An index of information content for genotype probabilities derived from segregation analysis

Genetics 1997. 145:479-483.

Kinghorn, B.P.
Department of Animal Science, University of New England, Armidale, NSW 2351, Australia.

ABSTRACT

A genotype probability index (GPI) is proposed to indicate the information content of genotype probabilities derived from a segregation analysis. Typically, some individuals are genotyped at a marker locus or a quantitative trait locus, and segregation analysis is used to make genotype inferences about ungenotyped relatives. Genotype probabilities for a 2-allele autosomal locus are plotted on a triangular surface. The GPI has a value of zero at the point corresponding to Hardy-Weinberg frequencies, and a value of 100 percent at the vertices of the triangle. Trigonometric functions are used to help calculate intermediate index values. It is proposed that such an index can be useful to help identify which ungenotyped individuals or loci should be genotyped in order to maximize the benefit/cost of genotyping operations.

INTRODUCTION

There is increasing activity to genotype individuals from pedigreed populations at marker loci and at quantitative trait loci (QTL). The value of this information can be extended for each locus by segregation analysis to provide probabilities of belonging to each genotype class for each ungenotyped individual (see, for example, Kerr and Kinghorn, 1996). The current paper proposes a genotype probability index (GPI), with one index value for each individual, which can be used to help identify which of these ungenotyped individuals should be genotyped in order to maximize utility of information from a further segregation analysis. As suggested by Kerr and Kinghorn (1996), such an iterative approach to genotyping can reduce costs considerably, by minimizing the number of individuals actually genotyped in the laboratory.

Utility of information may relate to factors such as confidence of QTL detection/mapping, accuracy of QTL effect estimation, or predicted merit of prospective progeny from a round of selection. The marginal utility of a genotyping is not simple to compute for such applications, as its realization generally depends on the resulting genotype. The proposed key role of a GPI is to help provide an alternative criterion for choosing individuals and loci to be genotyped.

An extreme application of this approach, of use where DNA preparation is cheap and genotyping from that stage is fast but expensive, is to set up the laboratory robotics to genotype one chosen individual and locus at a time. Segregation analysis is performed and GPI values calculated between each genotyping to help identify the next individual and locus which will likely add most to utility after genotyping and segregation analysis over the complete data set.

A more realistic scenario with current technology might be to prepare DNA samples from 50 or 100 chosen individuals in each cycle, rather than just one, then, for each individual, to genotype those loci with a sufficiently low GPI - those loci for which genotyping would add most to the utility of the total information gained.

For this type of application, desirable properties for a GPI are:

- Individuals whose genotype is known, whatever that genotype is, should have a GPI value of 100 percent.

- Individuals whose genotype probabilities are equal to the Hardy-Weinberg (H-W) frequencies should have a GPI value of zero. This would apply to ungenotyped individuals which are not linked by pedigree. Special consideration is required where estimated allele frequencies differ between different base groups in the data set.

- GPI value should be associated with utility in a relationship which is close to linear. Correlations between genotype probability and true genotype status, for each genotype, are taken here as indicators of utility.

This paper proposes a GPI which has these properties.

MATERIALS AND METHODS

Calculation of genotype probabilities: It is assumed that a 2-allele autosomal locus is segregating in the population of interest. Some individuals in the population have been DNA tested, or genotyped, and scored with values 0, 1, or 2 for genotypes *aa*, *Aa* and *AA* respectively, where *a* and *A* are the alleles segregating at the locus. Ungenotyped individuals are allocated a score code of '9'.

Segregation analysis was carried out following Kerr and Kinghorn (1996). Given direct DNA testing at the locus of interest, the penetrance values relating score (0, 1, 2, or 9) to true genotype (*aa*, *Aa* or *AA*) are quite simple to deduce, as shown in Table 1.

Table 1. Penetrance values. These are the probabilities of achieving each score (excluding 'non-score', 9) given the genotype shown. With scores coming from direct DNA testing the result here is quite simple.

