**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: [hemstrow@gmail.com](mailto:hemstrow@gmail.com)

\* Mark Christie; email: [christ99@purdue.edu](mailto:christ99@purdue.edu)

**This PDF includes:**

Supplementary Methods

Supplementary Example

Supplementary References

Supplementary Figures 1-6

**Supplementary Methods:**

## Coalescent simulations:

To test *S’* estimations with a realistic site frequency spectrum, we used the coalescent simulator *scrm* (Staab et al., 2015) with a relatively low per-bp mutation rate of 1x10-6, a per-bp recombination rate of 2x10-8, and Ne of 1,000 alongside a chromosome size of 1x106 bp to generate a relatively small number of variant sites for 5,500 gene copies. We then subset the resulting data randomly to 10,000 SNPs, then calculated *S’* and *S* via manual rarefaction for a range of γ and iteration numbers as described in the main text.

## Simulating inbred, outbred, and neutral populations:

To test the accuracy and precision of *S’*, projected *S*, and Watterson’s θ in either inbred or outbred populations, we used genotypic data from Hemstrom et al (2022) with no minor allele frequency filter, phased and imputed it using beagle 5.4 (Browning et al., 2018) with the default settings, then subset individuals to keep those from either North America or Queensland, Australia. The North American population was used as our inbred population.

To simulate outbreeding, we conducted forward simulations with a high rate of migration between the North American and Queensland populations. To do so, we randomly assigned sexes and paired individuals to produce a total of four offspring per pair. We allowed chromosomes to recombine by randomly placing recombination events, where was determined using a Poisson distribution with = 1 for an average of one recombination event per chromosome. 50% of offspring were then moved between populations and mating was repeated as above to generate two outbred populations. Density dependence was simulated by capping the number of parents at 50 in each generation. We used the outbred “North American” population as our outbred population. To simulate a neutral population, we then simulated the “North American” population as above without migration for another five generations.

# Supplementary Example:

The unbiased expected number of segregating sites after standardizing sample sizes, *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\_sites\_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\_rarefaction/main/data/example\_vcf.vcf",   
 destfile = "example\_vcf.vcf")

This VCF file contains a subset of data from Hemstrom et al. (Hemstrom et al., 2022) which includes genotypes 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), γ = *nmin*, where *nmin* 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 *nmin – 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:

# 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 *S’* and its variance , 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 (, and the variance of , , respectively.

# Supplementary References:

Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *The American Journal of Human Genetics*, *103*(3), 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015

Hemstrom, W. B., Freedman, M. G., Zalucki, M. P., Ramírez, S. R., & Miller, M. R. (2022). Population genetics of a recent range expansion and subsequent loss of migration in monarch butterflies. *Molecular Ecology*, *31*(17), 4544–4557. https://doi.org/10.1111/mec.16592

Hemstrom, W., & Jones, M. (2023). snpR: User friendly population genomics for SNP data sets with categorical metadata. *Molecular Ecology Resources*, *23*(4), 962–973. https://doi.org/10.1111/1755-0998.13721

Staab, P. R., Zhu, S., Metzler, D., & Lunter, G. (2015). scrm: Efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics*, *31*(10), 1680–1682.

# Supplementary Figures:

A graph of a number of data

Description automatically generated with medium confidence

**Figure S1:** The percentage of expected segregation probabilities 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.

