Theor Appl Genet (1996) 92:998–1002

J. I. Weller · G. R. Wiggans · P. M. VanRaden
M. Ron

Application of a canonical transformation to detection of quantitative trait loci with the aid of genetic markers in a multi-trait experiment

Received: 22 July 1994 / Accepted: 28 October 1995

Abstract Effects of individual quantitative trait loci (QTLs) can be isolated with the aid of linked genetic markers. Most studies have analyzed each marker or pair of linked markers separately for each trait included in the analysis. Thus, the number of contrasts tested can be quite large. The experimentwise type-I error can be readily derived from the nominal type-I error if all contrasts are statistically independent, but different traits are generally correlated. A new set of uncorrelated traits can be derived by application of a canonical transformation. The total number of effective traits will generally be less than the original set. An example is presented for DNA microsatellite D21S4, which is used as a marker for milk production traits of Israeli dairy cattle. This locus had significant effects on milk and protein production but not on fat. It had a significant effect on only one of the canonical variables that was highly correlated with both milk and protein, and this variable explained 82% of the total variance. Thus, it can be concluded that a single QTL is affecting both traits. The effects on the original traits could be derived by a reverse transformation of the effects on the canonical variable.

Key words Quantitative trait loci · Multi-trait analysis · Canonical transformation

Introduction

Genetic markers can be used to detect individual loci affecting quantitative traits (QTLs). While many different experimental designs have been suggested for self-breeding and crossbreeding organisms (reviewed by Soller 1990, 1991, 1994; Weller 1992), until 1980 the application of these techniques was limited by the scarcity of segregating markers in populations of interest. Recently, new classes of highly polymorphic DNA-level genetic markers have been developed and applied to QTL detection (Georges et al. 1995; Jacob et al. 1991; Paterson et al. 1988; Ron et al. 1994). Most of the studies were an analysis of each marker or pair of linked markers separately for each trait included in the analysis. For example, Weller et al. (1988) performed a separate analysis of variance for all 180 possible combinations of ten genetic markers and 18 quantitative traits. Thus, nine "significant" effects at the 5% level would be expected purely by chance. Various studies have noted the problem of multiple comparisons, especially with respect to multiple markers. The usual suggestion is to raise the "nominal" significance level so that the overall experiment type-I error is no greater than 5%. Lander and Botstein (1989) developed formulas for the relationship between "nominal" and experiment type-I errors for multiple markers for both "sparse map" and "dense map" situations. For a sparse map, markers are assumed to be uncorrelated, and the overall experiment type-I error, α, can be computed as: $\alpha = 1 - (1 - p)^n$, where p = "nominal" type-I error and n = number of markers. However, as the nominal type-I error is reduced, the type-II error is increased, and the statistical power is lowered. Jansen (1994) has considered this problem in detail with respect to QTL detection.

For several uncorrelated traits, the overall experiment type-I error, α, can be computed similarly with n = number of traits. However, this formula is not correct if some of the traits are correlated. Another problem in QTL detection arises when the same marker is shown to have a similar association with several correlated traits. If each trait is analyzed separately, it cannot be deduced whether these effects are due to one locus with correlated effects on these traits, or to several loci each affecting a different trait.

Communicated by L. D. Van Vleck

J. I. Weller (✉) · M. Ron
Institute of Animal Sciences, A.R.O., The Volcani Center, Bet Dagan, 50250 Israel

G. R. Wiggans · P. M. VanRaden
Animal Improvement Programs Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705-2350, USA

Korol et al. (1995) suggested a multivariate analysis for two correlated traits, and showed that under certain circumstances a multivariate analysis can be more powerful than single-trait analysis. In their method the number of parameters that must be estimated is a function of the number of traits analyzed. They did not attempt to analyze a situation of more than two traits, but the increase in complexity of the analysis would be dramatic. Thus, this method is not a practical alternative if the number of traits analyzed is large. In commercial animal populations genetic evaluations of many individuals for many correlated traits are required. In order to avoid the complexity of a very large multivariate analysis, several studies have proposed univariate analysis on uncorrelated traits derived by a canonical transformation of the original traits (reviewed by Ducrocq and Besbes 1993). The solutions for the original traits can be obtained by reverse transformation. James (1991) was the first to suggest the application of canonical transformation to QTL detection, but he did not elaborate on the approach.

