Ankara Üniv Vet Fak Derg, **60**, 209-212, 2013

Investigations on pleiotropy and genome wide association analyses by random effects using QTL-MAS 2010 dataset

Burak KARACAÖREN

Akdeniz Üniversitesi, Ziraat Fakültesi, Zootekni Bölümü, Antalya, Türkiye.

Summary: Recent advances in molecular genetics have provided hundreds of thousands of single nucleotide polymorphisms to detect mutations at the vicinity of genes related with quantitative traits. Breeding values could be used as response variable in mixed model framework to detect possible associations with genomic relationship matrix. It is known that most of quantitative traits are correlated which leads to construct of networks and pathways of genes due to pleiotropy. Hence the main aim of this paper is to a) detecting pleiotropy by principal component analyses methods b) prediction of genomic breeding values by ridge regression c) detecting associations based on predicted genomic breeding values obtained from b) using QTLMAS 2010 simulated dataset. Most of the Quantitative Trait Locus (QTLs) were located at chromosome 1 and 3. Highest correlation between true breeding value and predicted breeding value were obtained by Gaussian Kernel function as 0.557. To detect pleiotropy we used first and second principal components as response variable and success rates found to be 0.2727 and 0.1714 and error rates found to be 0.5952 to 0.6400 for first two principal component loadings respectively. Using genomic breeding values as response variable gave better success rate and lower error rate compared with when using raw phenotypes. We found that using the most heritable and variable component (first component) had higher change to detect pleiotropic genes using QTLMAS-2010 dataset.

Keywords: Breeding values, genome wide association analyses, genomic selection, pleiotropic genes.

Rassal etkiler kullanılarak yapılan genom tabanlı ilişki ve pleiotropi analizi için QTL-MAS 2010 veri seti üzerine incelemeler

Özet: Moleküler genetikteki son gelişmeler fenotipler ile ilişkili olabilen başkalaşımların: yüz binlerce tekil nükleotit polimorfizmi ile saptanmasına olanak tanımıştır. Damızlık değerlerin karışık modellerde cevap değişkeni olarak kullanılması ile genom tabanlı ilişkiler tespit edilebilir. Pleiotropi nedeni ile farklı fenotipler birbirleri ile bağıntılı olabilmekte ve böylece gen ağları oluşturulabilmektedir. Dolayısı ile bu çalışmanın amaçları a) pleiotropinin temel bileşenler analizi ile tespiti b) Ridge regresyonu kullanarak genomik damızlık değerlerin tahmini ve c) b)'den elde edilen damızlık değerler ile ilişki analizini benzeşim yolu ile elde edilmiş QTLMAS 2010 veri seti ile yapmaktır. Verimden sorumlu bölge (QTL)'lerin büyük çoğunluğu 1 ve 3. kromozomlarda bulundu. Gerçek ve tahmin edilen damızlık değerler arasındaki en yüksek korelasyon Gausçu çekirdek ile bulundu (0.557). Birinci ve ikinci temel bileşenler ile pleiotropi tespitinde başarı oranları 0.2717 ve 0.1714; hata oranları ise 0.5952 ve 0.6400 olarak bulundu. Genomik damızlık değerlerinin cevap değişkeni olarak kullanılması fenotiplerin kullanımına oranla daha yüksek başarı oranı ve daha düşük hata oranları verdi. Pleiotropik genlerin tespitinde kalıtım derecesi ve çeşitliliği en yüksek olan ilk temel bileşenin kullanılması QTLMAS 2010 veri seti için daha iyi sonuç vermiştir.

Anahtar sözcükler: Damızlık değerler, genom tabanlı ilişki incelemesi, genomik seçilim, pleiotropik genler.

Introduction

A mixed model is a mathematical model including both fixed and random effects and used in many applications including prediction of genomic breeding values (12) and prediction of longitudinal breeding values (9). After taken into account the random and fixed effects from observations remaining random effects could give useful information regarding underlying phenomena.

