NIASRA
NATIONAL INSTITUTE FOR APPLIED
STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

11-16

The Analysis of QTL and QTL x Treatment Experiments using Spatial Models for Marker Effects

Brian Cullis and Alison Smith

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522, Australia Phone +61 2 4221 5435, Fax +61 2 4221 4845.
Email: karink@uow.edu.au

Statistics for the Australian Grains Industry Technical Report Series

The analysis of QTL and QTL×treatment experiments using spatial models for marker effects

Brian Cullis and Alison Smith
National Institute for Applied Statistics and Research Australia
School of Mathematics and Applied Statistics
University of Wollongong

email: bcullis@uow.edu.au

January 14, 2016

1 Introduction

The continued increase in the availability of markers has led to much interest in their use in genetic improvement programs of crop species, such as wheat and maize. There is a large amount of literature on this topic and much of the focus has turned from marker assisted selection and the identification of quantitative trait loci (QTL) to genomic selection. The basic idea in marker-assisted selection is to exploit statistical dependencies (linkage disequilibrium, LD) existing in the joint distribution of marker and QTLs. Linkage disequilibrium between markers and QTL has two main objectives, and in some way these are not disjoint and not surprisingly the statistical models used in these two applications are similar. We refer to these objectives as (i) QTL analysis in which the aim is to infer genomic locations and effects (i.e. QTLs) which affect a (quantitative) trait and (ii) genomic selection in which the aim is to obtain predictions of genetic merit of individuals for selection as parents in a breeding program. Since the seminal paper of Meuwissen et al. (2001) there has been significant progress made in the second objective where there was a realisation that unravelling the genetic architecture of a trait via identification of (major) QTLs is not necessary for prediction of genetic merit. This concept built on the idea that a trait is the result of the influences of many, possibly small QTLs which would be very difficult if not impossible, to detect and hence routinely use within a breeding program.

A plethora of statistical methods have been developed for the first objective. Some of the so-called whole genome approaches are remarkably similar to the approaches used for genomic selection. Verbyla et al. (2007) presented a whole genome average interval mapping (WGAIM) approach for QTL analysis of a single trait in a single trial. They used an approach which was embedded within the framework of ridge regression or or so-called genomic BLUP (GBLUP), in which all the intervals on a linkage map are used simultaneously avoiding, in some sense, repeated genome scans to detect QTLs. Their approach uses forward selection, commencing by fitting a model similar to ridge regression (except based on intervals rather than markers) and then choosing putative QTLs using the concept of the alternative outlier model of Thompson (1985). Once an interval (or marker) is chosen it is then fitted in the model as a fixed effect and the process repeated until all significant QTLs have been identified and included in the model as fixed effects. Their method was shown to be much more powerful than composite interval mapping although there is a small increase in selecting false positives. Their approach has been implemented in the R statistical computing environment (R Core Team, 2015) in the package **wgaim** (Taylor & Verbyla, 2011).

Recently Verbyla et al. (2012) addressed two issues of WGAIM. Firstly they improved the efficiency of the analysis when the number of markers is large. Secondly, they considered the issue of (selection) bias involved in moving the selected QTL to the fixed effects. They addressed the first issue by considering a reformulation of the ridge regression model, which was similar to the approach originally proposed by VanRaden (2008), in which they fitted a model using a genomic relationship matrix to avoid inclusion of marker effects. Verbyla et al. (2012) used a variant of this idea which also avoided fitting marker effects directly and hence was found to be computationally efficient when the number of markers (r) exceeded the number of genotypes (m). They addressed the second issue of selection bias by fitting the set of selected intervals (markers) as random in the final step. We note however that this does not really fully address the issue of selection bias.

There has been an increasing interest in the use of so-called spatial models in both QTL analysis and genomic selection. Gianola et al. (2003) considered a range of alternative models for use in marker-assisted selection, which included an extension in which the model for marker effects included both chromosomal effects and within chromosomal deviations which were correlated according to a first-order autoregressive process. They extended the first-order autoregressive model to extend the implicit assumption that the markers are equally spaced (in a genetic sense), by considering the exponential model, which is the continuous-lag extension of the first-order autoregressive model. They noted that distances between markers could be based on physical units such as kilobases. They did not apply these models to real data-sets. Yang & Tempelman (2010) considered the use of the class of ante-dependence models for genomic selection. These models were popularised by (Pourahmadi, 1999) for the analysis of longitudinal data. Their approach was framed in a Bayesian context and they concluded that, on the basis of a simulation study that the models offered a "biologically reasonable and computationally tractable method to accommodate LD", and that the antedependence based model "should lead to measurably greater gains in accuracy of whole genome selection as greater levels of LD are attained between markers with newly developed SNP marker panels".

In a related approach there has also been interest in the use of spatial (and related) models for genomic selection, but rather than extending the ridge regression model for markers, various authors have considered alternative models to the genomic relationship matrix generated via the ridge regression model for markers (see for example, VanRaden (2008)). de Los Campos et al. (2009) considered the use of reproducing kernel Hilbert spaces regression (RKHS) for genomic selection and developed an approach based on the assumption that the additive genetic signal is a arbitrary function of the set of markers. The specification of the function is based on the class of semi-parametric regression models used in the smoothing splines literature and advocated by Green & Silverman (1994). Their choice of penalty function comes from the RKHS class of models (Wahba, 1990). Specifically they suggest use of the so-called gaussian kernel, which is a one parameter covariance model allowing for flexibility in the rate of decay of covariance as the "distance" between individuals increases. They do not provide a formal approach for estimation of this rate constant.

In a similar approach, Ober (2010) suggested the application of a high-dimensional kriging-extension to genomic selection. Their model is similar to the model of de Los Campos et al. (2009) but they choose the covariance model for the genomic relationship matrix to be based on the Matérn class of covariance functions (Stein, 1999). This model allows for more flexibility in capturing the functional dependency of the covariances on the (genetic) distance of individuals based on their SNP profiles. They compared their approach to GBLUP in a small simulation study. Their results suggest that there was little to choose between the spatial approach and conventional GBLUP.

There is a clear interest in the use of spatial models in genomic selection, but as yet these models have not been applied to the analysis of real data-sets. Spatial models have been proposed to model both the elements of the covariance between individuals and the covariance of marker effects within a chromosome. In this paper we will develop a general class of spatial models for the identification of the genetic architecture of complex traits in QTL mapping experiments by exploiting the existence of high LD between markers in modern marker panels. Our approach is a natural extension of the WGAIM approach, based on markers, not intervals, which uses the Matérn class of spatial covariance models. Stein (1999) advocates the use of the Matèrn model as a model which can be broadly and effectively employed to the problem of prediction in irregularly spaced spatial data. He argues that the Matérn model has much more flexibility then other models such as the class of smoothing splines or RKHS models which lead to a serious loss in efficiency. The additional flexibility of the Matérn class comes from the inclusion of the parameter which controls the so-called "smoothness" of the (gaussian) random field and he illustrates that many other covariance models are simply specific forms of a Matérn model in which the smoothness parameter is chosen a priori. Haskard et al. (2007) and Kammann & Wand (2003) have demonstrated the utility of the Matérn class for prediction in a spatial context.

The structure of the paper is as follows.