In this study the application of canonical transformation to QTL detection and analysis using genetic markers is described. An example is given using milk production traits of Israeli Holstein dairy cattle and a segregating DNA microsatellite.

Theory

For a given set of traits with known (co)variance matrix, a new set of traits can be derived by multiplication of the vector of the original traits by a matrix whose columns are the eigenvectors of the phenotypic (co)variance matrix. The resultant "canonical variables", which are linear functions of the original traits, are phenotypically uncorrelated. The number of canonical variables will be equal to the original number of traits. Generally, the original traits are standardized to unit variance by division by their standard deviations prior to computation of the eigenvectors. The (co)variance matrix among the traits is then equal to the correlation matrix, and the eigenvalues of the correlation matrix are equal to the coefficients of determination of the canonical variables for the overall variance. Thus, canonical variables with very low eigenvalues, relative to the sum of all eigenvalues, can be deleted from the analysis because they explain only a minuscule fraction of the variance of the original traits (Mardia et al. 1979).

The marker-linked effects can then be estimated on the canonical variables rather than the original traits, with three advantages. (1) Because the canonical variables are uncorrelated, the overall experiment significance level can be readily derived from the nominal significance level, as described above. (2) For highly correlated traits, it will generally be possible to reduce the number of variables analyzed because some of the eigenvalues will be very small. Thus, it will be possible to obtain the same experimentwise significance level with a less stringent nominal significance level. (3) Because the canonical variables are uncorrelated, a significant marker association with two traits is indicative of two linked QTL affecting the two variables.

Once significant effects are detected for the canonical variables, the effects on the original traits can be derived by the reverse transformation. That is, the inverse of the eigenvector matrix is multiplied by the vector of effects. It is assumed that all traits are recorded on all individuals, but removal of this restriction will be considered below.

Example calculation

The methodology will be illustrated using the results of Ron et al. (1994). Genetic evaluations were computed from the entire milk production recorded Israeli-Holstein population since 1985 for 305 day milk, fat, and protein yields (in kilograms) by a repeated measures animal model (Israel 1993; Wiggans et al. 1988). Only lactations with valid records for all three traits were included in the analysis. Production records of cows that were culled during the lactation, and records in progress at the time of data collection were extended to expected complete lactation production. Evaluations of fat and protein concentration were derived from the evaluations of total production (Israel 1993). Thus, estimated breeding values were derived for all animals for five production traits. Reliabilities of the genetic evaluations were estimated by the method of Misztal et al. (1991).

Blood or hair roots were collected from 151 cows, daughters of sire 783, in 11 Kibbutz herds chosen at random. Mean reliability of the cow evaluations was 52%. Direct polymerase chain reaction (PCR) analysis of 4–10 hair roots was performed as described by Ron et al. (1994). The cows were genotyped for microsatellite D21S4, which is located on chromosome 21 (Steffen et al. 1993). Sire 783 was heterozygous for alleles 18 and 21, where each allele is denoted by the number of TG repeats in the microsatellite core. Because dams of the cows were not genotyped, the origin of the alleles of the daughters could be determined only if a daughter genotype was different from that of her sire (Ron et al. 1993). Thus, allele 18 was inherited by 65 daughters, allele 21 by 50 daughters, and sire allele origin could not be determined for the remaining 36 daughters.

The differences in breeding values between the daughter groups, standard errors of the differences, and significance levels are given in Table 1. The results presented were collected to confirm a previous experiment in which sons of sire 783 were genotyped for ten loci, including D21S4, but only the latter locus was found to have a significant effect on production traits (Ron et al. 1994). Thus, for our experiment, single-sided significance values are appropriate. The differences between the two cow groups were 121 kg milk, 1.7 kg fat, and 2.7 kg protein, and in the same direction as in the

Table 1 The substitution effects and the *t*-test probabilities for alleles 18 and 21 of locus D21S4 among the daughters of sire 783

Trait (kg)	Substitution effect [a]	Standard error	Probability
Milk	121.31	60.57	0.029
Fat	1.69	2.02	0.203
Protein	2.75	1.63	0.047

[a] Difference in mean breeding values between 65 cows that inherited allele 18 and 50 cows that inherited allele 21

previous experiment. The average standard deviations for the estimated daughter breeding values were 326 kg milk, 10.7 kg fat, and 8.7 kg protein. Since breeding values are regressed, these standard deviations are less than the genetic standard deviations for this population given in Pasternak and Weller (1993). The allele substitution effects were significant by the t-test ($P < 0.05$) for milk and protein, but not for fat. The effects of this locus on percentage fat and protein were not significant (Ron et al. 1994).