A graph of a graph

Description automatically generated with medium confidence

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

A diagram of a target

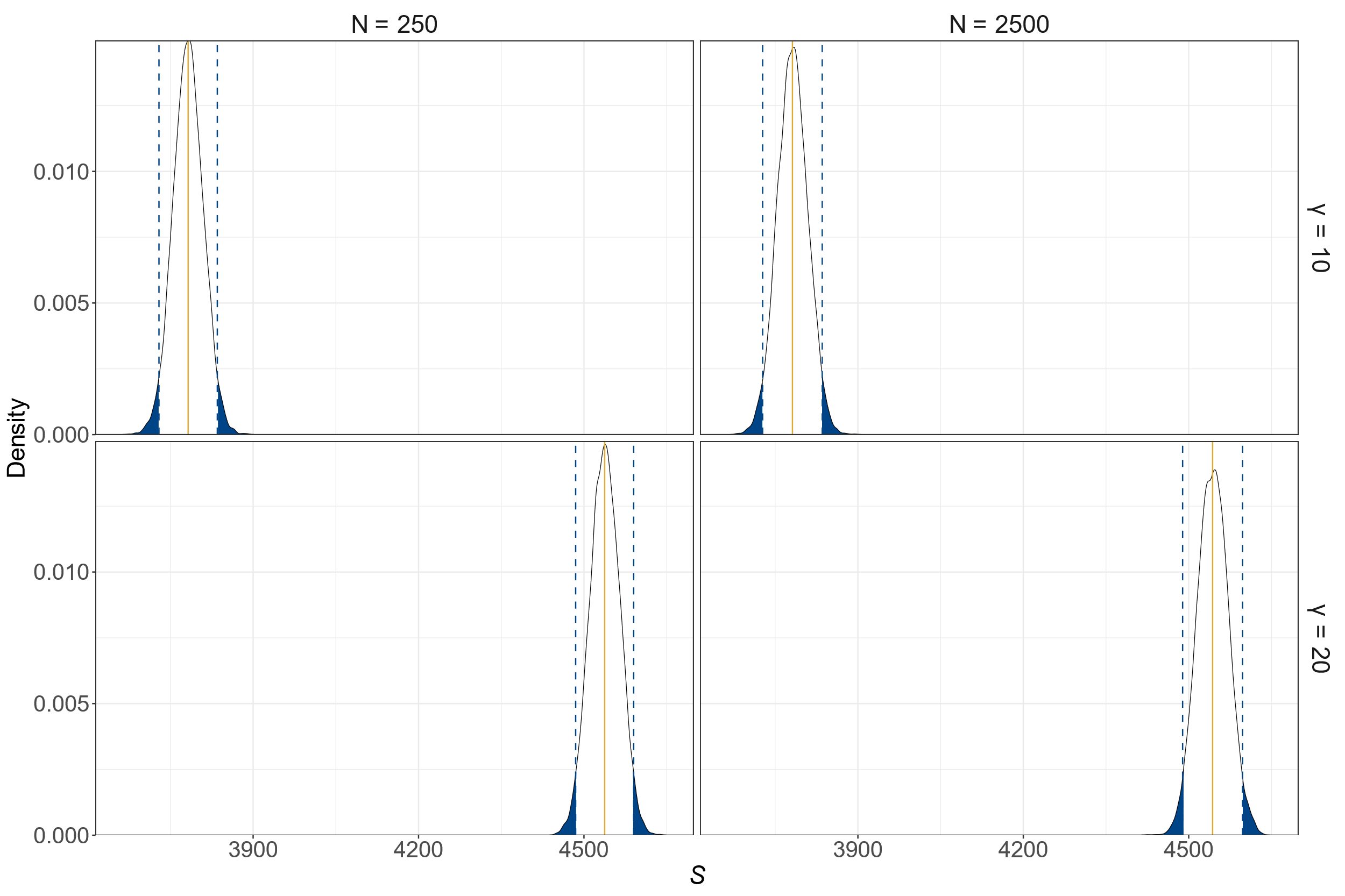
Description automatically generated with medium confidence

**Figure S3:** The expected number of segregating sites (*S’*, red) compared to the distributions of the number of segregating sites observed after rarefaction based on 100 simulations (black). *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.

