Original study

Empirical comparison of association and admixture mapping for body weight using F_2 mice data set

Burak Karacaören

Section of Biometry and Genetics , Department of Animal Science, Faculty of Agriculture, University of Akdeniz, Turkey

Abstract

Recent advances in molecular genetics have provided hundreds of thousands of single nucleotide polymorphisms (SNPs) to detect mutations in genes related with complex traits. Undetected shared ancestry within samples of individuals could lead to the detection of false genomic signals in association mapping. Pedigree-based relationship matrices or genomic relationship matrices could be used in a mixed model to predict and correct for genetic stratifications. Genotypic information of founder populations could also be used to explore patterns of inheritance for complex traits by admixture mapping. An F_2 population was created using M16 and ICR mouse lines for studying body weight at 8 weeks of age. Genotypes were collected for 1813 SNPs for each animal, including the founders. Bayesian residuals were used for population stratification in the admixture model. Association and admixture mapping detected similar genomic signals from chromosome 10. Our results provide empirical proof that gene flow from ancestral populations could be traced by admixture mapping with founder genotypes.

Corresponding author:
Burak Karacaören; email: burakkaracaoren@akdeniz.edu.tr
University of Akdeniz, Faculty of Agriculture, Department of Animal Science, Section of Biometry and Genetics, 07059 Antalya, Turkey

Introduction

Recent advances in molecular genetics have provided hundreds of thousands of single nucleotide polymorphisms (SNPs) to detect mutations in genes related with complex traits. Complex traits are often found among agricultural traits, which are affected not only by genes, but by the environment as well. Use of association mapping (Karacaören 2011) could lead to false genomic signals due to undetected shared ancestry within samples of individuals. Prediction and correction for genetic stratification could be achieved by use of pedigree-based relationship matrices or genomic relationship matrices in a mixed model. In addition, admixture mapping could be used to explore inheritance patterns of genotypic information among founder populations.

Admixed populations are formed by a mixture of two or more populations. Admixture mapping as described by Chakraborty & Weiss (1988) was used to detect genomic signals by correlating disease prevalence with the admixture proportions estimated by polymorphic markers in the admixed populations. Admixture mapping assumes that ancestral populations are genetically different and therefore have discordant gene frequencies as well as distinct associated phenotypes. If this assumption holds, the association between phenotypes and ancestral genotypic frequencies could be used to detect genetic factors related to complex traits. Hence, in the base admixed population, genetic factors may have higher frequencies on chromosome segments inherited from the ancestral population, which has higher disease variant frequency (Winkler *et al.* 2010).

Experimental genotypic information of human founders is not available. Numerous algorithms have been developed to predict ancestral genotypes under various scenarios (Winkler *et al.* 2010). While some theoretical models propose comparing genome-wide association to admixture mapping (Clarke & Whittemore 2007), we are not aware of any empirical study that does so. We hypothesize- that under optimal conditions the results of the two approaches will converge. We recognize that each methodology has different assumptions and tools to detect genomic signals, however we predict the results will be nearly the same, given optimal conditions.

Genomic investigations of complex traits such as body weight using model organisms may be useful for biological and agricultural research. Ancestral genotypes of founder populations were available for the F_2 mice (Ehsani *et al.* 2012) for body weight at 8 weeks of age. The objectives of this study were 1) to correct ancestral stratification in an admixture model similar to the GRAMMAR approach (Aulchenko *et al.* 2007) using a Bayesian model, 2), to use admixture mapping to detect body weight genes of F_2 mice using founder genotypes and 3) to detect genomic signals by association mapping (Endelman, 2011) and compare results with those obtained by use of admixture mapping.

Material and methods

Study population

An F_2 population (n=661) was created by crossing M16 (F0; n=12) and ICR (F0; n=12) mouse lines for body weight studies at 8 weeks of age (Allan *et al.* 2005, Ehsani *et al.* 2012). 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.

Statistical methods

Analyses were done using the R software (R Development Core Team 2007).

Bayesian genetic analysis
We propose to use a polygenic component with the admixture model to correct for population stratification. Our two-step model is similar to GRAMMAR (Aulchenko *et al.* 2007), but we employed a Bayesian model to obtain residuals instead of maximum likelihood methods. We used MCMCglmm (Hadfield 2010) to obtain conditional distributions for predicting parameters using Gibbs sampling in the mixed model equation stated above. We ran the model with 165 000 iterations using a 15 000-iteration burn-in period for body weight. To reduce auto-correlation, we sampled every 10th iteration. We tried different parameters of inverse Wishart prior distributions to obtain residuals.