The correlation matrix of the estimated breeding values for milk, fat, and protein yields is given in Table 2: all correlations are greater than 0.5, and milk and protein are the most highly correlated. These relationships are well-known from many studies (Pasternak and Weller 1993). The matrix of eigenvectors and the eigenvalues are given in Table 3. The eigenvalues indicate that the second and third variables explain more than 95% of the variance. Thus, variable one can be disregarded with virtually no loss of explanatory power.

Canonical variables were computed by first standardizing the production traits to unit variance by division by their standard deviations, and then multiplication of the matrix of standardized evaluations by the eigenvector matrix. The correlations between the original traits and the canonical variables are given in Table 4. Variable three is highly correlated with all three original traits. Variable two has a correlation of about 0.5 with fat and a lower correlation with the other original traits. The correlations between variable one and the other three traits are all less than 0.3. Thus, variable three includes the common variation among the three traits, especially milk and protein; variable two includes the variation in fat that is not correlated to the other traits; and variable one includes the residual variation.

Table 2 The correlation matrix of estimated cow breeding values for the three production traits

Trait	Milk	Fat	Protein
Milk	1.000	0.590	0.833
Fat	–	1.000	0.729
Protein	–	–	1.000

Table 3 The matrix of eigenvectors and the eigenvalues for the estimated breeding values of the production traits

	Eigenvectors			Total
	1	2	3	
Milk (kg)	0.5688	−0.5870	0.5762	
Fat (kg)	0.2683	0.7946	0.5447	
Protein (kg)	−0.7775	−0.1552	0.6094	
Eigenvalues	0.1385	0.4218	2.4398	3.0000
Proportion of total of eigenvalues	0.0462	0.1406	0.8133	1.0000

Table 4 The correlations between the original traits and the canonical variables

Canonical variable	Trait (kg)		
	Milk	Fat	Protein
1	0.2117	0.0999	−0.2893
2	−0.3812	0.5160	−0.1008
3	0.8999	0.8507	0.9519

Table 5 The allele substitution effects and t-test probabilities for locus D21S4 for the canonical variables

Canonical variable	Substitution effect [a]	Standard error	Probability
1	0.008	0.070	0.452
2	−0.143	0.122	0.122
3	0.492	0.291	0.047

[a] Difference in mean breeding values between 65 cows that inherited allele 18 and 50 cows that inherited allele 21

Substitution effects, their standards errors, and significance levels for the effect of locus D21S4 on the canonical variables are given in Table 5. This locus has a significant effect only on variable three, which is highly correlated with all three traits. The significance level was the same as for protein, which was slightly higher than for milk.

The allele substitution effects for the original traits can be derived by multiplication of the inverse of the eigenvector matrix by the vector of allele effects on the canonical variables. Nearly identical results to those presented in Table 1 are derived if the estimated effects for variable one are replaced with zeros.

Discussion

A systematic genome search for QTLs of economic importance in any of the major agricultural species requires genotyping hundreds of genetic markers and analysis of 10–20 quantitative traits (Georges et al. 1995; Paterson et al. 1988; Weller et al. 1988). Thus, the total number of comparisons will be huge if each marker-trait combination is analyzed separately. As the number of comparisons increases, the "nominal" type-I error per test must be decreased in order to obtain the required experiment-wise type-I error. For example, Georges et al. (1995) only considered contrasts with LOD scores > 3, i.e. a likelihood ratio of 1000. Although decreasing the type-I error decreases the probability of "false positives", it also decreases the statistical power to detect "true" effects. Furthermore, with many comparisons, the estimates of those effects deemed "significant" will be biased (Georges et al. 1995). These problems can only be alleviated by decreasing the number of comparisons.

