A SIMPLE METHOD TO DETECT A SINGLE GENE THAT DETERMINES A CATEGORICAL TRAIT WITH INCOMPLETE PENETRANCE

B.P. Kinghorn
Sygen Chair of Genetic Information Systems, University of New England, Armidale, NSW 2351

SUMMARY

This paper presents a simple method to infer mode of inheritance, penetrance values, allele frequency and genotype probabilities for a putative biallelic gene that affects a trait with all-or-none expression. Penetrance is assumed to be unaffected by other genes. The preliminary analysis tests the method in a variety of scenarios. Results can be unreliable for data generated with overdominance plus partial penetrance, but otherwise results were good, with mode of inheritance, allele frequency and penetrance estimated close to true values. The method will work for multiple mutually-exclusive phenotypic classes.

Keywords: Gene detection, segregation analysis, penetrance

INTRODUCTION

Experience shows that it is reasonable to suspect that one gene might play a major role in expression of traits that involve a distinct morphological effect or defect. Inspection of pedigreed data can sometimes be enough to identify likely modes of inheritance. However, the task can be complicated by imperfect penetrance – where a phenotypic effect does not always occur in carriers of 'offending' genotypes, and may occasionally occur in 'normal' genotypes. This paper presents a simple method to detect mode of inheritance for an all-or-none trait, and penetrance for each of the putative genotypes. The method assumes that penetrance is random with respect the rest of the genome, which may not be true.

MATERIALS AND METHODS

The method is based on segregation analysis (eg. Kerr and Kinghorn, 1996). With known mode of inheritance and penetrance values at a single biallelic locus, segregation analysis provides the probability for each individual of being each of the three possible genotypes. For simplicity, it is here assumed that there are just two categories of phenotype, non-affected and affected (y = 0 or 1). The analysis requires penetrance values as inputs – these are the probabilities of phenotype given genotype, $g(y \mid u)$, as in Table 1.

Table 1. Penetrance values for a simple recessive gene.

	Genotype, *u*		
Phenotype, *y*	*aa*	*Aa*	*AA*
0	1	1	0
1	0	0	1

Given penetrance values, segregation analysis uses information from self, plus information from known relatives, to calculate genotype probabilities. Now consider that in truth the penetrance for genotype *AA* is not full, with probabilities of say 10% and 90% for expressing phenotypes 0 and 1, respectively. True *AA* genotypes with phenotype 0 will tend to have positive genotype probabilities of being genotype *AA* because of information from relatives – an extreme example being where both parents are of phenotype 1.

Following segregation analysis, new estimates of penetrance values can be estimated for each of the three genotypes by summing over individuals the probability of belonging to that genotype within each phenotypic class, then dividing each result by the sum of these figures over phenotypes within genotype. This can be applied in cases where there are multiple mutually-exclusive phenotypic classes. However, it is simpler to present for two categories of phenotype (y = 0 or 1), as y can act as an indicator variable for phenotype 1. Penetrance values can thus be estimated from genotype probabilities and phenotypes as:

$$\hat{g}(y=1|u) \;=\; \frac{\sum_i y_i\, prob(u)_i}{\sum_i prob(u)_i} \qquad\qquad \hat{g}(y=0|u) \;=\; 1-\hat{g}(y=1|u)$$

However, these estimates will be heavily conditioned by the penetrance values used as input for the segregation analysis, and so the process should be iterated until convergence is approached. Allele frequency can also be re-estimated at each iteration from the genotype probabilities for foundation individuals:

$$\hat{p}_A = \frac{\sum\left(p_{Aa}+2p_{AA}\right)}{2\sum\left(p_{aa}+p_{Aa}+p_{AA}\right)}$$

Data simulation. Data were generated using module PopQTL of Genup (http://www-personal.une.edu.au/~bkinghor/genup.htm). Four sires were mated to 25 dams each year, with overlapping generations (2 to 4 years for sires and 2 to 7 years for dams). Adult survival was 90% per year. Parents were selected on a 4-trait selection index. Dams left singles, twins or no offspring with probabilities 0.25, 0.7 and 0.05 respectively. The population was bred for 5 years, giving a total of 13 years represented when the foundation population was included. A total of 1260 individuals were generated. A single biallelic gene *(a, A)* was simulated with an allele frequency of $p(A)$=½ or $p(A)$=¼. This gene had no effect on reproduction, animal selection or mate allocation.

A single population was used for all runs at each allele frequency, but with different patterns of phenotype generation. Genotypes were generated as u = 0, 1 or 2 copies of allele *A* carried. Phenotypes were generated as y = 0 or 1 (notionally non-affected and affected for some condition), depending on mode of inheritance and random chance. For example, for recessivity and 75% penetrance, individuals with u = 2 had a 75% probability of being y = 1, else y = 0.