	Genotype		
Score	***aa***	***Aa***	***AA***
0 (*aa*)	1	0	0
1 (*Aa*)	0	1	0
2 (*AA*)	0	0	1
9 (not scored)	1	1	1

An allele frequency estimate is required for the analysis, and this was initially taken as:

$$p(A) = \frac{n_{Aa} + 2n_{AA}}{2(n_{aa} + n_{Aa} + n_{AA})}.$$

- where n_{Aa} is the number of individuals genotyped as *Aa*, etc. This equation reflects allele frequency among genotyped individuals only. However, an estimate of allele frequency in the base population alone is required for segregation analysis, and so this was made by iterating segregation analyses until convergence of base population allele frequency estimated at the end of each iteration as:

$$\frac{\sum[p(Aa) + 2p(AA)]}{\sum[2(p(aa) + p(Aa) + p(AA))]}.$$

- with summation over base individuals only. The result of the segregation analysis is a probability for each individual of being each of the three genotypes (*p(aa)*, *p(Aa)* and *p(AA)*). For a number of individuals, these probabilities are exactly 0 or 1, either because they have been genotyped, or because genotype can be inferred exactly from relatives' genotypes. Mutation is assumed unimportant.

A genotype probability index (GPI): Figure 1 illustrates the plotting of genotype probabilities *p(aa)*, *p(Aa)*, *p(AA)* within an equilateral triangle on a 2-dimensional plane. Two dimensions are sufficient for two alleles at an autosomal locus, as the three probabilities must sum to one, and so contain just two degrees of freedom.

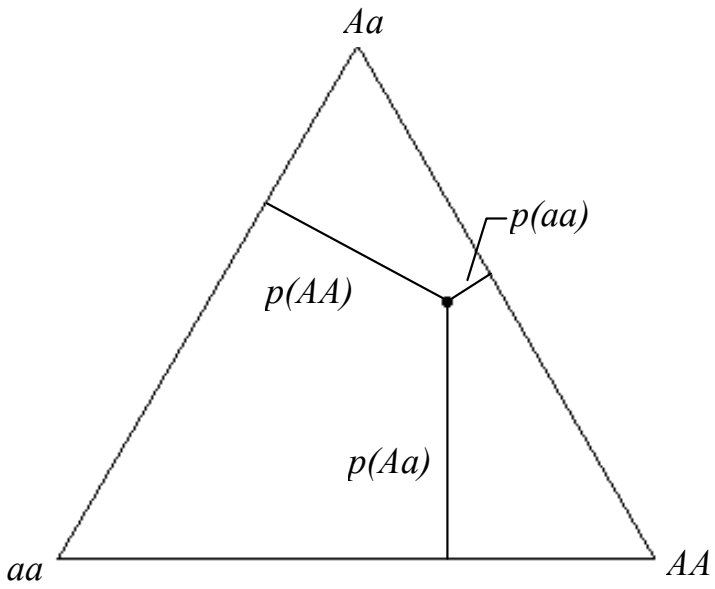


Figure 1. Genotype probabilities *p(aa)*=.1, *p(Aa)*=.5, *p(AA)*=.4 are plotted for a single individual. The vertices of the triangle represent confident genotyping for each of the three genotypes.

The index proposed can be defined with reference to the example in Figure 2. Individuals at the location corresponding to H-W probabilities (*p(aa)* = .5625, *p(Aa)* = .1875, and *p(aa)* = .0625, given *p(A)* = .25) have a GPI value of zero. Individuals at the vertices have GPI values of 100 percent. This would be true of all individuals lying on the outer circular contour, except that points outside the triangle are outside the parameter space as they involve probabilities either less than zero or greater than one, or both.