Recently VanRaden and Weller (1994) showed how a single analysis can be performed for each chromosome even if the experiment includes many segregating markers on each chromosome. Thus, for a single trait with many markers, the number of analyses is collapsed into a single independent analyses for each chromosome, which are independently distributed. When a canonical transformation is used, a multi-trait analysis with many markers can be collapsed into a single analysis for each combination of marked chromosome and canonical variable. Besides the very large reduction in the number of comparisons, each test will be statistically independent of all the others so that the overall experiment significance level can be readily calculated from a given nominal significance level.

A number of studies have suggested that statistical power per individual genotyped can be increased by genotyping only individuals with extreme genotypes for a quantitative trait (Lander and Botstein 1989; Lebowitz et al. 1987). With the phenotyping of many extreme individuals, the power to detect a segregating QTL per individual genotyped can be increased fourfold (Darvasi and Soller 1992). However, it is not clear how to apply this technique if more than one quantitative trait is considered. By a canonical transformation, it should be possible, first, to reduce the total number of traits under consideration, and then to rank the importance of the canonical variables based on their eigenvalues. Thus, even if many traits are considered, selective genotyping can still be used for two or three canonical variables with the highest eigenvalues.

For the example presented, the nominal significance level, p, required to obtain an overall significance level of $\alpha = 0.05$ will be: $p = 1 - \exp[(\log\ 0.95)/3] = 0.017$. None of the original three traits are significant at this nominal level. However, once the canonical transformation is applied, only the two variables with the highest eigenvalues need to be considered (Table 3). In this case, there are only two contrasts, and the overall significance level is 0.025. This is still less than the probability of 0.047 for canonical variable three, and the null hypothesis still cannot be rejected with an experiment-wise significance level of 0.05.

Both milk and protein were significant at the nominal 0.05 level, but only one of the canonical traits showed a nominally significant effect. Moreover, by application of the reverse transformation, it was shown that the effect of locus D21S4 on both traits can be explained by the effects associated with two canonical variables, of which only one was significant. Thus, it appears that only a single QTL is affecting both traits.

Although the original analysis also considered fat and protein concentration (Ron et al. 1994), the matrix of eigenvectors could not be computed for all five traits because of dependencies among the traits. Genetic evaluations for fat and protein concentration are direct functions of the other traits (Israel 1993).

The data presented are somewhat problematic in that the analysis was performed on predicted breeding values, which are regressed estimates of the true breeding values, and the variance of these values is a function of their reliability. Direct analysis of records was not practical because data were collected over many herds, the cows have multiple records, and individuals are related. Hoeschele and VanRaden (1993) suggested using "daughter yield deviations", for analysis via the "granddaughter design" (Weller et al. 1990). Similarly, in the current "daughter design" analysis it would be possible to analyze cow "yield deviations" that are corrected for fixed effects but are not regressed. However, the variance of yield deviations is still a function of the quantity of information available on each cow.

In the current analysis, the correlation matrix of the estimated breeding values was diagonalized. Since this matrix is slightly different from the genetic correlation matrix (Pasternak and Weller 1993), the genetic correlation matrix was not completely diagonalized. A possible alternative would be to compute genetic evaluations or yield deviations from a set of canonically transformed variables with both residual and genetic covariance matrices diagonalized. However, this can only be directly applied if the model includes no other random factors, all traits are recorded on all individuals, and all traits have the same incidence matrix for all fixed effects (Ducrocq and Besbes, 1993).

The example presented met the last two criteria, but the repeated records analysis included a permanent environmental effect in addition to the genetic and residual effects. Additional random factors can be accommodated by applying a compromise diagonalization, which is adequate if the (co)variance matrices of the random factors are similar (Lin and Smith 1990). Furthermore, Ducrocq and Besbes (1993) demonstrate that a canonical transformation can be applied even if all traits are not recorded on all individuals by replacing missing values with their expectations.

Acknowledgements This research was supported by a grant from the Binational Agricultural Research and Development Fund (BARD).

