Am J Hum Genet 39:137–144, 1986

Brief Communication

Bias of the Contribution of Single-Locus Effects to the Variance of a Quantitative Trait

ERIC BOERWINKLE[1] AND CHARLES F. SING

SUMMARY

Advances in our understanding of the physiology of many quantitative phenotypes combined with better measurement abilities is providing a means for pursuing a measured genotype approach to partitioning the phenotypic variance into the contribution of separate loci. The standard estimate of the contribution of a single locus to the phenotypic variance applied recently in the human genetics literature is a biased statistic. We compare the biased estimates from several published studies with biased corrected estimates to illustrate the general problem.

INTRODUCTION

Statistical evidence from commingling and segregation analyses that support the contribution of an unmeasured single locus to quantitative phenotypic variability has been obtained for several traits (for examples, [1–4]). Improved measurement methods, such as two-dimensional gel electrophoresis and restriction site polymorphisms, and an improved understanding of the biology of many phenotypes allow one to pursue an approach that uses measured genetic variability at a locus physiologically involved in the quantitative phenotype to define the effect of a single locus [5, 6]. This measured genotype approach directly assesses the frequencies and effects of genetic variability at loci whose contribution may not be large enough to be detected by an unmeasured genotype approach but is large enough to make a major biological contribution

Received September 16, 1985; revised January 22, 1986.

This work was supported in part by grants P01 CA 26803 from the National Cancer Institute, R01 HL24489 and R01 HL30428 from the National Institutes of Health.

[1] Both authors: Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109.

to the variability in the trait. Information about the frequencies and effects of alleles at separate loci provide information about the genetic architecture underlying quantitative phenotypic variability (i.e., the number of loci affecting the phenotype and the frequency and effects of alleles at these loci). Two measures of interest used to describe the impact of a single locus on a quantitative phenotype include the contribution by the locus to the variability in the phenotype and the proportion of the phenotypic variance attributable to variability at the locus. Estimators of these measures that simply substitute parameters of the model with parameter estimates have been recently proposed and applied in the human genetics literature [4, 7, 8–10]. We discuss here the bias of these estimators and investigate this bias for a sampling design that allows one to compute bias-corrected estimates. The effect of this bias on estimates of the impact of a single locus on variability of a quantitative phenotype for several published studies is also presented.

MODELING THE QUANTITATIVE PHENOTYPE

The model used here parameterizes the quantitative phenotype of the ith individual ($i = 1 \ldots n$) with genotype j (y_{ij}) as an additive combination of effects. Let y_{ij} be written

$$y_{ij} = \mu_j + G_i + E_i \ , \tag{1}$$

where μ_j is the mean of the jth genotype at a locus ($j = 1 \ldots J$) and is considered a fixed effect. G_i is the polygenic effect of the ith individual and is assumed to arise from the action of a large number of unlinked genes, each with small effects, acting additively and independently. E_i represents the totality of all nongenetic effects specific to the ith individual. G and E are random effects with expectation 0 and variance σ_G^2 and σ_E^2, respectively. The three components of the model are assumed to be uncorrelated and act additively. We also let f_j be the frequency of the jth genotype in the general population. f_j may be written in terms of allele frequencies if one assumes Hardy-Weinberg equilibrium. The parameters of the model to be estimated include the genotype frequencies (or the allele frequencies if one assumes Hardy-Weinberg equilibrium), the J genotype specific means, the within single-locus genotype polygenic variance, and environmental variance.

The total population variance of the random variable y is the sum of the variances of its independent parts. The contribution of the single locus to the variance of y is

$$V(\mu) = \sum_{j=1}^{J} f_j(\mu_j - \mu.)^2 \ , \tag{2}$$

where $\mu. = \Sigma_{j=1}^{J} f_j \mu_j$. The total phenotypic variance for the population is then given by

$$V(y_{ij}) = \sum_{j=1}^{J} f_j(\mu_j - \mu.)^2 + \sigma_G^2 + \sigma_E^2 \ . \tag{3}$$

