A GENETIC ALGORITHM TO INVESTIGATE GENOTYPING IN GROUPS

P.E. Macrossan and B.P. Kinghorn
Department of Animal Sciences, University of New England, Armidale, NSW 2351

SUMMARY
Innovative strategies are required to reduce the cost of DNA testing for both commercial use and research in agricultural species. Previous research has focused on maximising the utility of genotyping by prioritising animals for genotyping according to the whole-herd information gained by that genotyping. This is done under the assumption that animals are genotyped one at a time, with segregation analysis carried out after each genotyping. For logistic reasons, animals may have to be genotyped in groups rather than individually, and the best group of animals chosen will be expected to differ from the same sized group chosen when animals are genotyped singly. A genetic algorithm is used to investigate the problem of group genotypings in individual herds, with the focus on finding patterns in the evolved solutions from which to draw guidelines for group genotyping in practice.
Keywords: segregation analysis, group genotyping, genetic algorithm, differential evolution.

INTRODUCTION
Inferring the genotype of an animal from pedigree information has a potential role to play in reducing genotyping costs. Previous research by Kinghorn (1999) and Macrossan *et al.* (2002) has focused on developing a predictive index by which animals may be ranked for genotyping in order of the whole-herd information made available by that genotyping. The ranking index is based on a genotype probability index (*GPI*) for each individual derived from genotype probabilities from segregation analysis (Kerr and Kinghorn, 1996). Kinghorn (1999) showed that even a random selection of individuals to genotype will give a considerable increase in 'utility' for the first genotypings made, since the pedigree links between these animals and the remainder of the population will lead to genotype probabilities other than Hardy-Weinberg frequencies for much of the population.

This previous research has assumed that in each genotyping 'cycle', a single animal is chosen for genotyping based on an index that predicts utility to the whole population, and the results of this genotyping incorporated into a new round of segregation analysis. After this, the index is recalculated and the cycle repeated. This method assumes inexpensive DNA preparation, with robotics to genotype one chosen individual and locus at a time. However, a more realistic scenario might be to genotype a group of animals in each cycle (Kinghorn, 1997).

Thus, whilst the single animal genotyping meets the overall objective of minimising the total number of individuals genotyped and/or maximising the 'utility' of genotypings, the group genotypings will maximise the benefit/cost ratio, provided the group can be chosen 'intelligently'. Kinghorn (1999) showed that when genotyping in groups of the top 1, 10, 20 and 50 animals chosen on a ranking index based on the results of a single initial segregation analysis, the cumulated utility (average *GPI* of live individuals) decreases as group size increases. The value of genotyping an individual depends on its relationship links with the rest of the herd, and on whether or not its relatives are to be simultaneously genotyped.

An immense solution space exists for investigation into favourable attributes of animals for group genotypings, given complete segregation analysis information. For example, in a simulated herd of 930 live animals available for genotyping where a group of ten animals is to be chosen for genotyping (the figures chosen for this simulated study), the number of subsets of ten distinct animals that may be chosen is shown in Equation (1).

$$C(930,10) = \frac{930!}{(10!)(920!)} = 1.27x10^{23} \qquad (1)$$

In an exhaustive approach to the investigation, segregation analysis, a time-expensive algorithm, must be carried out on each subset of animals to determine which group is the most informative. This means that such methods for finding the solution in any particular herd are infeasible, and evolutionary methods of optimisation are ideally suited. This problem lends itself to investigation using a stochastic search algorithm such as a genetic algorithm (GA) (Holland, 1975).

A GA was used in this study, with the 'chromosomes' of the GA representing the identification numbers for a group of live animals to be genotyped in a single herd. Chromosomes were compared for their 'fitness', or the herd average GPI from segregation analysis with the chosen group of animals genotyped (i.e. their genotypes made available to the segregation analysis). Populations of such chromosomes were evolved over generations, using reproduction with modification, with segregation analysis performed for each chromosome at each generation, until convergence was reached. Segregation analysis dominates the run-time performance of both traditional search methods and the GA. By employing a GA, the total number of segregation analyses necessary is reduced from 1.27 x 10^{23} (see Equation (1)) to 3.0 x 10^{5} (3000 (number of generations) x 100 (number of chromosomes in the population)), a reduction in the order of 10^{18}. This reduces the time for the algorithm to run from 2 x 10^{16} years to 16 hours! (running on a Pentium 4 – 2.66GHz system).

The results of this simulated investigation into group genotypings is then used to extract patterns in the evolved solutions from which to draw guidelines for group genotyping in practice. Such patterns might include individual attributes of animals chosen for group genotypings, such as their age, estimated breeding value (EBV) and connectivity with the remainder of the live herd, or group dynamics such as group connectivity.

