Kafkas Univ Vet Fak Derg
18 (5): 857-860, 2012
DOI:10.9775/kvfd.2012.6661

Journal Home-Page: http://vetdergi.kafkas.edu.tr
Online Submission: http://vetdergikafkas.org

RESEARCH ARTICLE

Some Observations for Discordant Sib Pair Design Using QTL-MAS 2010 Dataset [1]

Burak KARACAÖREN * ✎

* Department of Animal Science, Faculty of Agriculture, University of Akdeniz, TR-07070 Antalya - TURKEY

[1] *The author acknowledges funding from TUBITAK (The Scientific and Technological Research Council of Turkey)*

Makale Kodu (Article Code): KVFD-2012-6661

Summary

Association mapping seeks for markers at vicinity of genes affecting on complex traits. Family based association studies extensively used in human genetics for mapping genes. Discordant Sib Pair (DSP) design has advantages in controlling population stratification. The main aim of this paper was to investigate relation between number of discordant sib pairs to association mapping using QTL-MAS 2010 simulated dataset. The pedigree included four generations with 2326 individuals for quantitative trait. The genome consisted of 10031 Single Nucleotide Polymorphisms (SNPs) distributed over 5 chromosomes. We used mixed model to perform genome wide association analyses. In addition count of marker alleles in the affected and unaffected sib pairs contrasted to detect putative association. Most of the QTLs were located at chromosome 1 and 3. Genome based heritability were found to be more accurate compared with traditional pedigree based estimates of heritability. Since error rate decreases with increasing number of sib pairs the increasing proportions of success rate is not surprising. It is much more likely to have larger number of sib pairs in animal genetics compared with human genetics. Hence we believe that discordant sib pair approach might be useful for association mapping in domestic species.

Keywords: *Genome wide association analysis, Single nucleotide Polymorphism, Sibling relations*

Ayrık Kardeşler Deneme Deseni İçin QTL-MAS 2010 Veri Setinden Elde Edilmiş Bazı Gözlemler

Özet

İlişki haritalamacılığı; genler civarında yeralan ve karmaşık verimleri etkileyen işaretleyicileri tesbit etmeyi amaçlar. Genlerin haritalanmasında aile tabanlı çalışmalar sıklıkla kullanılagelmiştir. Ayrık kardeşler deneme deseni populasyon tabakasını kontrol etmek için faydalıdır. Bu çalışmanın amacı QTL-MAS 2010 simule verisini kullanarak ayrık kardeş sayısının ilişki haritalamacılığı ile bağıntısını incelemektir. Pedigri dört kuşağı 2326 birey için içermektedir. Kalıtga(genom) 10031 tekil nükleotid polimorfizmin 5 kromozoma dağıtılması ile oluşturulmuştur. Biz ilişki haritalamacılığını karışık etkili modeller kullanarak tespit ettik. Buna ek olarak mümkün ilişkiyi bulmak için hasta ve sağlıklı kardeşlerin farklı allelerini saydık. Pek çok QTL 1. ve 3. kromozomlarda bulunuldu. Kalıtga tabanlı kalıtım derecesi, geleneksel pedigri tabanlı kalıtım derecesinden daha doğru bir şekilde tahmin edildi. Kardeş sayısı arttıkça hata oranı azaldı ve buna bağlı olarak başarı oranı arttı. Hayvan genetiği çalışmalarında aile başına kardeş sayısı insan genetiğine oranla daha fazla olabilmektedir. Bu gözlemlerden dolayı ayrık kardeşler deneme deseninin çiftlik hayvanlarının ilişki haritalamacılığında kullanışlı olabileceği sonucuna vardık.

Anahtar sözcükler: *Kalıtga tabanlı ilişki incelemesi, Tekil nükleotid polimorfizm, Kardeş ilişkileri*

INTRODUCTION

Association mapping seeks for markers at vicinity of genes that has impact on complex traits. Genomic structure of cases and controls are compared for dense set of markers to detect putative associations between marker and disease gene. However when there is population stratification at case-control samples false positive associations might occur [1].

