Investigations on Pleiotropy for Growth Related Traits based on Bayesian Residuals by Principal Component Analyses using F2 Mice Dataset

Burak Karacaören*[1]

[1,] Department of Animal Science, Faculty of Agriculture, Akdeniz University, TURKEY.

(burakkaracaoren@akdeniz.edu.tr)

ABSTRACT

Admixture mapping could be used to detect inheritance patterns of complex traits associated with Single Nucleotide Polymorphisms. Genetic stratification in the population could be detected and corrected using Bayesian residuals in admixture mapping. Prediction and use of principal components may lead to detection of pleiotropic genes. Ancestral genotypes of founder populations were available for the F_2 mice for growth related traits at 8 weeks of age. Genotypes of founders could be used to predict ancestry informativeness of the markers. The objective of this study was to detect pleiotropic genes for growth related traits using Bayesian residuals and principal component analyses by F2 admixture model under various ancestry informative levels of markers. The model identified that SNPs from chromosomes 6, 10 and 11 potentially have (suggestive) pleiotropic effects. As was expected higher ancestry informative levels lead to higher test statistics in the admixture model. However none of the SNPs were found statistically significant. Since Bayesian models could be used with non-normal and/or small datasets; employing Bayesian residuals in the admixture model instead of (restricted) maximum likelihood residuals may lead to advantages in practical applications. Principal component analyses could be useful to detect pleiotropic genes.

Keywords: Genomics, Bayesian Model, Gibbs Sampling, Admixture Mapping, Population Stratification, Genetic Analyses

1. INTRODUCTION

Admixture mapping could be used to detect inheritance patterns of complex traits associated with Single Nucleotide Polymorphisms (SNPs). [1] used residuals of the regression model based on maximum likelihood predictions of genetic variance components to account for population stratification in association mapping. Similarly genetic stratification in the population could be detected and corrected using Bayesian residuals in admixture mapping [2].

Most of the quantitative traits are found to be related each other due to common action of pleiotropic genes [3]. Prediction and use of principal components may lead to detection of pleiotropic genes. Ancestral genotypes of founder populations were available for the F_2 mice [4] for growth related traits (body weight, BW, feed intake, FI, and feed efficiency, FE,) at 8 weeks of age. Genotypes of founders could be used to predict ancestry informativeness of the SNPs. The objective of this study was to detect pleiotropic genes for ancestry informative markers using Bayesian residuals and principal component analyses by F2 admixture model under various ancestry informative levels of SNPs.

2. MATERIAL AND METHODS

An F_2 population (n=661) was created by crossing M16 (F0; n=12) and ICR (F0; n=12) mouse lines for growth related traits at 8 weeks of age [4, 5]. The M16 line was formed by selecting for rapid weight gain while the ICR line was used as random control. Genotypes were collected for 1813 SNPs for each animal.

2.1. Bayesian Analyses

We will use the notation and results of [2, 6]. For genetic analyses of traits following mixed model is normally used in animal breeding;

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z_a a} + \mathbf{e} \qquad (1)$$

where $\mathbf{y}$ is the vector of observations, $\boldsymbol{\beta}$ is the vector of fixed effects, $\mathbf{a}$ is the vector of breeding values, $\mathbf{X}$ and $\mathbf{Z_a}$ are design matrices and $\mathbf{e}$ is the vector of random residual effects. For the random effects it was assumed that

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

where σ_a^2 and σ_e^2 ; are genetic and error variances. $\mathbf{A}$ is the additive genetic relationship matrix for the animals; $\mathbf{I}$ is an identity matrix. $\mathbf{A}$ is estimated by both pedigree file and SNPs information[13]. In the following, we show general assumptions used in (1), based on Bayesian principles. Proportional joint posterior distribution without constant terms given in (2)