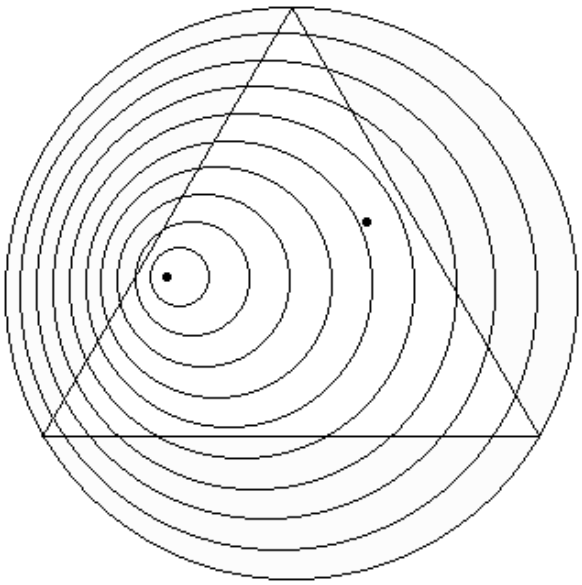


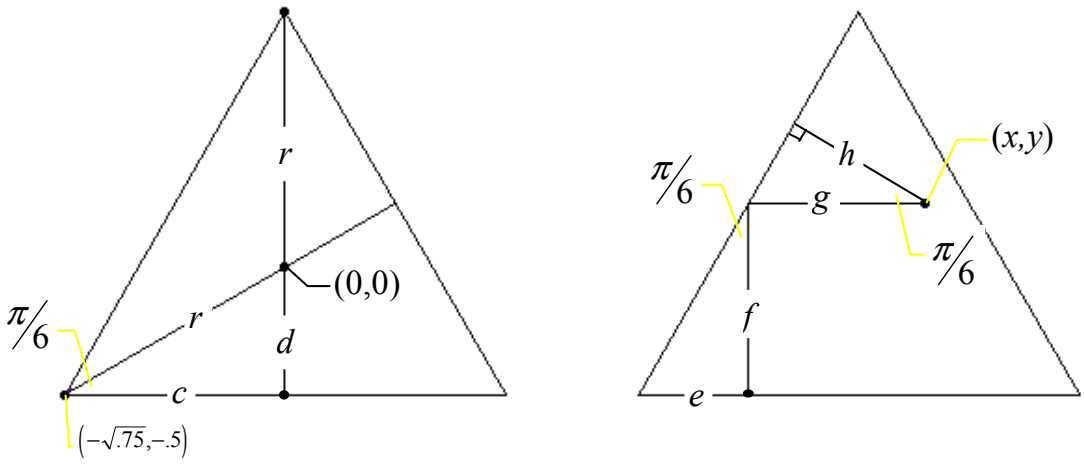
Figure 2. A conceptual description of the GPI. The dot to the left represents H-W genotype frequencies for *p(A)* = .25. Individuals at this location have a GPI value of zero. Individuals at the vertices have GPI values of 100 percent. The individual as plotted in figure 1 has a GPI value of just over 50 percent, as can be seen from counting the contours.

The index value for any individual, in percentage units, is defined as the percentage distance that the individual lies from the point of H-W probabilities to its projection on the outer contour. It can be seen that all linear projections originating at H-W probabilities involve linear gradients in index value. The individual at (.1, .5, .4) has a GPI of 51.8 percent, and it lies just beyond the fifth circular contour which represents GPI values of 50 percent.

Figures 3 and 4 aid description of how to calculate the GPI. Figure 3 shows that genotype probabilities *p(aa)*, *p(Aa)*, and *p(AA)* can be plotted on the triangular object using the following transformations:

$$x = \sqrt{.75}\left[-1+\left(p(Aa)+2\,p(AA)\right)\right]$$

$$y = -.5 + 1.5\,p(Aa)$$



3a. 3b.

Figure 3. Plotting genotype probabilities (point (x,y) in figure 3b.) on a 2-dimensional plane. The center of the triangle is defined as the origin. Following figure 2, r=1 is taken as the radius of the outer circle, corresponding to an GPI value of 100%. $\pi/6$ is an angle in radians. $\cos(\pi/6)=\frac{c}{r}=\sqrt{.75}$ and $\tan(\pi/6)=\frac{d}{c}=\sqrt{1/3}$ giving $c=\sqrt{.75}$ and $d=0.5$. The lower left vertex is thus at $\left(-\sqrt{.75},-.5\right)$, such that x deviates from $-\sqrt{.75}$ by e+g, and y deviates from -.5 by f. As $\tan(\pi/6)=\frac{e}{f}=\sqrt{1/3}$, $e=\sqrt{1/3}\,f$, and $f=(r+d).p(Aa)$ given that $p(Aa)$ is the proportion that f is of the height of the triangle. As $\cos(\pi/6)=\frac{h}{g}=\sqrt{.75}$, then $g=\frac{h}{\sqrt{.75}}$. Also $h=(r+d).p(AA)$. Thus $x=-\sqrt{.75}+e+g$ and $y=-.5+f$ gives:

$$x=-\sqrt{.75}+1.5\left[\sqrt{1/3}p(Aa)+\frac{p(AA)}{\sqrt{.75}}\right] \quad \text{and} \quad y=-.5+1.5p(Aa)$$

which simplify to:

$$x=\sqrt{.75}\left[-1+\left(p(Aa)+2\,p(AA)\right)\right] \quad \text{and} \quad y=-.5+1.5p(Aa)$$

