GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.139519/-/DC1

A Maximum-Likelihood Method to Correct for Allelic Dropout in Microsatellite Data with No Replicate Genotypes

Chaolong Wang, Kari B. Schroeder, and Noah A. Rosenberg

DOI: 10.1534/genetics.112.139519

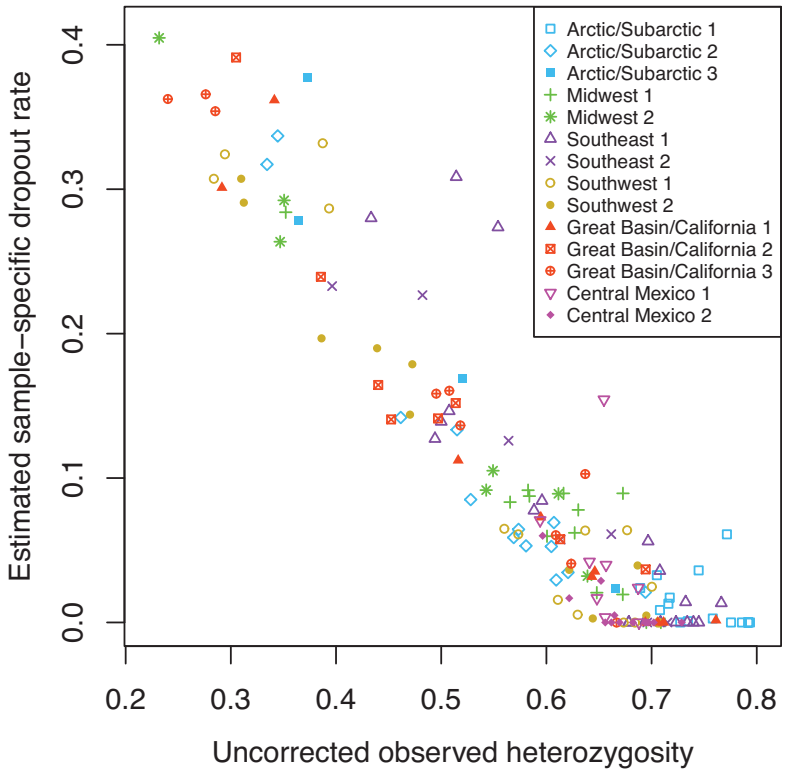

Figure S1: The estimated sample-specific dropout rate versus the observed heterozygosity before correcting for allelic dropout in the Native American data. For each individual, loci with both copies missing are excluded from the calculation of observed heterozygosity.