2 Statistical model for a simple QTL mapping experiment

We commence by considering the analysis of a simple QTL mapping experiment. By this we mean an experiment with only a single treatment factor, namely the genotypes from the mapping population, including parental and check varieties. Let $\boldsymbol{y}$ denote the $n \times 1$ vector of phenotypic data, where n is the number of observations in the experiment. We can write a model for the data vector as

$$\boldsymbol{y} = \boldsymbol{X}_p\boldsymbol{\tau}_p + \boldsymbol{Z}_g^*\boldsymbol{u}_g^* + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{\tau}$ is a vector of (incidental) fixed effects with associated design matrix $\boldsymbol{X}$; $\boldsymbol{u}_g^*$ is the $(m + m_0) \times 1$ vector of random total genetic effects corresponding to all genotypes, both those with marker data (m) and those without (m_0) marker data. The latter genotypes may include both parental and check varieties but also those DH lines which were genotyped but were discarded from the marker set during construction of the linkage map on the basis of either too many cross-overs or too much missing data. We consider the partition of both $\boldsymbol{u}_g^*$ and $\boldsymbol{Z}_g^*$ which is conformal with this classification. That is,

$$\boldsymbol{u}_g^* = (\boldsymbol{u}_{g_0}^\top \ \boldsymbol{u}_g^\top)^\top \qquad \text{and} \qquad \boldsymbol{Z}_g^* = [\boldsymbol{Z}_{g_0} \ \boldsymbol{Z}_g]$$

Equation 1 can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{u}_g + \boldsymbol{e} \tag{2}$$

where $\boldsymbol{X} = [\boldsymbol{X}_p \ \boldsymbol{Z}_{g0}]$ and $\boldsymbol{\tau} = (\boldsymbol{\tau}_p^\top \ \boldsymbol{u}_{g0}^\top)^\top$. This notation and form of the model is necessary as we fit the genetic effects for those genotypes without marker data as fixed effects to exclude their influence on the genetic analysis. This model is now extended to consider the partitioning of the total genetic effects for those genotypes with marker data into additive and residual genetic effects. The extension commences with the decomposition:

$$\boldsymbol{u}_g = \boldsymbol{u}_a + \boldsymbol{u}_e \tag{3}$$

where the two terms are the additive and residual genetic effects respectively. Given $\boldsymbol{M}$, the matrix of (SNP) marker data (assumed known, without missing values and columns ordered according to linkage groups and in map order within linkage groups) of size $m \times r$, then we consider a model for the additive genetic effects given by

$$\boldsymbol{u}_a = \boldsymbol{u}_m + \boldsymbol{u}_\epsilon, \quad \text{and} \quad \boldsymbol{u}_m = \boldsymbol{M}\boldsymbol{\alpha} \tag{4}$$

where $\boldsymbol{\alpha}$ is the vector of marker regression coefficients and $\boldsymbol{u}_\epsilon$ is the vector of lack of fit additive genetic effects. For simple mapping populations such as doubled haploid populations and recombinant inbred lines, genotypes are usually derived from a bi-parental cross of in-bred lines and so the lack of fit term can be assumed to be (effectively) zero. We note that the non-imputed values in $\boldsymbol{M}$ are coded as -1 and 1 and r is much larger than m for most of our applications. Extensions to non-inbred populations is possible.

Thus the model in equation 2 can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{M}\boldsymbol{\alpha} + \boldsymbol{Z}_g\boldsymbol{u}_e + \boldsymbol{e} \tag{5}$$

This is referred to as the marker model. The genotype model is given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{u}_m + \boldsymbol{Z}_g\boldsymbol{u}_e + \boldsymbol{e} \tag{6}$$

These models can be extended to include peripheral random effects which often are associated with blocking factors arising in the analysis of designed comparative experiments. We omit this extension for pedalogical reasons.

2.1 Variance models for random effects: simple QTL experiments

We assume that the variance matrices of the residual genetic effects and the residuals are given by

$$\begin{aligned} \operatorname{var}(\boldsymbol{u}_e) &= \sigma_e^2 \boldsymbol{I}_m \\ \operatorname{var}(\boldsymbol{e}) &= \boldsymbol{R} \end{aligned}$$

where typically the matrix $\boldsymbol{R}$ is a function of an (R-level) variance parameter vector.

The variance matrix for the random marker effects is given by

$$\operatorname{var}(\boldsymbol{\alpha}) = \sigma_\alpha^2 \boldsymbol{D} = \oplus_{i=1}^{c} \boldsymbol{D}_i = \boldsymbol{G}_{\alpha\alpha}, \quad \text{say} \tag{7}$$

where c is the number of chromosomes and $\boldsymbol{D}_i$ is an $r_i \times r_i$ variance matrix of within-chromosome marker effects for chromosome i, and r_i is the number of markers in chromosome i. Note that $r = \sum_{i=1}^{c} r_i$. The block diagonality of $\boldsymbol{D}$ in equation 7 assumes that the within chromosome marker effects are independent across chromosomes. We assume further that the elements of $\boldsymbol{D}_i$ are given by

$$d_{i;j,j'} = (\phi|t_{i;j,j'}|)^{\nu}\mathcal{K}_{\nu}(\phi|t_{i;j,j'}|) \tag{8}$$

where $\mathcal{K}_{\nu}$ is a modified Bessel function, ϕ is the range parameter of the process and $|t_{i;j,j'}|$ is the absolute value of the distance between markers j and j' on chromosome i. Markers can be assumed equidistant within a chromosome, or more often, in the absence of physical distances we use the map distance in centimorgans. Underlying this variance model is the assumption that $\boldsymbol{\alpha}$ is a realisation of a Gaussian random process at specific locations (along the genome).

For a given ν, the range parameter ϕ affects the rate of decay of the correlation function with increasing $|t_{i;j,j'}|$. The parameter $\nu > 0$ controls the analytic smoothness of the underlying (genetic) process, the process being $\lceil \nu \rceil - 1$ times mean-square differentiable, where $\lceil \nu \rceil$ is the smallest integer greater than or equal to ν (Stein, 1999, page 31). Larger ν correspond to smoother processes. We note that $\nu = \frac{1}{2}$ yields the exponential correlation function,

$$d_{i;j,j'} = \exp(-\phi|t_{i;j,j'}|), \tag{9}$$

while $\nu = 1$ yields Whittles elementary correlation function, (Webster & Oliver, 2001, page 119). When ν is of the form $h + \frac{1}{2}$, with h a nonnegative integer, the correlation function in equation 8 is of the form $\exp(-\phi|t_{i;j,j'}|)$ times a polynomial in $|t_{i;j,j'}|$ of degree h. Kammann & Wand (2003) use the model where $\nu = \frac{3}{2}$, in which case

$$d_{i;j,j'} = \exp(-\phi|t_{i;j,j'}|)(1 + \phi|t_{i;j,j'}|) \tag{10}$$

and they term this model a "geo-additive" model. It has the advantage of being computationally simple to differentiate and gives rise to a process which is once differentiable.

It follows from equation 4 that the variance matrix for $\boldsymbol{u}_m$ is given by $\sigma_{\alpha}^2\boldsymbol{M}\boldsymbol{D}\boldsymbol{M}^{\top} = \sigma_{\alpha}^2\boldsymbol{K} = \boldsymbol{G}_{mm}$, say, where we call the matrix $\boldsymbol{K}$ or order $m \times m$, matern-genomic relationship. This matrix is dense, but is relatively cheap to compute given the block diagonal form for $\boldsymbol{D}$.

