A SIMPLE GENERALISATION OF KINGHORN'S GENOTYPIC PROBABILITY INDEX

Bruce Tier
Animal Genetics and Breeding Unit[1], University of New England, Armidale NSW 2351

SUMMARY
Using trigonometry, Kinghorn (1997) developed the genotype probability index as a measure of the amount of information contained in genotype probabilities predicted by segregation analysis for loci with only three genotypes. Percy (2005) represented the problem algebraically and generalised it for loci with any number of genotypes. This paper describes a general and simpler method for calculating the genotype probability index for loci with multiple ordered or unorderd genotypes which requires nothing more than solving a single quadratic equation.

INTRODUCTION
Kinghorn (1997) introduced the idea of a Genotype Probability Index (GPI) to describe the amount of information contained in the genotype probabilities inferred through segregation analysis, and to measure the change in amount of information resulting from genotyping additional individuals in the population. GPIs range from no knowledge (GPI=0) to complete knowledge (GPI=1), with increasing amounts of information resulting in higher GPIs. It was initially described for a two-allele locus with three unordered genotypes. Geometry and trigonometry were used to represent the problem and calculate the GPI respectively. Recently, Percy (2005) generalised the GPI to multiple alleles and multiple ordered genotypes using algebra. This paper describes a simplified generalisation for calculating the GPI for loci with any number of alleles and genotypes.

MATERIALS AND METHODS
With two alleles there are three possible unordered genotypes – AA, Aa and aa – with probabilities p (AA), p(Aa) and p(aa), respectively. Because the probabilities must sum to unity (ie. p(AA)+P(Aa)+p (aa) = 1) it is possible to represent them with two parameters. More generally with a alleles there can be g genotypes. g might be a^2 or $(a^2+a)/2$ if the genotypes are ordered or unordered respectively. With g genotypes there are g-1 independent probabilities.

2 Alleles. Kinghorn (1997) chose a point within an equilateral triangle with unit height to jointly describe the probabilities (see figure 1a). Vertices represent certainty for genotypes; the perpendicular distance from a side represents the genotype probabilities for the opposite vertex.

The quantity of information contained in an individual's set of probabilities – the GPI – was defined as a ratio of distances. The numerator was the distance from a reference point representing the state of zero knowledge to the point representing the individual's probability and the denominator was the distance between the zero reference point and a point representing complete certainty, along the line containing the individual's probability. Kinghorn (1997) chose Hardy-Weinberg equilibrium in the

[1] AGBU is a joint venture of the NSW Department of Primary Industries and the University of New England

base population to represent the state of zero knowledge. This is represented by the point r in figure 1b.

A set of points – the circumference of the circle containing the three vertices of the triangle (see figure 1b) – was chosen to represent states of complete knowledge. The GPI for an individual was the proportion of the distance from the point of zero knowledge to the point representing complete knowledge along a ray drawn from the zero reference point (r, in figure 1b), through the probability point (p) to the point where the ray intersects the circle (q). Point p in figure 1b represents an individual with probabilities p(AA)=0.40, p(Aa)=0.50, P(aa)=0.10. The zero reference point (r) is found at p(AA)=0.09, p(Aa)=0.42, and p(aa)=0.49 – the point of Hardy-Weinberg equilibrium probabilities in a base population with allele frequencies for alleles A and a of 0.3 and 0.7 respectively. The GPI for this individual is the ratio of the distance rp to rq, or 0.477.

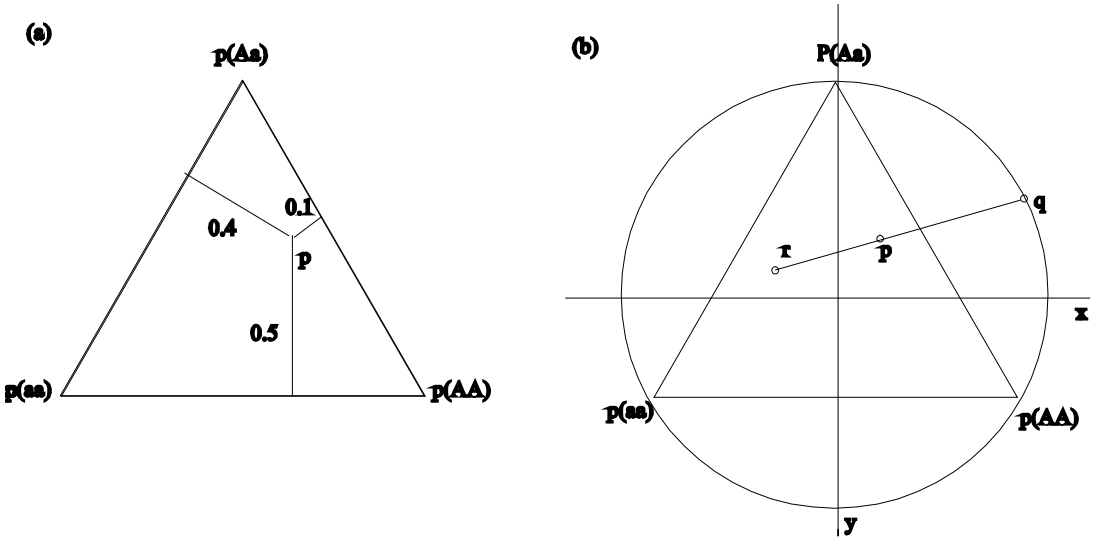

Figure 1. (a) Geometrical representation of the GPI within the equilateral triangle of unit height of the GPI for an individual (p) with probabilities of being AA, Aa or aa of 0.4, 0.5 and 0.1 respectively. Distances from the opposite sides are the probability of the genotypes represented by the vertices. (b) Calculating the GPI in two dimensions (x and y). The circle's circumference represents certainty and the points p and r represent the genotype probabilities of an individual and the state of zero knowledge described in the text respectively. The GPI is the ratio of the distance rp to rq.