Recent advances in molecular genetics have provided hundreds of thousands of Single Nucleotide Polymorphisms (SNPs) to detect mutations at the vicinity of genes related with quantitative traits. Undetected shared ancestry within samples of animals could lead to detect false genomic signals in association mapping (8). Although pedigree based relationship matrix could be used to introduce ancestral correlations into the mixed model equations; genomic relationship matrix could also be used to accomplish the aim. Remaining solutions: breeding values could be used as response variable in mixed model framework to detect possible associations.

It is known that most of quantitative traits are correlated which leads to construct of networks and pathways of genes (6). Such phenotypic correlations can arise from pleiotropic effect of genes. Principal component analyses (4, 7) could be used to discover loadings that combines underlying relationships hence pleiotropy among different quantitative traits.

Hence the main aim of this paper is to a) detecting pleiotropy by principal component analyses methods b) prediction of genomic breeding values by ridge regression (3) c) detecting associations based on predicted genomic breeding values obtained from b) using QTLMAS 2010 simulated dataset(16).

Material and Methods

Data: The pedigree included four generations with 2326 animals for quantitative trait. Additional last generation (*n*=900) were also simulated without phenotypes but with genotypes to predict genomic breeding values conditional to their ancestors (training population). The number of population founders was 20 (5 males and 15 females). Each female mated only once and gave birth approximately 30 progeny. Generations were forced to be nearly discrete hence over-lapping. To investigate to pleiotropy two genetically correlated traits were simulated. The genome consisted of 10031 Single Nucleotide Polymorphisms (SNPs) distributed over 5 chromosomes. The two major QTL positions were simulated on chromosome 3 and a set of other intermediate QTL positions were simulated on chromosome 1 and 2. Set of other QTL positions were simulated on chromosome 1 with tiny effects and lastly there was no QTL located at chromosome 5. More details about the dataset could be found at (Szydlowski and Paczynska, 2011).

Genome Wide Association Analyses and Predicting Genomic Breeding Values: We used mixed model to perform genome wide association analyses (3);

$$\mathbf{y} = \mathbf{X}\mathrm{b} + \mathbf{Z}\mathrm{a} + \mathrm{e} \qquad (1)$$

where **y** contains the observations, b is the fixed effects, a is the additive genetic effect, matrices **X** and **Z** are incidence matrices, and e is a vector containing residuals.

$$Var\begin{pmatrix} a \\ e \end{pmatrix} \sim N\left[\mathbf{0}; \begin{pmatrix} \mathbf{A}\sigma_a^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_e^2 \end{pmatrix}\right],$$

For the random effects, it is assumed that A is the coefficient of coancestry obtained from genotype of animals; I is an identity matrix, σ^2_a is the additive genetic variance and σ^2_e is the residual variance. Two criteria were used to compare the association results by different sampling schemes; the success rate (ratio of mapped QTL to the total number of simulated QTL) and the error rate (ratio of false positives to the number of reported positions) as was defined by (13). We judged mapped QTLs by if they were located within 1Mb distance from true QTL position.

We used ridge regression best linear unbiased prediction (RRBLUP) to predict genomic breeding values with different kernel functions (3): Gaussian and Exponential models. In RRBLUP each marker assumed to has same variance on the trait concerned. Gaussian model assumes that distances between genotypes could be scaled in [0 1] interval and measured by Euclidean metric. Exponential model has also same assumptions with Gaussian model but not in quadratic form.

Principal component analyses: Principal components analyses used to orthoganize the phenotypic space;

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p \\ &\vdots \\ y_p &= a_{p1}x_1 + a_{p2}x_2 + \ldots + a_{pp}x_p \end{aligned}$$

with the coefficients (*a*) being chosen so that $y_1, y_2, \ldots, y_p$ account for most of the explanatory proportions of the total variance of the original variables, $x_1, x_2, \ldots, x_p$, (4). We used linear combinations of multivariate phenotypes (loadings) for detecting pleiotropy (17). Loadings were used to estimate heritabilites and to detect associations.

Results

Quality Control: We excluded 263 SNPs due to minor allele frequency <1%, leaving 9768 SNPs in the analyses. We excluded 8 individuals with too high Identity By State (IBS) (>95%) leaving 2318 individuals in the dataset. We estimated heritability as 0.4239 and 0.4017 using first and second principal component loadings based on mixed model (1) using genomic coancestry matrix (1). Explanatory proportions and eigenvalues (in brackets) were 0.6240 (1.2484) and 0.376 (0.7516) for first and second principal components respectively.

