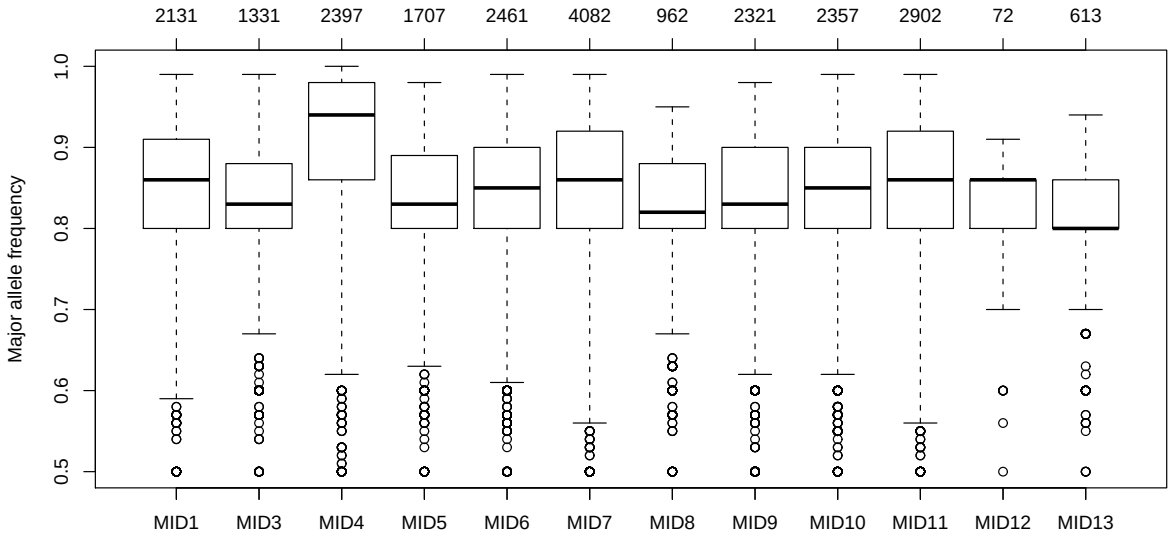
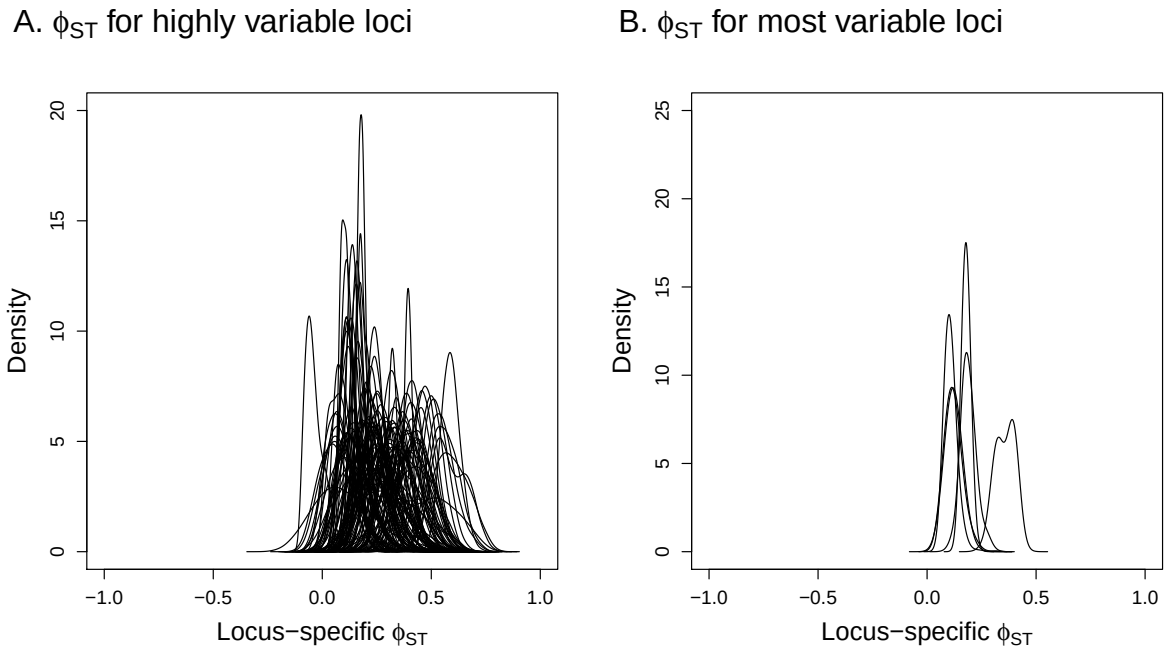
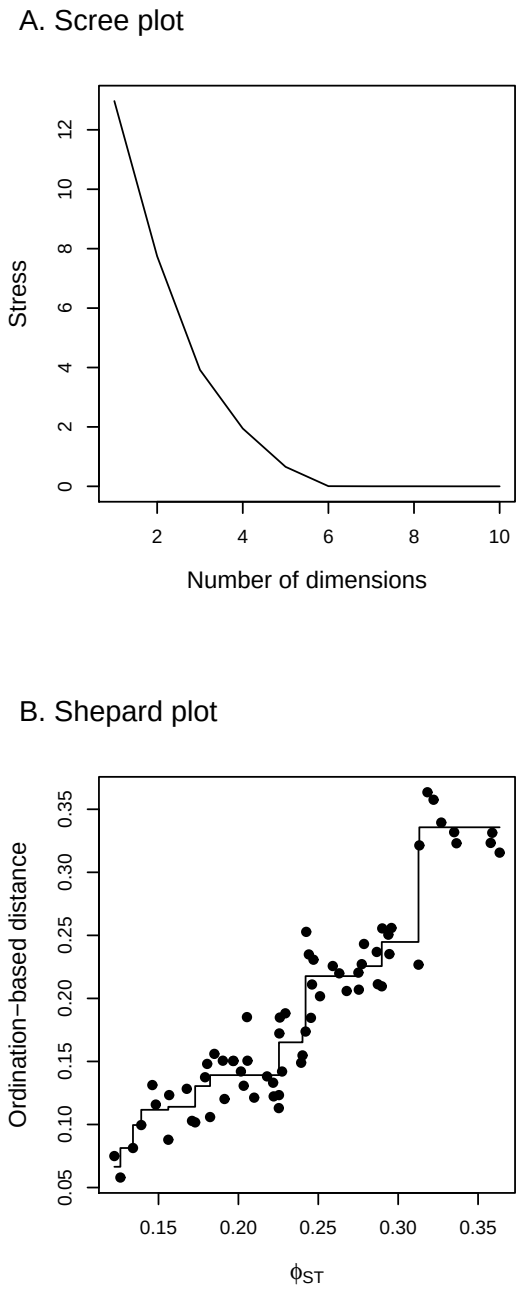
Supplementary Figure 1 Boxplots depicting the distribution of major allele frequencies for each population (MID). The number of SNPs per population with 5× coverage are given above each boxplot. Major allele frequencies were estimated by maximum likelihood while accounting for sequence errors.