Family based association studies extensively used in human genetics for mapping genes. Discordant Sib Pair (DSP) design has advantages in controlling population

✎ **İletişim (Correspondence)**
☎ +90 242 2274560
✉ burakkaracaoren@akdeniz.edu.tr

stratification [2]. In DSP design sib pairs from each family has been used as cases and controls. With this design, reflected association must be causal since the frequency of disease alleles in cases should be higher than frequency of disease alleles in controls within relatively homogenized pair samples.

Boehnke and Langefeld [2] suggested to use one affected and one unaffected sib per family in DSP design. Sampling more than one discordant sib pair per family has practical difficulties in human genetics research. However it is possible to have multiple sib pairs per family in domestic species. Hence the main aim of this paper is to investigate relation between number of discordant sib pairs to association mapping using QTL-MAS 2010 [3] simulated dataset.

MATERIAL and METHODS

Data

The pedigree included four generations with 2326 individuals for quantitative trait. The number of population founders were 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 overlapping. 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 [3]. We subsampled the dataset by selecting different number of discordant sib pairs from each family (*n*=2, 4, 6, 8, 16, 24, 28).

Genome Wide Association Analyses

We used mixed model to perform genome wide association analyses [4,5];

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e} \qquad (1)$$

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 [6]. We judged mapped QTLs by if they were located within 1Mb distance from true QTL position.

Discordant Sip Pair Analyses

Count of marker alleles in the affected and unaffected sib pairs could be contrasted to detect putative association *(Table 1)*; these counts may be of all alleles present in the sibs *(scheme 1)* or may contrast those different alleles in the two sibs *(scheme 2)*.

Table 1. *Allele-Counting Schemes for discordant sib pairs*
Tablo 1. *Ayrık kardeşler için allel sayım yöntemleri*

Case	Sib Genotypes		Alleles Counted			
			Scheme 1		Scheme 2	
1	11	11	1,1	1,1	...	...
2	11	12	1,1	1,2	1	2
3	11	22	1,1	2,2	1,1	2,2

NOTE 1 and 2 represent distinct alleles at the marker locus, Adapted from Boehnke and Langefeld [2]

Pearson homogeneity statistic could be estimated from 2xm table via following formula;

$$T^1 = \sum_{j=1}^{m} \frac{(n_{1j} - n_{2j})^2}{n_{1j} + n_{2j}}$$

where n_{ij} stands for counted alleles among cases and controls, *i*=1, 2 for cases and controls, respectively and *j*=1…*m* (number of alleles).

Test statistics from counting schemes may have different distributions due to dependency of sib-pairs, hence permutation tests could be used to asses the significance of the tests. In the DSP case, we randomly interchanged the affection statues of the sibs, under the null hypothesis of no association, that the approach allows data to be permuted equally likely [2]. We switched or not switched the phenotype labels of each DSP independently with probability ½ to obtain permutations. We applied this procedure to two of top markers found by full association model with 100.000 number of permutations.

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 0.42 based on mixed model (1) using genomic coancestry matrix ^{5}.

Association Analyses

A genome wide association analyses were conducted by generalized least squares method using (1) by different sampling schemes. 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.

QTLs were mapped with different success rate and error rates based on different sampling schemes. For each sampling scheme (and full genome wide data) success rate and error rates were calculated and compared. Success rate ranged from 0.19 to 0.32 and the error rate ranged from 0.23 to 0.38. Although there is a tendency of higher success rate and lower error rate by increasing sampling size, this trend was not truly linear function of it.

Discordant Sib Pair Analyses

Table 3 presents genotype and allele counts for top two markers by counting all alleles *(scheme 1)* or discordant alleles *(scheme 2)*. There was good agreement for both markers using both allele counting schemes and whole genome wide association analyses. The results showed that both GWA using all individuals and samples of discordant sip pairs gave similar results. Contrasting alleles that are discordant between sib pairs *(scheme 2)* also increased the association test statistics compared with test statistics obtained by all alleles *(scheme) (Table 4).* Hence evidence for association was much higher when using discordant alleles instead of using all alleles. We evaluated the P values using Monte Carlo simulations *(Table 4)* based on 100.000 permutations of the data. Permutated P-values show agreement with full genome wide association results. Stronger association were observed for marker 4480 compared with marker 913.