Association Analyses: A genome wide association analyses were conducted by generalized least squares method using model (1) and we used genomic breeding values and principal component loadings as response variables. Most of the QTLs were located at chromosome 1 and 3. We did not detect any QTL on chromosome 5 which is indicative of the model perform well in terms of false positives. Success rate found to be 0.3240 and error rate found to be 0.1714 based on model (1) using genomic breeding values as response variable. To detect pleiotropy we used first and second principal components as response variable and success rates found to be 0.2727 and 0.1714 and error rates found to be 0.5952 to 0.6400 for first two principal component loadings respectively.

Predicted Genomic Breeding Values: We used RRBLUP to predict genomic breeding values of animals and correlation between true and predicted genomic breeding values were found to be 0.703. To predict genomic breeding values of individuals of last generation we used Gauss and Exponential kernel functions (Table 1). Highest correlation between true breeding value and predicted breeding value were obtained by Gaussian Kernel function as 0.557.

Table 1: Correlations between true and predicted breeding values with different kernel functions: ridge regression, Gaussian and exponential models for last generation.
Tablo 1: Son kuşak için değişik çekirdek işleçleri ile(ridge regresyon, Gausçu, ve üssel modeller) elde edilen gerçek ve tahmin edilmiş damızlık değerler arasındaki korelasyonlar.

	Ridge Regression	Gaussian	Exponential
True Breeding Value	0.553	0.557	0.504
Ridge Regression		0.996	0.930
Gaussian			0.940

Discussion

We combined quantitative and qualitative (binary) traits using a multivariate data reduction method to obtain linear combinations of them. We assumed that underlying hidden structure could be correlated with common genes hence pleiotropy. (14) found that principal component of the quantitative phenotypes and the residual of a logistic regression of the binary phenotypes may be an optimal method to combine quantitative and binary trait to reduce the dimension from multivariate to univariate dataset.

Quantitative trait was simulated with 0.39 heritability and binary trait was simulated with 0.52 heritability whereas we estimated the heritability to be 0.4239 and 0.4017 by first and second principal component loadings using genomic coefficient matrix. Estimates of heritability of first component were found to have slightly higher compared with the estimates of heritability of second components. Again first principal component had higher explanatory proportion (0.6240) compared with the second one (0.3760). Success rate were found higher (0.2727) using first principal component compared with second one (0.1714) and again error rate were found smaller using first principal component (0.5952) compared with when second principal component used as response variable (0.6400). (2) did not able to detect pleiotropic genes using principal component analyses. However they noted that failure of principal component analyses to detect pleiotropic genes might related with sampling from population and/or trait-specific genes. Although we found that using the most heritable and variable component (first component) had higher change to detect pleiotropic genes; (11) suggested to use multivariate regression to include number of principal components. (17) found that detecting probability of pleiotropy would be much higher when residuals from principal component analyses used when environmental covariates exist.

We used RRBLUP to predict breeding values of animals and correlation between true and predicted breeding values were found to be 0.703. RRBLUP assumes that all markers have effect on trait with equal variances. However Bayes type methods (15) assume different priors hence different marker effects over genome. Since, in reality, most of traits affected by limited number of loci with various effect sizes Bayes type model could give better results compared with RRBLUP.

Using genomic breeding values as response variable gave better success rate (0.3240) and lower error rate (0.1714) compared with when using raw phenotypes (10) (success rate 0.3000; error rate 0.2900) in association model. Generalized least square method with estimated breeding values gave better success rate (0.3240) and lower error rate (0.1714) compared with our previous model; GRAMMAR(Genome-wide rapid association using mixed model and regression) using pedigree information(8) with phenotypes (success rate 0.1400; error rate 0.4400). (5) suggested using predicted breeding values as response variable for genomic prediction.