3 Statistical model for a factorial QTL mapping experiment

Here we extend the models for the analysis of a simple QTL mapping experiment to the analysis of a QTL mapping experiment with a factorial treatment structure. These experiments are often conducted to determine the genetic architecture of the tolerance or resistance of crops to a range of abiotic and biotic stresses. Recent examples include Linsell et al. (2014) and Genc et al. (2010), and the treatment structure of these experiments usually involves the factorial combination of the genotypes with a treatment with two levels, namely a control ('-') and a stress ('+'). The most common approach to the analysis of these experiments is to undertake a two-stage approach, forming differences or ratios of the '+' and the '-' treatment means for each genotype, thence subjecting these to a QTL analysis. This approach is piecemeal and results in a loss of information. Our approach is to extend the approach outlined in the previous section by jointly modelling the variance of the treatment × genotype effects using a particular form of the factor analytic models suggested by Smith et al. (2001) for the analysis of multi-environment trials.

Our model is of the same form as equations 5 and 6 except additional fixed effects are included in $\boldsymbol{\tau}$. These effects represent the saturated factorial structure between the factor associated with those genotypes without marker data and the treatment factor. This includes, by default the main effect of the treatment factor and hence the vector $\boldsymbol{u}_g$ represents the total genetic effect nested within treatments. For brevity we refer to the latter as the genotype by treatment total genetic effects. The vector $\boldsymbol{u}_g$ is $2m \times 1$ with the elements ordered genotypes within treatments, and hence $\boldsymbol{u}_g = (\boldsymbol{u}_{g_-}^\top \ \boldsymbol{u}_{g_+}^\top)^\top$ where the two sub-vectors are the effects for the '-' and '+' respectively. As before we consider the decomposition of $\boldsymbol{u}_g$ given by

$$\boldsymbol{u}_g = \begin{bmatrix} \boldsymbol{u}_{g_-} \\ \boldsymbol{u}_{g_+} \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_{a_-} \\ \boldsymbol{u}_{a_+} \end{bmatrix} + \begin{bmatrix} \boldsymbol{u}_{e_-} \\ \boldsymbol{u}_{e_+} \end{bmatrix}$$

and further

$$\begin{aligned} \boldsymbol{u}_a = \begin{bmatrix} \boldsymbol{u}_{a_-} \\ \boldsymbol{u}_{a_+} \end{bmatrix} &= \begin{bmatrix} \boldsymbol{M} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_- \\ \boldsymbol{\alpha}_+ \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{u}_{m_-} \\ \boldsymbol{u}_{m_+} \end{bmatrix} = \boldsymbol{u}_m \end{aligned} \tag{11}$$

3.1 Variance models for random effects: factorial QTL experiments

We model the vector of marker by treatment effects and residual genetic by treatment effects using a constrained factor analytic variance model of order 1. This model has three parameters and therefore has the same number of parameters as an unstructured variance matrix for two "traits" (i.e. treatments), but it permits a natural and biologically meaningful interpretation. It has the added advantage of dealing with non-positive definite variance matrices. The so-called extended factor analytic models were introduced by Thompson et al. (2003) as a computationally efficient alternative to the approach presented by Smith et al. (2001). The model we consider here is a sub-class of these models, in which one of the specific variances is set to zero. The regression form of the model for marker by treatment effects is given by

$$\begin{aligned} \boldsymbol{\alpha} &= \begin{bmatrix} \lambda_{\alpha_-} \boldsymbol{f}_\alpha \\ \lambda_{\alpha_+} \boldsymbol{f}_\alpha \end{bmatrix} + \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{\delta}_\alpha \end{bmatrix} \\ &= (\boldsymbol{\lambda}_\alpha \otimes \boldsymbol{I}_r) \boldsymbol{f}_\alpha + \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{\delta}_\alpha \end{bmatrix} \end{aligned}$$

where $\boldsymbol{\lambda}_\alpha = (\lambda_{\alpha-}\ \lambda_{\alpha+})^\top$ and we assume that

$$\text{var}\begin{pmatrix} \boldsymbol{f}_\alpha \\ \boldsymbol{\delta}_\alpha \end{pmatrix} = \begin{bmatrix} \boldsymbol{G}_{f_\alpha f_\alpha} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{G}_{\delta_\alpha \delta_\alpha} \end{bmatrix}$$

where $\boldsymbol{G}_{f_\alpha f_\alpha} = \boldsymbol{D}$ and $\boldsymbol{G}_{\delta_\alpha \delta_\alpha} = \psi_\alpha \boldsymbol{D}$. Hence it follows that

$$\text{var}\,(\boldsymbol{\alpha}) = \boldsymbol{G}_{T_\alpha} \otimes \boldsymbol{D} = \boldsymbol{G}_{\alpha\alpha} \tag{12}$$

where

$$\boldsymbol{G}_{T_\alpha} = \boldsymbol{\lambda}_\alpha \boldsymbol{\lambda}_\alpha^\top + \text{diag}\,(0, \psi_\alpha)$$

This regression representation of the model admits a simple interpretation. The vector $\boldsymbol{f}_\alpha$ represents the pleiotropic marker effects across treatments. The effects for the `'-'` and `'+'` treatments are scaled by $\lambda_{\alpha-}$ and $\lambda_{\alpha+}$ respectively. The vector $\boldsymbol{\delta}_\alpha$ represents the deviations from the additive genetic regression of $\boldsymbol{\alpha}_+$ on $\boldsymbol{\alpha}_-$. Hence this term captures the non-pleiotropic effects associated with the `'+'` treatment which are independent of the `'-'` treatment. Specifically this additive genetic regression is given by

$$\boldsymbol{\alpha}_+ = \frac{\lambda_{\alpha_+}}{\lambda_{\alpha_-}} \boldsymbol{\alpha}_- + \boldsymbol{\delta}_\alpha \tag{13}$$

Consistent with how this model is fitted, for either $\boldsymbol{\alpha}$ (or $\boldsymbol{u}_m$ or $\boldsymbol{u}_e$), we refer to it as a `rr(trt) + diag(+)` model which is an abbreviation for a reduced rank model of order one, which is a factor model of order 1 for two traits where both specific variances are set to zero, plus a scaled identity or default variance matrix for the marker effects for the stress or '+' treatment. This parameterisation is computationally efficient, numerically stable and defaults to a positive semi-definite variance matrix (i.e a 2 × 2 matrix of order one) when $\psi_\alpha = 0$. When $\psi_\alpha = 0$ then $\boldsymbol{u}_{m-}$ and $\boldsymbol{u}_{m+}$ are perfectly correlated.

It is straightforward to extend the marker regression model to the genotype regression model

$$\boldsymbol{u}_m = (\boldsymbol{\lambda}_\alpha \otimes \boldsymbol{I}_m)\boldsymbol{f}_m + \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\delta}_m \end{bmatrix} \tag{14}$$

where

$$\boldsymbol{f}_m = \boldsymbol{M}\boldsymbol{f}_\alpha \qquad \text{and} \qquad \boldsymbol{\delta}_m = \boldsymbol{M}\boldsymbol{\delta}_\alpha \tag{15}$$

Hence

$$\begin{aligned} \text{var}\,(\boldsymbol{f}_m) &= \boldsymbol{K} \\ \text{var}\,(\boldsymbol{\delta}_m) &= \psi_\alpha \boldsymbol{K} \\ \text{var}\,(\boldsymbol{u}_m) &= \boldsymbol{G}_{T_\alpha} \otimes \boldsymbol{K} \\ &= \boldsymbol{G}_{mm}, \quad \text{say} \end{aligned}$$

and this is termed the `rr(Trt) + diag(+)` variance model for (additive) genetic effects (i.e. $\boldsymbol{u}_m$).

