A New Powerful Method For Genome-wide Association Mapping Using Local Genealogies In A Mixed Model

G. Sahana[*], T.Mailund[À], M.S. Lund[*] and B. Guldbrandtsen[*]

Introduction

Local genealogies are genealogies for the longest chromosome region around a marker that do not require recurrent mutation or recombination (Mailund et al. 2006). A local genealogy-based genome-wide association mapping approach, Blossoc, was introduced by Mailund et al. (2006) for a case-control study and extended by Besenbacher et al. (2009) for complex traits. Ledur et al. (2009) used Blossoc for QTL mapping using the dataset simulated for QTLMAS-XII workshop by Lund et al. (2009). Blossoc was found to perform best out of several approaches used to analyze the common dataset (Crooks et al. 2009). However, the method developed can be applied only for unrelated samples which are not always available in human genetics and generally unavailable in livestock populations. Ledur et al. (2009) corrected the data for the pedigree and then analyzed as if it was a population sample of unrelated individuals. Precorrecting the phenotypes for pedigree, however, is not an optimal solution because it reduces power to identify QTL (Aulchenko et al. 2007). Therefore, a method which can utilize the genealogy information and simultaneously account for multiple levels of relatedness among individuals is expected to have more power when the samples actually come from a complex pedigree. Yu et al. (2006) present a unified mixed model method for association mapping that accounts for multiple levels of relatedness. The mixed model approach has the added advantage of being flexible, as it can be applied to both family-based and population samples.

We present here a new method, **GENMIX** (**gen**ealogy based **mix**ed model), which uses local genealogies in a variance component based model for association mapping for complex traits, thereby combining a genealogy-based approach with a mixed model.

Material and methods

Simulated data. The data simulated for QTLMAS XII workshop was used to test efficiency of the new method (details in Lund et al. 2009). In brief, a historic population was created by 100 founder individuals. For each of the subsequent 50 generations, 50 males and 50 females were produced by randomly sampling parents from the previous generation. The recorded pedigree had 4,665 individuals from four generations. The base generation of recorded pedigree had 15 males and 150 females. Each of the subsequent three generations had 750 males and 750 females. Phased biallelic marker genotypes were given at 0.1 cM intervals for six chromosomes, each 100 cM long. Each chromosome had 1000 markers. Phase was known without error. Fifty biallelic QTL were simulated on five chromosomes. No QTL was

[*]Aarhus University, Facultry of Agricultural Sciences, Genetics and Biotecnology, DK-8830 Tjele, Denmark
[À]Aarhus University, Bioinformatics Research Centre, DK-8000 Aarhus C, Denmark

simulated on chromosome 6, making it a control for false positives. The individuals genetic value was sum total effect of these 50 QTL. A normally distributed error term was added to give a genetic variance of 0.3 times the phenotypic variance. The average effect of allele substitution for the QTLs varied from <0.001 to 0.61. The minor allele frequencies of the QTLs ranged from 0.04 to 0.47 except one QTL which was, by chance, fixed in the base population. Details of these QTLs locations and their effects can be seen in Crooks et al. (2009).

Genealogy based mixed-model (GENMIX). In contrast to regular genome-wide association studies where phenotypic differences either are associated with single markers or with groups of markers organized in to haplo-groups in a non-stratified fashion, here phenotypes were associated using a hierarchical approach. Both grouping of markers into haplo-groups and clustering of observed haplotypes was done based on local genealogies (Mailund *et al.*, 2006). This method identifies the widest possible region surrounding a marker that allows construction of a genealogy forming a bifurcating tree without either recurrent mutation or recombination, in other words it satisfies the four-gamete condition of Hudson and Kaplan (1985). Each bifurcation in the binary tree corresponds to one bi-allelic marker. Splitting the tree at the top generates two clusters of haplotypes. Splitting the tree at any other node generates three clusters: one above the split point and two corresponding to the two branches below. For the analyses presented in this paper we split the tree at the top (one set of two clusters), the second level (two sets of three clusters) and at the third level (four sets of three clusters). Successively each clustering of haplotypes was included as a random effect in the model for analysis:

$$y_i = \mu + a_i + q_{h1i} + q_{h2i} + e_i$$

where y_i is the phenotype of individual i, μ is the population mean, a_i is the additive polygenic effect with $E(a_i)=0$ and $Var(a)=\mathbf{A}\sigma_a^2$, $\mathbf{A}$ is the numerator relationship matrix and σ_a^2 is the additive polygenic variance; q_{h1i} and q_{h2i} are two haplotypes effects of individual i where, $h1_i$ and $h2_i$ can take values 11, 12, 13, 22, 23, and 33 and $Var(q_{11}, q_{12}, q_{13}, q_{22}, q_{23}, q_{33})=\sigma_h^2 I$, σ_h^2 is the haplotype variance, I is the identity matrix; and e_i is a random residual. The local genealogies were constructed using the software Blossoc (*http://www.daimi.au.dk/~mailund/Blossoc/*) and variance component analysis was carried out using the software DMU (*http://www.dmu.agrsci.dk/*). The significance of the SNP association was tested using likelihood ratio test and the significant threshold was fixed at genome-wide 5% level after Bonferroni correction for multiple testing.

Assessment of QTL detectability. Crooks et al. (2009) examined the detectability of the simulated QTLs in this data when the actual QTL genotypes were fitted in a multiple regression model. Fifteen of the 50 simulated QTLs were significant in multiple regression and were named as major-QTL (**M-QTL**), while the undetected QTLs were called secondary-QTL (**S-QTL**). In our study we declared a M-QTL detected by GENMIX method, if a SNP showed genome-wide significant association within ±1 cM of the true QTL location. The SNP with strongest association within the defined region was taken as the putative QTL location. The S-QTLs were taken as detected if there was a significant SNP within ±1 cM of S-QTL location and there were no M-QTL within ±10 cM of the S-QTL

location. If there were more than one S-QTL within this ±10 cM region, then only the S-QTL with the biggest effect was considered as detected. This was done to exclude spurious effects due to linkage with another QTL of large effect.

Results and discussion

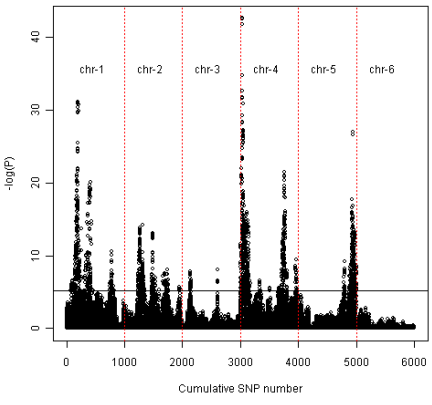
GENMIX detected 13 out 15 simulated M-QTL (Table 1). It was also able to identify two S-QTL. Figure 1 shows association signal across the whole genome. The distance between the simulated QTLs location and the strongest associated SNP ranged between 0-0.6 cM. There were no false positive results at 5% Bonferroni corrected significance threshold (Figure 1).

Table 1: QTLs identified by GENMIX

Name[a]	Chr[b]	Loc[c]	Effect[d]	MAF[f]	SNP[g]	-log$_{10}$(P)	Dist[h]
M1	1	20.00	0.62	0.28	19.7	31.04	0.3
S1	1	31.87	0.01	0.44	32.1	7.89	0.2
M2	1	40.00	0.56	0.07	40.2	19.75	0.2
M3	1	77.23	0.37	0.29	77.8	10.59	0.6
M4	2	27.41	0.35	0.44	27.1	13.82	0.3
M5	2	30.00	0.33	0.21	30.1	8.90	0.1
M6	2	48.62	0.37	0.40	48.8	13.04	0.2
M7	2	74.91	0.50	0.18	74.5	8.06	0.4
M8	3	14.91	0.30	0.40	14.5	6.70	0.4
M9	3	60.00	0.68	0.07	60.1	8.11	0.1
M10	4	3.21	0.61	0.39	3.3	42.59	0.1
M12	4	76.06	0.58	0.41	76.5	21.49	0.4
M13	4	96.49	0.29	0.19	96.5	9.47	0.0
S33	5	80.00	0.08	0.11	80.2	9.22	0.2
M15	5	93.49	0.75	0.26	93.5	26.98	0.0

[a]Names used by Crooks et al. (2009), M for major QTLs and S for secondary QTLs. [b]Chromosome. [c]simulated location for the QTL in cM. [d]Average effect of allele substitution (absolute value). [f]Minor allele frequency. [g]Position of the strongest associated SNP. [h]Distance from the real QTL location in *cM*.