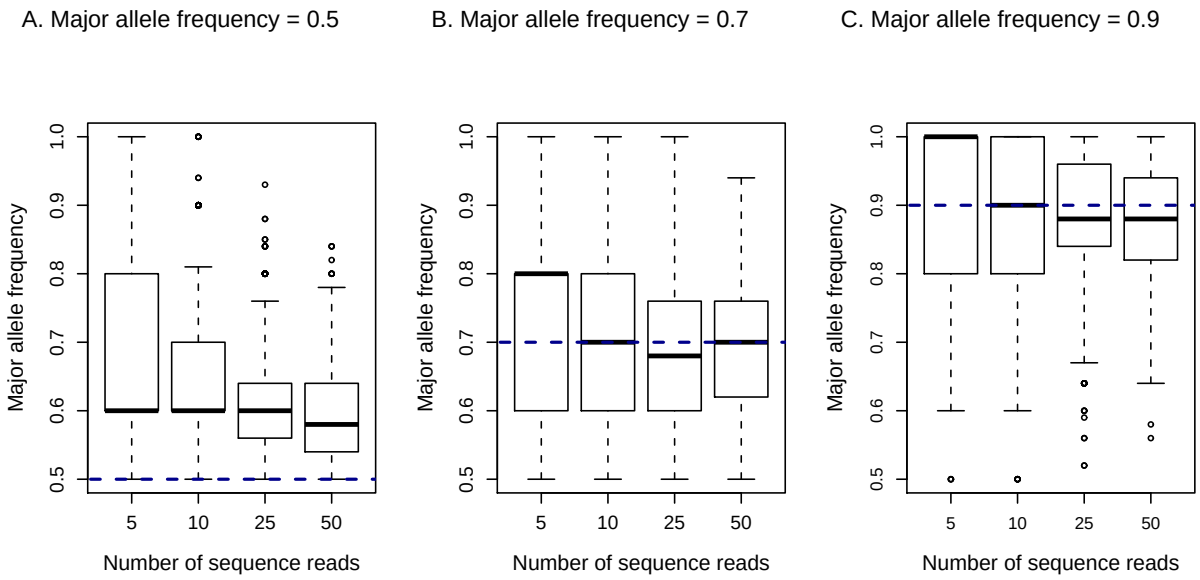
Supplementary Figure 2 Posterior probability distributions for locus-specific ϕ_{ST} for 125 loci with more $\geq$ 25 haplotypes (A) and six loci with $\geq$ 90 haplotypes. Posterior probability distributions were estimated from 24,000 MCMC iterations and were smoothed using a Gaussian kernel density function. These loci posses levels of variation that are similar to levels of variation for loci generally used in population genetic analyses.



