Simulation of Effects of Dominance on Estimates of Population Genetic Diversity and Differentiation

**K. V Krutovskii, S. Y. Erofeeva,
J. E. Aagaard, and S. H. Strauss**

The advent of PCR-based molecular markers has led to a rapid expansion in studies describing the levels and distribution of genetic variation among populations at the DNA level. Randomly amplified polymorphic DNA (RAPD; Williams et al. 1990) and amplified fragment length polymorphism (AFLP; Vos et al. 1995) markers are now commonly used in population genetic studies (e.g., Aagaard et al. 1998; Isabel et al. 1995; Liu and Furnier 1993; Mosseler et al. 1992; Peakall et al. 1995; Szmidt et al. 1996; Travis et al. 1996; Wu et al., in press). However, these PCR-based markers have limitations compared to allozymes, which had been the prevalent means for population studies prior to the use of PCR. At the majority of RAPD and AFLP loci the dominant allele masks the presence of the null allele in heterozygotes when assaying diploid tissues (e.g., about 97%-98%; Krutovskii et al. 1998), thus sampling variance for dominant allele frequencies is typically greater than that for codominant alleles (Lynch and Milligan 1994). The frequencies of null and dominant alleles are inferred from the frequency of null allele homozygotes; the precision of their estimation thus depends on mating system assumptions and is strongly affected by the sample size. Empirical studies have also suggested that dominant markers can bias estimates of genetic diversity and differentiation among populations (e.g., Isabel et al. 1995; Szmidt et al. 1996).

Although RAPD markers have proved to be useful for population studies, and their gross patterns of diversity usually agree with that of allozymes, the levels of genetic variation, differentiation, and fine-scale genetic structures often differ (e.g., Baruffi et al. 1995; Dawson et al. 1996; Heun et al. 1994; Lann(!r-Herrera et al. 1996; Latta and Mitton 1997; le Corre et al. 1997; Liu and Furnier 1993; Peakall et al. 1995; Puterka et al. 1993). To help assess whether these differences are biological or a simple consequence of the dominance and biallelism of RAPD and AFLP markers, we developed a dominance simulation program, DOMSIM, that transforms codominant population data into a biallelic dominant dataset. The program then estimates population genetic statistics with which dominant and codominant markers can be directly compared. We use data from a widespread North American conifer, Douglas-fir [Pseu*dotsuga menziesii* (Mirb.) Franco], and three California closed-cone pine species to illustrate the program's function. The test simulation suggests that dominant biallelic markers, such as RAPDs, can strongly underestimate population diversity but can still reasonably estimate population differentiation (G5,), if sample sizes are larger than about 30 individuals.

Program Functions

The program DOMSIM uses multiallelic datasets with a maximum number of six alleles per locus for which population allele frequencies are defined. Assuming HardyWeinberg equilibrium and no linkage among loci, the program generates *N* basic populations (Nm_ = 20) of up to 1,000 individuals each with multilocus genotypes that maintain the specified allele frequencies within populations. A total of S subpopulations (Smu = 400) of *n* individuals (n = 10-200) are then drawn with replacement for each of the *N* populations. The sampling is done in two different ways: by sampling subpopulations of size *n* with replacement directly from the initially generated basic population, and by resampling subpopulations of size *n* with replacement within the first sampled subpopulation of n individuals (bootstrap resampling). Population genetic parameters *(HS, HT,* and *GS,)* are calculated for each cycle of resampling in three ways. First, for a codominant dataset, calculations are made considering all alleles and genotypes present in the subpopulations. Second, the same subpopulations and data are used to simulate a dominant biallelic dataset by randomly selecting one allele as dominant, with the rest treated as recessive to it. The synthetic null allele frequency is then calculated from the null homorygote frequencies assuming HardyWeinberg equilibrium. Average parameters and their variance are calculated for each set of S subpopulations. Gene diversity is evaluated using *HS* and *HT,* either unmodified (Nei 1973) or modified (Nei and Chesser 1983) for the sample size. Genetic differentiation is evaluated via ew (Weir and Cockerham 1984) and G, parameters that are either unmodified (Nei 1973), modified for the sample size (Nei and Chesser 1983), or modified for both the sample size and population number (Nei 1986). Finally, null allele frequencies are corrected for dominance using Lynch and Milligan's (1994) equation 2a, and their asymptotically unbiased estimate of *FS,* recommended for dominant markers is also calculated following equation 14a.

