19-Jun-2024

Dear Dr. Hemstrom,

Regarding: manuscript # JOH-2024-037, "Unbiased estimation of the number of segregating sites across unequal sample sizes"

I have received two reviewers' comments and the recommendation of the Associate Editor concerning your manuscript. The AE and reviewers found the paper to have limitations that preclude my accepting it for publication in its current form. The AE's comments and the reviews are included below for your information.

Based on these reviews, I am rejecting the current manuscript, but encourage you to undertake major revision for resubmission as a new manuscript, in accordance with the suggestions made by the reviewers and the AE. If you choose to resubmit, please be sure to document the changes that you have made in a detailed cover letter.

Thank you for considering the Journal of Heredity for the publication of your research.

Sincerely,

Prof William J Murphy

Editor, Journal of Heredity

theaga@theaga.org

AE's Comments to Author: This ms was reviewed by two highly-qualified reviewers. Although both found some strong points about the manuscript, and it fits well within the purview of the journal, both also had some substantial concerns. One shared concern was that the simulations should consider a more realistic allele frequency spectrum. A more serious concern raised by #2 is that the authors have not provided a compelling argument for the necessity or practical usefulness of the new method. Given this, it seems best to reject this version but without prejudice and allow resubmission of a new manuscript, providing the authors can address reviewer concerns.

Reviewer: 1

Comments to the Author

L18. I'm not a fan of the first line of the abstract as it's based on people (geneticists) doing something; that's not a strong argument I suggest changing this. The second line actually, while requiring modificatin, is how I would start.

We agree and we have now re-arranged and edited this portion of the abstract to have a stronger start (lines X-XX).

L37. natural selection acts on phenotypes, not genotypes. So those diversity indicators need to reflect phenotypic differences.

We’ve added a clarification that functional genetic diversity is what is linked to population persistence, with an additional note explaining why neutral genetic diversity continues to be a useful metric (Lines 35-40).

L55; I see this cites the original, but don't we typically refere to Tajima's theata as pi?

We added a note clarifying that Tajima’s theta is often referred to as Tajima’s pi (lines 56-57).

L57-61;should you also mention the selection-based explanation? Genome-wide I get is demography, but we see this alot on windows /loci with the goal of selection inferences.

We’ve added a sentence mentioning that selective sweeps and balancing selection can also change allele frequencies/the site frequency spectra (Lines 63-65).

L62-70 this first line is very microsatellite specific. WGS/SNP data we only really ever have 2 alleles; you say this but it's not clear in that first line as it implies it's widely used. I'd rework this paragraph a little in that regard.

We’ve edited these lines for clarity (Lines XXX-XXX).

L75; delete "much"

L71-80. One option here is to project the SFS to a desired sample size (dadi does this); can you a) let the reader know this is an option and b) comment on problems / concerns if those exist. And again, 78-80 is really a ustat issue.

We’ve added quite a bit of discussion of SFS bases estimators of S, such as projection with dadi’s approach (lines XXX-XXX; XXX-XXX). More specifically, we have also added a comparison to dadi’s projection approach in Figure X, described in lines XXX-XXX. Primarily estimates of S from projection tend to underestimate the number of segregating sites which should persist following rarefaction. Secondarily, projection assumes HWE, and is thus prone to bias in cases where populations are out or in-bred: they will overestimate S when populations are inbred and underestimate when populations are outbred (after accounting for the overall underestimate).

Lines 87-163; I can follow but don't have formal training to to comment on the equations.

L164; empirical to some might imply data not generated by simulation (that's initially what I thought).

We’ve clarified our phrasing here (Lines XXX)!

L167; so MAF filters are often employed that would throw out at least half of these simulations (i.e. 0.05 common). Can the authors comment as to why more moderate frequencies were not simulated? Is the idea to encourage not using MAF filters? That's usually done becaue of sequencing concerns in my experience, not allele frequency questions. Would these estimates vary if we had MAF in 30-49%, or how do they perform?

We’ve extended the MAF range for our simulation (up to 50%, where we recognize that the term MAF is a bit ambiguous). We generally do recommend a permissive (but not absent) MAF filter for estimating S, since removing rare variants can erase signals of selection and demography (see Hemstrom *et al* 2024, now cited on this topic on lines XXX-XXX).

L209-214; again, some approaches project to sample size - how does this compare? This seems like it's valuable as I'm not interested in the Kalinowski approach, rather the Dadi projection (see https://dadi.readthedocs.io/en/latest/user-guide/importing-data/) of the SFS approach (and it's analogues)

We’ve added a comparison to dadi’s projection approach (see justification above; lines XXX-XXX; Figure X).

Reviewer: 2

Comments to the Author

I enjoyed and appreciated the opportunity to review the manuscript "Unbiased estimation of the number of segregating sites across unequal sample sizes" by Drs. Hemstrom and Christie.

The authors note that there is not currently a method available to project, S, the number of segregating sites from a current sample size to a smaller sample size in the same way as can be done, for example, via rarefaction to obtain the allelic richness. They argue that such a method would be useful, and then they provide a way of doing it.