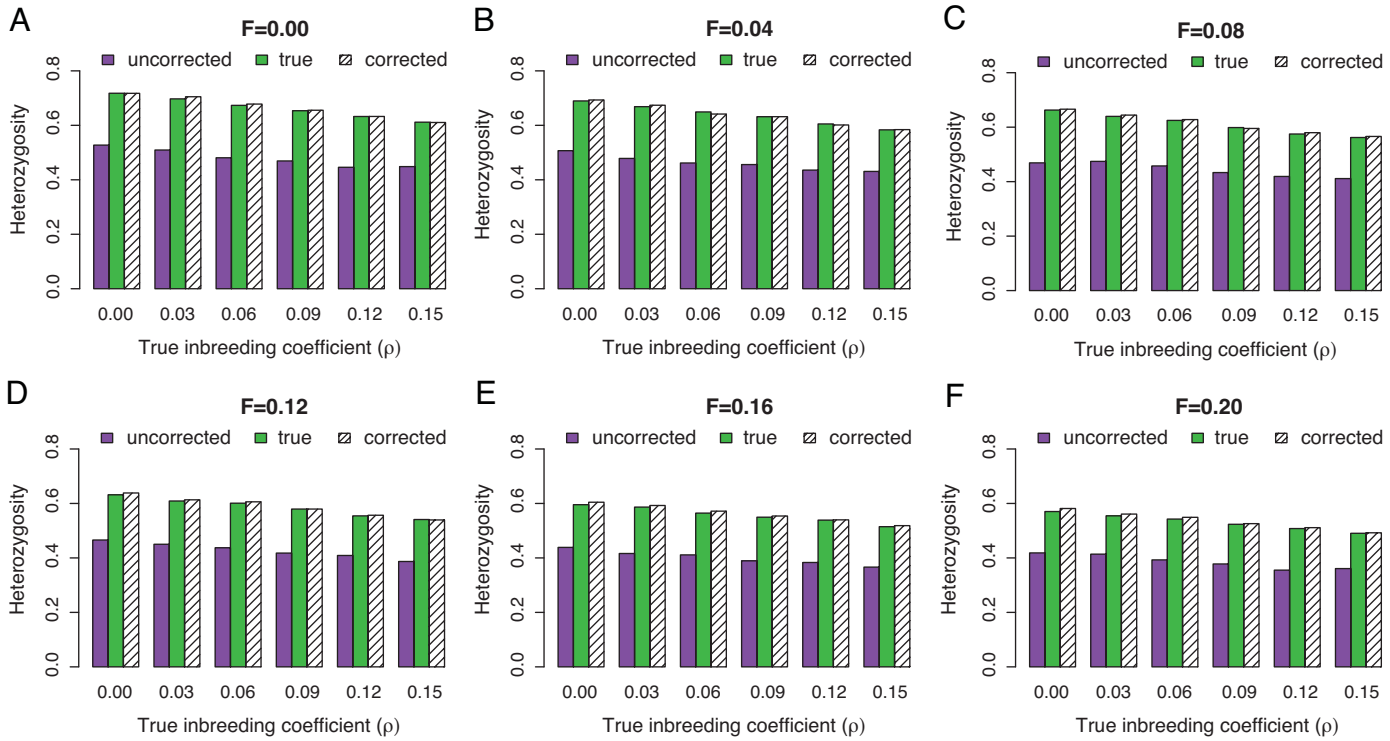

Figure S2: Correcting the underestimation of observed heterozygosity for simulated data with population structure. In each panel, a purple bar indicates the uncorrected observed heterozygosity averaged across all individuals in a simulated data set after applying allelic dropout; a green bar indicates the "true" observed heterozygosity averaged across all individuals in the same simulated data set before applying allelic dropout; and a striped black bar indicates the corrected observed heterozygosity averaged across all individuals and across 100 imputed data sets. The x-axis indicates values of the inbreeding coefficient that were set for different simulations. Different panels correspond to different values of the F parameter in the F-model for simulating structured populations. (A) $F = 0$; (B) $F = 0.04$; (C) $F = 0.08$; (D) $F = 0.12$; (E) $F = 0.16$; (F) $F = 0.20$.