The proportion of $V(y_{ij})$ attributable to variability at the single locus is defined as

$$\frac{\sum_{j=1}^{J} f_j(\mu_j - \mu_\cdot)^2}{\sum_{j=1}^{J} f_j(\mu_j - \mu_\cdot)^2 + \sigma_G^2 + \sigma_E^2} . \tag{4}$$

ESTIMATING THE CONTRIBUTION OF A SINGLE LOCUS TO THE PHENOTYPIC VARIANCE

Estimation of variance components and their ratios have received considerable attention in the statistics literature [11–15]. The simplest design that allows one to estimate single-locus frequencies and effects is equivalent to a single-factor design with a sample of unrelated individuals. The classifications are determined by the single-locus genotypes and the within class variance is $\sigma^2 = \sigma_G^2 + \sigma_E^2$. For this design, an unbiased estimator of the variance attributable to the measured locus, is given by

$$S^2 = \sum_{j=1}^{J} \hat{f}_j (\hat{\mu}_j - \hat{\mu}_\cdot)^2 - \frac{J-1}{n} \hat{\sigma}^2 , \tag{5}$$

where $\hat{}$ over the parameters denotes their usual unbiased estimators [15]. This follows because the expectation of the estimate of the variance contribution by the measured locus that simply substitutes parameters with parameter estimates is

$$E\left[\sum_{j=1}^{J} \hat{f}_j (\hat{\mu}_j - \hat{\mu}_\cdot)^2 \right] = \sum_{j=1}^{J} f_j(\mu_j - \mu_\cdot)^2 + \frac{J-1}{n} \sigma^2 . \tag{6}$$

Using equation (5) as the unbiased estimate of the contribution of the single locus to the phenotypic variance, an estimate of the proportion of the phenotypic variance attributable to the locus becomes

$$\frac{\sum_{j=1}^{J} \hat{f}_j (\hat{\mu}_j - \hat{\mu}_\cdot)^2 - [(J-1)/n] \hat{\sigma}^2}{\sum_{j=1}^{J} \hat{f}_j (\hat{\mu}_j - \hat{\mu}_\cdot)^2 + \hat{\sigma}^2 (1 - (J-1)/n)} . \tag{7}$$

Considering the single-locus effects as random rather than fixed, this estimator is equivalent to the standard estimator of the intraclass correlation coefficient [16] and is consistent [13].

The contribution of the measured locus to the variance of the quantitative

phenotype and the proportion of the variance attributable to variability at the locus have been estimated by replacing the parameters in equations (2) and (4) with their usual parameter estimates. The bias of this estimator of the contribution of the locus to the phenotypic variance is obtained from equation (6) and is equal to $[(J - 1)/n]\ \sigma^2$. This estimator of the proportion of the phenotypic variance attributable to the locus is similar to Elston et al.'s [4] and Lange et al.'s [7] "percent genetic variance." This estimator will, on the average, overestimate the proportion of the phenotypic variance attributable to the measured locus.

Figure 1 is a graph of the magnitude of the bias as a function of the total sample size for varying values of σ^2 and J. The bias is presented as a proportion of the variance contribution by the single locus which is considered to be 1.0. For example, with $n = 100$, $J = 6$, and the within mode variance equal to 2.0, that is, twice the variance contribution by the single locus, the bias of the estimate of the single-locus contribution to the phenotypic variance is 0.10 of the true single-locus contribution. The size of the bias decreases rapidly as the sample size increases and is expected to be small as long as the number of observations is large relative to the number of genotypic classes and the within mode variance. The bias is larger for a three-allele polymorphic system ($J = 6$) than a two-allele polymorphic system ($J = 3$) and, for a given number of alleles, is larger for larger values of the within mode variance. The size of the bias in relation to the single-locus contribution is appreciable if the within single-locus genotype variance component is large relative to the single-locus effects and the sample size is not large. In general, the bias is relatively low for each of the

