Supporting information for: Unbiased estimation of the number of segregating sites across unequal sample sizes

William Hemstrom*, Mark R. Christie*

Corresponding author information:
* William Hemstrom; email: whemstro@purdue.edu
* Mark Christie; email: christ99@purdue.edu

This PDF includes:

Supplementary Example

Supplementary Figures 1-3

Supplementary Example:

The unbiased expected number of segregating sites after standardizing sample sizes, $E(N_S)$, can be calculated with γ automatically set to the smallest sample size across all sample groups automatically across multiple sample groups using the R package “`snpR`” (W. Hemstrom & Jones, 2023). First, install `snpR` if needed (this can be skipped if already installed) and load it:

```
# Install and load snpR. Comment out first two lines if not needed.

install.packages("remotes")
remotes::install_github("hemstrow/snpR")
library(snpR)
```

Next, download the example data:

```
# download the metadata (if copy/pasting, check line breaks)
meta <- read.table(url("https://raw.githubusercontent.com/ChristieLab/seg_sit
es_rarefaction/main/data/example_vcf.vcf"),
                   header = TRUE)

# download the vcf (if copy/pasting, check line breaks)
download.file("https://raw.githubusercontent.com/ChristieLab/seg_sites_rarefa
ction/main/data/example_vcf.vcf",
              destfile = "example_vcf.vcf")
```

This VCF file contains a subset of data from Hemstrom et al. (W. B. Hemstrom et al., 2022) which includes gennotypes for 1,000 SNP loci from five populations. The VCF file can be loaded in alongside the metadata into a single object using “`read_vcf`”:

```
monarchs <- read_vcf("example_vcf.vcf", sample.meta = meta)
```

Population size information can be accessed using “`summarize_facets`” by referring to the column name in the metadata read in earlier (“pop”). “Facets” in `snpR` refer to any metadata column in the data (including both sample metadata, like we read in above, and locus metadata if supplied). We can view the number of individuals per population using “`summarize_facets`”:

```
summarize_facets(monarchs, "pop")
```

Running this will show the number of samples per population.

The expected number of segregating sites per population can be calculated using the function “calc_seg_sites”, supplying the object we imported above and naming the facet which contains our population information. The rarefaction level (γ) per locus will be automatically calculated according to the argument “*g*”. If “*g*” is zero (the default), $\gamma = n_{min}$, where n_{min} is the smallest sample size across all populations for each locus after accounting for missing data. On the other hand, if g < 0, γ will be set to $n_{min} - g$ and if g > 0, γ will be set to *g*. Either way, the result can be fetched with “get.snpR.stats”, referring to the object, facet, and statistic we are fetching:

```r
# g = 0, gamma = Nmin
monarchs <- calc_seg_sites(monarchs, "pop", g = 0)
get.snpR.stats(monarchs, "pop", "seg_sites")$weighted.means

# g = -1, gamma = Nmin - 1
monarchs <- calc_seg_sites(monarchs, "pop", g = -1)
get.snpR.stats(monarchs, "pop", "seg_sites")$weighted.means

# g = 10, gamma = 10
monarchs <- calc_seg_sites(monarchs, "pop", g = 10)
get.snpR.stats(monarchs, "pop", "seg_sites")$weighted.means
```

Note that the addition of “$weighted.means” to the end of each “get.snpR.stats” means that we are fetching the mean values specifically, not the per-locus data. We can fetch that instead by using “$single”, referring to the statistics for each single locus. The mean results will contain the columns “seg_sites” and “seg_sites_var” containing $E(N_S)$ and its variance $\sigma^2_{E(N_S)}$, respectively. The per-locus results will contain the columns “g_prob_seg”, “prob_seg”, and “prob_seg_var” which note γ, the probability the site segregates at γ in a specific population ($\hat{P}(S_{ij})$), and the variance of $\hat{P}(S_{ij})$, $\sigma^2_{\hat{P}(S_{ij})}$, respectively.

Supplementary Figures:

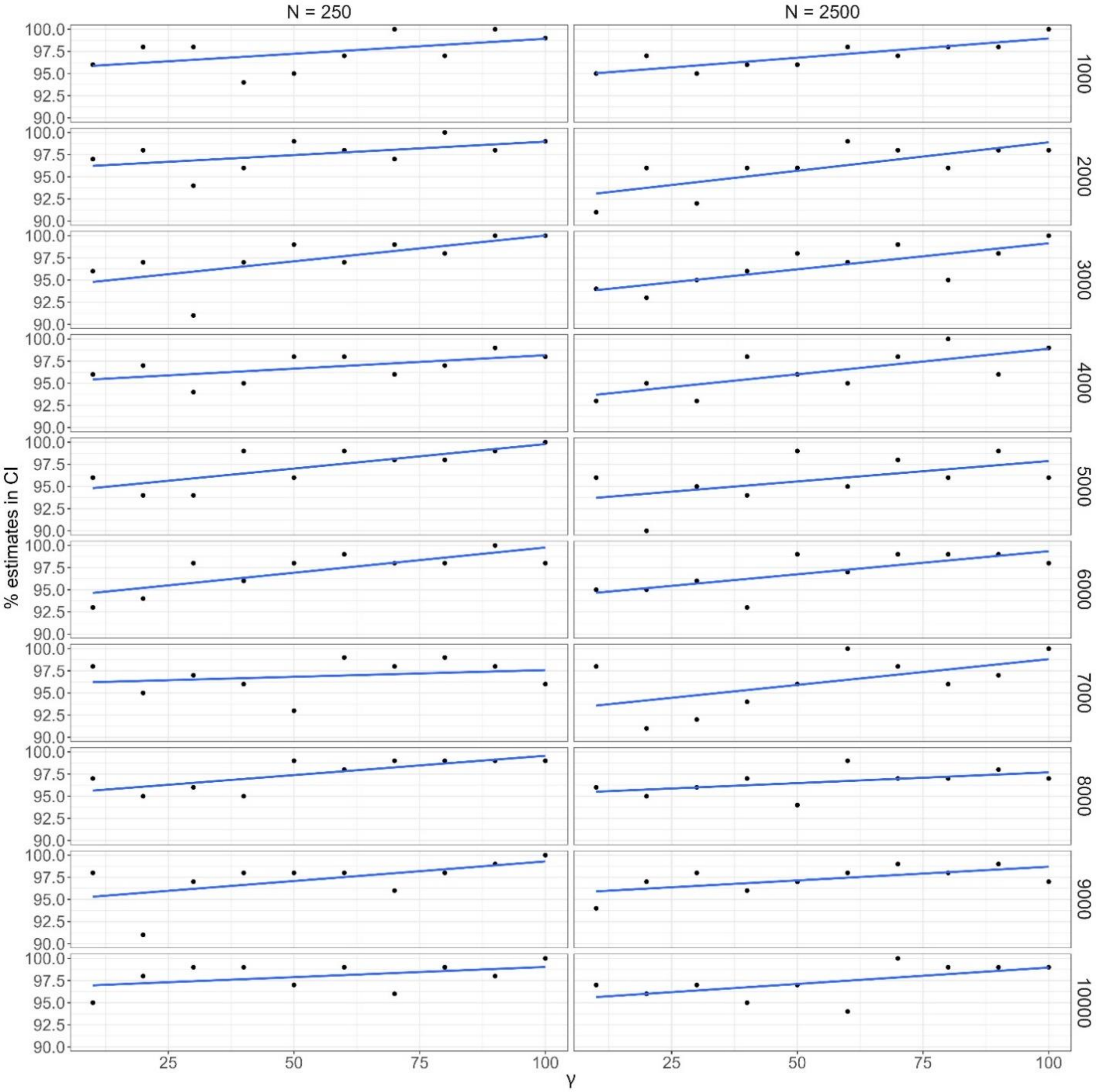

Figure S1: The percentage of expected segregation probabilities $\hat{P}(S_{iq})$ for which the probability a loci was segregating was inside the 95% confidence interval derived from simulations for different γ values and numbers of simulations for both a population of size 250 and 2,500.