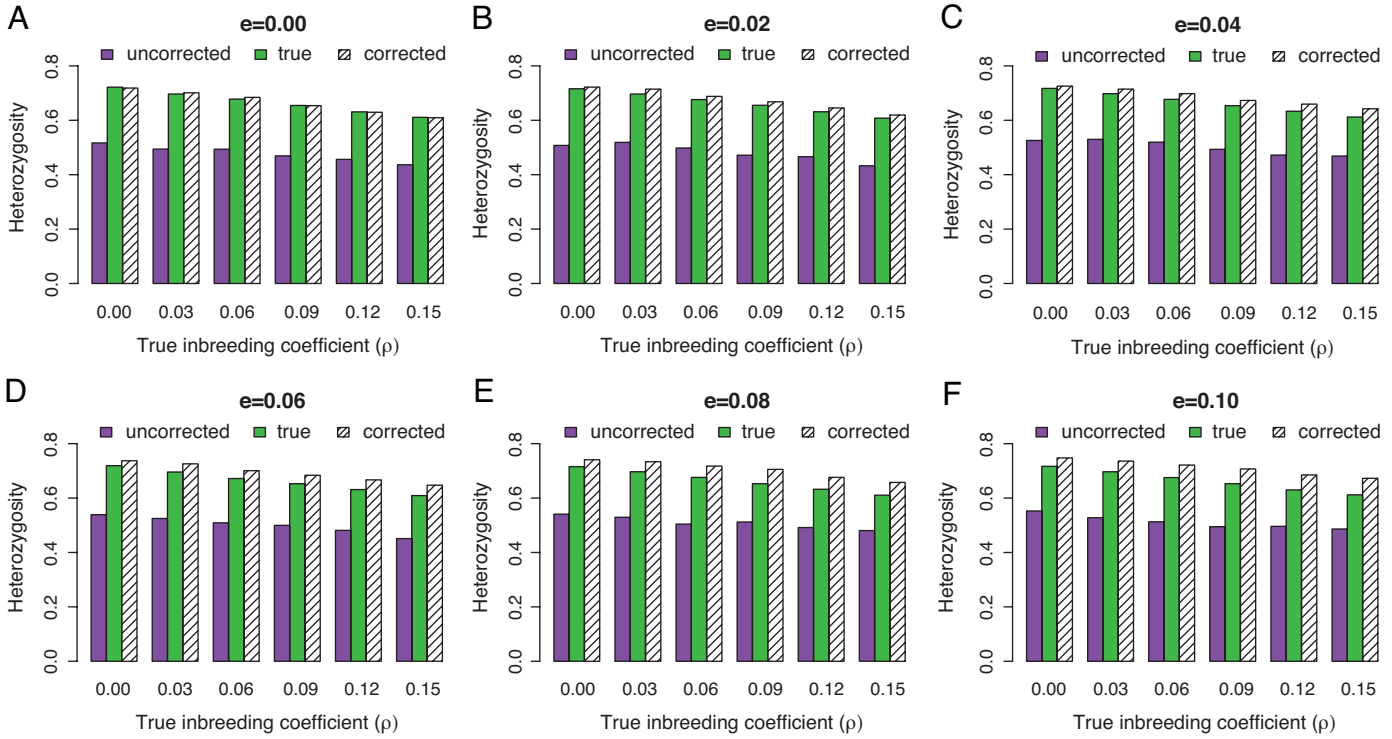

Figure S3: Correcting the underestimation of observed heterozygosity for simulated data with genotyping errors other than allelic dropout. In each panel, a purple bar indicates the uncorrected observed heterozygosity averaged across all individuals in a simulated data set after applying allelic dropout; a green bar indicates the "true" observed heterozygosity averaged across all individuals in the same simulated data set before applying allelic dropout and before introducing genotyping errors; and a striped black bar indicates the corrected observed heterozygosity averaged across all individuals and across 100 imputed data sets. The x-axis indicates values of the inbreeding coefficient that were set for different simulations. Different panels correspond to different levels of simulated genotyping errors that come from sources other than allelic dropout. (A) $e = 0$; (B) $e = 0.02$; (C) $e = 0.04$; (D) $e = 0.06$; (E) $e = 0.08$; (F) $e = 0.10$.