Algebra. While trigonometry can be used to calculate the GPI, Percy (2005) illustrated how this problem could be transferred to the Cartesian plane and solved using simultaneous equations. Figure 1b illustrates this representation. Kinghorn's equilateral triangle is constructed inside the unit circle ($x^2+y^2=1$), and the points represented by the probabilities are used to determine their location within

the triangle. The vertices of the equilateral triangle are at (0,1), (-√3/2),-1/2) and (√3/2),-1/2). The location of the point representing the probabilities on the plane can be calculated using the equations x=√3*(p(AA)-p(aa))/2 and y=1.5*p(Aa)-0.5. The perpendicular distances from the sides to the point are still the probabilities.

The equation for the line between the zero reference point and the individual's probability is computed, as is the point at which it intersects the circumference of the circle. The relative distance between these pairs of points is readily calculated. For the individual shown in Figure 1 the points representing its probability, the zero and certain reference points are at (0.26, 0.25), (-0.35,0.13) and (0.92, 0.38) and the equation of the line connecting them is y=0.198x+0.199.

Many other transformations are possible, for example one could choose the height of the triangle to be unity and place its base on the line y=0, then the vertices would be at (-1/√3,0), (0,1) and (1/√3,0) and the equation of the circle would be $x^2+(y-1/3)^2=4/9$.

Multiple genotypes. With the problem expressed algebraically Percy (2005) generalised the calculation of the GPI, firstly for three alleles and six unordered genotypes and then to any number of unordered genotypes. He continued to consider the problem in the reduced space of the g-1 dimensions, where g is the number of genotypes. The probability of one genotype was discarded as it could be computed from the others. Percy (2005) introduced to the discussion of the GPI the use of vectors to describe probability locations in space. His method was based on the use of hyper-dimensional 'equilateral triangles' and is too complex to be repeated here.

Complete dimensions. Let all genotype probabilities for an individual be contained in the column vector **p**. **p** describes a point in g-dimensional space. We know that the values in **p** sum to unity (**p'1**=1). Similarly, the genotype probabilities for the reference point of zero knowledge are described by the vector **r**. Now the direction away from the reference point is described by the vector **p-r**, and the equation

$$\mathbf{w}=\mathbf{r}+t(\mathbf{p}-\mathbf{r}),$$

describes the line connecting the points described by **p** and **r** (after Percy 2005). Note that **w'1** = 1 and (**p-r**)'**1**=0. The distance along this vector starting at **r** and passing through **p** is given by t, in units of distance **p-r**.

In g-dimensional space, there is set of vectors **x** that fit the equation **x'x**=1 and describe the surface of the unit radius hypersphere centred on the origin. As emphasised above, in two dimensions Kinghorn (1997) ascribes the circumference of the unit circle as the state of complete knowledge. Its equivalent in multidimensional space is the surface of this hypersphere.

Let **q** be the point where the line described by the vector **w** intersects the perimeter of the unit hyper-sphere. At that point **q'q=1**, so with **q**=**r**+t(**p-r**),

$$(\mathbf{r}+t(\mathbf{p}-\mathbf{r}))'(\mathbf{r}+t(\mathbf{p}-\mathbf{r})) = \mathbf{r'r} + 2t\mathbf{r'}(\mathbf{p}-\mathbf{r}) + t^2(\mathbf{p}-\mathbf{r})'(\mathbf{p}-\mathbf{r}) = 1.$$

This is a simple quadratic equation that can be solved for t ($t > 0$). The ratio of the distance traveled from zero knowledge to the current genotype probability (**p**-**r**) to the state of complete knowledge (**q**-**r**) is Kinghorn's GPI, and is equal to 1/t.

Equivalence to two dimensions. The equivalence to the method described by Kinghorn (1997) for the case of two alleles and three genotypes can be seen by considering the plane containing the vertices of the equilateral triangle at the points (0,0,1), (0,1,0) and (1,0,0) in three-dimensional space. This plane is described by the equation **p'1**=1 and contains both the triangle – and hence all possible sets of probabilities - and the circle where the sphere **p'p**=1 intersects that plane.

CONCLUSION

A simple method for computing Kinghorn's GPI for loci with any number of ordered or unordered genotypes has been presented. Complete sets of genotype probabilities are stored in vectors. GPIs are calculated by solving a single quadratic equation.

REFERENCES

Kinghorn B.P. (1997). *Genet*. **145**:479.
Percy A. (2005) *J. Anim. Breed. Genet*. (Submitted).