Installing and Running the Program

The program DOMSIM is written in FORTRAN-77 (simulation routines) and in LabWindows CVI (interface routines). The source code file, domsimd.f, was compiled using Microsoft FORTRAN Power Station Compiler version 1.0. DOMSIM runs on IBM PCs and compatibles under MS Windows 95 and NT for 32-bit operating environments. To install the program run the compressed self-extracting file domsimpr.exe which can be downloaded from the web site http://www.fsl.orst.edu/tgerc/ protocol.htm. It will automatically decom press five files domsimd.f, domsim.001, domsim.002, read.me, and setup.exe. Next, run the setup file and follow the instructions on your screen during installation. Run the program by either clicking the icon or executing the program file domsim.exe. A read.me file contains additional instructions for installing and running the program.

Input and Output Files

The input format is an ASCII file similar to GeneStat input files (Lewis 1994), but does not require population, locus, and allele names, and there should be no empty lines. An example (sample.dat) and brief help, which explicitly explains an input file structure, are provided with the program. The output file has all the parameters calculated for each resampled and bootstrap set, their average values, and standard deviations.

Examples of Simulation Based on Allozyme Data in Douglas-fir and California Closed-Cone Pines

In order to facilitate comparisons between dominant and codominant markers, and to help understand the effects of RAPD dominance and biallelism on our studies of genetic diversity and differentiation in Douglas-fir (Aagaard et al. 1998) and the California closed-cone pines (Wu et al., in press), we simulated dominance and biallelism in these two allozyme datasets (Li and Adams 1989; Wu et al., in press). The first allozyme dataset included six populations of three races of Douglas-fircoastal, north interior, and south interior-with two populations per race. The second one included four, five, and three populations of *Pinus attenuata, P. muricato,* and P. *radiata,* respectively. These populations are described in detail elsewhere (Aagaard et al. 1998; Wu et al., in press). From allozyme allele frequencies within populations we generated simulated populations of 1,000 individuals each, and a total of 400 subpopulations of *n* individuals were drawn with replacement from each of the populations. The program also performed 400 bootstrap resamplings using a subpopulation of size *n*. Population genetic parameters (H5, H,, G57, 9w, and FST) were then calculated for each set of 400 subpopulations in the three ways described above. We varied the number of individuals (rt) within the subsamples from 10 to 200 to bracket the range of sample sizes that might reasonably be employed in population studies, and the sample size of 30-50 trees per population that was used in our RAPD studies (Aagaard et al. 1998; Wu et al., in press). The results of the simulations are summarized in Figures 1 and 2. The simulations showed that diversity measurements (HS and H,) were likely to be underestimated by dominant biallelic markers approximately twofold regardless of sample size.