We used 3 different models to predict breeding values of last generation without phenotypes (Table 1). All 3 models had higher correlations within each other (ranged from 0.9300-0.9960). Highest correlation with true breeding values was obtained by Gaussian kernel (0.5570) although Ridge Regression with realized relation matrix had similar correlations (0.5530). (3) also obtained similar results for predicted breeding values using RRBLUP and Gaussian kernel in maize dataset.

References

1. **Aulchenko YS, Ripke S, Isaacs A, van Dujin, CM** (2007): *GenABEL: An R library for genome-wide association analysis*. Bioinformatics, **23**, 1294-1296.
2. **Bensen JT, Lange LA, Langefeld CD, Chang BL, Bleecker ER, Meyers DA, Xu J** (2003): *Exploring pleiotropy using principal components*. BMC Genet, **4**,S53.
3. **Endelman JB** (2011): *Ridge regression and other kernels for genomic selection with R package rrBLUP*. Plant Gen, **4**,250–255.
4. **Everitt BS, Landau S, Leese M** (2001)*: Cluster Analysis*. National Academy Press, Washington, DC.
5. **Guo G, Lund MS, Zhang Y, Su G** (2010): *Comparison between genomic predictions using daughter yield deviation and conventional estimated breeding value as response variables*. J Anim Breed Genet, **127**,423–432.
6. **Hill GW, Zhang SX** (2012)*: On the pleiotropic structure of the genotype-phenotype map and the evolvability of complex organisms*. Genetics, 3 Jan 2012(doi: 10.1534/genetics.111.135681).

7. **Karacaören B, Kadarmideen H** (2008): *Principal component and clustering analyses of functional traits in swiss dairy cattle. Turk. J. Vet. Anim. Sci,* **32**, 163-167.
8. **Karacaören B, Silander T, Alvarez-Castro MJ, Haley CS, de Koning DJ** (2011): *Association analyses of the MAS-QTL dataset using GRAMMAR, principal components and Bayesian network methodologies*. BMC Proc, **5** (Suppl 3), S8
9. **Karacaören B, Janss L LG, Kadarmideen HN** (2012): *Predicting breeding values in animals by kalman filter: application to body condition scores in dairy cattle*. Kafkas Univ Vet Fak Derg, **18**, 627-632.
10. **Karacaören B** (2012): *Some observations for discordant sib pair design using QTL-MAS 2010 dataset*. Kafkas Univ Vet Fak Derg, **18**:857-860.
11. **Mei H, Chen W, Dellinger A, He J, Wang M, Yau C, Srinivasan SR, Berenson GS** (2010): *Principal-component-based multivariate regression for genetic association studies of metabolic syndrome components*. BMC Genetics, **11**:100.
12. **Meuwissen THE, Hayes BJ, Goddard ME** (2001): *Prediction of total genetic value using genome wide dense marker maps*. Genetics,**157**,1819–1829.
13. **Mucha S, Pszczola M, Strabel T, Wolc A, Pacynska P, Szydlowski M** (2011): *Comparison of analyses of the QTLMAS XIV common dataset. II: QTL analysis*. BMC Proc, **5** (Suppl 3), S2.
14. **Mukhopadhyay I, Saha S, Ghosh S** *(*2011)*: Integrating binary traits with quantitative phenotypes for association mapping of multivariate phenotypes.* BMC Proc, **5***,*S73.
15. **Pszczola M, Strabel T, Wolc A, Mucha S, Szydlowski M** (2011): *Comparison of analyses of the QTLMAS XIV common dataset. I: genomic selection*. BMC Proc, **5**(Suppl 3),S1.
16. **Szydlowski M, Paczynska P** (2011) *QTLMAS 2010: Simulated dataset*. BMC Proc, **5** (Suppl 3), S3.
17. **Wang X, Kammerer CM, Anderson S, Lu J, Feingold E** (2009*): A comparison of principal component analysis and factor analysis strategies for uncovering pleiotropic factors*. Genet Epidemiolgy, **33**, 325-331.

Geliş tarihi: 19.07.2012 / Kabul tarihi: 15.02.2013

Address for correspondence:
Burak Karacaören
Akdeniz Üniversitesi, Ziraat Fakültesi,
Zootekni Bölümü, 07059, Antalya, Türkiye.
e-mail: burakkaracaoren@akdeniz.edu.tr