At first glance, a simple solution to the problem would be to randomly downsample individuals to the desired sample size, multiple times, and then, for each randomly downsampled data set, calculate S. There would be two challenges associated with such an approach. First, if one wishes to compare S between different populations/collections, the goal might be to downsample to a particular number of sampled gene copies, consistent across samples. This can't be done easily by randomly sampling whole individuals when data are missing across the sites in each sample. Second, doing repeated random sampling like that could take a very long time for a data set with a large number of polymorphic sites.

As an alternative, the authors propose executing the rarefaction by using the hypergeometric distribution. For each site they propose downsampling without replacement from the individuals (that don't have missing data at the site) to the desired sample size. Based on the hypergeometric distribution, it is fairly straightforward to calculate the probability that such downsampling will create a sample in which the site is fixed for one allele or the other. With such probabilities in hand, it is straightforward to calculate the expected number of segregating sites (it is merely the sum over sites of the expected values for each site). Doing this provides an approach to calculating the expected number of segregating sites upon downsampling. This approach appears correct and is much more efficient than taking the mean of a large number of randomly downsampled samples.

I'll pause here to note that I support the authors' approach of downsampling each site by individuals, rather than by downsampling gene copies as a way to incorporate the effect of non-independence of allelic type within individuals (i.e. lack of H-W equilibrium). I think it could also be explained/justified a little more easily to a general audience by stating that the unit of sampling at each site is the individual, just as it is in the original sample collection effort, so it makes sense to do it that way.

The authors then propose an estimate of the variance for the downsampled S. To do this, they assume that presence or absence of polymorphism in the downsampled sample is independent across all sites, and they simply take the sum of the Bernoulli variance (for each site) across all sites.

Following this they conduct a simulation study to show that their approach gives an unbiased estimate of the expected number of segregating sites after downsampling, and that the confidence intervals calculated from the estimated variances have accurate coverage.

Finally, in the discussion, the authors sketch out situations in which having an estimate of S is valuable. They note briefly that their approach allows a way to standardize the metric S across different sample groups (of different sizes) and across different studies; however most of the focus of this section is on the utility of S for comparisons to other estimators of diversity, particularly, as done with Tajima's D, comparing Watterson's \theta (estimated from S) with \theta estimated from the average heterozygosity.

I have a few queries and comments that I describe below.

1. While standardizing S across different sample sizes might help compare the value of the summary statistic across those samples with different sample sizes, there doesn't seem to be any mention in the paper that, under the standard neutral coalescent, it is well known how S changes with sample size for a given value of \theta (2 \* ploidy \* neutral mutation rate \* Ne). For many applications, the important parameter being estimated is \theta, rather than the summary statistic S, and Watterson's estimator accounts for sample size by dividing S by the expected total tree length under the neutral coalescent with constant population size. So, the authors might want to justify more why they are focusing on S instead of upon \theta\_S.

Related to this is the statement in the discussion, "In cases where biologically important conclusions can be drawn from the difference between heterozygosity or other allele frequency-based estimators of diversity and the number of segregating sites, it is particularly important that the latter is properly calculated across sample groups, since failing to correct for differences in sample size or data missingness could mask biologically interesting signals of demographic history and obscure a critical facet of overall genomic diversity." I'm not sure that I completely agree with this. Because Watterson's estimate, \theta\_S, explicitly takes account of sample size, I'm not sure that it is essential to make sure that sample sizes are standardized when you are, for example, calculating Tajima's D. In fact, I suspect rather strongly that if you had ten populations, all of them having had a recent expansion, and you had sample sizes of 100 individuals in the first 9 and of 10 individuals in the last 1, that you would have a much better chance of correctly rejecting the null hypothesis of the neutral coalescent with constant population size if you perform the test on the original sample size for the first 9, rather than performing the test after downsampling to only 10 individuals. So, I am not convinced that downsampling in this context has great application except possibly to show that a non-significant result for population 10 might be due to reduced power because of small sample size. But that doesn't, in itself, justify downsampling from the first 9 populations for estimating Tajima's D. Thus, I think the authors need to be more careful about justifying why and when one would actually want to do this sort of downsampling/rarefaction for S. As it stands now, I don't think they have suitably justified the procedure.

This is an excellent point! We extended the background section to include more information on theta estimates (lines XXX-XXX), then added a substantial section validating the use of rarefaction over either Watterson’s Theta estimator or linear combinations of projected SFSs between lines XXX and ZZZ. Principally, Watterson’s Theta and projection approaches all assume (typically implicitly) HWE within loci given that they assume alleles at loci are independent. This is very often not true, particularly in species of conservation concern when populations have any substantial degree of in-(or out-)breeding. In inbred populations, assuming HWE/allelic independence will result in over-estimating S after subsampling; in outbred populations, assuming HWE/allelic independence will result in under-estimating S. We demonstrate this pattern by comparing the proportion of segregating sites estimated by SFS projection to reduced sample sizes, Watterson’s theta estimates, and E(Ns) estimates to estimates generated by simulated downsampling in the new Figure X. The bias in populations with even slightly negative or positive FIS values is quite notable.