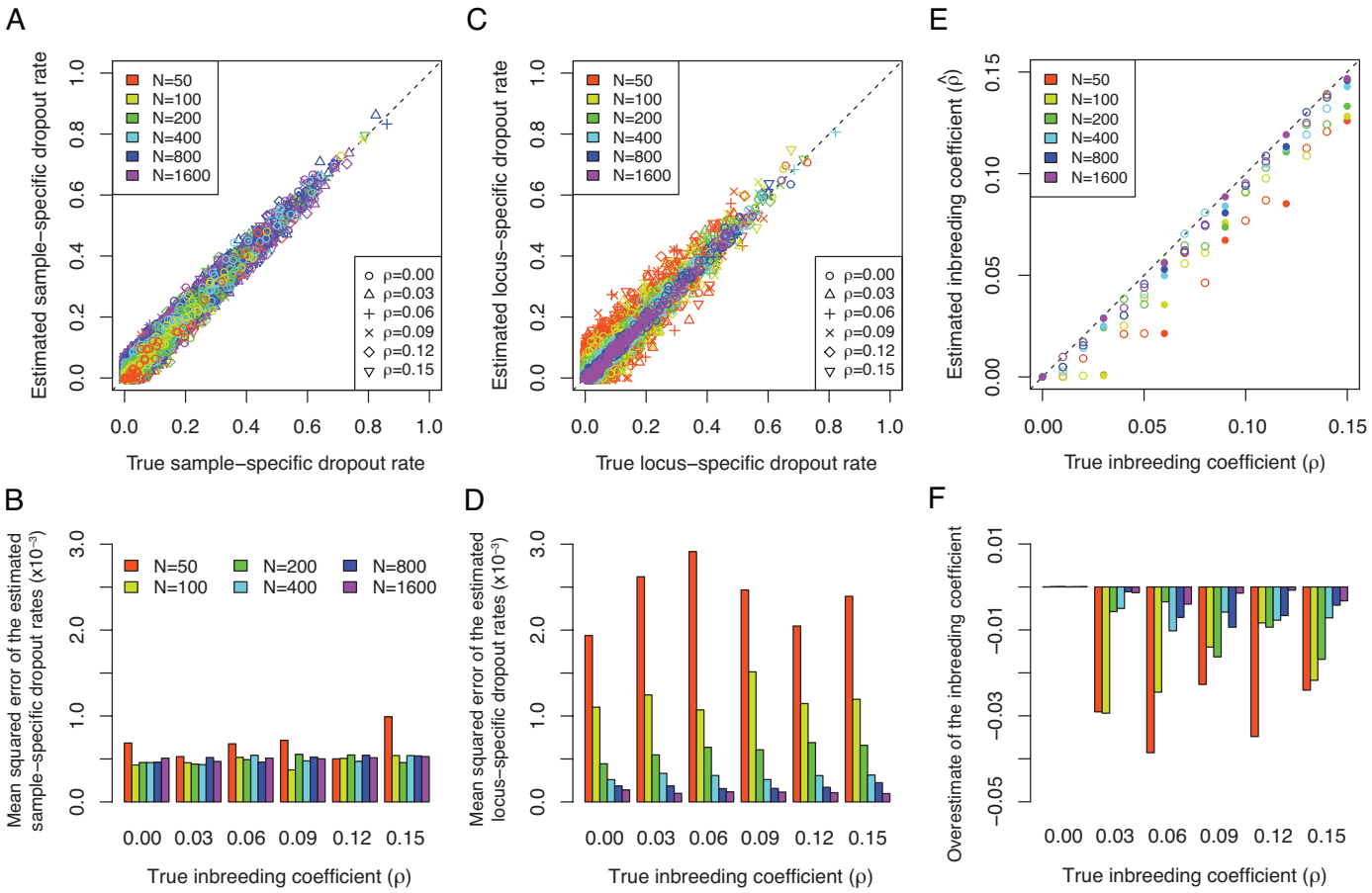

Figure S4: Estimated dropout rates and inbreeding coefficients for simulated data with different numbers of individuals and the same number of loci ($L = 250$). Each data set was simulated with no population structure and no genotyping errors other than allelic dropout. (A) Comparison of the estimated sample-specific dropout rates and the assumed true sample-specific dropout rates. (B) Mean squared errors across all the estimated sample-specific dropout rates for each of the 36 data sets shown in panel A. (C) Comparison of the estimated locus-specific dropout rates and the assumed true locus-specific dropout rates. (D) Mean squared errors across all the estimated locus-specific dropout rates for each of the 36 data sets shown in panel C. (E) Comparison of the estimated inbreeding coefficient and the assumed true inbreeding coefficient, in which each point corresponds to one of 96 simulated data sets. The 36 solid points correspond to the simulated data sets shown in the other panels (A, B, C, D, and F). (F) Overestimation of the inbreeding coefficient, calculated by subtracting the assumed true inbreeding coefficient from the estimated inbreeding coefficient, or $\hat{\rho} - \rho$.