Similarly applying the same model to the residual genetic by treatments effects, $\boldsymbol{u}_e$, the residual genetic regression model is

$$\boldsymbol{u}_e = (\boldsymbol{\lambda}_e \otimes \boldsymbol{I}_m)\boldsymbol{f}_e + \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{\delta}_e \end{bmatrix}$$

where

$$\begin{aligned} \text{var}\,(\boldsymbol{f}_e) &= \boldsymbol{I}_m \\ \text{var}\,(\boldsymbol{\delta}_e) &= \psi_e \\ \text{var}\,(\boldsymbol{u}_e) &= \boldsymbol{G}_{T_e} \otimes \boldsymbol{I}_m \\ &= \boldsymbol{G}_{ee}, \quad \text{say} \end{aligned}$$

where

$$\boldsymbol{G}_{T_e} = \boldsymbol{\lambda}_e \boldsymbol{\lambda}_e^\top + \text{diag}\,(0, \psi_e)$$

and this is termed the `rr(Trt) + diag(+)` variance model for the (residual) genetic effects (i.e. $\boldsymbol{u}_e$).

To complete this section, we now form the full model based on the two genetic regression models for $\boldsymbol{u}_m$ and $\boldsymbol{u}_{\boldsymbol{e}}$. This formulation is the most computationally efficient and numerically stable (Thompson et al., 2003). Substitution of equations 14 and 15 for $\boldsymbol{u}_m$ and $\boldsymbol{u}_{\boldsymbol{e}}$ respectively into equation 6 gives

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X\tau} + \boldsymbol{Z}_{f_\alpha}\boldsymbol{f}_m + \boldsymbol{Z}_{g+}\boldsymbol{\delta}_m + \boldsymbol{Z}_{f_e}\boldsymbol{f}_e + \boldsymbol{Z}_{g+}\boldsymbol{\delta}_e + \boldsymbol{e} \\ &= \boldsymbol{W\beta} + \boldsymbol{e} \end{aligned} \tag{16}$$

where $\boldsymbol{Z}_g = [\boldsymbol{Z}_{g-} \;\; \boldsymbol{Z}_{g+}]$, $\boldsymbol{\beta} = (\boldsymbol{\tau}^\top, \boldsymbol{f}_m^\top, \boldsymbol{\delta}_m^\top, \boldsymbol{f}_e^\top, \boldsymbol{\delta}_e^\top)^\top$, $\boldsymbol{W} = [\boldsymbol{X} \;\; \boldsymbol{Z}_{f_\alpha} \;\; \boldsymbol{Z}_{g+} \;\; \boldsymbol{Z}_{f_e} \;\; \boldsymbol{Z}_{g+}]$, $\boldsymbol{Z}_{f_\alpha} = \boldsymbol{Z}_g(\boldsymbol{\lambda}_\alpha \otimes \boldsymbol{I}_m)$, $\boldsymbol{Z}_{f_e} = \boldsymbol{Z}_g(\boldsymbol{\lambda}_e \otimes \boldsymbol{I}_m)$ and

$$\text{var}\begin{pmatrix} \boldsymbol{f}_m \\ \boldsymbol{\delta}_m \\ \boldsymbol{f}_e \\ \boldsymbol{\delta}_e \end{pmatrix} = \text{diag}\,(\boldsymbol{K}, \psi_\alpha \boldsymbol{K}, \boldsymbol{I}_m, \psi_e \boldsymbol{I}_m) = \boldsymbol{G}_{rr}$$

This model is similar to the model in equation (4) of Thompson et al. (2003).

4 Estimation and Prediction

Since r is usually large and much greater than m, we prefer to fit the genotype joint regression model and thence obtain predictions and (model-based) prediction error variances of the set of genetic effects in the marker joint regression model from those obtained in fitting the genotype joint regression model as a post processing step. Details are only provided for the factorial QTL experiment, as the results for the simple QTL experiment can be inferred from these.

The first step in fitting the genotype joint regression model is the estimation of variance parameters, the most common method being Residual Maximum Likelihood (REML, Patterson & Thompson (1971)). This usually involves an iterative process. In this paper we use the Average Information (AI) algorithm (Gilmour et al., 1995) as implemented in the R package **ASReml-R** (Butler et al., 2009). Haskard (2005) presents details for the fitting the Matérn model and this has been implemented in **ASReml-R**. This required evaluation of the derivatives of the residual likelihood with respect to ϕ and ν. The differential for ϕ is relatively straightforward to compute analytically, though more complex for ν, unless ν is of the form $h + \frac{1}{2}$, with h a nonnegative integer. Following Haskard (2005) we avoid occasional numerical problems, and use numerical methods to obtain the differentials with respect to ν. This issue is mostly avoided through our modelling strategy (see section 6). We have found that, given sensible starting values and sequential model building, the AI algorithm to obtain REML estimates of the variance parameters performs well for most cases, again this is aided by our approach to modelling.

However, one potential obstacle to routine use of REML is the burden in computing the likelihood, score and AI matrix for large numbers of markers. Unlike general spatial problems, however, we can exploit the block diagonality of $\boldsymbol{D}$, significantly reducing the computational load to the inversion of $r_i \times r_i$ matrices.

Given estimates of the variance parameters we obtain Empirical Best Linear Unbiassed Estimates (E-BLUEs) of the fixed effects and Empirical Best Linear Unbiassed Predictions (E-BLUPs) of the random effects, E- denoting that we replace all variance parameters with their REML estimates. In particular, our interest centres on the E-BLUPs of the genetic for both genotypes and markers. The E-BLUPs for the terms in the genotype regressions can be obtained from the solutions to the mixed model equations. The mixed model equations for the genotype joint regression model given by equation 16 are

$$\boldsymbol{C}\tilde{\boldsymbol{\beta}} = \boldsymbol{W}^\top \boldsymbol{R}^{-1}\boldsymbol{y}$$

where $\tilde{\boldsymbol{\beta}} = (\hat{\boldsymbol{\tau}}^\top, \tilde{\boldsymbol{f}}_m^\top, \tilde{\boldsymbol{\delta}}_m^\top, \tilde{\boldsymbol{f}}_e^\top, \tilde{\boldsymbol{\delta}}_e^\top)^\top$, $\boldsymbol{C} = \boldsymbol{W}^\top \boldsymbol{R}^{-1}\boldsymbol{W} + \boldsymbol{G}^*$ and

$$\boldsymbol{G}^* = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}_{rr}^{-1} \end{bmatrix}$$

The (model-based) prediction error variances of the $\tilde{\boldsymbol{f}}_m$ and $\tilde{\boldsymbol{\delta}}_m$ are given by pev $\left(\tilde{\boldsymbol{f}}_m\right)$ and pev $\left(\tilde{\boldsymbol{\delta}}_m\right)$, where pev () is a matrix function which extracts the block diagonal matrix of $\boldsymbol{C}^{-1}$ relating to its vector argument. It is then straightforward to obtain the E-BLUP of $\tilde{\boldsymbol{u}}_m$ and its associated prediction error variance matrix from these quantities (including the prediction error covariance between $\tilde{\boldsymbol{f}}_m$ and $\tilde{\boldsymbol{\delta}}_m$). This is achieved in **ASReml-R** as a post-processing procedure using the `predict.asreml` method. This implements the strategies outlined in Gilmour et al. (2004).