$$f\left(y\middle|b,a,\sigma_a^2,\sigma_e^2\right) \propto \left(\sigma_e^2\right)^{-\frac{1}{2}N} \exp\left(-\frac{1}{2}\left(\mathbf{y}-\mathbf{Xb}-\mathbf{Z_a a}\right)'\left(\mathbf{y}-\mathbf{Xb}-\mathbf{Z_a a}\right)/\sigma_e^2\right)$$
$$\mathrm{x}\left(\sigma_a^2\right)^{-\frac{1}{2}N_a} \exp\left(-\frac{1}{2}\mathbf{a}'\mathbf{A^{-1}a}\ /\sigma_a^2\right)\left[\prod_{i=1}^{N}\left(\sigma_a^2\right)^{-\frac{1}{2}N_a} \exp\left(-\frac{1}{2}\left(\mathbf{a}\ \right)'\mathbf{A^{-1}}\left(\mathbf{a}\ \right)/\sigma_a^2\right)\right]$$

$$\mathrm{x}\left(\sigma_e^2\right)^{-\left(\frac{\nu_e}{2}+1\right)} \exp\left[-\frac{\nu_e S_e}{2\sigma_e^2}\right]\left(\sigma_a^2\right)^{-\left(\frac{\nu_a}{2}+1\right)} \exp\left[-\frac{\nu_a S_a}{2\sigma_a^2}\right] \quad . \qquad (2)$$

Last line of (2) are product of density of scaled inverted chi-square distributions assumed prior for variance parameters. The markov chain was run for 1.000 iterations with a burn-in period of 100. Details of the model (2) with conditional distributions for Gibbs sampler could be found in [2].

2.2. 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, $x_1, x_2, \ldots, x_p$, [7]. We used linear combinations of multivariate phenotypes (loadings) for detecting pleiotropy. We used residuals of (1) using BW, FE, and FI as response variable to obtain principal components.

2.3. Admixture Model

We will use the notation and results of [2, 8]. The F_2 population has two founder populations: M16 and ICR. A chromosome of a random mouse from the base population will have a mixture of inherited segments from the two founder populations in mosaic form. If ∇_M is the number of copies of the allele at locus τ originating from the M16 population passed from the mother, and ∇_F is the number of copies passed from the father, then ∇ can be equal only to 0 or 1. Over many generations, an allele will asymptotically have two possible fates: it will either be fixed or eliminated. However, for a specific number of generations, allele behavior could be modeled using Markov models. If π is the probability of passing a copy sourced from population M16 through random mating, $\nabla = \nabla_M + \nabla_F$ will have a binomial distribution, $\nabla \sim B(2, \pi)$. [8] showed that under intercross experimental design $\pi = 1/2$. Letting $y_1, y_2, \ldots, y_n$, denote of the residuals obtained from (1) and Z the test statistic for additive gene effects used to detect QTL using phenotypes from n (F2=661) unrelated mice,

$$Z = \frac{\sum_{i=1}^{n}\{(y_i - \bar{y})(\nabla_i - 2\pi)\}}{\left[n\hat{\sigma}_y^2 2\pi(1-\pi)^2\right]^{1/2}}$$

with sample of variance, $\hat{\sigma}_y^2$,and mean of top principal component, $\bar{y}$, (obtained for BW, FE or FI). We assumed that markers were distributed in equal spaces along the genome. We refer readers to [2, 8] for a detailed discussion of admixture mapping using the intercross model. Due to the large numbers of markers, we used 3.71 (LOD = 3.0) as the threshold for correcting multiple hypothesis testing.

3. RESULTS AND DISCUSSION

The means of BW, FI and FE was 38.81 gr, 187.26 gr and 0.049 (n=661) respectively. The Kolmogrov-Smirnow test confirmed normality for FI (P>0.15) and FE (P = 0.72). Although it appears as if BW is normally distributed, the Kolmogrov-Smirnow test did not confirm normality at P < 0.05. The effect of sex was included as a fixed effect in subsequent analyses. Maximum likelihood estimates of heritabilites for all traits did not converge due to small sampling size (n=661). We used full (unselected) data (n=1163 for BW and n=748 for FI and FE) to estimate genomic heritabilities [9]. Again estimate of heritability for BW was not available due to non-normality. Estimate of heritability for FI and FE was found to be 0.11 and 0.18 without sex effect and 0.18 and 0.28 with sex effect respectively. Estimates of genomic heritability for FI and FE were found as 0.28 and 0.24 by [4] as similar to our results. We excluded 6 individuals with too low call rate (<0.10) and 2 individuals too high identity by state (>95%) leaving 653 individuals in the dataset. We excluded 11 markers due to minor allele frequency (<1%) leaving 1802 SNPs in the analyses. Mean heterozygosity estimated to be 0.42 (0.04) and mean identity by state estimated to be 0.69 (0.03). Although SNPs used in this study were found to be highly informative; the association model and its results should be confirmed with a denser set of markers. Explanatory proportions and eigenvalues (in brackets) were 0.730 (2.1892), 0.225 (0.6742) and 0.046 (0.1366) for first, second and third principal components respectively. We used multidimensional scaling with genomic relationship matrix