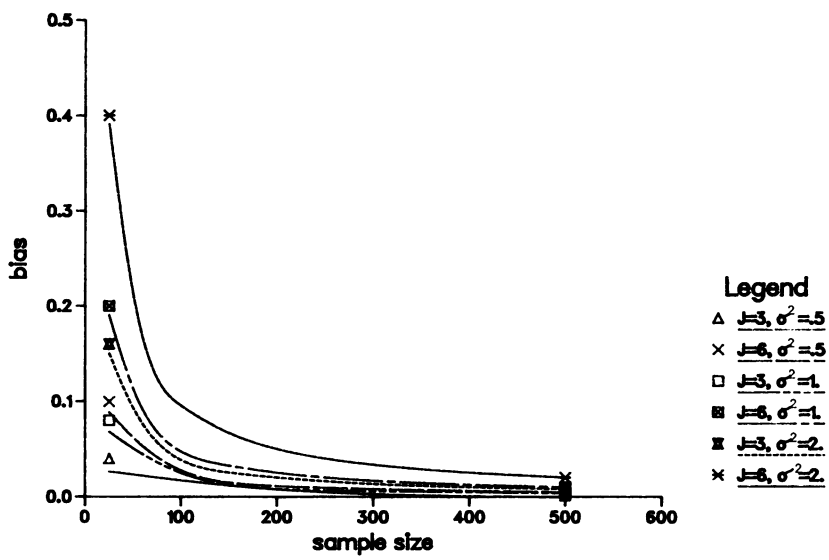

FIG. 1.—Bias of the estimate of the single-locus contribution to the total phenotypic variance. The bias is considered for varying values of the within genotype variance (σ^2) and the number of single-locus genotypes (J). The bias is presented as a proportion of the variance contribution by the single locus, which is considered to be 1.

cases considered here when the sample size is at least 150 individuals. That the bias is a problem in fact and not merely in principle is documented by the studies addressed here and elsewhere that have not obtained this sample size.

For more complicated models that include the effects of other loci or effects due to a common environment, the estimate of the contribution to the total phenotypic variance attributable by a single locus that substitute parameters with parameter estimates into the definition is also biased. It is unrealistic to anticipate all of the possible models of interest and attempt to explore the consequences of using the biased estimator in each case. In the general case, the bias is a corollary of Jensen's inequality [17]. Unfortunately, Jensen's inequality does not yield general information about the magnitude of the inequality. This must be obtained for each model and for each sampling design considered.

EXAMPLES

Table 1 compares the biased and unbiased estimates of the measured single-locus contribution to the total phenotypic variance using published results. The apolipoprotein (Apo) E-cholesterol studies [9, 18] investigate the effect of the three-allele *Apo E* polymorphism on total serum cholesterol. The Gc study [8] investigates the effect of the common two-allele *Gc* protein isoelectric focusing polymorphism on serum Gc concentration. The peptidase A study [19] examines the effect of the electrophoretic polymorphism in *peptidase A* on peptidase

TABLE 1

COMPARISON OF THE BIASED ESTIMATE OF THE VARIANCE CONTRIBUTION BY THE MEASURED SINGLE LOCUS [$V(\mu)$] WITH THE UNBIASED ESTIMATE [$V(\mu')$]; THEIR RATIOS TO THE ESTIMATED TOTAL PHENOTYPIC VARIANCE ARE ALSO INCLUDED

Study	No.	$V(\mu)^1$	$\frac{V(\mu)^2}{V(y)}$	$V(\mu')^3$	$\frac{V(\mu')^4}{V(y')}$
Apo E-cholesterol [9]	102	64.91	.087	31.46	.044
Apo E-cholesterol [18]	1,000	26.05	.023	20.53	.018
Gc [8]	89	3.93	.228	3.63	.214
Peptidase A [19]	79	13,428.0	.572	13,174.0	.568
RCAP [20]	275	646.15	.651	639.86	.648
RFLP-apo A II [21]	87	3.34	.115	2.75	.096

1. $\sum_{j=1}^{J} \hat{f}_j(\hat{\mu}_j - \hat{\mu}_\cdot)^2$

2. $\dfrac{\sum_{j=1}^{J} \hat{f}_j(\hat{\mu}_j - \hat{\mu}_\cdot)^2}{\sum_{j=1}^{J} f_j(\hat{\mu}_j - \hat{\mu}_\cdot)^2 + \hat{\sigma}^2}$