We now present a summary of the results required to obtain the E-BLUPS and prediction error variances of the effects of interest in the marker joint regression model as functions of the effects of interest in the genotype joint regression model. A sketch of the proof of these results is given in the appendix. In the following if we define

$$\boldsymbol{D}_{mm\cdot} = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{M}^\top \boldsymbol{K}^{-1}\boldsymbol{M}\boldsymbol{D}$$

then

$$\begin{aligned}
\tilde{\boldsymbol{f}}_\alpha &= \boldsymbol{D}\boldsymbol{M}^\top \boldsymbol{K}^{-1}\tilde{\boldsymbol{f}}_m \\
\text{pev}\left(\tilde{\boldsymbol{f}}_\alpha\right) &= \boldsymbol{D}_{mm\cdot} + \boldsymbol{D}\boldsymbol{M}^\top \boldsymbol{K}^{-1}\text{pev}\left(\tilde{\boldsymbol{f}}_m\right)\boldsymbol{K}^{-1}\boldsymbol{M}\boldsymbol{D} \\
\tilde{\boldsymbol{\delta}}_\alpha &= \boldsymbol{D}\boldsymbol{M}^\top \boldsymbol{K}^{-1}\tilde{\boldsymbol{\delta}}_m \\
\text{pev}\left(\tilde{\boldsymbol{\delta}}_\alpha\right) &= \psi_\alpha \boldsymbol{D}_{mm\cdot} + \boldsymbol{D}\boldsymbol{M}^\top \boldsymbol{K}^{-1}\text{pev}\left(\tilde{\boldsymbol{\delta}}_m\right)\boldsymbol{K}^{-1}\boldsymbol{M}\boldsymbol{D} \\
\tilde{\boldsymbol{\alpha}}_- &= \boldsymbol{D}\boldsymbol{M}^\top \boldsymbol{K}^{-1}\tilde{\boldsymbol{u}}_{m_-} \\
\text{pev}\left(\tilde{\boldsymbol{\alpha}}_-\right) &= \lambda^2_{\alpha_-}\boldsymbol{D}_{mm\cdot} + \boldsymbol{D}\boldsymbol{M}^\top \boldsymbol{K}^{-1}\text{pev}\left(\tilde{\boldsymbol{u}}_{m_-}\right)\boldsymbol{K}^{-1}\boldsymbol{M}\boldsymbol{D} \\
\tilde{\boldsymbol{\alpha}}_+ &= \boldsymbol{D}\boldsymbol{M}^\top \boldsymbol{K}^{-1}\tilde{\boldsymbol{u}}_{m_+} \\
\text{pev}\left(\tilde{\boldsymbol{\alpha}}_+\right) &= (\lambda^2_{\alpha_+} + \psi_\alpha)\boldsymbol{D}_{mm\cdot} + \boldsymbol{D}\boldsymbol{M}^\top \boldsymbol{K}^{-1}\text{pev}\left(\tilde{\boldsymbol{u}}_{m_+}\right)\boldsymbol{K}^{-1}\boldsymbol{M}\boldsymbol{D}
\end{aligned}$$

5 Inference and examination of genomic marker profiles

Need to add in the chromosome score tests and then the way to look at the turning points within those chromosomes selected from scores.

6 Example

6.1 Phenotypic data and experimental design

Borg et al. (2015) describe the analysis of a QTL factorial experiment on wheat. We present a brief overview here. The aim of the experiment was to investigate whether phenotyping for osmotic stress tolerance could be a valid surrogate screening method for drought tolerance in wheat. A total of 168 genotypes (166 doubled haploid (DH) lines and the two parents; Cranbrook and Halberd) were grown in two successive runs within a glasshouse. Full details concerning the DH population can be found in Kammholz et al. (2001). The glasshouse contained four hydroponic tubs which were placed on benches. Each hydroponic tub could hold a maximum of 144 pots arranged in a 16 row by 9 column rectangular array. The key practical constraint was the assignment of genotypes to maturity groups, which were placed to be located within in so-called maturity blocks (`Matblks`) and these maturity blocks were systematically located across the tubs (and benches). Each genotype was classified into one of five maturity groups; very quick, quick, moderate, slow and very slow. This classification was based on days until flowering observed in field trials conducted at Yanco and Narrabri experimental stations in New South Wales. Whenever maturity blocks were adjacent within a tub at least one row of plots was left vacant between them. The genotypes were allocated (or randomised) to a set of three row-adjacent pots within their respective maturity block. The osmotic treatment was applied to two randomly chosen pots in each set and the remaining pot was the control. Tillers at the right stage within the pots assigned to the '+' treatment were tagged and these pots were then placed into tanks located in the same glasshouse which contained varying concentrations of NaCl^-. The pots were exposed to increasing concentrations over a 6 day period after which time they were returned to their original position within the maturity group. A sample of spikes from the control pots were also tagged for identification and measurement at the completion of the experiment.

Spike grain number (SGN) was recorded on tagged tillers from each pot. The number of tillers tagged varied per pot depending on the number of tillers at the appropriate developmental stage during tagging. At the time of measurement, the identity of the pot within each set of three pots was not recorded. The tiller was identified (non-uniquely) by the genotype, the treatment and the run. There was only one set of three pots for each genotype per run.

The experiment design is non-standard and is complicated by the following issues. Firstly, the same randomisation of genotypes to sets of three row-adjacent pots was used for each run. Secondly, the maturity blocks were located in the same position for each run and lastly, we do not know the allocation of treatments to pots within each set of three row-adjacent pots. The design is strictly unreplicated for genotypes as the experimental unit (EU) for genotypes is the set of three row-adjacent pots, which is invariant across runs and within maturity blocks. The EU for the treatment by genotype combination is a pot within a set of three row-adjacent pots, however we do not know if the allocation of treatments changed between runs and we cannot identify the allocation of treatments to pots. Fortuitously as a result of the genetic analysis, genetic clones were identified among the DH lines. (A genetic clone is defined to be a pair (or in some cases up to 4) of DH lines which have identical matching of alleles for greater than 99.5%. This meant that there was (minimal) p-artial replication of the genotypes (Cullis et al., 2006) for each run.

6.2 Genotypic data and map construction

A total of 165 lines were genotyped using a 90K SNP chip containing gene-associated SNPs that provided dense coverage of the wheat genome (Wang et al., 2014). These markers were combined with three phenological markers which were available for this population. There were a total of 16231 markers available for linkage map construction. The consensus map of Wang et al. (2014) was used as a reference map during construction. We used the R package **ASMap** (Taylor & Butler, 2014) for map construction.

The final linkage map had a total of 15601 which was thinned to 1383 non-redundant markers for the QTL analysis. The number of markers per chromosome ranged from 14 to 111 with a median number of 74. The overall length of the map (using the Kosambi distance measure) was 3867, individual lengths ranging from 95 to 234. There were a total of 9 genotypes excluded from the QTL analysis, which included the two parents. A total of 143 genotypes were included and some of these genotypes being groups of genetic clones.