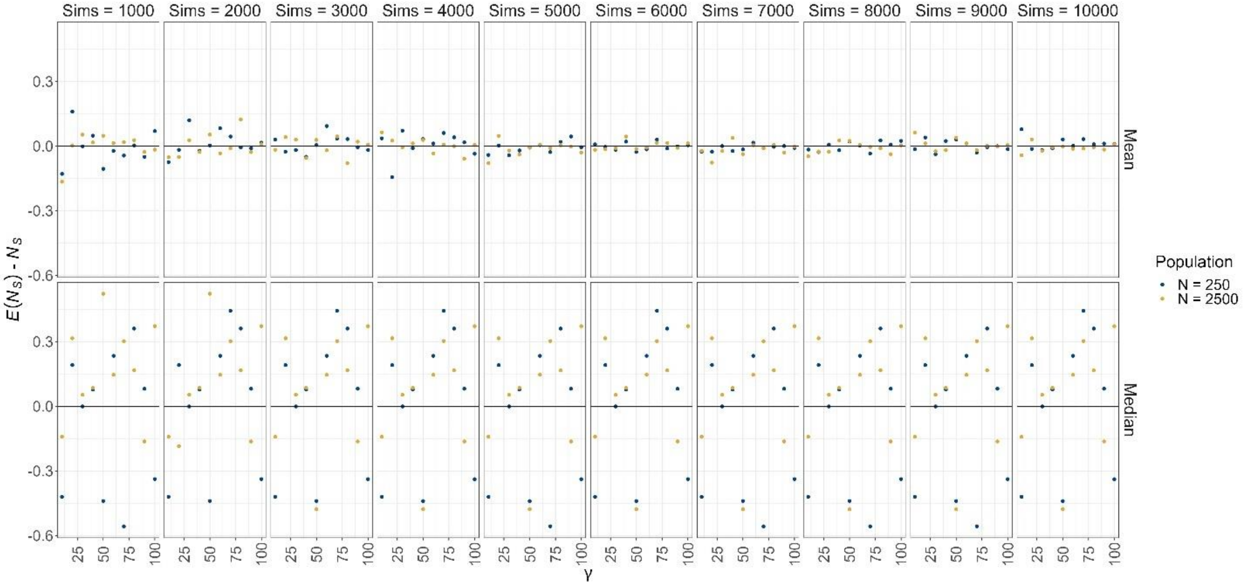

Figure S2: The difference between the expected number of segregating sites ($E(N_S)$) and the mean and median observed number from simulations across different numbers of simulations and γ values for populations of both size 250 and 2,500.

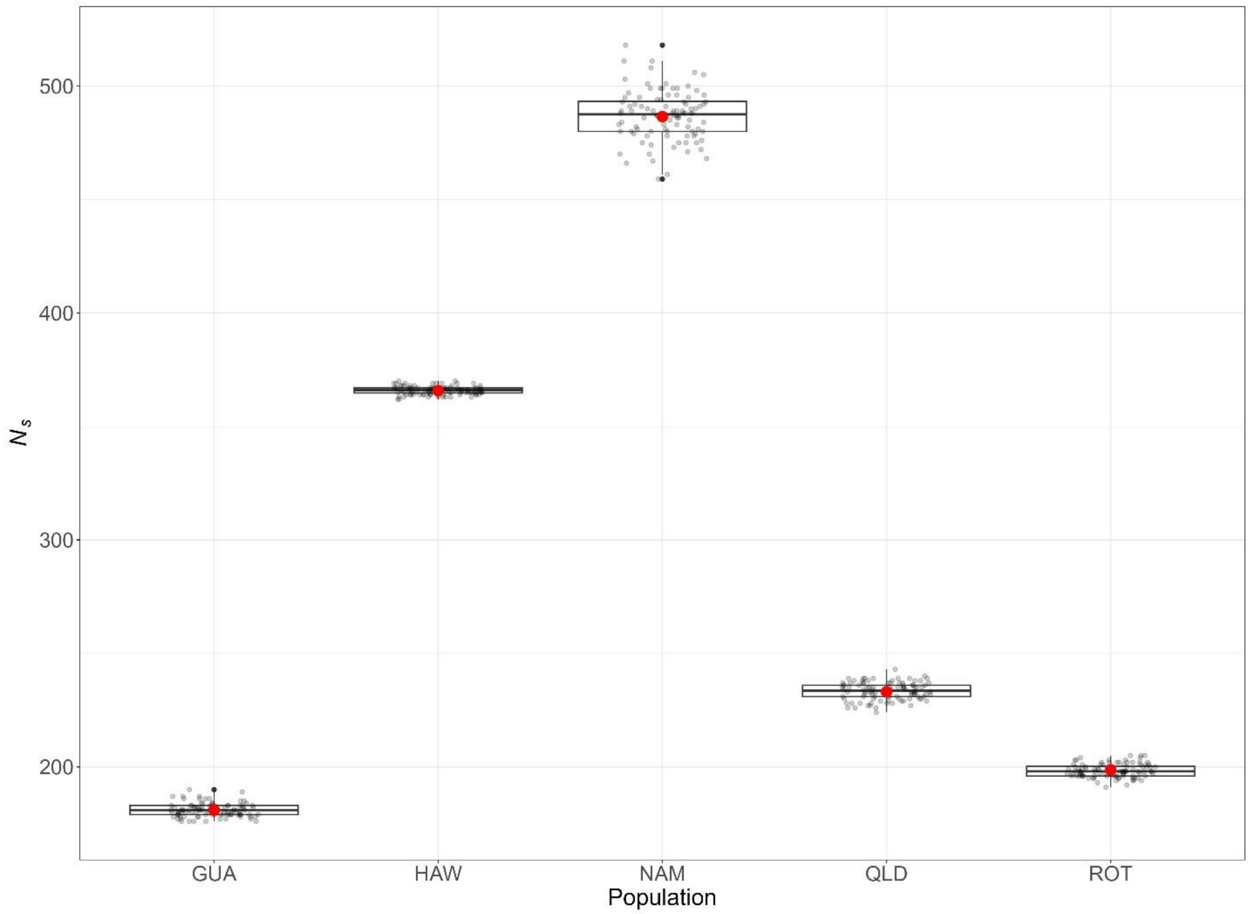

Figure S3: The expected number of segregating sites ($E(N_S)$, red) compared to the distributions of the number of segregating sites observed after rarefaction based on 100 simulations (black). $E(N_S)$ values were obtained using the “calc_seg_sites” function in snpR using “g = 0”, which sets γ equal to the minimum sample size across all populations for each locus.