Supplementary Figure 3 Scree plot (A) and Shepard plot (B) from NMDS analysis. The former shows the relationship between the number of dimensions used for NMDS and the sum of squared differences between the ordination-based distances and the distances predicted by regression (stress). The latter depicts the relationship between genome-level pair-wise ϕ_{ST} and ordination-based distances (points) as well as the predicted values from regressing the latter on the former (solid line). These plots suggest that the three dimensions used for NMDS captured the structure of the pair-wise ϕ_{ST} well.



Supplementary Figure 4 Boxplots depicting the distribution of major allele frequencies for simulated data. Dashed blue lines denote the major allele frequency used for each set of simulations.



Supplementary document S1

We conducted simulations to determine whether allele frequency estimates using Eqn. 1 were unbiased and robust to low coverage. We simulated allele counts by first sampling 30 gene copies (the number sampled in our empirical study) from a population with a major allele frequency π_i and then sampling 5, 10, 25, or 50 sequence reads from the sampled gene copies with a sequence error probability of 0.005 (approximately the mean sequence error probability estimated from our empirical data). We conducted simulations with π_i set to 0.5, 0.7, and 0.9. For each combination of major population allele frequency and read number, we generated 1000 replicate data sets. We then obtained maximum likelihood estimates of π_i for each replicate data set using Eqn. 1, as described in the manuscript for our empirical data. The distribution of maximum likelihood estimates for π_i under each set of simulation parameters is shown in Sup. Fig. 4. We detected a clear tendency to overestimate the population major allele frequency when the true value of π_i was 0.5. This result was not surprising as our estimate of the major allele frequency must always be ≥ 0.5, otherwise the alternative allele would be defined as the major allele. For $\pi = 0.7$ or 0.9 the distributions of maximum likelihood estimates of π_i were generally centered on the true value, although with only 5 sequence reads the median estimated value tended to be higher than the simulated value for π. This discrepeancy largely reflects the fact that maximum likelihood estimates for π_i with five sequence reads and few sequence errors will generally be 0.6, 0.8, or 1 (i.e., finer discrimination is not possible with five reads), and thus are unlikely to yield a median value of 0.7 or 0.9 (the means where closer to the simulated values).