Figure 1: Results of association mapping using GENMIX method for QTLMAS-XII data. The horizontal line (Y=5.92) is the genome-wide significant threshold at 5% after Bonferroni multiple testing correction. The vertical dotted lines show the chromosome boundaries.



The QTLMAS-XII dataset was analyzed for QTL mapping with a number of approaches usually applied in data generated with a complex pedigree structure. The methods were

combined linkage disequilibrium and linkage approach, Bayesian linkage analysis, Blossoc, multiple regression LD analysis (Crooks et al. 2009). Among all these approaches, Blossoc had highest power for detecting QTL. Therefore, we compared our method with Blossoc. Blossoc identified 11 M-QTL. GENMIX was able to detect all those 11 M-QTL plus two additional M-QTL. GENMIX also identified two S-QTL. The nearest M-QTL from these S-QTL was 11.8 cM away. Thus, these results are unlikely to be due to spurious linkage with M-QTL. The precision of QTL position estimates were comparable for both the methods. None of these two methods reported any false positive.

GENMIX combines advantages from two powerful association mapping methods: Blossoc and mixed model approaches. Blossoc performs better than single-marker analysis, since haplotype approaches can combine sets of common markers to define a rare haplotype in strong LD with a rare causative variance (Besenbacher et al. 2009). The mixed-model approach allows us to incorporate multiple levels of relatedness in the model instead of pre-correcting the data for pedigree. Future large-scale association studies will analyze thousands of samples from multiple populations in an effort to detect common genetic variants of weak effect (Ioannidis et al. 2009). GENMIX provides a powerful approach to analyze such combined data. The computer time required to analyses a chromosome with 1000 marker using GENMIX was ~2.5 h in a IBM HS22 blade servers equipped with one Intel Xeon X5570 2.93 GHz CPU and 48 GB RAM. Therefore, it is possible to carry out genome-wide association studies using multiple clusters in reasonable time.

Conclusion

We have presented a very powerful new method of association mapping combining Local genealogy information in a mixed model. The method was tested using QTLMAS-XII data which was earlier analyzed using several association mapping approaches. The power of GENMIX was better than the best method in the QTLMAS-XII workshop. GENMIX allows flexible modeling where both family and population relationships can be included in the model. It is a very powerful complement to currently available methods for association mapping.

The work was funded by the Danish Agency for Science, Technology and Innovation, FTP grant no. 09-065751.

References

Aulchenko, Y.S., de Koning, D.J., and Haley, C. (2007). *Genet.* 177:577-585.

Besenbacher, S., Mailund, T., and Schierup, M.H. (2009). *Genet.*, 181:747-753.

Crooks L., Sahana G., de Koning D.J. et al. (2009). *BMC Proc.*, 2009, 3(Suppl 1):S2.

Ledur, M.C., Navarro, N., and Rerez-Enciso, M. (2009). *BMC Proc.*, 2009, 3(Suppl 1):S9.

Hudson, R.R., and Kaplan N.L. (1985). Genet. 111:147-164

Ioannidis, J.P.A., Thomas, G., and Daly, M.J. (2009). *Nature Rev. Genet*. 10:318-329.

Lund, M.S., Sahana G., de Koning D.J. et al. (2009). *BMC Proc.*, 2009, 3(Suppl 1):S1.

Mailund, T., Besenbacher, S., and Schierup, M.H. (2006). *BMC Bioinformatics,* 7:454

Yu, J., Pressoir, G., Briggs, W.H. et al. (2006). *Nature Genet.*, 38:203-208.