COMPARISON OF SEGREGATION ANALYSIS FOR NORMAL AND BINARY DISEASES TRAITS

H. Ilahi and H.N. Kadarmideen
Statistical Animal Genetics Group, Institute of Animal Sciences, Swiss Federal Institute of Technology, ETH Zurich CH-8092, Switzerland

SUMMARY

Segregation analysis was performed on normal and binary data in order to compare accuracy and bias of parameters estimated on different scales. Data simulated and analysed were: data on the underlying normally distributed liability (NDL) and binary data created by truncating NDL data based on two thresholds corresponding to two incidences. The parameters estimated under mixed inheritance (H_1) for normal trait were similar to the true values of parameters used in the simulation. The major gene variance was however underestimated. On the other hand, the estimated parameters under polygenic inheritance (H_0) were overestimated, especially for the genetic and the permanent environmental variances. Under H_0, for binary trait and for both incidences, the estimates of heritabilities and repeatabilities were the same and high. However, under H_1, these estimates were very low and slightly higher for 40% incidence. Using a low incidence (15%), the results show an overestimation of the unfavorable genotype A_1A_1 frequency and underestimation of A_1A_2 and A_2A_2 genotypes frequencies. In the case of high incidence (40%) however, there is an overestimation of the favorable genotype A_2A_2 frequency and underestimation of unfavorable genotype A_1A_1 frequency. For the normal trait, the estimated heritabilities and repeatabilities were lowered from polygenic H_0 to mixed inheritance H_1. However, for binary trait these estimates for both incidences were dramatically lowered from H_0 to H_1. Following these preliminary results, it could be concluded that power for detecting major gene is higher for NDL than 0/1 data and estimates are more biased for 0/1 than NDL data.

Keywords : Normal trait, binary trait, incidence, major gene, mixed model, segregation analysis

INTRODUCTION

In domestic animal populations, genetic analyses of quantitative traits have been thoroughly addressed for traits whose phenotypes are controlled by many genes (polygene) each having a small effect, and follow a continuous distribution (normal). However, in many cases phenotypes and especially disease traits are expressed in two or more categories, representing binary and categorical traits, respectively. Several of these traits are controlled by few genes with a large effect (major genes or quantitative trait loci, QTLs) and polygene. The objective of this study was to compare the segregation analysis applied to trait on the normal versus binary scale with two incidences (15 and 40%) using simulated data. In a first step we test if the transformation of normally distributed liability data to binary data has an influence (bias) on the estimation of genetic parameters of the population. In a second step we compare the segregation analysis for binary traits using two different incidences.

MATERIALS AND METHODS

Simulation of the normally distributed liability (NDL) data. The data were simulated using a mixed inheritance model (polygene + major gene) and according to a hierarchical and balanced family structure: one population consists of 20 sire families with 100 dams per sire, which resulted in 2000 dams. A total of 3 records/phenotypes per dam was simulated (i.e. 300 records per sire). Therefore the total number of records in the population is 6000. We assumed more than one phenotype per dam, because in the most cases, livestock data sets are consisted of repeated records. The phenotypic data were simulated as follows: $y_{ij} = m_i + a_{ij} + pe_{ij} + e_{ij}$ where y_{ij} is the phenotype, m_i is the effect of the i^{th} genotype at major gene, a_{ij} is the polygenic effect of the j^{th} individual bearing the i^{th} genotype, $a_{ij} \sim N(0, s_g^2)$, pe_{ij} is the permanent environmental effect, $pe_{ij} \sim N(0, s_{pe}^2)$ and e_{ij} is the residual effect, $e_{ij} \sim N(0, s_e^2)$, where s_g^2, s_{pe}^2 and s_e^2 are polygenic, permanent environmental and residual variances, respectively. The single major gene is assumed to be an additive, biallelic (A_1 and A_2), autosomal locus with Mendelian transmission probabilities. We consider here that $p_1 = 0.6$ and $p_2 (= 1 - p_1)$ are the frequencies of alleles A_1 and A_2. Three genotypes can be encountered: A_1A_1, A_1A_2 and A_2A_2, with a frequency of p_1^2, $2p_1p_2$ and p_2^2, respectively. The A_2 allele is assumed to increase the trait value, and is called the favorable allele. Further, we assume no dominance and the additive allele effect a was 3.7 phenotypic standard deviation units of the trait. The phenotypic data were simulated using heritability h^2 of 0.41 and repeatability r of 0.52. The genotype of the offspring was determined according the Mendelian transmission probabilities. The polygenic effect of the offspring was determined as the summation of the mean of the parents' polygenic effect and the Mendelian sampling effect. The true values of parameters (major gene and polygene) used in the simulation of the population are illustrated in the Table 1.