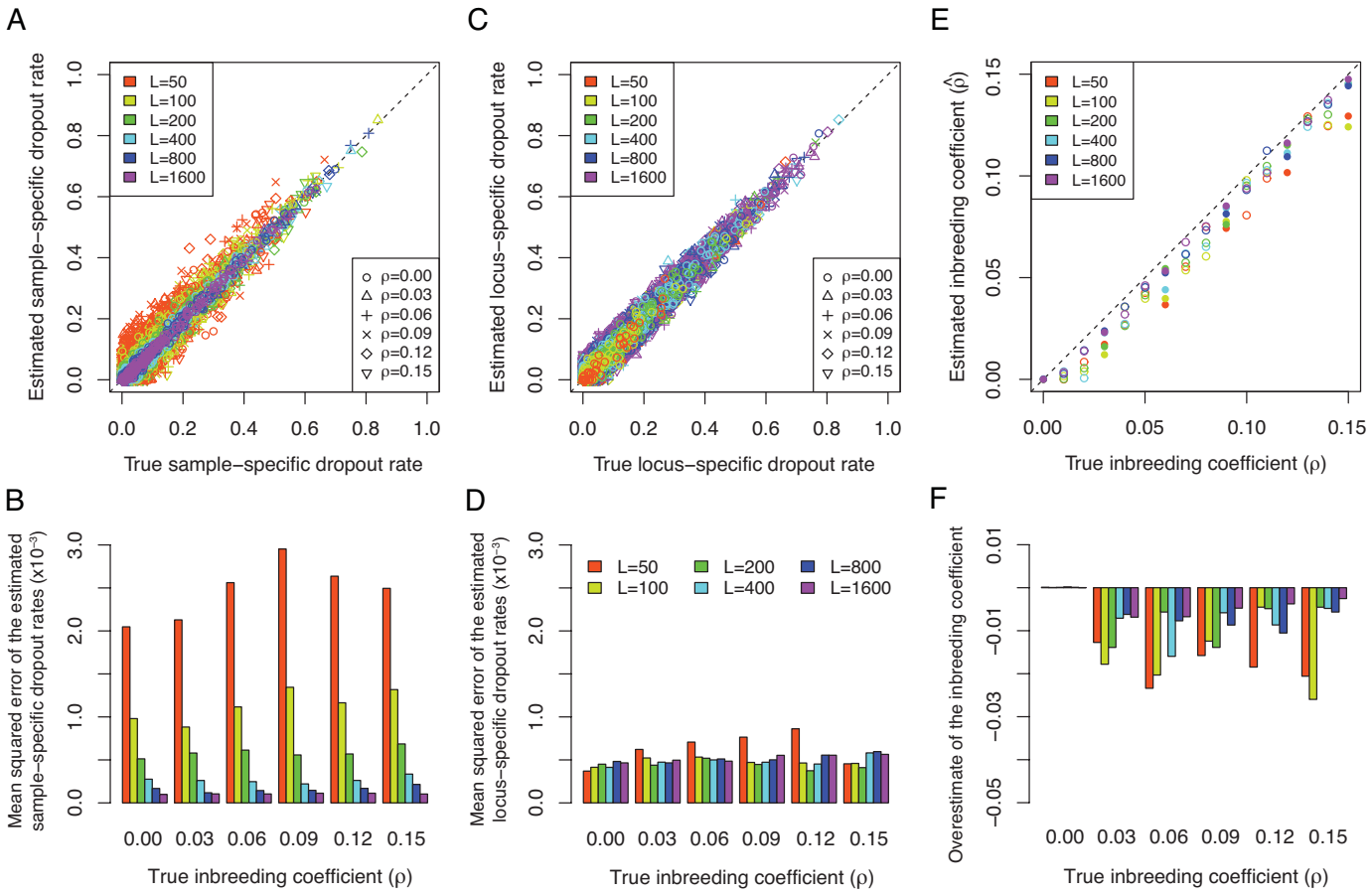

Figure S5: Estimated dropout rates and inbreeding coefficients for simulated data with different numbers of loci and the same number of individuals ($N = 250$). The allele frequencies for the loci were sampled with replacement from the MLEs of the Native American data. Each data set was simulated with no population structure and no genotyping errors other than allelic dropout. (A) Comparison of the estimated sample-specific dropout rates and the assumed true sample-specific dropout rates. (B) Mean squared errors across all the estimated sample-specific dropout rates for each of the 36 data sets shown in panel A. (C) Comparison of the estimated locus-specific dropout rates and the assumed true locus-specific dropout rates. (D) Mean squared errors across all the estimated locus-specific dropout rates for each of the 36 data sets shown in panel C. (E) Comparison of the estimated inbreeding coefficient and the assumed true inbreeding coefficient, in which each point corresponds to one of 96 simulated data sets. The 36 solid points correspond to the simulated data sets shown in the other panels (A, B, C, D, and F). (F) Overestimation of the inbreeding coefficient, calculated by subtracting the assumed true inbreeding coefficient from the estimated inbreeding coefficient, or $\hat{\rho} - \rho$.