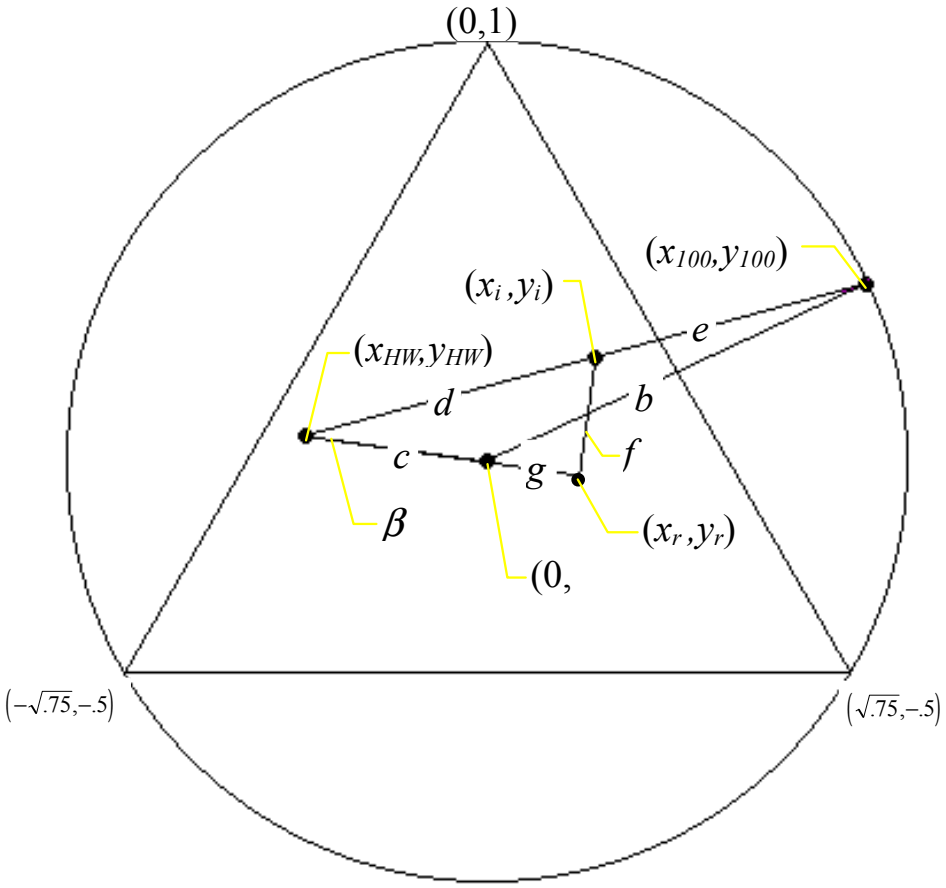


Figure 4. Calculation of the GPI for individual *i*. Subscript *HW* relates to Hardy-Weinberg genotype frequencies. Point (x_{100}, y_{100}) is the projection for individual *i*, in this case, to a point constituting full (100%) information. Point (x_r, y_r) is described in the text. Angle β is made by lines *c* and *d*. See text for further details.

With reference to figure 4, $|(x_{HW},y_{HW})\text{-}(x_{100},y_{100})|$ relates to 100 percentage units, by definition, given that points on the circle crossing the three vertices constitute full information. Thus the GPI for individual *i* is:

$$\text{GPI}_i = 100 * \frac{|(x_{HW},y_{HW})\text{-}(x_i,y_i)|}{|(x_{HW},y_{HW})\text{-}(x_{100},y_{100})|} = 100 * \frac{d}{d+e}.$$

From Pythagoras' theorem: $d = \sqrt{(x_i - x_{HW})^2 + (y_i - y_{HW})^2}$,

and $c = \sqrt{(x_{HW} + y_{HW})^2}$.

Also, $b = 1$, the radius of the outer circle.

Following the law of cosines: $b^2 = c^2 + (d+e)^2 - 2c(d+e)\cos(\beta)$

which leads to $d + e = c.\cos(\beta) \pm \sqrt{c^2.\cos(\beta)^2 - (c^2 - b^2)}$

Experience shows that $d + e = c.\cos(\beta) + \sqrt{c^2.\cos(\beta)^2 - (c^2 - b^2)}$ should always be used.