RESULTS

Table 2 shows that the method gives generally good results for recessive and dominant gene action. At lower penetrance values there are fewer affected individuals, giving greater error. In some cases the results were the ‘mirror image’ of those actually shown, as the method has no information to differentiate between the two alleles. Ordering to show allele *A* more associated with the effect was carried out for simpler presentation.

Table 2. Results from using the method on simulated data at *p*=0.5

Conditions generated		Number affected	Estimated allele frequency	Estimated penetrance (%)		
Underlying mode	Penetrance (%)			*aa*	*Aa*	*AA*
Recessive	100	249	0.498	0	0	100.0
	75	190	0.506	0	1.1	75.3
	50	122	0.490	0	0	49.9
	25	63	0.487	0	0	23.6
	10	25	0.605	0	0	5.3
Dominant	100	863	0.486	0	100.0	100.0
	75	640	0.493	1.2	77.8	65.2
	50	412	0.541	0.1	52.1	43.9
	25	195	0.648	1.2	29.3	18.6
	10	80	0.500	0	12.7	0

Table 3 shows results from further analyses at *p*(*A*) = 0.25 and 0.5. For these runs, a variety of penetrance value sets were used to generate phenotypes involving some unusual patterns of inheritance. In some cases, two sets of results are shown, derived from two different sets of starting values for allele frequency and penetrance. Starting value sets were generally (*p*/2, 30%, 40%, 50%). For data set (0.25, 100%, 0%, 100%) the two sets of results are mirror images, following reasons given above. However, for other cases, ambiguous results show that the method, at least as presented here, is not fully robust. In particular, either data generated or starting values that reflect overdominance plus partial penetrance can lead to incorrect results. Three such cases are shown, for data generated at (0.5, 50%, 100%, 50%), (0.5, 25%, 75%, 25%) and (0.25, 50%, 0%, 50%).

DISCUSSION

The method worked generally well in this preliminary study. The inheritance conditions under which incorrect results could be found were extreme. Moreover, in each of these cases, sensible results were found using an alternative set of starting values.

The method was fast to implement, typically ten seconds per run on the data described (1.4GHz processor). It is also applicable to large populations. The segregation analysis component was applied to an industry data set of over 2.3 million cattle, running in about ten minutes.

The method should be tested for its utility in scenarios under which modifier genes, including polygenes, can have an effect on penetrance. If results from such study are appropriate, then the method proposed could be used to help indicate possible modes of inheritance, together with genotype probabilities, that can be useful in setting up more efficient gene mapping experiments. An alternative approach is to use a parameter sampling method (for example, Tier and Henshall 2001) that fits polygenic effects on liability to express given phenotypes.

Table 3. Further results from using the method on simulated data at *p*=0.5 and *p*= 0.25 (Data generating parameters are in normal font, and parameter estimates are in italics)

	Penetrance			
Allele Frequency	*aa*	*Aa*	*AA*	Number affected
0.5	0	50	100	
0.485	*1.4*	*47.1*	*100*	*519*
0.5	25	50	75	
.461	*24.9*	*46.1*	*87.6*	*562*
0.5	0	100	0	
*0.489, 0.511**	*0*	*100*	*0*	*614*
0.5	100	0	100	
*0.489, 0.511**	*100*	*0*	*100*	*646*
0.5	50	0	50	
*0.496, 0.504**	*49.2*	*0*	*49.2*	*319*
0.5	50	100	50	
0.546	*49.5*	*99.9*	*50.7*	*933*
0.468	*53.4*	*87.5*	*79.7*	*933*
0.5	25	75	25	
0.480	*31.8*	*68.2*	*19.8*	*590*
0.517	*63.5*	*21.9*	*80.7*	*590*
0.25	0	50	100	
0.247	*0*	*42.6*	*100*	*266*
0.25	25	50	75	
0.14	*24.9*	*59.9*	*78.4*	*424*
0.25	0	100	0	
0.24	*0*	*100*	*4.2*	*438*
0.25	100	0	100	
0.24	*100*	*0*	*95.8*	*822*
0.76	*95.8*	*0*	*100*	*822*
0.25	50	0	50	
0.27	*50.2*	*0*	*52.6*	*394*
0.29	*22.1*	*33.5*	*86.5*	*394*
0.25	50	100	50	
0.31	*41.2*	*94.3*	*79.3*	*832*
0.25	25	75	25	
0.19	*23.7*	*80.1*	*18.7*	*509*

* Oscillating

ACKNOWLEDGEMENT

The author thanks Richard Kerr and Julius van der Werf for discussions.

REFERENCES

Kerr, R.J. and Kinghorn, B.P. (1996) *J. Anim. Breed. Genet.* 113:457.

Tier, B. and Henshall, J.M. (2001) *Gen. Sel. Evol.* **33**:587.