#21: A New Unbiased Estimator of Gene Diversity for Samples Containing Related, Inbred, and Non-Diploid Individuals with Improved Variance

Alexandre M. Harris, Michael DeGiorgio

Gene diversity, or expected heterozygosity, is a common statistic for assessing genetic variation within populations. The accurate estimation of this statistic depends on factors including sample size and independence among allele copies in the sample. Dependence among allele copies can arise when individuals in the sample are related to one another, or when individuals are inbred. The original unbiased estimator (H^), which uses sample proportions as estimates of allele frequencies, was first introduced by Nei, and underestimates the true diversity of the population in samples containing relatives. DeGiorgio et al. previously developed a generalized version (H~) of this estimator to handle related and inbred individuals, and derived its exact variance. Though unbiased with relatives, H~ has an increase in variance relative to the Nei estimate. To address this, we introduce a new unbiased estimator of gene diversity (H~_new) in samples containing related or inbred individuals, which employs the best linear unbiased estimator (BLUE) of allele frequencies rather than the sample proportion. The BLUE has the advantage over the maximum likelihood estimator of allele frequencies in terms of computational efficiency, while still providing smaller variance than other linear unbiased estimators, such as sample proportion. H~_new retains the unbiased properties of H~ in samples with relatives, but has smaller variance and therefore smaller mean squared error. Interestingly, the theoretical variance for H~_new takes the same form as the variance of H~, except that individual measurements are weighted differently. We examine the properties of H~_new relative to six alternative estimators using both theory and simulations, and apply our estimator to a global human microsatellite dataset of 5,795 individuals at 645 loci. H~_new outperforms other estimators in simulation-based and theoretical comparisons, yielding the lowest mean squared error for gene diversity measurements, and similarly surpasses the others when applied to empirical data. We additionally developed an R package to compute this estimator from genomic and pedigree data, providing researchers with the ability to obtain accurate and precise estimates of genetic diversity in any type of sample.