://doi.org/10.7717/peerj.10886

Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, *76*(10), 5269–5273. https://doi.org/10.1073/pnas.76.10.5269

Paradis, E. (2010). Pegas: An R package for population genetics with an integrated–modular approach. *Bioinformatics*, *26*(3), 419–420. https://doi.org/10.1093/bioinformatics/btp696

Petit-Marty, N., Vázquez-Luis, M., & Hendriks, I. E. (2021). Use of the nucleotide diversity in COI mitochondrial gene as an early diagnostic of conservation status of animal species. *Conservation Letters*, *14*(1), e12756. https://doi.org/10.1111/conl.12756

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E., & Lercher, M. J. (2014). PopGenome: An efficient swiss Army knife for population genomic analyses in R. *Molecular Biology and Evolution*, *31*(7), 1929–1936. https://doi.org/10.1093/molbev/msu136

R Development Core Team. (2009). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org

Rockman, M. V. (2012). The Qtn program and the alleles that matter for evolution: All That's gold does not glitter. *Evolution*, *66*(1), 1–17. https://doi.org/10.1111/j.1558-5646.2011.01486.x

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., & Sánchez-Gracia, A. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology and Evolution*, *34*(12), 3299–3302. https://doi.org/10.1093/molbev/msx248

Subramanian, S. (2019). Population size influences the type of nucleotide variations in humans. *BMC Genetics*, *20*(1), 93. https://doi.org/10.1186/s12863-019-0798-9

Tatarinova, T. V., Chekalin, E., Nikolsky, Y., Bruskin, S., Chebotarov, D., McNally, K. L., & Alexandrov, N. (2016). Nucleotide diversity analysis highlights functionally important genomic regions. *Scientific Reports*, *6*(1), 35730. https://doi.org/10.1038/srep35730

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., & RStudio. (2021). ggplot2: Create elegant data visualisations using the grammar of graphics (version 3.3.5). https://CRAN.R-project.org/package=ggplot2

Wilson, A. C., Cann, R. L., Carr, S. M., George, M., Gyllensten, U. B., Helm-Bychowski, K. M., Higuchi, R. G., Palumbi, S. R., Prager, E. M., Sage, R. D., & Stoneking, M. (1985). Mitochondrial DNA and two perspectives on evolutionary genetics. *Biological Journal of the Linnean Society*, *26*(4), 375–400. https://doi.org/10.1111/j.1095-8312.1985.tb02048.x

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Konopiński, M. K. (2023). Average weighted nucleotide diversity is more precise than pixy in estimating the true value of π from sequence sets containing missing data. *Molecular Ecology Resources*, *23*, 348–354. https://doi.org/10.1111/1755-0998.13707