6.3 Linear mixed model for the base-line model

The complex and multi-strata nature of the experimental design, necessitated use of a base-line mixed model which contained many non-genetic terms. See Borg et al. (2015) for details of the derivation of this base-line model. Using the extended model formulae syntax of Butler et al. (2009) after Wilkinson & Rogers (1973), this model is given by

```
fixed    = ~ Run*Gdrop*Trt
random   = ~ rr(Trt):Gkeep + at(Trt,‘+’):Gkeep + Matblk + Tub + Mplot +
             Run:Matblk + Run:Tub + Gkeep:Run + Gkeep:Trt:Run +
             Run:Mplot
residual = ~ idv(units)
```

The terms of interest are `rr(Trt):Gkeep + at(Trt,‘+’):Gkeep`, which represent the total genetic effects for the genotype × treatment effects. The variance matrix for these effects is the model described in section 3.1, as the `rr(trt) + diag(+)` model. We exclude the additive genetic effects from this model, as this model serves as the reference model for the subsequent QTL analysis. The set of models presented in section 6.4 include all of the terms which were fitted in the base-line model. A shifted power transformation was used to improve the normality of the SGN data.

6.4 Model fitting

Table 1 presents a summary of the four models fitted to the SGN data for the CxH factorial QTL experiment. The first model we fit is the base-line model which does not consider the decomposition of the total genetic effects into additive and residual genetic effects. The model denoted 'QTL-ide' includes a pair of terms to model the additive genotype × treatment effects. This model assumes that the marker effects are independent, that is, the standard WGAIM or GBLUP variance model. The REMLLRT for testing H_0 that $\boldsymbol{\lambda}_a = \mathbf{0}$ and $\psi_a = 0$ using this model, was 21.73, which, when compared to the reference distribution of a mixture of χ^2 variates (Stram & Lee, 1994) gave a p-value <0.001. Thus there is strong evidence for additive genetic variance due to the markers in the genotype × treatment effects.

The next model we fit is the correlated marker (effects) QTL model, which considers the vector $\boldsymbol{\alpha}$ as a realisation of a Gaussian process with correlation function given by equation 9. This model has an additional two variance parameters. Our approach is to examine the fit of the Matérn model for a range of values of ν. It is likely that direct REML estimation of ν will be problematic (see for example Stein (1999) and Haskard (2005)). We chose $\nu = 0.5$ and $\nu = 1.5$, the former being the exponential form for the correlation function and the latter is the geo-additive model used by Kammann & Wand (2003). The Matérn model with $\nu = 1.5$ resulted in a marginally better fit and hence we choose $\nu = 1.5$. This correlated marker effects model results in a marker additive genetic process being once differentiable, which has practical and biological advantages which we will exploit in the identification of the putative QTLs.

Table 1 indicates that this model gave only a modest improvement in fit compared to the fit from the QTL-ide model. This is not surprising, though given the substantial amount of non-genetic variation in the data, and the moderate sample size. The mean accuracies of the key genetic effects from the base-line model demonstrates this point. The mean accuracies of the E-BLUPs of the total genetic effects for the '-' and '+' genotype × treatment effects were 0.439 and 0.536 respectively. The mean accuracy for the E-BLUPs of the deviations from the (total) genetic regression of $\boldsymbol{u}_{e_+}$ on $\boldsymbol{u}_{e_-}$ (i.e. of $\tilde{\boldsymbol{\delta}}_e$ was only 0.163.

Table 1: Summary of residual maximum likelihood (REML) and REML estimates of the key variance parameters for the four models fitted to the CxH factorial QTL experiment. Note that ϕ and ψ_α are fixed at zero for models QTL-ide and QTL-MATf respectively.

	logl	$\hat{\lambda}_{\alpha_-}$	$\hat{\lambda}_{\alpha_+}$	$\hat{\psi}_\alpha$	$\hat{\phi}$	$\hat{\lambda}_{e_-}$	$\hat{\lambda}_{e_+}$	$\hat{\psi}_e$
Base	-2282.25					0.542	1.198	0.341
QTL-ide	-2271.38	0.805	0.703	0.510	0.000	0.231	1.004	0.000
QTL-MAT	-2269.73	0.777	0.637	0.449	0.055	0.287	1.089	0.000
QTL-MATf	-2272.33	0.783	0.633	0.000	0.035	0.251	1.251	0.000

The REMLLRT for testing H$_0$ that $\phi = 0$ (as ν is redundant when $\phi = 0$) was 3.299. An approximate p-value for this statistic is 0.0693. There is only mild evidence to reject H$_0$ using this test.

The final model we fit is the model which constrains the variance for the residuals from the (additive) genetic regression of $\boldsymbol{\alpha}_+$ on $\boldsymbol{\alpha}_-$ (or equivalently $\boldsymbol{u}_{m_+}$ on $\boldsymbol{u}_{m_-}$) to zero. This model is referred to as model 'QTL-MATf' model in table 1. The REMLLRT for testing H$_0$ that $\psi = 0$ was 5.203 with an approximate p-value <0.001. Hence we strongly reject H$_0$, concluding that there is significant deviation from the additive genetic regression of $\boldsymbol{\alpha}_+$ on $\boldsymbol{\alpha}_-$, supporting further investigation of the additive genetic (marker) profiles to identify genomic regions of interest.

Another interesting feature of this modelling approach can be seen by closer examination of the REML estimates of the variance parameters associated with the residual genetic effects. The REML estimates of $\lambda^2_{e_-}$ were 0.293 and 0.082 for the Base and QTL-MAT models respectively. On the other hand the REML estimates of ψ_{e_-} were 0.341 and 0 for the Base and QTL-MAT models respectively. That is, inclusion of the additive effects in the model(s) resulted in some reduction in the residual (polygenic and pleiotropic) genetic variance for the overall genotype effects, but a substantial reduction in the residual (polygenic and specific) genetic variance for the deviations from the residual genetic regression by inclusion of the markers.

Table 2 presents a summary of the model based accuracies for the E-BLUPs of genetic effects for the QTL-ide and QTL-MAT models respectively. There is a large increase in accuracy for the additive genetic effects associated with the most important trait, namely $\boldsymbol{\delta}_{\boldsymbol{\alpha}}$.

Table 2: Summary of model based accuracies for the E-BLUPs of key genetic effects for two models fitted to the CxH factorial QTL experiment

	QTL-ide	QTL-MAT
$\mathrm{acc}(\tilde{\boldsymbol{\delta}}_m)$	0.384	0.494
$\mathrm{acc}(\tilde{\boldsymbol{u}}_{m_-})$	0.495	0.609
$\mathrm{acc}(\tilde{\boldsymbol{u}}_{m_+})$	0.501	0.593
$\mathrm{acc}(\tilde{\boldsymbol{\delta}}_a)$	0.016	0.264
$\mathrm{acc}(\tilde{\boldsymbol{\alpha}}_-)$	0.023	0.345
$\mathrm{acc}(\tilde{\boldsymbol{\alpha}}_+)$	0.023	0.333

6.5 Identification of putative QTLs

A key step we consider before proceeding with formal identification of putative QTLs, based on the fit of the QTL-MAT model, is to examine the set of additive genetic effects in more detail across the genome. We propose a graphical exploration of the genome-wide additive genetic profiles of the effects associated with the additive genetic regression for both the marker × treatment and genotype × treatment effects. This is analogous to the genome-wide scan conducted in CIM and we have found that it can provide useful informal information in determining regions of interest for a particular trait.