Simulation of binary (0/1) data. The liability models for analysis of binary data were first proposed by (Wright 1934) and have been thoroughly investigated (e.g. Kadarmideen *et al.* 2000 applied liability models to QTL mapping). The simulated normal trait was standardized using the average, m and the standard deviation, s_P of the trait as: $y^* = (y - m)/s_P$, where y^* is the standardized normal data with $N(0,1)$. Then, based on the liability concepts, the y^* could be transformed into binary data as follows: *If* $y^* > t$ *then* $y^b = 1$, and *if* $y^* \le t$ *then* $y^b = 0$, where t is the threshold point. Here y^b taking value of '1' could be considered as diseased and '0' as healthy, thus representing liability model for complex diseases. The values for thresholds t were chosen in such a way that it represents two scenarios: a less common disease with 15% and more common disease with 40% incidence. Therefore the corresponding values of t were: $t = 1.036$ for 15% and $t = 0.253$ for 40% (Falconer and Mackay, 1996). Both the underlying normally distributed liability data (NDL) and binary (0/1) data resulting from truncating the same NDL data were kept for segregation analyses.

Statistical analyses. There were 3 types of data sets. The original NDL data and two binary data sets with 15% and 40% incidences. Same segregation analysis method was performed on all the normal and binary data sets. Simulations and analyses were replicated 100 times for each combination of parameters. Different values of parameters were used as initial values for the calculations of the estimated parameters. The segregation analysis method used in this study was based on the comparison of the likelihoods under 2 inheritance hypotheses (Le Roy *et al.*, 1990, Ilahi *et al.*, 2000 and Bodin *et al.*, 2002):

Mixed inheritance hypothesis (H_1). This model describes the genetic transmission of the simulated trait by polygenic effects and a single major gene effect. The parameters to be estimated are: the mean of each genotype($m_{A_1A_1}$, $m_{A_1A_2}$, $m_{A_2A_2}$), the three variance components (s_g^2, s_{pe}^2, s_e^2) and the genotypic frequencies $f(A_1A_1)$, $f(A_2A_2)$ and $(f(A_1A_2) = 1 - f(A_1A_1) - f(A_2A_2))$. These estimated parameters allowed the computation of the 'residual' heritability, and the repeatability.

Polygenic inheritance hypothesis (H_0). This model, which is a sub-model of the H_1 mixed inheritance hypothesis, is given by $m_{A_1A_1} = m_{A_1A_2} = m_{A_2A_2} = m$. In this case the parameters to be estimated are: m, s_g^2, s_{pe}^2, s_e^2 from which we can compute the 'total' heritability, and the repeatability. The likelihoods ℓ_0 and ℓ_1 were computed respectively for both hypotheses H_0 and H_1, the likelihood ratio is given by $LR = 2\log(\ell_0 / \ell_1)$. This likelihood ratio is compared to the value of c_d^2 with degrees of freedom d equal the difference in number of parameters between the mixed and polygenic inheritance hypotheses (Le Roy *et al.* 1990, Kadarmideen *et al.* 2000). In this analysis, $d = 4$. The estimation of parameters maximising the likelihoods was carried out using the Gauss-Hermit quadrature (D01BAF) and optimization (E04JBF) subroutines of the NAG Fortran Library with a quasi-Newton algorithm in which the derivatives were estimated by finite differences.