***Table 2.** Success rate and error rates for genome wide association (GWA) with all individuals and different sampling schemes for discordant sib pairs (DSP) (n=2, 4, 6, 8, 16, 24, 28).*

***Tablo 2.** Bütün bireylere ait kalıtga ilişkisi (GWA) ve değişik örneklemli ayrık kardeşler(DSP) için (n=2, 4, 6, 8, 16, 24, 28) başarı ve hata oranları.*

Sampling Scheme	Success Rate	Error Rate
GWA	0.30	0.29
DSP (*n*=2)	0.22	0.23
DSP (*n*=4)	0.19	0.35
DSP (*n*=6)	0.22	0.36
DSP (*n*=8)	0.19	0.38
DSP (*n*=16)	0.27	0.36
DSP (*n*=24)	0.30	0.26
DSP (*n*=28)	0.32	0.28

***Table 3a1.** Genotype counts for markers 913*
***Tablo 3a1.** 913 numaralı işaretleyiciye ait genotip sayımları*

DSP Genotype Count=913			
Unaffected-Sib Genotype	Affected-Sib Genotype		
	AA	AB	BB
AA	2	8	3
AB	2	18	19
BB	0	5	18

***Table 3a2.** Genotype counts for markers 4480*
***Tablo 3a2.** 4480 numaralı işaretleyiciye ait genotip sayımları*

DSP Genotype Count=4480			
Unaffected-Sib Genotype	Affected-Sib Genotype		
	AA	AB	BB
AA	3	13	5
AB	3	29	12
BB	0	1	9

***Table 3b1.** Allele counts for markers 913*
***Tablo 3b1.** 913 numaralı işaretleyiciye ait allel sayımları*

DSP Allele Count=913		
Counting Scheme	A	B
All Alleles (scheme 1)		
Affected Sibs	65	85
Unaffected Sibs	39	111
Discordant Alleles (scheme 2)		
Affected Sibs	35	26
Unaffected Sibs	15	38

***Table 3b2.** Allele counts for markers 4480*
***Tablo 3b2.** 4480 numaralı işaretleyiciye ait allel sayımları*

DSP Allele Count=4480		
Counting Scheme	A	B
All Alleles (scheme 1)		
Affected Sibs	86	64
Unaffected Sibs	55	95
Discordant Alleles (scheme 2)		
Affected Sibs	38	16
Unaffected Sibs	17	36

***Table 4.** Test statistics, for all alleles, DSP alleles and whole genome wide association analyses (-log(P))for different markers*

***Tablo 4.** Bütün alleler, DSP allel sayımı ve bütüncül kalıtga tabanlı ilişki (-log(P)) analizlerine ait farklı işaretleyiciler için test istatistikleri*

Markers	All Alleles	DSP Allele Count	-log(P)
Marker 913	7.16 (0.01099)	10.25 (0.00241)	4.037788
Marker 4480	12.86 (0.00036)	15.71 (0.00003)	11.50308

DISCUSSION

Quantitative trait was simulated with 0.39 heritability whereas we estimated to be 0.42 by genomic coefficient matrix and 0.58 by pedigree based relationship matrix [7]. Genome based heritability were found to be more accurate compared with traditional pedigree based estimates of heritability. One reason is marker based heritabilities able to capture Mendelian sampling variation within families which is not possible by pedigree information [8]. Generalized least square method gave better success rate (0.30) and lower error rate (0.29) compared with our previous model; GRAMMAR (Genome-wide rapid association using mixed model and regression) using pedigree information [7] (success rate 0.14; error rate 0.44). Again significant difference between the two models in terms of accuracy of association mapping could be explained by Mendelian sampling. However as similar to results of GRAMMAR we localized only additive genes by generalized least square method; neither epistatic QTLs nor imprinted QTLs were detected.