Angle β is the only undescribed part of the above. It can be evaluated as follows:

Point (x_r,y_r), shown in figure 4, is chosen to lie on the projection of (x_{HW},y_{HW}) - (0,0) and to make a right angle (x_{HW},y_{HW}) - (x_r,y_r) - (x_i,y_i):

$$x_r = \frac{x_i + y_i \dfrac{y_{HW}}{x_{HW}}}{1 + \left[\dfrac{y_{HW}}{x_{HW}}\right]^2} \qquad \text{and} \qquad y_r = x_r \left[\frac{y_{HW}}{x_{HW}}\right]$$

As $\tan(\beta) = \dfrac{f}{c+g}$, then

$$\beta = \arctan\left(\frac{f}{c+g}\right) = \arctan\left(\frac{\sqrt{(x_i - x_r)^2 + (y_i - y_r)^2}}{\sqrt{(x_{HW} - x_r)^2 + (y_{HW} - y_r)^2}}\right)$$

To generalize for all locations of (x_{HW},y_{HW}) and (x_i,y_i), it is necessary that if $x_i/x_{HW} > 1$, then the value of β as calculated above should be replaced by $\pi - \beta$.

A test of the GPI: Simulated populations were generated, each consisting of five discrete generations of 20 sires mated to 10 dams each to give four progeny per dam (5,20,10,4, giving a total of 4000 individuals). A single locus was simulated with genotypes in the base population generated from $p(A) = 0.25$. Twenty percent of individuals were chosen at random to be genotyped (ie. their genotype was revealed to the analysis). Segregation analysis was carried out following Kerr and Kinghorn (1996), as described above.

Figure 5 illustrates the results from one replicate of a smaller population (5,5,5,4, giving a total of 500 individuals). Realized base population allele frequency was $p(A)$=0.246. Individuals which have been confidently genotyped, whether by DNA test or by inference from relatives, are plotted at the appropriate vertices of the triangles. It can be seen that for other individuals, the segregation analysis tends to give probabilities towards true genotypic status.

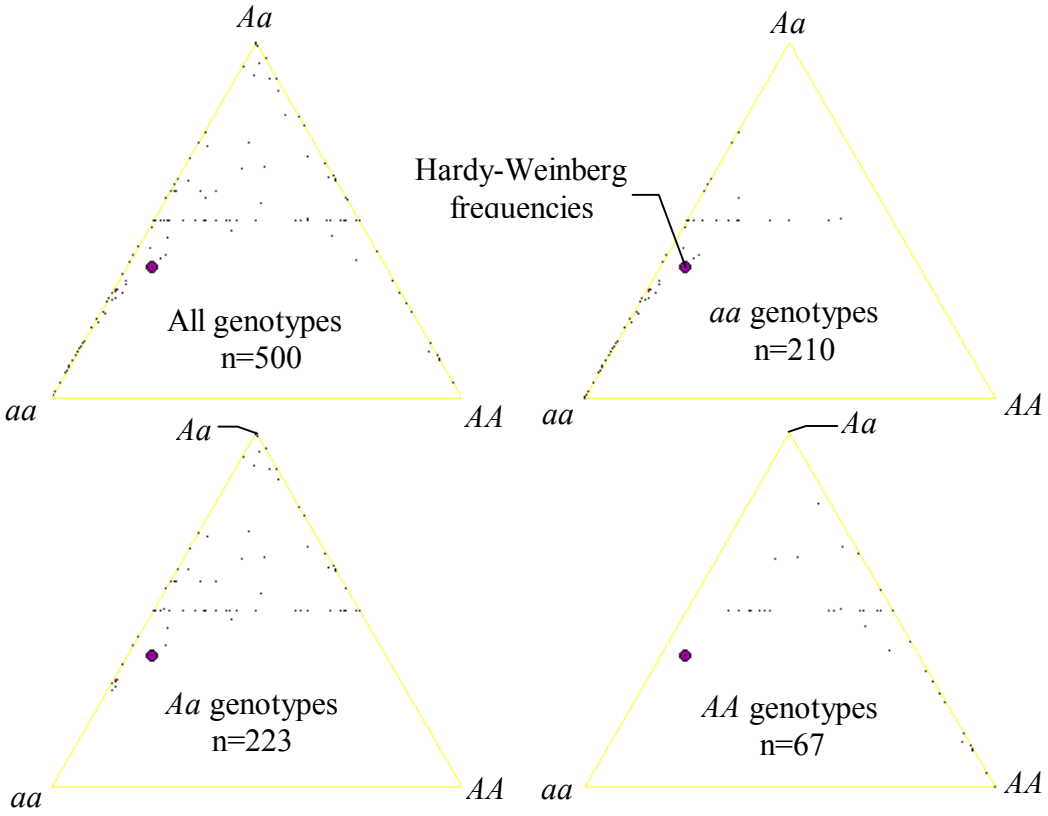


Figure 5. Example result from a segregation analysis run. Each point within each triangle represents the genotype probabilities for an individual. The top left triangle contains all individuals, and the other triangles contain individuals which are *truly* of genotypes indicated.