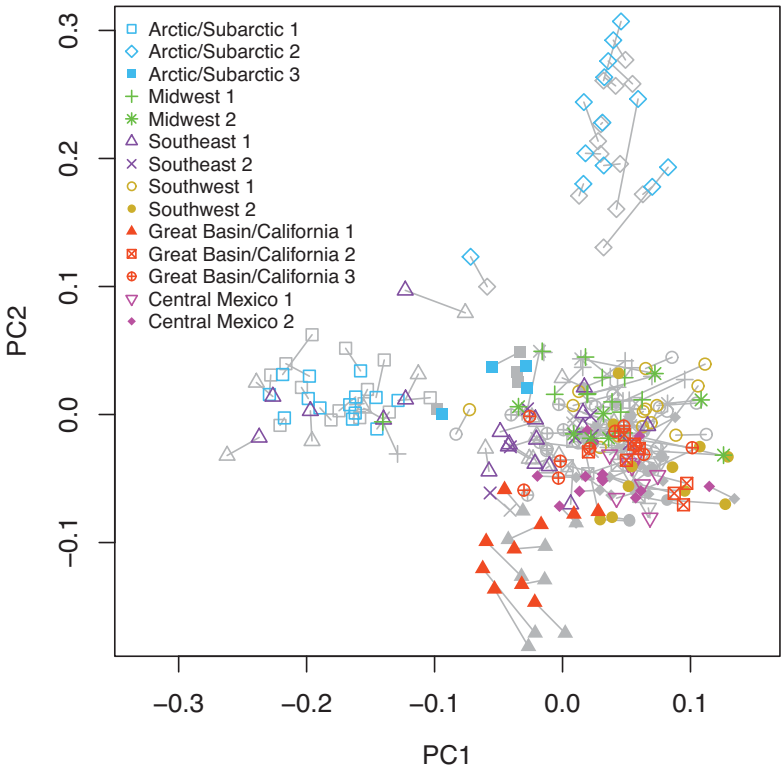

Figure S6: Multidimensional scaling (MDS) analysis of the Native American data. The results of MDS analysis on the original microsatellite data are shown by colored points, with the x-axis corresponding to the first principal coordinate and the y-axis corresponding to the second principal coordinate. The results of MDS analysis on one set of imputed microsatellite data are displayed with gray points, Procrustes-transformed to best match the results from the original data (*Stat. Appl. Genet. Mol. Biol.* 13: 9, 2010). Each pair of corresponding points is connected by a gray line. The allele-sharing distance matrices calculated from the original data, averaging across loci and ignoring loci for which one or both individuals was missing, and from one set of imputed data (after correcting for allelic dropout) were used as the input to the *cmdscale* function in *R*.

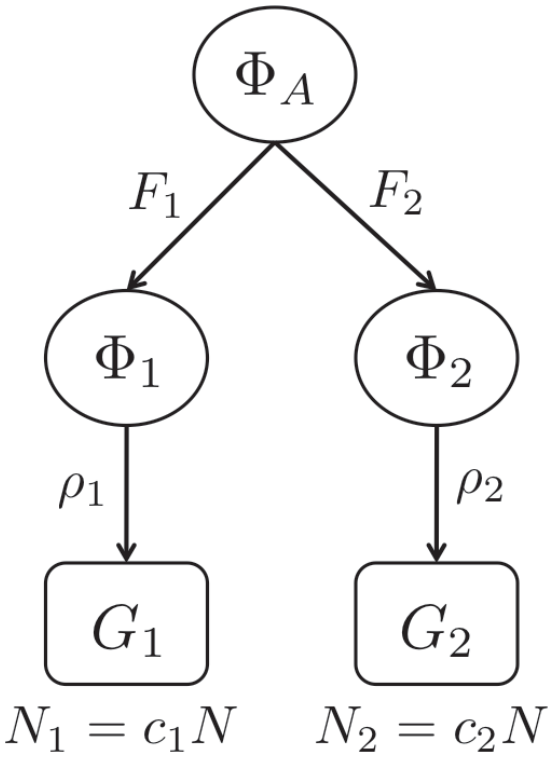

Figure S7: Illustration of a structured population with two subpopulations, under the F model. Φ_A denotes the allele frequencies of a common ancestral population of the two subpopulations. Φ_1 and Φ_2 are allele frequencies of the two subpopulations. The F parameter and the inbreeding coefficient for subpopulation j are F_j and ρ_j, respectively ($j = 1, 2$). In the pooled genotype data of N individuals, c_1 is the proportion sampled from subpopulation 1, producing genotype data G_1, $c_2 = 1 - c_1$ is the proportion sampled from subpopulation 2, producing genotype data G_2.