RESULTS AND DISCUSSION

The results of parameter estimates by segregation analyses for normal and binary traits under both polygenic and mixed inheritance models are given in Tables 1 and 2, respectively. The mean of the likelihood ratio (LR) for the normal trait, comparing mixed and polygenic models was about 165, greatly exceeding 13.3, the tabulated value of c_4^2 distribution at 1% significance level. This has confirmed the true mixed genetic determinism of the simulated trait. Using normal trait, the estimated parameters under mixed inheritance (H_1) were similar to the true values of parameters used in the simulation. The major gene variance was however underestimated (Table 1). On the other hand, the estimated parameters under polygenic inheritance (H_0) were overestimated, especially for the genetic and the permanent environmental variances. This is explained by the genetic model used in the simulation of data set: the major gene has a large additive effect on the trait. Moreover, under H_0, the major gene effect was not taken into account to explain the genetic variability of the analysed trait, which resulted in overestimation of genetic and permanent environmental variances.

For binary trait with both incidences (15 and 40%) under H_0, the estimates of heritabilities and repeatabilities were the same and high. In the case of H_1, however, the segregation analysis method used in this study did not allow the estimation of the permanent environmental variance. This may be due to the loss of genetic variability and information when normal distributed data were truncated to 0/1 binary form (Kadarmideen *et al*. 2000). In a recent study Miyake *et al*. (2002) using segregation analyses for binary traits, have also found similar problems in the estimation of variance components and to obtain a good convergence to true values. Using a low incidence (15%), the results of segregation analysis for binary trait show an overestimation of the unfavorable genotype A_1A_1 frequency and underestimation of A_1A_2 and A_2A_2 genotypes fequencies. In the case of high incidence (40%) however, there is an overestimation of the favorable genotype A_2A_2 frequency and underestimation of unfavorable genotype A_1A_1 frequency, (Table 2). This corresponds to earlier findings that statistical power is lower and bias is higher for low incidence than for intermediate incidence (Kadarmideen *et al*. 2000). For the normal trait, the estimated heritabilities and repeatabilities were lowered from H_0 to H_1, from 0.54 to 0.38 and from 0.80 to 0.51, respectively. This was expected and due to the taking into account of major gene effect in H_1. However, in the binary trait these estimates for both incidences were dramatically lowered from H_0 to H_1. It decreased from 0.38 to 0.01 and from 0.60 to 0.01 for 15% incidence, and from 0.39 to 0.012 and 0.60 to 0.012 for 40% incidence, respectively. We can observe that the estimated of residual variance did not change from H_0 to H_1, an underestimation of genetic variance especially for binary trait with low incidence and non estimation of the permanent environmental variance. This indicated that the parameters estimated on the binary scale are biased. This paper showed the possibility of applying segregation analysis to binary traits with intermediate incidence under mixed inheritance. However, more research is needed to apply and to investigate more appropriate statistical methods and softwares to detect major genes segregating in binary or categorical traits. The method used in this study for segregation analysis for binary traits show a weakness on the estimation of all the parameters and to give an expected likelihood values in both polygenic and mixed inheritance models of the population. Further analysis and other alternative methods using Bayesian methodology (e.g. Janss *et al*. 1995, 1998) are required.

REFERENCES

Bodin, L., SanCristobal-Gaudy, M., Lecerf, F., Mulsant, P., Bibé, B., Lajous, D., Belloc, J.P., Eychenne, F., Amigues, Y. and Elsen, J.M. (2002) *Genet. Sel. Evol.* **34**:447.

Falconer, D.S. and Mackay, T.F.C. (1996) Introduction to quantitative genetics 4th ed. Long. London.

Ilahi, H., Manfredi, E., Chastin, P., Monod, F., Elsen, J.M. and Le Roy, P. (2000) *Genet. Res.* **75**:315.

Janss, L.L.G. (1998). 6th WCGAPL. **27**:459.

Janss, L.L.G. and Van Arendonk, J.A.M. (1995) *Theor. Appl. Genet.* **91**:1137.

