Association Genetics in a Simulated Barley Population

Katrin MacKenzie and Christine Hackett
Biomathematics and Statistics Scotland

Introduction

This work investigates strategies for dealing with population structure in a simulated barley association mapping population. Barley is an inbreeding plant with a selfing rate of more than 90%. This leads to almost completely homozygous accessions. Commonly used approaches for locating traits, such as the transmission disequilibrium test, cannot be applied to such populations. Accounting for population structure by using marker information to estimate relatedness between each pair of cultivars is an alternative strategy and 3 approaches are investigated in this work. The naïve approach is also studied.

Simulated Population

48 landraces from the Mapping Adaptation of Barley to Drought Environments (MABDE) project were used as founders for the simulation. These came from 5 geographically distinct regions. 500 generations of random mating with a selfing rate of 0.92 were followed by several rounds of selection on three simulated traits, using different selection criteria in each region.

Initially, individuals for the selection panel were chosen within regions, generating what we denote as 'old cultivars'. This was followed by selection across regions to produce 'modern cultivars'. Landraces (generated after random mating) were also included in the selection panel.

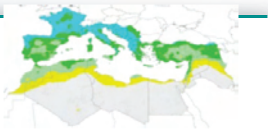
811 DaRT markers were available on the original 48 landraces and were simulated from these for the association panel.

10 association panels were simulated and 10 traits were generated for each panel from markers selected at random.

Association panel 5 regions -

20 landraces per region 40 'old' cultivars per region 40 'modern' cultivars per region

Replicated field trial, 3 complete blocks



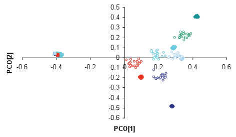


Figure 1. Principal coordinate plot based on marker data from a simulated association mapping population shows marked substructure, colours represent the 5 regions, symbols represent landraces (circles), old cultivars (squares) and modern cultivars (diamonds).

Simulated traits

Each trait was simulated as follows:

$y_{ij} = m + ad_i + g_i + e_{ij}$

where y_{ij} is the trait value for genotype i in block j, m is the mean, a is the marker effect, d = 0 if the marker is absent and 1 if it is present, g_i is the polygenic effect for genotype i, g ~ $N(0,\mathbf{A}\sigma_A^2)$, and e_{ij} is the residual term, e_{ij} are IID, $N(0,\sigma^2)$.

m is set to zero and $\sigma^2 = 1$. A is the additive genetic relationship matrix estimated from information about inheritance of markers from founders.

σ_A^2 and marker effect chosen according to 4 scenarios:

1 Heritability = 50%, Marker effect = 50%
2 Heritability = 50%, Marker effect = 25%
3 Heritability = 25%, Marker effect = 50%
4 Heritability = 25%, Marker effect = 25%

Model for Analysis

$Y = \mu + a_p + g_i + b_j + p_k + e_{ijkp}$

- a_p – effect of allele p
- g_i – genetic effect for line i
- b_j – effect of block j
- Fixed Effects - Marker data a_p
- Random Effects – Genotype g_i, Block b_j
- Kinship matrix (**K**) used for variance of random effect *g*, Var(*g*) = $2\mathbf{K}\sigma^2$
  The following statement was used in Genstat to include the kinship matrix in the mixed models:

```
VCOMPONENTS [fixed=SNP] random=ID+Genotype+Block
VSTRUCTURE [term=Genotype] model=fix; inverse=k
REML trait
```

Methods for Accounting for Population Structure

1 Do nothing (Naïve approach)

2 Kinship matrix – Ritland coefficients (Ritland, 1996), estimated from a random sample as

$F_{ij} = (Q_{ij} - Q_m)/(1 - Q_m)$

where Q_{ij} is the probability of identity by state for random genes for individuals i and j, Q_m is the average probability of identity by state for genes from random individuals.

3 Kinship matrix K_T – Stich et al. (2008), a development of the approach of Bernardo (1993).

$K_{ij} = (S_{ij} - 1)/(1 - T_{ij})$

where S_{ij} is the proportion of identical markers for lines i and j, and T_{ij} = P(IBS | not IBD). Stich et al. (2008) estimates T by minimising deviance from REML.

Results

Figures 2a and b shows –log10(p-value) for fitting a trait to 811 markers. The difference between accounting for population structure (Figure 2a (using K_T)) and failing to do so (Figure 2b) is clear. Both models locate the trait locus but the model with no correction for population structure has many false positives. The circle indicates the trait locus.

Figure 2(a) KT accounts for population structure

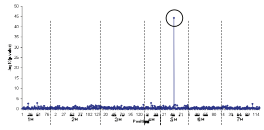


Figure 2(b) No population structure

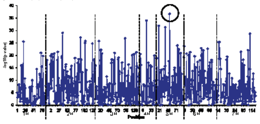


Figure 3 shows the total number of false positives (lying more than 3cM from the trait locus) among 117 markers on 1H ($P < 0.001$) for each scenario. The models with no population structure (None) shows a high number of false positives. The Ritland measure of kinship has slightly fewer false positive than the K_T measure.

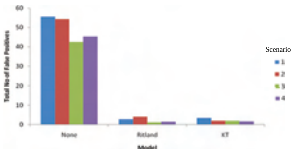


Figure 4 shows the proportion of true trait loci detected ($P < 0.001$) for the 3 models and the 4 different scenarios. The naïve approach detects more of the trait loci than the other approaches. The K_T kinship method detects more traits than the Ritland method when marker effects are low.

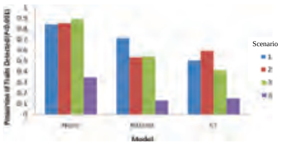
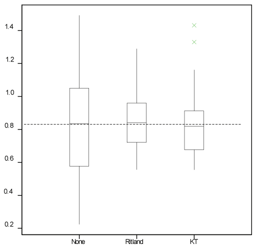


Figure 5 shows boxplots of the effects at the trait locus for the different models with heritability = 50% and marker effect of 50%. All three models estimate the effect well but the naïve model has greater variation than either of the methods which include a kinship matrix. The dashed line shows the simulated marker effect.



Discussion

Unless population structure is taken properly into account, association mapping of quantitative traits can give high numbers of false positive associations.

Including a kinship matrix in the model was found to be a better approach with both kinship measures giving similar performance.

References

Ritland, K., 1996. Estimators for pairwise relatedness and individual inbreeding coefficients. Genet. Res., Camb. **67**: 175-185.

Stich, B., Möhring, J., Piepho, H-P., Heckenberger, M., Buckler, E.S., Melchinger, A.E., 2008. Comparison of Mixed-Model Approaches for Association Mapping. Genetics **178**: 1745-1754.

Acknowledgements

Bill Thomas, SCRI
Jordi Comadran, SCRI
The MABDE project was funded by the European Union-INCO-MED program (ICA3-CT2002-10026).
BioSS and SCRI receives grant in aid from the Scottish Executive Environmental and Rural Affairs Department