Admixture mapping
We will use the notation and results of Siegmund & Yakir (2007). 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 χ_M is the number of copies of the allele at locus τ originating from the M16 population passed from the mother, and χ_F is the number of copies passed from the father, then χ 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 behaviour could be modelled using Markov models. If π is the probability of passing a copy sourced from population M16 through random mating, $\chi = \chi_M + \chi_F$ will have a binomial distribution, $\chi \sim B(2, \pi)$. Siegmund & Yakir (2007) showed that under intercross experimental design $\pi = 1/2$. Letting Y denote an $n \times 1$ vector of the observations and Z the test statistic for additive gene effects used to detect QTL using phenotypes $y_1, y_2, \ldots, y_n$, from n (F_2=661) unrelated mice,

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

with sample variance $\hat{\sigma}_y^2$ and average phenotype $\bar{y}$. We assumed that markers were distributed equispaced along the genome (∇=10cM). We refer readers to Siegmund & Yakir (2007) 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.

Association mapping
We used a mixed model to perform genome-wide association analyses (Endelman 2011):

$$y = Xb + Za + e \tag{2}$$

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

$$Var\begin{pmatrix} a \\ e \end{pmatrix} \sim N\left[0; \begin{pmatrix} A\sigma_a^2 & 0 \\ 0 & I\sigma_e^2 \end{pmatrix} \right] \tag{3}$$

For random effects it is assumed that A is the coefficient of coancestry obtained from the genotype of animals; I is an identity matrix, σ_a^2 is the additive genetic variance and σ_e^2 is the residual variance. We used a threshold of LOD score 3.0 to detect a genomic signal in association mapping as recommended by Siegmund & Yakir (2007) for multiple hypothesis testing correction. This equals to $[3 \times 4.6]^{1/2} \approx 3.71$ for z statistics.

Results

The means of body weight at 8 weeks, was 37.89 g (n=661) .While visually it appears as if body weight is normally distributed, the Kolmogrov-Smirnow test did not confirm normality at $P<0.01$. The effect of sex was statistically significant ($P<0.001$) and was included as a fixed effect in subsequent analyses.

Admixture mapping

We performed genome-wide admixture mapping using Bayesian residuals of body weight with ancestral information for two distinct founder populations using 3.71 as the Z threshold value (LOD=3) for multiple hypothesis testing. Figure 1 (Wickham 2009) shows that many SNPs influence body weight, however after multiple hypothesis testing correction, a total of 7 SNPs were significant. We found strong signals from chromosome 10. The effect of SNP rs13480530 on chromosome 10 was significant (Z=4.22, $P<2.44E-05$) for body weight, explaining approximately 3 % of the phenotypic variance.

We estimated admixture proportions (%AP) of the two founder populations and tabulated the SNPs in order of greatest influence in Table 1. SNP rs3707772 explained 2.3 % of the variation in body weight and is estimated to fully explain variation at the ancestral level.
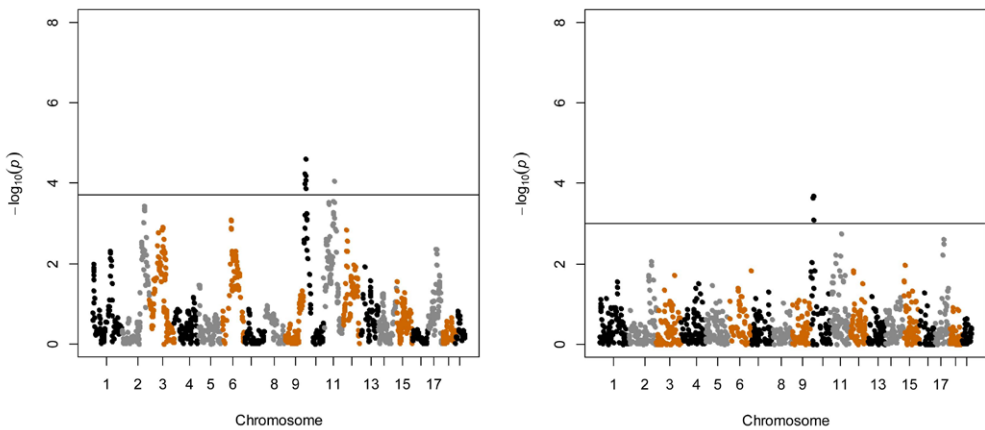

Figure 1
Manhattan plots for admixture mapping (left) and for association mapping (right) of body weight. LOD score thresholds of 3.71 and 3.0 are indicated by horizontal lines respectively.

Table 1
Descriptive statistics for top five most significant SNPs from admixture mapping

Chr	SNP	Pos (bp)	Z value (P val)	Corrected Z (P val)	% AP	% ph var
10	rs13480530	20206351	4.22 (2.44E-05)	3.91 (9.40E-05)	0.79	3.0
10	mCV23854739	14708850	4.03 (5.78E-05)	3.58 (0.000343)	0.62	2.0
10	rs3677768	17356586	4.01 (6.30E-05)	3.69 (0.000227)	0.50	0.5
10	rs13480514	16741683	3.94 (8.19E-05)	3.57 (0.000352)	0.17	0.0
11	rs3707772	70064591	3.93 (8.67E-05)	3.91 (9.05E-05)	1.0	2.3

Association mapping

Genome-wide association analyses were conducted by generalized least squares method using the mixed model equation. We used a genomic relationship matrix from the SNP data in the mixed model to correct for ancestral stratification (Table 2). Again, we found strong signals from chromosome 10. Three SNPs located on chromosomes 10 and 11 had effects on body weight after correcting for stratification using a genomic coancestry matrix (LOD>3).