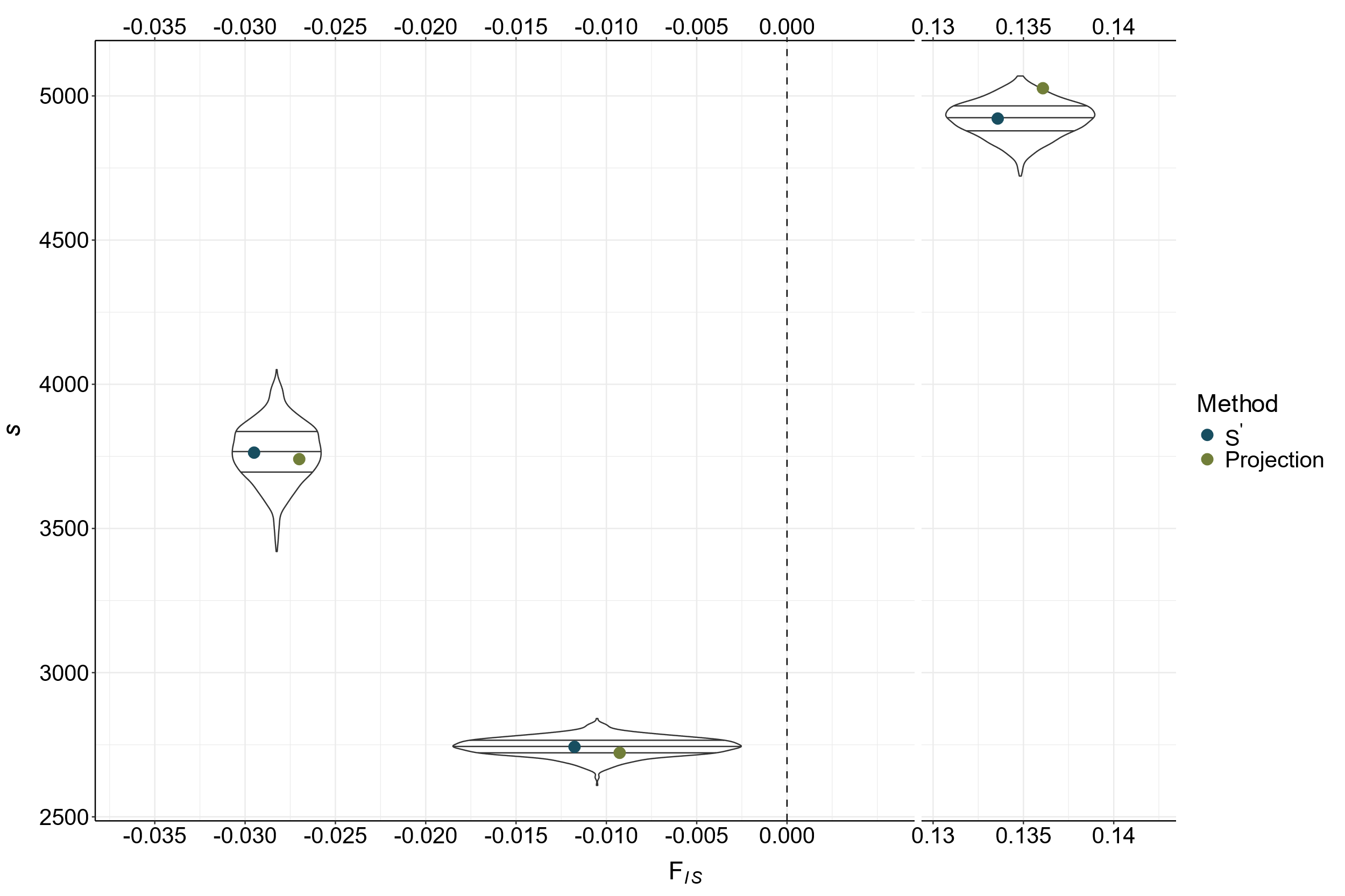
A graph of a function

Description automatically generated

**Figure S4:** The expected probabilities of observing a segregating site at each locus () for loci generated with the coalescent simulator *scrm*. Probabilities are derived from Equation 4 for each locus for population sizes of 250 and 2500, rarefacted to sample sizes of either 10 or 100 (corresponding to = 10 or 100). Since loci were not simulated in HWP, points vary to a small degree due to variation in genotype frequencies for a given minor allele frequency. Each locus had a random, independent percentage of missing data between 0% and 30%. Points are colored depending on estimates were contained within a 95% confidence intervals (marked with error bars) based on 10,000 simulated rarefaction trials for each minor allele frequency at each sample size. Compare to main text Figure 1, which was generated with known, pre-defined allele frequencies.



**Figure S5:** The distribution of the total number of segregating sites (*S*) observed for 10,000 replicate simulated trials, rarefacted to either 10 or 200 samples (corresponding to either = 10 or 20, respectively) for starting population sizes of either *N* = 250 or *N* = 2500. The mathematically expected number of segregating sites (*S’*)and 95% prediction are shown with solid yellow and dashed blue lines, respectively, for each distribution. Data was generated with the coalescent simulator *scrm*; compare to Figure 3 in the main text, which was generated with known, pre-defined allele frequencies.



**Figure S6:** Comparison between the number of segregating sites (*S*) calculated via projection (green), *S’* (blue), and observed, rarefacted *S* (black distributions, sampled 1,000 times) following projection or rarefaction to 25 individuals (or γ = 25 for *S’*) for outbred, neutral, or inbred populations (left to right, with corresponding *F*IS values). Points (not distributions) are offset on the *x* axis for clarity—each value is plotted for the specific *F*IS of the corresponding population.