3. $\sum_{j=1}^{J} \hat{f}_j(\hat{\mu}_j - \hat{\mu}_\cdot)^2 - [(J-1)/n]\hat{\sigma}^2$

4. $\dfrac{\sum_{j=1}^{J} \hat{f}_j(\hat{\mu}_j - \hat{\mu}_\cdot)^2 - [(J-1)/n]\,\sigma^2}{\sum_{j=1}^{J} \hat{f}_j(\hat{\mu}_j - \hat{\mu}_\cdot)^2 + \hat{\sigma}^2\,[1 - (J-1)/n]}$

A activity. A red cell acid phosphotase study (RCAP) [20] examines the effect of a polymorphism in the *RCAP* protein on the activity of RCAP. Furthermore, the Apo A-II study [21] investigates the effect of a restriction site polymorphism 3′ to the *Apo A-II* structural gene on the levels of serum Apo A-II. The first two entries into the table give the biased estimates of the variance contribution by the single locus and of the proportion of the total phenotypic variance attributable to variability in the single locus. The third entry in the table is the unbiased estimate of the variance contribution of the respective single locus given in equation (5). The fourth entry is the estimator given in equation (7) of the proportion of the phenotypic variance attributable to the measured locus. As expected from examination of the formula, the bias may be substantial if the within class variability is large relative to the single gene effects and the number of genotype classes is large relative to the sample size. For the statistics considered here, the contribution to the phenotypic variability by the measured locus was reduced by an average of 16% when the bias was accounted for. The contribution of the *Apo E* polymorphism to total serum cholesterol variability was reduced by approximately one-half in the Sing and Davignon study [9] when the bias was accounted for. The large drop can be attributable to the relatively small number of individuals (no. = 102) in relation to the number of genotypic classes at the *Apo E* locus ($J = 6$) and the large within mode variance. This large drop in the contribution to the total phenotypic variance should not be expected as long as the measured locus effects are at least moderate with respect to the residual variance component and the measured genotype locus is a two-allele polymorphic system. Clearly, these biases need to be accounted for when partitioning the phenotypic variability using measured genotype information.

DISCUSSION

The question of how best to partition the phenotypic variability into the contribution of separate loci is yet to be adequately resolved. The statistics presented here are not the only methods to measure the variance contribution by a measured locus. For example, Hopper et al. [22] suggest parameterizing the single-gene effects as a random effect that would yield direct estimates of a variance component. Gold et al. [23] suggested the use of a minimum percent misclassification probability. One could also interpret the R^2 statistic from the analysis of variance to be an estimator of this proportion. Daiger et al. [8] used the parameterization presented by Falconer [24] as an estimator of this proportion. Finally, if the data are collected from a sample of related individuals and the effects of the single locus and residual polygenic effects are estimated simultaneously, the reduction in the polygenic variance component when the measured genotype effect is added to the model will yield an estimate of this proportion [6]. More work is needed to compare the statistical properties of these and other estimators and to explore their appropriate use.

Advances in our understanding of the physiology of many quantitative phenotypes combined with better measurement abilities such as two-dimensional gel electrophoresis, restriction site polymorphisms, and monoclonal antibodies

are providing abundant measured genotype information for many quantitative phenotypes. Using this information, the loci contributing to quantitative phenotype variability can be identified and their effects estimated. As the frequencies and effects of more loci in a system are estimated, the unmeasured polygenic random component will get smaller and fundamentally important questions about the genetic architecture of quantitative phenotypic variability can begin to be addressed.

ACKNOWLEDGMENTS

We thank Dr. Alan Robertson at the Institute of Animal Genetics in Edinburgh for his stimulating conversations concerning problems in estimating the contribution of a single locus to the variance of a quantitative phenotype.

