Heredity (2013) 111, 86–87

www.nature.com/hdy

Heredity

LETTER TO THE EDITOR

Controlling type 1 error rates in genome-wide association studies in plants

Heredity (2013) **111,** 86–87; doi:10.1038/hdy.2012.101; published online 28 November 2012

I would like to thank the authors of Mueller *et al.* (2011) for their valuable contribution to the area of genome-wide association mapping. In their paper, they present a strategy for controlling the type 1 error rate (probability of making a false positive) when multiple correlated hypothesis tests are being performed. This issue lies at the heart of all genome-wide analyses.

Association mapping techniques for the analysis of data collected from highly structured populations are best cast within a linear mixed model framework (Yu *et al.,* 2006; Zhao *et al.,* 2007). Here, a separate linear mixed model is constructed for each marker locus where the marker locus is treated as a fixed effect. Maximum likelihood or residual maximum likelihood estimates are then obtained. From these estimates, a hypothesis test is performed to access the significance of the marker effect in the model. A test statistic is calculated and its significance is a measure of the strength of association between the marker locus and quantitative trait. However, this process does not take into account that in genome-wide studies, a large number of hypothesis tests are being performed. Without adjustment, the type 1 error rate is inflated.

Mueller *et al.* (2011) present a resampling approach for this problem. It is based on the Wald statistic. Briefly, the Wald statistic (W_j) for testing the significance of a fixed marker effect can be expressed as a quadratic form $W_j = \hat{\omega}'_j M_{jj}^{-1} \hat{\omega}_j$ where $\hat{\omega}_j = L_j \hat{\tau}_j$ and $M_{jj} = L_j (X'_j H_{jj}^{-1} X_j)^{-1} L'_j$. Here, L_j is a matrix of zeroes and ones such that $H_0 : L_j \tau_j = 0$, $\hat{\tau}_j$ is a vector of fixed marker effects estimates, X_j is a design matrix for the fixed effects in the linear mixed model and contains the marker genotypes for the *j*th locus and H_{jj} is a submatrix of the variance matrix H. H is calculated from the variance component estimates obtained from fitting the linear mixed model to the phenotypic data, assuming there is no association between any of the marker loci and trait. The authors show that the joint distribution of $(\hat{\omega}_1, \hat{\omega}_2, \ldots, \hat{\omega}_m)$, under the joint null hypothesis of no marker–trait associations, is a multivariate normal distribution with vector mean zero and variance matrix with elements $M_{ij} = L_i \left(X'_i H_{ii}^{-1} X_i\right)^{-1} X'_i H_{ii}^{-1} H_{ij} H_{jj}^{-1} X_j (X'_j H_{jj}^{-1} X_j)^{-1} L'_j$ where M_{ij} is the covariance between $\hat{\omega}_i$ and $\hat{\omega}_j$. Vector samples are drawn randomly from this multivariate normal and used to calculate, empirically, a genome-wide threshold.

Despite the relative simplicity of their resampling approach, I did encounter practical issues when implementing it for the analysis of real data that were obtained from an association study in wheat (unpublished). In the study, phenotypic and genotypic data were collected on 186 different cultivars yielding 1890 data records. The genotypic data consist of genotypes recorded from 2100 single-nucleotide polymorphisms (SNPs). These SNPS had been mapped in a previous study (Huang *et al.,* 2012). As M_{ij} is calculated for all unique pairings of the marker loci, this meant $2100 \times 2101/2 = 2\,206\,050$ M_{ij} calculations are required to form the variance matrix. Also, on a locus by locus basis, Mueller removes plants whose marker genotypes are missing from the analysis. This results in the dimension of H_{ii}, H_{ij}, H_{jj}, X_i, X_j varying across a genome with the amount and pattern of missing marker data. To form the variance matrix for the joint distribution of $\hat{\omega}_j$, a large number of matrix inversions are required. This is a significant computational overhead to Mueller's resampling approach.

Yet, with a few simple changes, I found their approach to be an effective and efficient way of controlling the type 1 error rate in my genome-wide study. First, instead of calculating M_{ij} across the entire genome, I calculated M_{ij} within linkage groups. The joint distribution of ω_j becomes the product of L lower dimensioned multivariate distributions where L is the number of linkage groups.

Here, I am assuming unlinked loci are not correlated strongly which I have found to be true generally. For the wheat study that has 21 linkage groups, this reduced the number of covariance calculations from 2 206 050 to 156 538. For those situations where a marker map is not available, loci could be grouped, based on their linkage disequilibrium, with clustering techniques such as partitioning about medoids.