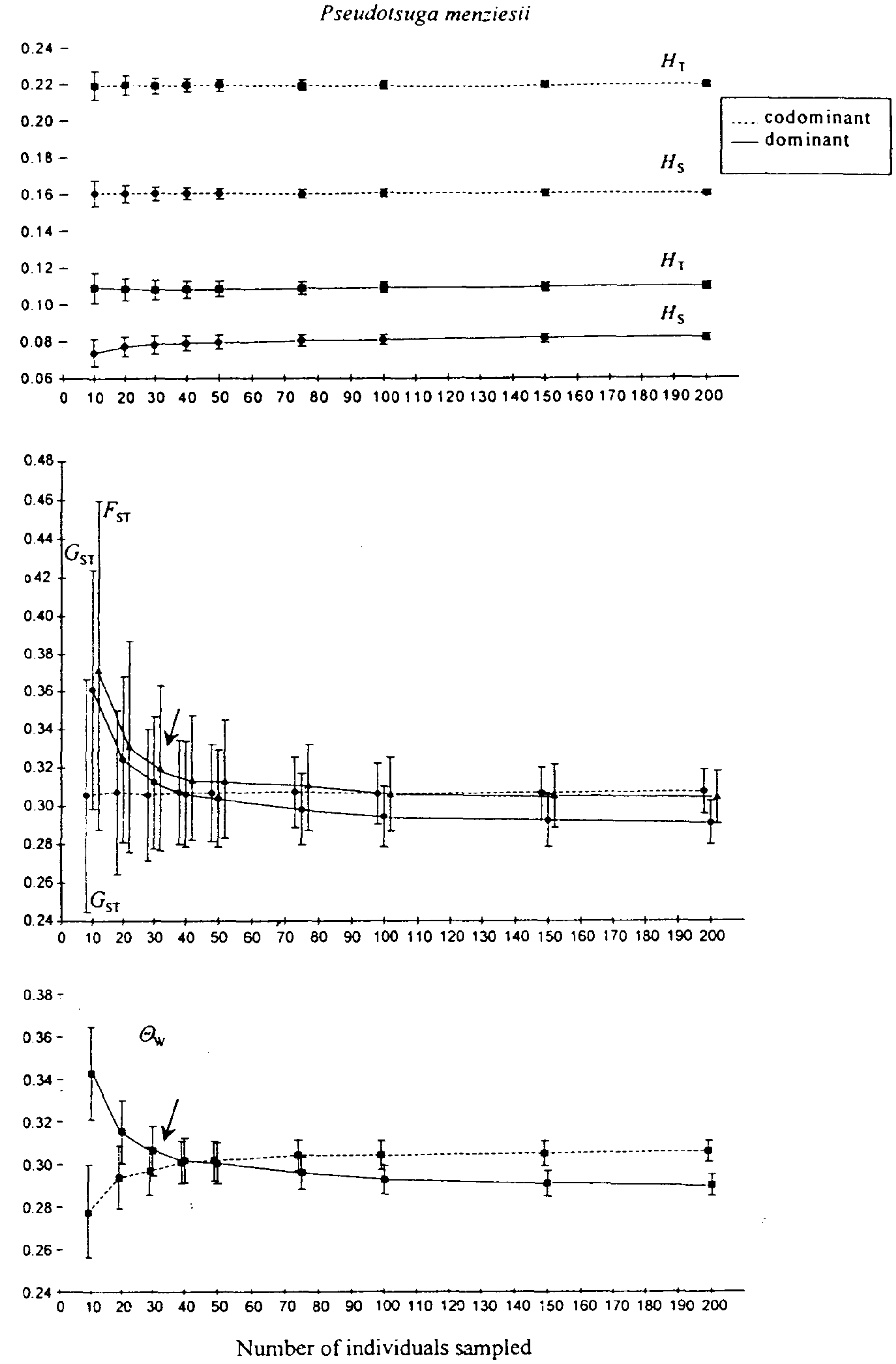

Figure 1. Levea of diversity and differentiation for codominant, multia!lelic allozymes versus biailelic, dominant markers, as simulated from an allozyme dataset from Douglas-fir studied with varying sample sizes. Standard deviations (error bars) were calculated from the variance among 400 bootstrap subsamples and represent the var;ance due to res3rnpiing of individuals at each level of sampling from the master population of 1.000 individuals. The arrow shows the population sample size between 30 and 40 needed to eliminate the tendency for overestimation of population differentiation caused by dominance and biallelism.

When 30 or more diploid individuals per population were sampled, there was little effect on differentiation estimates (GST, 6„,