[9] to investigate population stratification. Visual inspection of the plot did not show clearly separable clusters (Figure 1).

We combined BW, FI, and FE traits using a principal component model to obtain linear combinations of them. We assumed that underlying hidden structure could be correlated with common genes hence pleiotropy. We used the pedigree file in the admixture model to correct for ancestral stratification. Table 1 and Table 2 identified that SNPs from chromosomes 6, 10 and 11 potentially have (suggestive) pleiotropic effects. As was expected higher ancestry informative levels lead to higher test statistics in the admixture model (Table 1 and Table 2). However none of the SNPs were found statistically significant.

The first principal component had higher explanatory proportion (0.730) compared with the second one (0.225). [3] found higher success rate to detect pleiotropic effects using first principal component compared with second one in a simulated dataset. [10] suggested to use residuals with principal component analyses for detecting pleiotropic genes when environmental covariates exist.

Table 1. Test statistics of SNPs for first loading of principal components for pleiotropy using pedigree model and residuals.

Chromosome	Name	SNP	Position (base pair)	Ancestry Information Level	Z-value
11	rs3707772	1139	70452035	0.75	0.514
				0.90	3.336
				1.00	3.470
10	rs3685588	1026	10705123	0.75	0.191
				0.90	2.900
				1.00	2.913
10	mCV25163613	1025	6888336	0.75	0.397
				0.90	2.803
				1.00	2.803
10	rs4228098	1023	8493375	0.75	0.467
				0.90	2.639
				1.00	2.636
10	mCV25374719	1024	8325975	0.75	0.635
				0.90	2.634
				1.00	2.634

Table 2. Test statistics of SNPs for second loading of principal components for pleiotropy using pedigree model and residuals.

Chromosome	Name	SNP	Position (base pair)	Ancestry Information Level	Z-value
6	rs3706039	661	77051716	0.75	0.936
				0.90	3.120
				1.00	3.116
6	rs6412632	662	78689920	0.75	1.603
				0.90	3.098
				1.00	3.093
6	rs4226013	663	79558217	0.75	1.314
				0.90	3.100
				1.00	3.092
6	rs6259164	664	77051716	0.75	1.008
				0.90	3.100
				1.00	3.090
6	rs4226002	657	77066714	0.75	2.260
				0.90	3.091
				1.00	3.090