Table 2
Descriptive statistics for top five most significant SNPs from association mapping

Chr	SNP	Pos (bp)	LOD (P val)	% AP	% ph var
10	rs6217650	24664975	3.69 (0.000206)	0.875	1.6
10	rs13480530	20223487	3.64 (0.000229)	0.79	3.0
10	rs3679120	22640984	3.09 (0.000811)	0.54	1.0
11	rs3707772	70265436	2.75 (0.001776)	1.0	2.3
17	rs3715723	58810327	2.61 (0.002469)	0.29	2.0

Discussion

Although SNPs used in this study were found to be highly informative; the F_2 admixture model and its results should be confirmed with a denser set of markers due to undetected stratification. Based on previous studies (Flint & Eleazar 2012) we expect that power will increase proportionate to the density of markers. There are several ways to improve the admixture model. For example, while we implicitly took into account the F_2 structure using a distribution test, the admixture model may still suffer from population stratification. To address this, we used Bayesian residuals in the model (Table 1, corrected Z values). The number of significant SNPs was reduced from 7 to 3 using Bayesian genetic covariances. Because Bayesian residuals were normally distributed ($P>0.01$), employing a two-step approach is recommended when using association mapping.

Tables 1 and 2 show that adjacent SNPs demonstrate linkage disequilibrium based on chromosomal locations. In the admixture model we assumed that founder marker alleles are in linkage equilibrium, however this assumption cannot be validated. Linkage disequilibrium could be accounted for by principal component regression models (Karacaören et al. 2011).

The effect of SNP rs13480530 on chromosome 10 was significant (Z=4.22, P<0.001) for body weight, explaining about 3% of the phenotypic variance, as shown in Table 1. This SNP lies near the phosphodiesterase 7B gene. Phosphodiesterase 3B is associated with leptin signalling

in the hypothalamus and controlling homeostasis of food intake and body weight (Sahu *et al.* 2011). This may indicate that the phosphodiesterase 7B gene could be relevant to body weight.

Table 1 compares the admixture proportions (%AM) with Z values. Our results confirm that discordant ancestral gene frequencies lead to higher Z values. In fact, admixture mapping assumes that ancestral populations are genetically different and therefore have discordant gene frequencies and distinct associated phenotypes as well. If this assumption holds, an association between phenotypes and ancestral genotypic frequencies could be used to detect genetic factors related to complex traits. However linkage disequilibrium between SNPs could be one reason why there are non linear trends between ancestry admixture proportions and Z values in Table 1.

The null hypothesis cannot be rejected (LOD>3); association and admixture mapping detected similar genomic signals. Our results provide empirical proof that gene flow from ancestral populations could be traced by admixture mapping with founder genotypes.

Acknowledgements

References

Allan MF, Eisen EJ, Pomp D (2005) Genomic Mapping of Direct and Correlated Responses to Long-Term Selection for Rapid Growth Rate in Mice. Genetics 170, 1863-1877

Aulchenko YS, de Koning DJ, Haley C (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

Chakraborty R, Weiss KM (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc Natl Acad Sci U S A 85, 9119-9123

Clarke G, Whittemore AS (2007) Comparison of admixture and association mapping in admixed families. Genet Epidemiol 31, 763-775

Ehsani A, Sørensen P, Pomp D, Allan M, Janss L (2012) Inferring genetic architecture of complex traits using Bayesian integrative analysis of genome and transcriptome data. BMC Genomics 13, 456

Endelman JB (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. The Plant Genome 4, 250-255

Flint J, Eskin E (2012) Genome-wide association studies in mice. Nat Rev Genet 13, 807-817

Hadfield JD (2010) MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. J Stat Softw 33, 1-22

Karacaören B, Silander T, Álvarez-Castro J, Haley CS, de Koning DJ (2011) Association analyses of the MAS-QTL data set using grammar, principal components and Bayesian network methodologies. BMC Proc 5 (Suppl.), S8

R Development Core Team (2007) R: A language and environmental for statistical computing. Vienna: R Foundation for statistical computing, Vienna, Austria

Sahu M, Litvin DG, Sahu A (2011) Phosphodiesterase-3B is expressed in proopiomelanocortin and neuropeptide Y neurons in the mouse hypothalamus. Neurosci Lett 505, 93-97

Siegmund DO, Yakir B (2007) The statistics of gene mapping. Springer Publishing Company, New York, NY, USA

Wickham H (2009) ggplot2: elegant graphics for data analysis. Springer, New York, USA

Winkler CA, Nelson GW, Smith MW (2010) Admixture Mapping Comes of Age. Annu Rev Genomics Hum Genet 11, 65-89