References

Darvasi A, Soller M (1992) Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. Theor Appl Genet 85:353–359

Ducrocq V, Besbes B (1993) Solution of multiple trait animal models with missing data on some traits. J Anim Breed Genet 110: 81–92

Georges M, Nielson D, Mackinnon MJ, Mishra A, Okimoto R, Pasquino AT, Sargent LS, Sorensen A, Steele MR, Zhao X, Womack JE, Hoeschele I (1995) Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. Genetics 139:907–920

Hoeschele I, VanRaden PM (1993) Bayesian analysis of linkage between genetic markers and quantitative trait loci. II. Combining prior knowledge with experimental evidence. Theor Appl Genet 85: 946–952

Israel C (1993) Animal model: a memory efficient computational strategy applied to the genetic evaluation of the Israeli dairy cattle population. M.Sc. thesis, The Hebrew University, Jerusalem

Jacob HJ, Lindpainter K, Lincoln SE, Kusumi K, Bunker RK, Mao Y-P, Ganten D, Dzau VJ, Lander ES (1991) Genetic mapping of a gene causing hypertension in the stroke-prone spontaneously hypertensive rat. Cell 67:213–224

James JW (1991) Estimation of relations between genetic markers and quantitative traits. Erasmus Course, Wageningen Agricultural University, Wageningen, The Netherlands

Jansen RC (1994) Controlling the Type I and Type II errors in mapping quantitative trait loci. Genetics 138:871–881

Korol AB, Ronin YI, Kirzhner VM (1995) Interval mapping of quantitative trait loci employing correlated trait complexes. Genetics 140:1137–1147

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121: 185–199

Lebowitz RJ, Soller M, Beckmann JS (1987) Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. Theor Appl Genet 73: 556–562

Lin CY, Smith SP (1990) Transformation of multitrait to unitrait mixed model analysis of data with multiple random effects. J Dairy Sci 73:2494–2502

Mardia KV, Kent JT, Bibby JM (1979) Multivariate analysis. Academic Press, London

Misztal I, Lawlor TJ, Short TH, Wiggans GR (1991) Continuous genetic evaluation of Holsteins for type. J Dairy Sci 74:2001–2009

Pasternak H, Weller JI (1993) Optimum linear indices for nonlinear profit functions. Anim Prod 55:43–50

Paterson AH, Lander ES, Hewitt J, Peterson S, Tanksely SD (1988) Resolution of quantitative traits into Mendelian factors by using a complete RFLP linkage map. Nature 335:721–726

Ron M, Band M, Wyler A, Weller JI (1993) Unequivocal determination of sire allele origin for multiallelic microsatellites when only the sire and progeny are genotyped. Anim Genet 24:171–176

Ron M, Band M, Yanai A, Weller JI (1994) Mapping quantitative trait loci with DNA microsatellites in a commercial dairy cattle population. Anim Genet 25:259–264

Soller M (1990) Genetic mapping of the bovine genome using DNA-level markers with particular attention to loci affecting quantitative traits of economic importance. J Dairy Sci 73:2628–2646

Soller M (1991) Mapping quantitative trait loci affecting traits of economic importance in animal populations using molecular markers. In: Schook LB, Lewin HA, McLaren DG (eds) Gene mapping: strategies, techniques and applications. Marcel Dekker, New York, pp 21–49

Soller M (1994) Marker-assisted selection – an overview. Anim Biotechnol 5:193–207

Steffen P, Eggen A, Dietz AB, Womack JE, Stranzinger G, Fried R (1993) Isolation and mapping of polymorphic microsatellites in cattle. Anim Genet 24:121–124

VanRaden PM, Weller JI (1994) A simple method to locate and estimate effects of individual genes with a saturated genetic marker map. J Dairy Sci 72 [Suppl 1]: 249

Weller JI (1992) Statistical methodologies for mapping and analysis of quantitative trait loci. In: Beckmann JS, Osborn TC (eds) Plant genomes: methods for genetic and physical mapping. Kluwer Academic, Dordrecht, pp 181–207

Weller JI, Kashi Y, Soller M (1990) Power of "daughter" and "granddaughter"designs for genetic mapping of quantitative traits in dairy cattle using genetic markers. J Dairy Sci 73:2525–2537

Weller JI, Soller M, Brody T (1988) Linkage analysis of quantitative traits in an interspecific cross of tomato (*L. esculentum* × *L. pimpinellifolium*) by means of genetic markers. Genetics 118: 329–339

Wiggans GR, Misztal I, Van Vleck LD (1988) Implementation of an animal model for genetic evaluation of dairy cattle in the United States. J Dairy Sci 71 [Suppl 2]: 54–69