In their manuscript, Konopiński (2022) focus on two methods of averaging per-site nucleotide diversities (hereafter “pi”) over the length of a sequence or genomic window.

The first method is the one employed by pixy (Korunes & Samuk, 2020). This method, as outlined in the pixy manuscript, involves summing the individual numerator (number of pairwise differences) and denominators (number of comparisons made) of each site-level pi, and then computing the total value for the sequence by dividing the sum of the numerators by the sum of the denominators. This metric is thus a regular *weighted average* of the site-level nucleotide diversities, as sites with fewer samples (genotyped individuals) will contribute proportionally less to the overall average compared to sites with more individuals. When writing pixy, this choice was deliberately made, as we specifically want to properly weight sample sizes across sequences with variable amounts of missing data at each site.

Konopiński (2022) explores a separate method for averaging across a sequence/window, which they refer to as “weighted pi”. This metric is a “average of ratios”, in that the nucleotide diversity ratio (differences / comparisons) is pre-computed for each site and then summed and divided by the number of sites. As such, all sites contribute equally to the final estimate regardless of the amount of underlying data. As such, by most regular definitions, this is in fact an *unweighted average*. To reduce confusion, I will be referring to this method as “unweighted average pi”.

Here, I will discuss several points. First, I discuss why from first principles a weighted average (as used by pixy) is preferable to an unweighted average in the presence of missing data. Secondly, I will address the claim that the use of an unweighted average results in an increase in precision when averaging pi across a sequence/window. Finally, I will address the claim that the use of an unweighted estimator results in improved accuracy in an empirical data set.

**Weighted vs. unweighted averages**

Generally speaking, when averaging any kind of value across multiple groups, the contribution of data from each subgroup is weighted by the number of samples in the group. For example, if one was measuring the overall height in two groups of people, one with n=2 people, and another with n=100 people, we would sum all of the heights together and divide by 102. Unless there was a some very specific reason to do so, we would not compute the group level means individually, and then average them. This would give undue weight to the group with n=2 people.

This is specifically important when considering variances of estimates. If all subgroups have identical sample sizes, the sampling distribution of weighted and unweighted means are identical (Meier, 1953). However, when sample sizes vary among subgroups, weighted means have wider sampling distributions than unweighted means(Meier, 1953; Cochran, 1977; Gatz & Smith, 1995). This should make intuitive sense: variation in the group-level degrees of freedom due to differences in sample size should cause greater uncertainty in the estimator. Treating all groups the same (e.g. via an unweighted mean) causes this error to not be propagated, and is thus variance is incorrected deflated.

This increase in sampling variance might seem like a detriment when the “true” mean pi is known. However, when the true mean is not known, as is the case when employing estimators to actual data, a decrease in variance can result in overconfidence in a particular incorrect value of pi.

**Increases in precision**

Konopiński (2022)’s first claim is that an unweighted average of pi has higher precision than the weighted average employed by pixy. As discussed above, the sampling distribution of a weighted mean will always be wider than an unweighted mean. This increase in variation isdesired, because it correctly reflects the uncertainty in the estimate introduced by variation in sample sizes among sites.

Konopiński (2022) also claims that the weighted mean employed by pixy can result in misestimation of average pi (i.e. a bias), based on simulations performed by dropping data from an empirical data set. The author does not provide a statistical explanation for this phenomenon.

This result seems to be tied to a specific distribution of missing data employed by Konopiński (2022). This distribution of missing data was inferred from a separate dataset generated by the authors, which is unpublished. This bias was not apparent in Konopiński (2022)’s coalescent simulations, and the effect is not visible when employing a uniform distribution of missing data on the same data (Figure 1). Using simple comparisons of sequences with differing amounts of diversity and uniform missing data, this effect is also not visible (Figure 2). If this apparent bias was a general property of the difference between the two methods of averaging, presumably we should observe it regardless of the specific distribution of missing data. This suggests more work needs to be done to explore the behavior of the unweighted average before it is adopted as an alternative method.

**Other notes**

Konopiński (2022) notes that their method does not require the addition of invariant sites, as long as the sequence length was known. The issue is, when computing pi using a VCF, invariant sites must be included in order to properly infer correct sequence length (the number of sequenced sites). The inclusion of variability in sample size among sites (weights) also requires the preservation of all invariant sites in the file – they cannot be summarized into a single length value. So, the inclusion of invariant sites will remain an important practice when calculating pi from sequencing data.

**Conclusion**

In conclusion, the use of a weighted average, as employed by pixy, in line with statistical best practices for averaging summary statistics over multiple sites. In our opinion, it is important to propagate the uncertainty in the estimates of pi by use of a weighted average. Ultimately, what constitutes the best averaging method will be up to investigator.

**Data Availability**

All code and data used in this study are available at ww.github.com/ksamuk/weighted\_pi

**Data Availability**

I thank Dr. Maciej Konopiński for corresponding with me about this result and providing supplementary data.

**References**

Cochran, W. G. (1977). Sampling Techniques (3rd ed.). Nashville, TN: John Wiley & Sons. [ISBN](https://en.wikipedia.org/wiki/ISBN_(identifier)) [978-0-471-16240-7](https://en.wikipedia.org/wiki/Special:BookSources/978-0-471-16240-7)

Meier, P. (1953). Variance of a Weighted Mean. *Biometrics*, *9*(1), 59–73.

Gatz, D. F., & Smith, L. (1995). The standard error of a weighted mean concentration—I. Bootstrapping vs other methods. *Atmospheric Environment*, *29*(11), 1185–1193.

Korunes, K. L., & Samuk, K. (2020). pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. In *Mol. Ecol. Resour.* (Issue 4, p. 2020.06.27.175091). https://doi.org/10.1101/2020.06.27.175091