Two other phenomena can be seen in figure 5:

- There are no points plotted at the lower middle of the triangles. This blank area is below the line scribed by all possible points of H-W equilibrium, and it exists because of the properties of transmission probabilities. However, it can be filled under some circumstances other than from use of segregation analysis, for example by the allocation $p(aa)$=0.8, $p(Aa)$=0, and $p(AA)$=0.2 from a gel reading which clearly indicates homozygosity (one band) but with no evidence as to which allele is involved, no information from relatives, and a prior frequency for allele *A* of $p(A)$=0.2.

- There is a horizontal line of points corresponding to $p(Aa)$=0.5 exactly. This probability occurs for all progeny of known heterozygotes which are not parents and are not genotyped.

To test the properties of the GPI, 1000 replicate populations of 4000 individuals each were generated, as described above. GPI values were calculated for all individuals, and correlations of genotype probability and true genotype status (values of 0 for false or 1 for true) were calculated separately for each genotype, for each replicate, for each ten-percent range of values of the GPI. The averages of these correlations are plotted against GPI in figure 6.

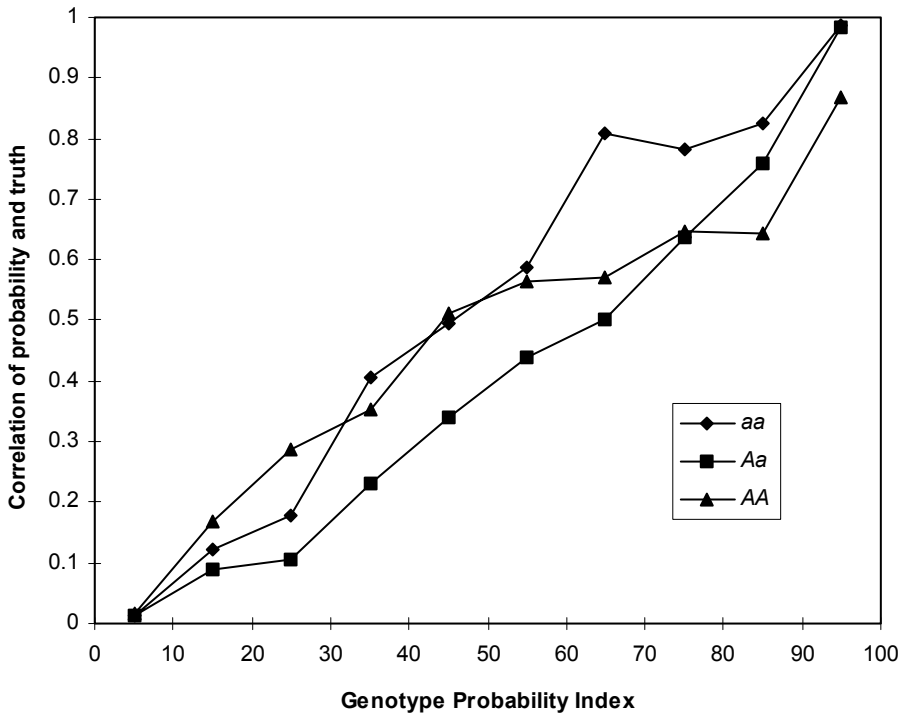


Figure 6. Correlations between genotype probability and true genotype status (0 or 1), for three genotypes, *AA*, *Aa* and *aa*. Correlations were calculated within each ten-percent range of values of the GPI, and plotted against the mid-range value. Each point is the average of results from 1000 simulated populations.

This correlation between true and predicted genotype status is taken as a simple measure of utility of genotype probabilities. It can be seen from figure 6 that for each genotype, there is an approximately linear relationship between GPI and correlation. The departures from linearity are moderately repeatable, as the standard errors of the points in figure 6 average just 0.003, with a maximum value of 0.009. These departures from linearity may be due to properties of the segregation analysis and/or the GPI, in conjunction with population structure.