Kadarmideen, H.N., Janss, L.L.G., Dekkers, J.C.M. (2000) *Genet. Res*. **76**:305.

Le Roy, P., Naveau, J., Elsen, JM. and Sellier, P. (1990) *Genet. Res.* **55**:33.

Miayke, T., Sasaki, Y., Dolf, G. and Gaillard, C. (2002) 7th WCGAPL. Communication N° 21-05.

Wright, S. (1934) *Genetics.* **19**:506.

Table 1. True values of parameters and parameter estimates by segregation analyses for normal trait (averages and standard deviations of 100 replicates)

Parameters	True values	Polygenic inheritance (H_0)	Mixed inheritance (H_1)
m	0	-0.10 ($\pm$ 0.28)	-
$m_{A_1A_1}$	-3.50	-	-3.47 ($\pm$ 0.20)
$m_{A_1A_2}$	0	-	0.04 ($\pm$ 0.18)
$m_{A_2A_2}$	3.50	-	3.55 ($\pm$ 0.19)
$f(A_1A_1)$	0.36	-	0.35 ($\pm$ 0.05)
$f(A_1A_2)$	0.48	-	0.48 ($\pm$ 0.09)
$f(A_2A_2)$	0.16	-	0.17 ($\pm$ 0.03)
s^2_q	1.44	4.74 ($\pm$ 0.77)	1.30 ($\pm$ 0.26)
s^2_{pe}	0.36	2.38 ($\pm$ 0.66)	0.45 ($\pm$ 0.18)
s^2_e	1.69	1.68 ($\pm$ 0.03)	1.68 ($\pm$ 0.03)
s^2_m	5.88	-	3.45 (($\pm$ 0.80)
Heritability	0.41	0.54 ($\pm$ 0.07)	0.38 ($\pm$ 0.06)
Repeatability	0.52	0.80 ($\pm$ 0.01)	0.51 ($\pm$ 0.02)

Table 2. Parameter estimates by segregation analyses for binary trait using two incidences (average and standard deviations of 100 replicates)

	Incidence= 15%		Incidence= 40%	
Parameters	Polygenic inher. (H_0)	Mixed inheri. (H_1)	Polygenic inher. (H_0)	Mixed inheri. (H_1)
m	0.190 ($\pm$ 0.022)	-	0.410 ($\pm$ 0.072)	-
$m_{A_1A_1}$	-	0.023 ($\pm$ 0.004)	-	0.053 ($\pm$ 0.023)
$m_{A_1A_2}$	-	0.050 ($\pm$ 0.005)	-	0.096 ($\pm$ 0.018)
$m_{A_2A_2}$	-	0.852 ($\pm$ 0.011)	-	0.882 ($\pm$ 0.019)
$f(A_1A_1)$	-	0.53 ($\pm$ 0.08)	-	0.19 ($\pm$ 0.09)
$f(A_1A_2)$	-	0.38 ($\pm$ 0.10)	-	0.51 ($\pm$ 0.14)
$f(A_2A_2)$	-	0.09 ($\pm$ 0.05	-	0.30 ($\pm$ 0.12)
s^2_q	0.051 ($\pm$ 0.010)	0.0005 ($\pm$ 0.000)	0.091 ($\pm$ 0.021)	0.001 ($\pm$ 0.00)
s^2_{pe}	0.030 ($\pm$ 0.010)	0.000	0.050 ($\pm$ 0.018)	0.00
s^2_e	0.053 ($\pm$ 0.003)	0.049 ($\pm$ 0.002)	0.091 ($\pm$ 0.004)	0.084 ($\pm$ 0.009)
s^2_m	-	0.050 ($\pm$ 0.010)	-	0.090 ($\pm$ 0.022)
Heritability	0.38 ($\pm$ 0.07)	0.010 ($\pm$ 0.004)	0.39 ($\pm$ 0.08)	0.012 ($\pm$ 0.007)
Repeatability	0.60 ($\pm$ 0.02)	0.010 ($\pm$ 0.004)	0.60 ($\pm$ 0.02)	0.012 ($\pm$ 0.007)