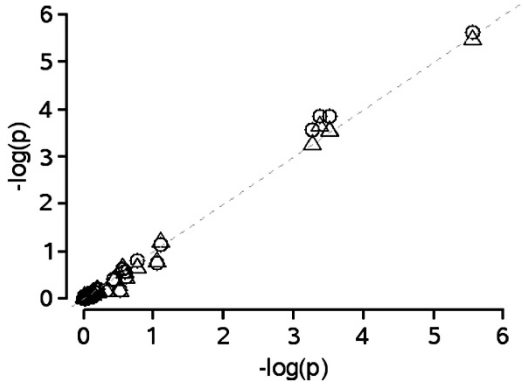

Figure 1 A comparison of three different approaches for performing genome-wide association mapping and adjusting for multiple testing. Circles (triangles) denote the –log adjusted *P*-values from a single-stage analysis with raw *P*-values adjusted with Mueller's resampling approach verse the –log adjusted *P*-values from a single-stage analysis (weighted two-stage analysis) with raw *P*-values adjusted with my modified resampling approach. Symbols on the diagonal mean the two approaches are giving equivalent adjusted *P*-values. This figure shows that the three approaches give similar results despite there being orders of magnitude differences in computing times.

Second, much of the computational burden in implementing Mueller's resampling approach stems from removing data on plants with missing marker genotypes. I avoid this by replacing missing genotypes with suitable values. Since in our study, data are recorded on SNPs on inbreds, marker loci are treated as covariates in the linear mixed models. The marker data for a locus are normalized and missing genotypes replaced with zeroes. If the marker locus was multi-allelic, then it would be treated as a categorical variable in the model. A new factor level is assigned to the missing genotypes. For our analyses, the covariance calculation simplified to $M_{ij} = L\left(X_i' H^{-1} X_i\right)^{-1} X_i' H^{-1} X_j \left(X_j' H^{-1} X_j\right)^{-1} L'$ where H^{-1} need only be calculated once and reused for all M_{ij} and $\left(X_i' H^{-1} X_i\right)^{-1}$ is calculated for $i = 1, 2, \ldots, m$. With these changes, the computing time is reduced from 133 days to 36 h. I was also able to make further computational savings by performing the linear mixed model analyses as a weighted two-stage approach (Smith *et al.*, 2001). Here, the analysis was completed in 76 min. This significant drop in computing time is due to the simple form of the second-stage model that allows rapid calculation of the M_{ij}.

Note that it is not the absolute times that are of importance here but the relative times. The absolute times are based on prototype code written in the R language. Calculations are distributed across 10 computer processors via functionality contained in the R package snow (Rossini *et al.*, 2007). Significant reductions in the absolute computing times should be able to be achieved with better optimized code. However, the relative times will change little.

To examine the impact of the above described changes, I selected, randomly, 400 SNPs. This subsetting of the available marker data was done for computational expediency. The strength of association, adjusted for multiple testing, for these SNPs were then assessed with: (1) Single-stage analyses with the raw *P*-values adjusted for multiple testing with Mueller's resampling approach, (2) Single-stage analyses with the raw *P*-values adjusted for multiple testing with my modified resampling approach and (3) weighted two-stage analyses with the raw *P*-values adjusted with my modified resampling approach. As Figure 1 shows, I found very little difference in the results for the three approaches.

Muller's resampling approach is a simple and statistically sound procedure for controlling, empirically, the type 1 error rate in genome-wide studies. With the changes suggested above to improve its implementation, it should become the approach of choice.

DATA ARCHIVING

There were no data to deposit.

CONFLICT OF INTEREST

The author declares no conflict of interest.

AW George[1,2]
[1]*Division of Mathematics, Informatics, and Statistics, CSIRO, Brisbane, Queensland, Australia and*
[2]*Food Futures National Research Flagship, CSIRO, Acton, ACT, Australia*
E-mail: andrew.george@csiro.au

Huang BE, George AW, Forrest KL, Kilian A, Hayden MJ, Morell MK *et al.* (2012). A multiparent advanced generation inter-cross population for genetic analysis in wheat. *J Plant Biotechnol* **10**: 826–839.

Mueller BU, Stich B, Piepho HP (2011). A general method for controlling the genome-wide type I error rate in linkage and association mapping experiments in plants. *Heredity* **106**: 825–831.

Rossini AJ, Tierney L, Li N (2007). Simple parallel statistical computing in R. *J Comput Graph Stat* **16**: 399–420.

Smith A, Cullis B, Gilmour A (2001). The analysis of crop variety evaluation data in Australia. *Aust NZ J Stat* **43**: 129–145.

Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF *et al.* (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**: 203–208.

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C *et al.* (2007). An Arabidopsis example of association mapping in structured samples. *Plos Genet* **3**: e4.