REFERENCES

1. BOERWINKLE E, TURNER ST, SING CF: The role of the genetics of sodium lithium countertransport in the determination of blood pressure variability in the population at large, in *The Red Cell: Sixth Ann Arbor Conference,* edited by BREWER GJ, New York, Alan R. Liss, 1984
2. MOLL PP, SING CF, WILLIAMS RR, MAO SJT, KOTTKE BA: The genetic determination of plasma apolipoprotein A-I levels measured by radioimmunoassay: a study of high-risk pedigrees. *Am J Hum Genet* 38:361–372, 1986
3. LALOUEL JM, DARLU P, HENROTTE JG, RAO DC: Genetic regulation of plasma and red blood cell magnesium concentrations in man. II. Segregation analysis. *Am J Hum Genet* 35:938–950, 1983
4. ELSTON RC, NAMBOODIRI KK, NINO HV, POLLITZER WS: Studies on blood and urine glucose in Seminole Indians: indications for segregation of a major gene. *Am J Hum Genet* 26:13–34, 1974
5. BOERWINKLE E: The use of measured genotype information in the genetic analysis of the quantitative phenotype. I. Models and methods. Ph.D. thesis, Ann Arbor, Univ. of Michigan, 1985
6. BOERWINKLE E, CHAKRABORTY R, SING CF: The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and methods. *Ann Hum Genet* 50:181–194, 1986
7. LANGE K, SPENCE MA, FRANK MB: Application of the LOD method to the detection of a linkage between a quantitative trait and a qualitative marker: a simulation experiment. *Am J Hum Genet* 28:167–173, 1976
8. DAIGER SP, MILLER M, CHAKRABORTY R: Heritability of quantitative variation at the group-specific component (Gc) locus. *Am J Hum Genet* 36:663–676, 1984
9. SING CF, DAVIGNON J: Role of apolipoprotein E polymorphism in determining normal plasma lipid and lipoprotein variation. *Am J Hum Genet* 37:268–285, 1985
10. HANIS CL, CHAKRABORTY R, HEWITT-EMITT D: How much of variability in apolipoprotein A-II concentrations is explained by polymorphism adjacent to A-II gene? *Lancet* 1:1339–1340, 1985
11. FISHER RA: *Statistical Methods for Research Workers*. New York, Hadner, 1958
12. DANIELS HE: The estimation of components of variance. *J R Stat Assoc Suppl* 6:186–187, 1939
13. DONNER A, KOVAL JJ: The analysis of intraclass correlation in multiple samples. *Ann Hum Genet* 49:75–82, 1985
14. SCHEFFE H: *The Analysis of Variance*. New York, John Wiley, 1959
15. NETER J, WASSERMAN W, KUTNOR MH: *Applied Linear Statistical Models*. Homewood, Ill, R. D. Irwin, 1985
16. SNEDECOR GW, COCHRAN WG: *Statistical Methods*. Ames, Iowa, Iowa Univ. Press, 1980

17. RAO CR: *Linear Statistical Inference and Its Applications*. New York, John Wiley, 1965
18. MENZEL HJ, KLADETSKY RG, ASSMAN G: Apolipoprotein E polymorphism and coronary artery disease. *Arteriosclerosis* 3:310–322, 1983
19. LEWIS WHP: Common polymorphism of peptidase A. Electrophoretic variants associated with quantitative variation of red cell levels. *Ann Hum Genet* 36:267, 1973
20. SPENCER N, HOPKINSON DA, HARRIS H: Quantitative differences and gene dosage in the human red cell acid phosphatase polymorphism. *Nature* 201:299, 1964
21. SCOTT J, KNOTT TJ, PRIESTLEY LM, ET AL.: High-density lipoprotein composition is altered by a common DNA polymorphism adjacent to apoprotein A II gene in man. *Lancet* 771–773, 1985
22. HOPPER JL, TAIT BD, PROPERT DN, MATHEWS JD: Genetic analysis of systolic blood pressure in Melbourne families. *Clin Exp Pharmacol Physiol* 9:247–252, 1982
23. GOLD RJM, MAAG UR, MEAL JL, SCRIVER CR: The use of biochemical data in screening for mutant alleles and in genetic counselling. *Ann Hum Genet* 37:315–326, 1974
24. FALCONER DS: *Introduction to Quantitative Genetics*. New York, Ronald Press, 1960

MEETING: *International Congress of Human Genetics,* September 22–26, 1986, West Berlin.