Sampling finite populations using supersaturated designs

Kent M. Eskridge[1], **Xiaojuan Hao, Juan Diego Hernandez Jarquin and George L. Graef**

[1] *University of Nebraska - Lincoln, USA, keskridge1@unl.edu*

Keywords: fractional factorials, genetic resources, genomic selection

Supersaturated designs are two-level factorial designs useful for screening a large number of factors (p) with a limited number of runs (n) where $n < p$. Substantial work has been done on constructing supersaturated designs for various optimality criteria when there are no restrictions on the treatment combinations to be considered. However, little work has been done on identifying optimal supersaturated designs when candidate design points must be selected from a restricted set of treatment combinations. Constructing supersaturated designs for such finite populations is an important problem in genetics. For example, a large number of p genetic markers may be available on N ($< p$) individuals where the interest is to assess how the markers are related to some end point variable such as disease level or yield. Often the individuals' end point variables are not available and collecting such data on all N individuals is too costly resulting in the need to use the genetic information to identify the n ($< N$) most informative individuals for which the endpoint data should be obtained. The objective of this work is to develop and compare several different criteria in identifying supersaturated designs comprised of the n most informative individuals from a finite population of N individuals where $N < p$ and to compare these criteria in their ability to (1) maximize genetic variance among individuals and (2) identify important markers when predicting end point variables based on the resulting models. The methods are applied to the USDA Germplasm Resource Information System (GRIN) database of ∼20,000 soybean accessions using 50K SNP genetic markers. Multi-environment field trial evidence indicates substantial differences between the methods in terms of genetic information and the predictive ability of the resulting models.