MATERIALS AND METHODS

Base animals of both sexes were simulated with a single biallelic gene, although with no effect on the animal's *EBV*, and a base population frequency of 0.1 in each of five "herds" or test populations. Each herd consisted of approximately 930 live individuals in a pedigree containing 3500 individuals. Animals were mated at the ratio of 1:40 (male:female) for ten years, with one age class of breeding males and four age classes of breeding females, selection on *EBV* across age groups and 10% adult mortality. Using a GA, a group of ten live animals, representing approximately 1.08% of the live population, was chosen for genotyping based on the optimisation of the objective function, in this case the average *GPI* of live individuals in the herd after group genotyping. Optimisation of the GA was carried out using a form of Differential Evolution (DE) adapted from Storn and Price (1997), with the population size of solutions set at 100, a weighting factor of F=0.8 and a crossover constant of CR=0.5, in line with their latest recommendations (www.icsi.berkeley.edu/~storn/code.html). The

DE was run for 3000 generations, with convergence normally reached at around 2000 generations. Ten replications of the DE were carried out from different starting seeds.

RESULTS AND DISCUSSION

Table 1 shows the average results of segregation analysis over ten runs when the group of ten live animals DNA tested is chosen using four different methods. The first method, DE, uses full segregation analysis information. The second method, Maximum *CON*, selects live animals with the highest *CON* (where *CON* is defined as the average of the numerator relationship with live animals in the herd). Maximum *CON* is a predictive method that functions without the use of segregation analysis information (the situation in practice). The final two methods are included for comparison; a group of live males selected on *EBV* for breeding purposes, and a random choice of live animals. The results in Table 1 indicate that the average herd *GPI* after segregation analysis when the group of animals genotyped are chosen using DE with full segregation analysis information is significantly higher ($p<0.001$) than that of the other three 'blind' selection methods, as expected. However, the predictive method Maximum *CON* clearly outperforms random selection of animals for genotyping, a promising result for the development of successful predictive methods in practice.

Table 1. Means ($\bar{x}$) and standard errors (s) of the herd average GPI with a group of ten animals genotyped using four different methods to select animals for genotyping

	Choice of animals to genotype based on			
Statistic	DE	Maximum *CON**	Live selected males	Random
$\bar{x}$	23.86	15.13	9.13	10.35
s	0.99	-	1.50	2.32

**CON* is defined as the average of the numerator relationship with live animals in the herd

Table 2 compares *YOB* (Year of birth; 0 for foundation animals, 10 for the youngest generation), *CON*, and *EBV* of animals chosen for genotyping using DE compared with the herd averages for live animals from the same herd. All animals chosen by DE were from the younger year groups and had significantly higher values for *CON*. Animals having more connections to the herd will yield more information from their genotyping, and younger animals are expected to be more related to the herd. Animals chosen by DE also demonstrated significantly higher *EBV*s than the herd average. Although the simulated biallelic gene had no effect on the animal's *EBV*, animals were selected for breeding purposes on the basis of *EBV* so that animals with a higher *EBV* tend to be more related to the herd.

The value of genotyping an individual either singly or as part of a group depends, amongst other things, on its relationship links with other genotyped animals and the rest of the herd. A detailed study of the animals chosen for genotyping within a single herd would be informative, particularly in terms of the herd structure, the amount of inbreeding present, and the presence or absence of marriage loops. For the sake of completeness similar experiments should be run for gene frequencies 0.2, 0.3, 0.4 and 0.5 in order to cover the range of situations that may occur in practice (the results for frequencies 0.6 to 0.9 will mirror those for 0.4 to 0.1).

Table 2. Percentage of males, and means ($\overline{x}$) and standard errors (s) of *YOB*, *CON* and *EBV* for animals selected for genotyping using DE compared with the figures for the live herd

	Statistic	Animals chosen using DE	All live animals
CON	$\overline{x}$	0.035	0.025
	s	0.007	0.009
YOB	$\overline{x}$	9.77	8.89
	s	0.68	1.30
EBV	$\overline{x}$	35.9	26.8
	s	11.8	10.5

Another animal attribute that may contribute favourably to the identification of animals to be genotyped in a group is the estimated number (probability) of favourable alleles (*EFA*) carried by any individual. In this simulated study, the group was chosen before the commencement of genotyping, so that there is no knowledge, other than H-W equilibrium frequencies, of the *EFA* of any animal. Genotyping one or a number of animals before the group is chosen will supply individual *EFA*s, but these will be affected by the pedigree relationships of the particular individuals genotyped.

A key practical contrast is among a predictive index for: (1) groups, (2) individuals, but selecting group-at-a-time, and (3) individuals, genotyping between each individual. This paper concentrated on the development of (1), using (2) with different indices for comparative purposes. A comparison of (1) with (3) will reveal the real compromise in going from individual genotypings to group genotypings. In conclusion, the results presented herein confirm the intuitive expectations that animals chosen for group genotypings should come from the younger age-groups, have a high connectivity to the remainder of the herd, whilst having little relationship to the other animals in the group.

REFERENCES

Holland, J. H. (1975). *Adaptation in Natural and Artificial systems.* Ann Arbor, MI., University of Michigan Press.

Kerr, R. J. and B. P. Kinghorn (1996). *Journal of Animal Breeding and Genetics* **113**: 457.

Kinghorn, B. P. (1997). *Genetics*, **145**: 479.

Kinghorn, B. P. (1999). *Journal of Animal Breeding and Genetics* **116**: 175.

Macrossan, P. E., Kinghorn, B.P. and Abbass, H.A. (2002). *7th World Congress of Genetics Applied to Livestock Production,* Montpellier, August 17th to 23rd, 2002.

Storn, R. and Price, K. (1997). *Journal of Global Optimization* **11**: 341.