DISCUSSION

It can be seen from figure 2 that the GPI proposed assumes a linear gradient of utility in any direction from the point of H-W equilibrium. This may not be fully appropriate. For example, the difference in utility between $p(AA)$=0.99 and $p(AA)$=1 could be very high where there is a commercial need to be able to guarantee the genotype of certain individuals. Moreover, it is possible to compose a GPI based on ellipses rather than circles. As an example, an ellipse with an aspect ratio greater than one (i.e. horizontally- and vertically-symmetrical, with its vertical radius greater than its horizontal radius) would put more emphasis on the value of being able to eliminate one of the homozygotes as a possible genotype.

Such deviations from linearity and unit aspect ratio could well be appropriate in certain circumstances - but these circumstances seem likely to relate to somewhat subjective criteria, and possibly population structure, making integration into a theoretical framework difficult and/or arbitrary. It is proposed that the assumptions of linearity and unit aspect ratio made here will be generally acceptable, but that care should be taken to ensure that the properties of the GPI are appropriate to the prevailing task.

The GPI developed in this paper handles just two alleles at one locus. A GPI for multiple alleles can be defined in a similar manner - the percentage distance that the individual lies from the point of H-W probabilities to its projection on the outer multidimensional contour which represents full information. Whereas two alleles can be plotted on a 2-dimesional surface, as in this paper, n alleles require $n(n+1)/2-1$ dimensions, making visualization somewhat difficult. However, an algebraic solution to an equivalent multiple allele GPI seems quite possible. This would prove useful for work with multi-allelic genetic markers.

The current development is based on single-locus segregation analysis. Extension to multi-locus segregation analysis should prove to be of value. In this case, inference about an individual's genotype at a given locus can be augmented by information about its known genotypes, and/or those of its relatives, at linked loci. This means that GPI's will be generally higher.

The key use of GPI's proposed here is to decide which individuals to genotype by DNA test in successive cycles of segregation analysis, ranking of individuals, and DNA testing of chosen individuals (Kerr and Kinghorn, 1996). Use of information is improved by having a smaller number of individuals DNA tested in each cycle, but this must balance with the efficiency of the testing operation. The objective in such an exercise is to maximize the benefit/cost of the genotyping operation, for example by minimizing the total number of individuals genotyped, over all cycles, for a given utility of the overall results. For multiple locus applications, the procedure can be extended to rank all possible genotypings - i.e. combinations of loci and individuals.

A number of factors can influence the ranking process. Considering the single locus case for simplicity, an individual's ranking criterion for DNA testing might be influenced by:

1. The amount of information contained in the individual's genotype probabilities, as indicated by the proposed GPI.

2. The extent to which the individual is, or is likely to become, connected to the rest of the population, and thus provide useful information about its relatives. This might be influenced by estimated breeding value (*EBV*, affecting the likelihood of being selected for breeding) and by sex (males featuring more strongly due to higher fecundity).

3. The relative importance of gaining information on past, present and future members of the population. For example, QTL detection and effect estimation is supported by genotype information on past individuals, but marker assisted selection is generally not.

4. The number and relationship of relatives which come to be DNA tested in the next cycle. The value of genotyping an individual generally decreases as more of its relatives are genotyped.

Development of ranking criteria which can accommodate such issues is not attempted in this paper. However, it is proposed that the GPI described, and its extensions, can play a key role in such development. For illustration, one candidate criterion is given here, derived intuitively as appropriate for ranking juveniles for genotyping a QTL of commercial value:

$$Ranking\ criterion_i = \frac{prog_i . EBV_i^x}{y + GPI_i}$$

where $prog_i$ is the predicted number of progeny of individual i, if selected, probably based on the sex of i alone, x is chosen to modify the influence of EBV, whose coefficient of variance is somewhat low compared to other components, and y is chosen to give appropriate influence to the GPI. This criterion would favor individuals which currently have poor QTL information, and which are likely to provide more information to the population through future genetic links.

ACKNOWLEDGMENTS

This method in this paper was conceived by the author while he was a visiting scientist at Newsham Hybrids (USA) Inc. The author would like to thank Tianlin Wang and Richard Kerr for valuable discussions. The author thanks the Twynam Pastoral Company which funds his position.

LITERATURE CITED

Kerr, R.J., and B.P. Kinghorn, 1996: An efficient algorithm for segregation analysis in large populations. J. Anim. Breed. Genet. In Press.