Secondly, we note that Watterson’s theta calculations require a count of the number of gene copies considered (K). K is not necessarily straightforward to determine in the presence of missing data, since K is not consistent in this case for each segregating site; our rarefaction approach avoids this problem allowing *g* (the number of subsampled individuals/genotypes) to vary at each locus without introducing bias. We have added a mention of this on lines XXX-XXX.

2. To compare with other estimators of theta, should you downsample those as well? I suspect not, since the variance in the denominator for most neutrality tests would account for different sample sizes. But, this is something that is worth considering and discussing. This sort of gets us to the question of, "If we are calculating Tajima's D after downsampling, as proposed by the authors, how shall we go about calculating the variance of the difference of the two theta estimators?" Here, it seems that the presentation in the manuscript is a little bit incomplete---the authors provide an estimate of the variance of S, but what is really needed is an estimate of the variance of theta as estimated by the newly downsampled S. I suppose it would look something like:

Var(theta) = E[Var(theta | S\_downsample)] + Var[E(theta | S\_downsample)]

where theta is the estimate of theta made from S\_downsample (the estimate of S from the downsampling method). The first term is known from population genetics under neutrality and the second term can be calculated from the variance of S\_downsample provided by the authors. At any rate, if any users are ever going to want to compare theta from the downsampled S, to another estimator of theta I think they would have to account for the extra variance created by the downsampling, so this would need to be expanded in the paper.

We agree, and have added a short section mentioning how one might go about using our E(Ns) estimate to calculate Watterson’s theta (lines XXX-XXX) and show that theta estimates based on this value do not show the bias associated with allelic non-independence that raw theta values do in Figure X. We likewise describe the variance of a Watterson’s theta estimate derived from our estimator of S rather than the raw value of S given more or less the exact equation presented here (lines XXX-XXX).

3. The simulations could be more faithful to how the data arrive. Rather then just simulating alleles at various fixed frequencies, it would probably make more sense (especially if we wish to test whether the method has certain advantages for comparing theta, rather than, simply, S, in practical situations of unequal sample sizes) to simulate data from the coalescent. The program msprime could do that on a whole-genome or single-chromosome scale quite quickly.

We chose to simulate data in this way rather than use a traditional coalescent simulator to very clearly demonstrate how variance in minor allele frequency and, in particular, missing data affect estimates of P(S). To show that our method produces reliable results with a neutral SFS, we have now additionally included results from a coalescent simulation (run via *scrm*) in Figure SX. We added a description of this in lines XXX-XXX.

4. There is a considerable amount of work on estimating theta (from various linear combinations of the site frequency spectrum). The current work could and should be more generally described in that context. A good entry point for that is Achaz 2009, Genetics 183: 249–258.

We added a sentence describing more of the background on estimating theta from the SFS to the introduction as part of the new discussion on theta estimators in general (lines)

5. It is not quite clear what the target data type is for this methodology. I imagine it could be useful for genotype data from RAD seq, but would not be useful for low-coverage, whole genome sequencing data. In this regard, the authors seem to overlook a large amount of work on estimating the SFS (and hence any estimator of theta, including S, that is a linear combination thereof) from low coverage whole genome sequences (See Korneliussen et al. BMC Bioinformatics 2013, 14:289). That work actually also addresses a number of issues mentioned in this paper, most importantly the issue of missing genotypes. In that work a recursion is implemented to sum over missing individuals at each site. This is a different approach than downsampling to the lowest sample size and it would be good to know which performs better, or at least to discuss these other approaches. Regardless, the authors should be more explicit about what type of data they are targeting with their method, and why existing approaches for inferring various unbiased estimators of theta are less appropriate than their downsampling approach in such cases.

Given that this method relies on called genotypes, we have added a section clarifying that this method should probably not be used with lcWGS approaches when coverage is too low to accurately call genotypes (this method is intended for slightly higher coverage data and above, which can include moderate coverage WGS data (7x+ or so to enable calling with reasonable certainty) and have cited several additional papers which suggest methods for estimating the SFS (and thus theta or S) in such cases (lines XXX-XXX).

In summary, the authors have correctly implemented a hypergeometric rarefaction scheme for the number of segregating sites, but have not yet provided a good rationale for why it should be done. Additionally, they have not yet provided a complete picture for how the downsampled estimates should be used in the context of comparing different downsampled theta estimators and testing whether the differences are beyond what is expected under neutrality with constant population size (since expressions for the variance of theta itself after downsampling are not provided). A focus on estimating theta, rather than simply computing the summary statistic S would bring the paper in line with population genetic theory in this realm, and doing so would benefit from simulations with greater realism. Finally, actual empirical or simulated examples showing that the downsampling provides a tangible benefit that can't be obtained by other approaches for handling missing data would make this work more compelling.

We thank the reviewers for their insightful comments and hope that this draft 1) better summarizes the existing state of the methods for theta and S estimation and 2) better clarifies the additional utility of this method, particularly in cases where populations are not in HWE at neutral loci.