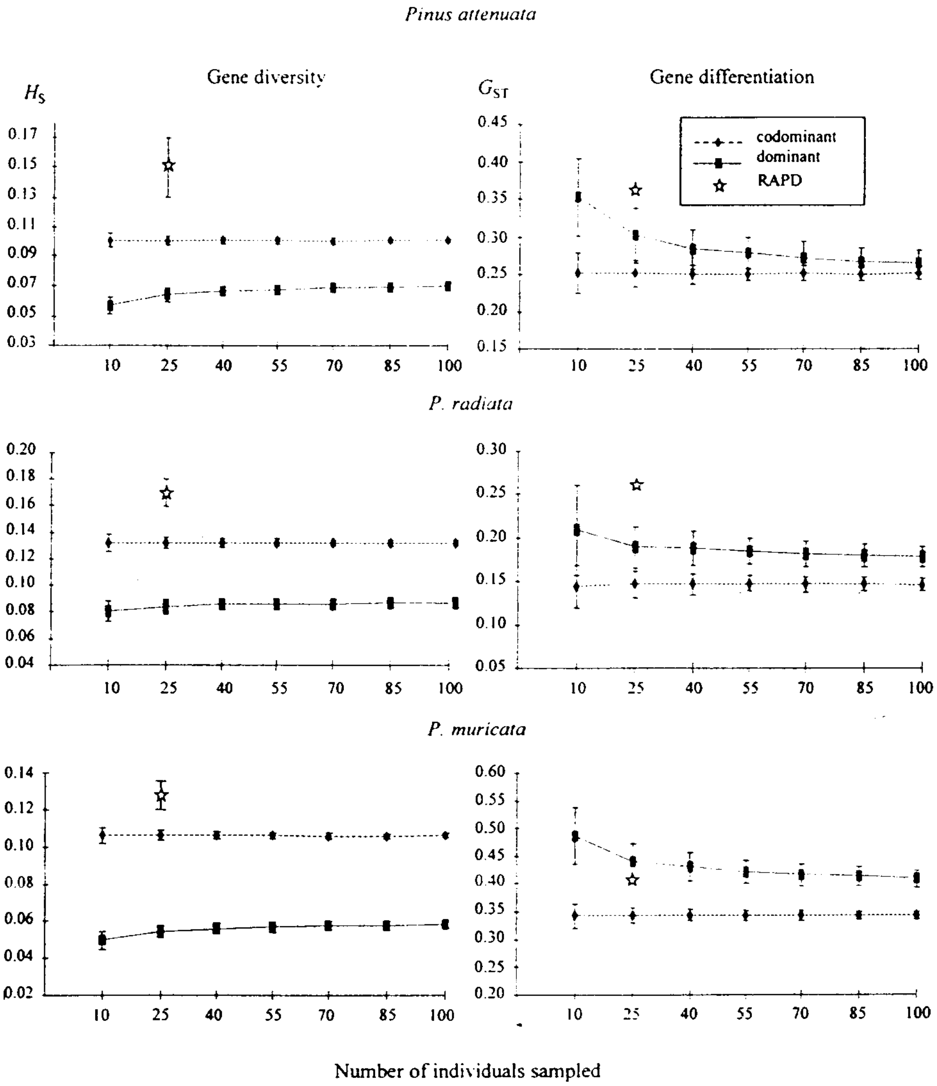

Figure 2. Genetic diversity (H_S) and differentiation (G_{ST}; Nei 1986) values averaged over populations of each California closed-cone pine species for codominant multiallelic allozyme and dominant biallelic markers simulated in the samples of different sizes. Standard deviations (error bars) were calculated from the variance among 400 bootstrap subsamples simulated for each population of each species. Observed RAPD values are also shown as a star.

and FST) in Douglas-fir. However, though still very similar to the estimates for codominant markers, the estimates for the simulated dominant markers began to diverge slightly but significantly downward at large population sizes in Douglas-fir. In the California closed-cone pines, the estimates for the simulated dominant markers converge toward the estimates for codominant multiallelic markers at large population sizes, but were always significantly higher (Wu et al., in press). Our simulations were in close agreement with our empirical studies of Douglas-fir where, despite dominance and biallelism of RAPD markers, we have found that RAPDs and allozymes exhibit similar levels of differentiation at the population and race levels with adequate sample sizes (Aagaard et al. 1998). However, the California closed-cone pine allozyme data showed that the larger sample sizes than we employed in our RAPD study (Wu et al., in press) are desirable for a fair comparison of RAPD and allozyme data. Finally, despite the expectation of much reduced diversity for dominant biallelic markers predicted by the simulations, our RAPD data gave higher estimates of diversity than did allozymes in both Douglas-fir (Aagaard et al. 1998) and the California closed-cone pines (Wu et al., in press). This suggests that RAPD markers may have much higher intrinsic genetic diversity than do allozyme markers. Our results demonstrate the importance of simulations to help compare and interpret the results of population studies with dominant markers.

From the Departments of Forest Science (Aagaard. Krutovskii, and Strauss) and the College of Oceanic and Atmospheric Sciences (Erofeeva), Oregon State University. Corvallis. OR 91331-7501. We thank Tom Adams for providing allozyme data and Vladislav Erofeev (or help with computer software. This work was supported in part by NSF grants DEB 9300083 and BSR 895702 to S.H.S. The dominance simulation program (DOMSIM) is available for public use and can be downloaded as a self-extracting file domsimpr.exe from the TGERC web site: http:/www.fsl.orst.edu!tgerc/protocol.htm Address correspondence to Dr. K. V Krutovskii at the address above or e-mail: krutovskiik@fsl.orst.edu.

References

.4agaard JE. Krutovskii KV, and Strauss SH, 1998. RAPDs and allozymes exhibit similar levels of diversity and differentiation among populations and races of Douglas-fir. Heredity 81:69-78.

Baruffi L. Damiani G, Guglielmino CR, Bandi C, Malacrida AR, and Gasperi G. 1995. Polymorphism within and between populations of *Cerotitis copituto*: comparison between RAPD and multilocus enzyme electrophoresis data. Heredity 74:425-437.