Another useful graphical tool is to plot the E-BLUPs of $\boldsymbol{u}_{m_+}$ against the E-BLUPs of $\boldsymbol{u}_{m_-}$. This plot allows identification of genotypes which may provide information to support the genomic regions identified in the previous plot, and those identified using the formal approaches we present later in this section.

Figure 1 presents the genome-wide trellis plot of the E-BLUPs of $\boldsymbol{\alpha}_-$ and $\boldsymbol{\delta}_\alpha$ for each of the 21 chromosomes from the QTL-MAT model. This plot provides a useful overview of the contrasting genetic profiles for the pleiotropic and specific effects associated with both treatments and the deviations from the genetic regression respectively. Our focus is on the latter. There appears to be little evidence of putative QTLs on the D genome, while there appears to be regions of interest on chromosomes 1A, 3B and 5A. The trough in the profile on 5A is close to a peak for $\boldsymbol{\alpha}_-$. There appears to be some separation between these, although this warrants further formal examination.

Figure 2 presents the genome-wide trellis plot of the E-BLUPs of $\boldsymbol{\alpha}_-$ and $\boldsymbol{\delta}_\alpha$ for each of the 21 chromosomes from the QTL-ide model. There is good agreement between this plot and the plot for the QTL-MAT model apart from the (assumed) smoothness of the process for the QTL-MAT model.

Figure 3 presents the scatter plot of the E-BLUPs of $\boldsymbol{u}_{m_+}$ against the E-BLUPs of $\boldsymbol{u}_{m_-}$. We refer to these E-BLUPs as mGBLUPs, as they are analogous to the GBLUPs obtained from the QTL-ide model. The genetic regression accounts for 47.5 of the total variation in $\boldsymbol{u}_{m_+}$. Four DH lines are labelled on the plot. These correspond to DH lines which exhibit large (absolute) deviations from the genetic regression.

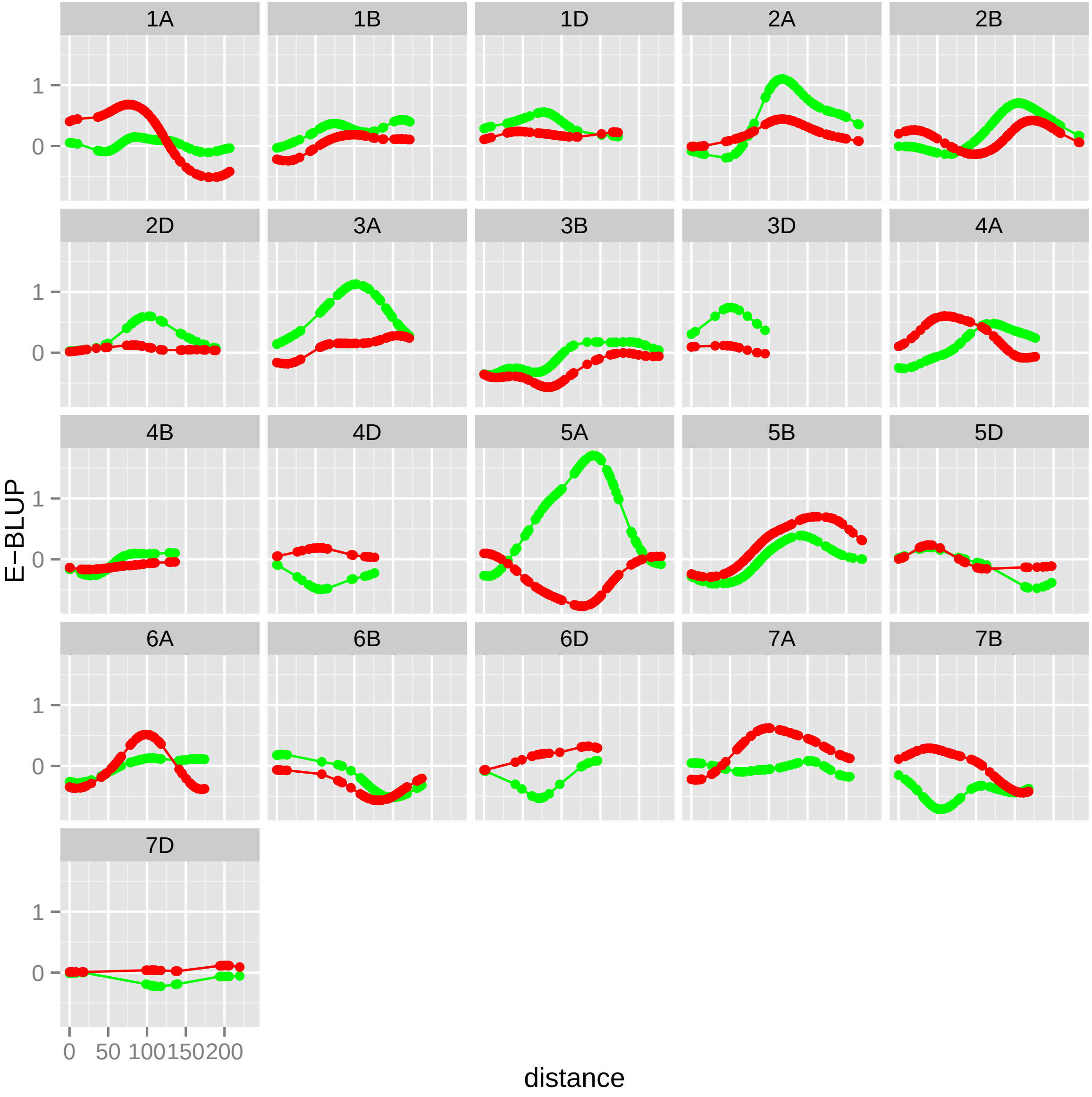

Figure 1: Genome-wide trellis plot of the E-BLUPs of $\boldsymbol{\alpha}_{-}$ (green) and $\boldsymbol{\delta}_{\alpha}$ (red) for each of the 21 chromosomes from the QTL-MAT model.

Table 3 presents the chromosome contributions to the mGBLUP for the four DH lines which were labelled in figure 3, while figure 4 presents the marker profiles for these DH lines in a trellis plot across the 21 chromosomes as ancillary information to assist in the interpretation of these contributions. Although there is some agreement in the relative magnitude of the absolute contributions for each of the four lines across the 21 chromosomes, it is clear that there may be (minor) QTLs present on chromosomes other than 1A, 3B and 5A.