Increasing the number of affected and unaffected sibs per family improved success rate. Kerber et al.[9] found similar results in their simulation study. A similar trend also found in error rate *(Table 2)*. However instead of using the total population (*n*=2326) samples of 24(*n*=1800) discordant sip pairs per family started to give better success and error rate.

We used two different schemes to count alleles, using all alleles *(scheme 1)* and discordant alleles within each sib pairs *(scheme 2)*. Since sib pairs can share none, one or both alleles at a specific chromosomal position such a counting scheme is possible. For both markers test statistics were found larger using *scheme 2*. Boehnke and Langefeld [2] compared different test statistics under various experimental designs using simulated datasets and concluded that discordant-alleles test *(scheme 2)* was the most powerful one. Due to dependency among sip pairs usual chi square tables cannot be used to assess significance of test statistics. Therefore we used permutation test. In order to obtain distribution of permutations we randomly permuted affection status of cases and controls 100.000 times. Associated P values were given in *Table 4*. In practice many more permutations may be needed according to desired accuracy for test statistics.

Our results are based on single simulated data set therefore it is not possible to derive analytical conclusions. However due to heavy computation cost for DSP permutations it is not easy to obtain results for multiple whole genome wide data sets. Observed non linear trends in *Table 2* for success and error rates might be associated with this problem. Probably if we used replicated datasets; success and error rates would converge to linear function of sampling size.

Although DSP design has been proposed for human genetics; area of interest could easily be extended to domestic species as well. Karacaören et al.[10] detected genomic signals using different approaches included discordant sib pair test for boar taint in pigs. As was demonstrated in this simulation study; since error rate decreases with increasing number of sib pairs the increasing proportions of success rate is not surprising. It is much more likely to have larger number of sib pairs in animal genetics compared with human genetics. Hence we believe that discordant sib pair approach might be useful for association mapping in domestic species.

ACKNOWLEDGEMENTS

The author wishes to acknowledge useful discussions with Dr Chris Haley and Dr Dirk Jan de Koning on discordant sib pair design.

REFERENCES

1. Reich DE, Goldstein DB: Detecting associations in case-control studies while correcting for population stratification. *Genet Epidemiol*, 20, 4-16, 2001.

2. Boehnke M, Langefelf CD: Genetic association mapping based on discordant sib pairs: The discordant-alleles test. *Am J Hum Genet*, 62 (4): 950-961, 1998.

3. Szydlowski M, Paczynska P: QTLMAS 2010: Simulated dataset. *BMC Proc*, 5 (Suppl 3): S3, 2011.

4. Endelman JB: Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Gen*, 4, 250-255, 2011.

5. Aulchenko YS, Ripke S, Isaacs A, van Dujin, CM: GenABEL: An R library for genome-wide association analysis. *Bioinformatics*, 23 (10): 1294-1296, 2007.

6. Mucha S, Pszczola M, Strabel T, Wolc A, Pacynska P, Szydlowski M: Comparison of analyses of the QTLMAS XIV common dataset. II: QTL analysis. *BMC Proc*, 5 (Suppl 3): S2, 2011.

7. Karacaören B, Silander T, Alvarez-Castro MJ, Haley CS, de Koning DJ: Association analyses of the MAS-QTL dataset using GRAMMAR, principal components and Bayesian network methodologies. *BMC Proc*, 5 (Suppl 3): S8, 2011.

8. Krag K, Janss LL, Shariati MM, Buitenhuis AJ: Heritability estimation based on small sample size using SNP markers. *9th World Congress on Genetics Applied to Livestock Production, 5th August 2010, Leipzig, Germany*, pp. 3-80, 2010.

9. Kerber RA, Amos CI, Yeap BY, Finkelstein DM, Thomas DC: Design considerations in a sib-pair study of linkage for suspectibility loci in cancer. *BMC Med Genet*, 9, 64, 2008.

10. Karacaören B, de Koning DJ, Velander I, Petersen S, Haley CS, Archibald AL: Alternative association analyses on boar taint using discordant sib pairs experimental design. *9th World Congress on Genetics Applied to Livestock Production, 5th August 2010, Leipzig, Germany*, ID743, 2010.