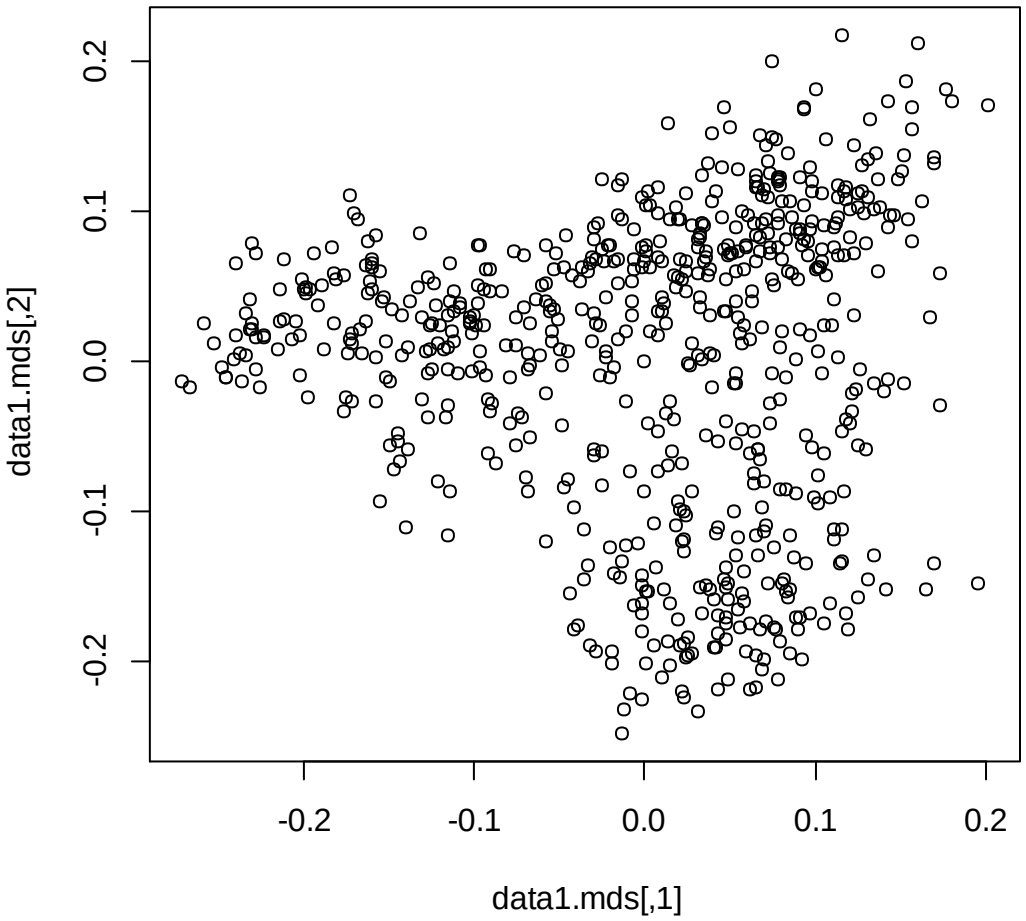


Figure 1. Multidimensional scaling plot based on first two genomic principal components.

4. CONCLUSION

Since Bayesian models could be used with non-normal and/or small datasets; employing Bayesian residuals in the admixture model instead of (restricted) maximum likelihood residuals may lead to advantages in practical applications. Principal component analyses could be useful to detect pleiotropic genes.

ACKNOWLEDGMENT

We would like to thank the Scientific and Technological Research Council of Turkey [TUBİTAK-TOVAG-113O799] for its financial support.

REFERENCES

[1] S. Y. Aulchenko, D.J. de Koning, C. S. Haley, 2007. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*. 177: 577-585.

[2] B. Karacaören, 2014. Admixture mapping of growth related traits in F2 mice dataset using ancestry informative markers. *Journal of Bioinformatics and Computational Biology* (in press).
[3] B. Karacaören, 2013. Investigations on pleiotropy and genome wide association analyses by random effects using QTL-MAS 2010 dataset. *Ankara Üniversitesi Veteriner Fakültesi Dergisi*. 60:209-212.
[4] A. Ehsani, P. SørenseP, D. Pomp, M. Allan, L. Janss, 2012. Inferring genetic architecture of complex traits using Bayesian integrative analysis of genome and transcriptome data. *BMC genomics* 13: 456.
[5] M.F. Allan,E. J. Eisen, D. Pomp, 2005. Genomic mapping of direct and correlated responses to long-term selection for rapid growth rate in mice. *Genetics* 170: 1863-1877.
[6] R. Mrode, R. Thompson, Linear Models for the Prediction of Animal Breeding Values, UK: CABI press, 2005.
[7] B.S Everitt, S. Landau, M. Leese, Cluster Analysis. National Academy Press, Washington, DC, 2001.
[8] D.O. Siegmund, B. Yakir B, The statistics of gene mapping, Springer Publishing Company, 2010.
[9] Y. S. Aulchenko, S. Ripke, A. Isaacs, C. M. van Dujin CM, 2007 GenABEL: An R library for genome-wide association analysis. *Bioinformatics.* 23 (10): 1294-1296.
[10] X. Wang, C. M. Kammerer, S. Anderson, J. Lu, E. Feingold, 2009. A comparison of principal component analysis and factor analysis strategies for uncovering pleiotropic factors. *Genetics Epidemiology*. 33 (4):325-331.