Dawson fK. Simons AJ. Waugh R, and Powell W, 1996. Diversity and genetic differentiation among subpopulations Cliricidin septum revealed by PCR-based assays. Heredity 74:10-18.

Heun M. Murphy JP, and Phillips TD. 1994. A comparison of RAPD and isozyme analyses for determining the genetic relationships among *Auena s*terilis L. accessions. Theor Appl Genet 87:689-696.

Isabel N, Beaulieu J, and Bousquet J, 1995. Complete congruence between gene diversity estimates derived from genotypic data at enzyme and random amplified polymorphic DNA loci in black spruce. Proc Nat] Acad Sci USA 92:6369-6373.

Krutovskii KV. Vollmer SS, Sorensen FC, Adams WT. Knapp SJ, and Strauss SH. 1998. RAPD genome maps of Douglas-fir. J Hered 89:191-205.

Lanner-Herrera C. Gustalsson M, Falt AS, and Bryngelsson T, 1996. Diversity of wild Brassica oleraceae as estimated by isozyme and RAPD analysis. Genet Resources Crop Evol 43:13-23.

Latta RG and Mitton JB. 1997. A comparison of population differentiation across four classes of gene marker in limber pine (!'inus llexilis James). Genetics 146: 1153-1163.

le Corre V, fhrmolin-Lapegue S, and Kremer A, 1997. Genetic variation at allozyme and RAPD loci in sessile oak Querrus *petraea (Matt.)* Liebl.: the role of history and geography. Moll Ecol 6:519-529.

Lewis PO. 1994. GeneStat-PC 3.3. **Raleigh, North** Carolina: Department of Statistics. North Carolina State University.

Li P and Adams NT, 1989. Range-wide patterns of allozyme variation in Douglas-fir (PseudotsugQ merr.;iesir). Can J For Res 19:149-161.

Liu Z and Furrier GR, 1993. Comparison of allozyme, RFLP, and RAPD markers for revealing genetic variation within and between trembling aspen and bigtooth aspen. Theor Appl Genet 87:97-105.

Lunch M and Milligan BG, 1994. Analysis of population genetic structure with RAPD markers. Mol F.col 3:9199.

Mosseler A, Egger KN, and Hughes GA. 1992. Low levels of genetic diversity in red pine confirmed by random amplified polymorphic DNA markers. Can J For Res 22: 1332-1337.

Nei M. 1973. Analysis of gene diversity in subdivided populations. Proc Nail Acad Sci USA 70:3321-3323.

Nei M. 1986. Definition and estimation of fixation indices. Evolution 40:643-645.

Nei M and Chesser RK, 1983. Estimation of fixation indices and gene diversities. Ann Hum Genet 47:253-259.

Peakall R, Smouse PE, and Huff DR. 1995. Evolutionary implications of allozyme and RAPD variation in diploid populations of dioecious bu(falograss *Buchloe dactyio*ides. Mol Eeol 4:135-147.

Puterka CJ, Black N WC. Steiner WM, and Burton RL, 1993 Genetic variation and phylogenetic relationships among worldwide collections of the Russian wheat aphid, *Diuraphis noxia* (Mordvilko), inferred from allo ryme and RAPD-PCR markers. Heredity 70:60418.

Szmidt AE. Wang X, and Lu M, 1996. Empirical assessment of allozyme and RAPD variation in *Pinus* syluesms (L.) using haploid tissue analysis. Heredity 76:412-420.

Travis SE, Maschinski J, and Keim P, 1996. An analysis of genetic variation in *Asrrogalus* cremnophyfax var. cremrtophylar, a critically endangered plant, using AF1P markers. Mol Ecol 5:735-745.

Vos P, Hogers R, Sleeker M, Reijans M, van de Lee T Hornes M, Frijters A, Pot J. Peleman J. Kuiper M, and Zabeau M, 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407414.

Weir BS and Cockerham CC, 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358-1370.

Williams JG. Kubelik AR, Livak KJ, Rafalski JA, and Tingey SV, 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18:6531-6535.

Wu J, Krutovskii KV, and Strauss SH, in press. Nuclear DNA diversity, population differentiation and phylogenetic relationships in the California closedtone pines based on RAPD and allozyme markers. Genome.

Received April 12, 1998
Accepted February 18.1999
Corresponding Editor: Robert Angus