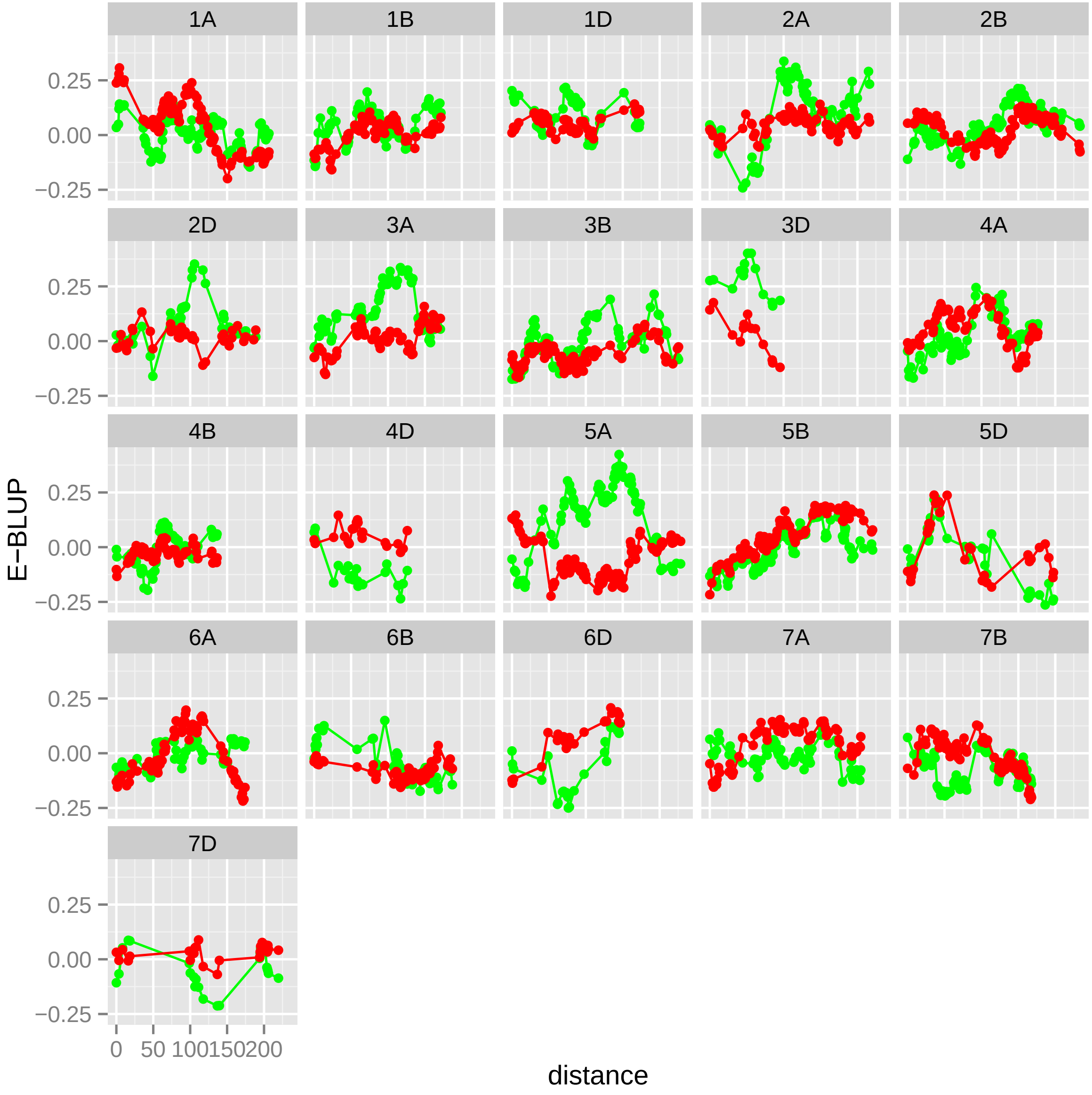

Figure 2: Genome-wide trellis plot of the E-BLUPs of $\boldsymbol{\alpha}_{-}$ (green) and $\boldsymbol{\delta}_{\alpha}$ (red) for each of the 21 chromosomes from the QTL-ide model.

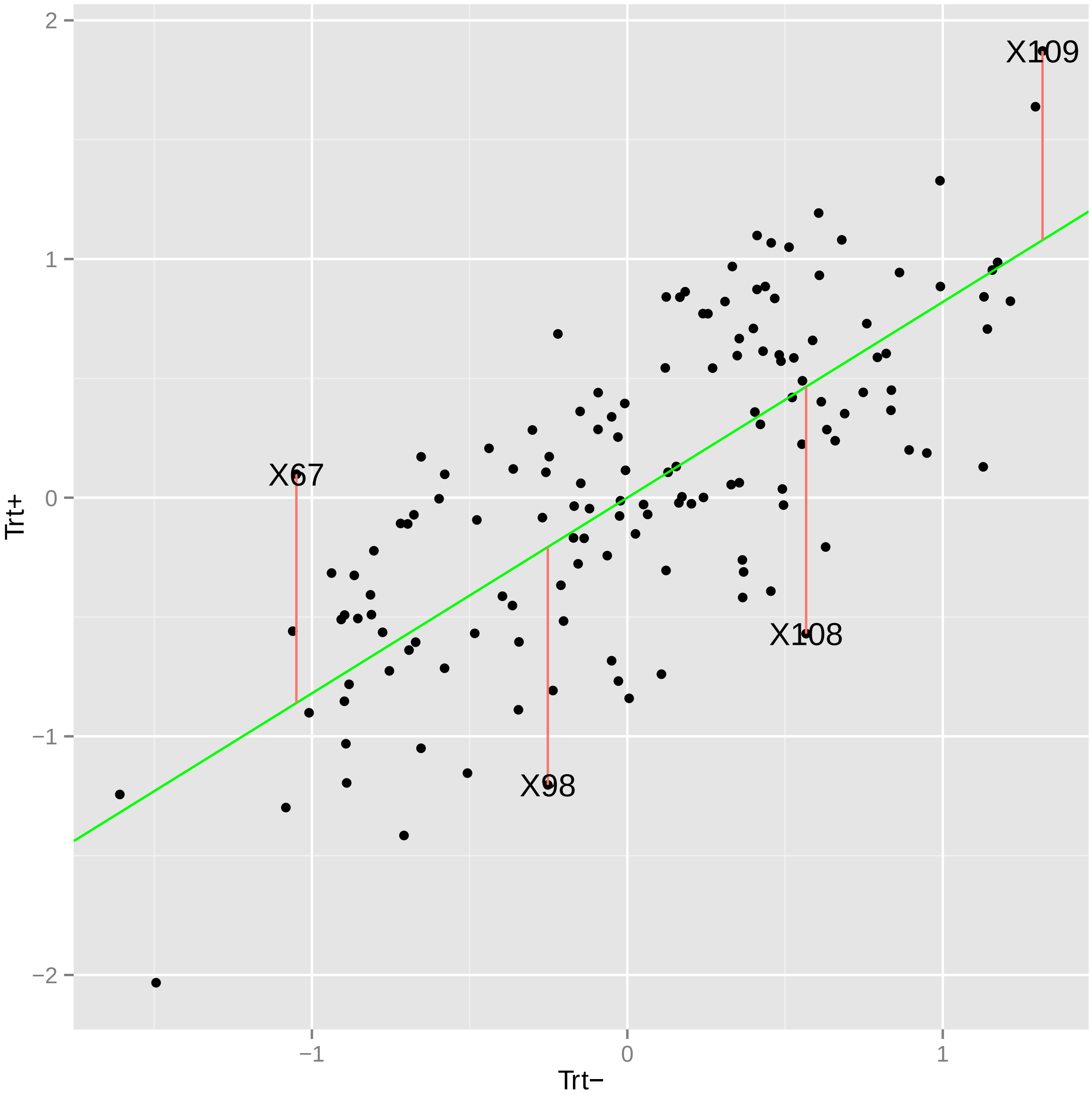

Figure 3: Scatter plot of the E-BLUPs of $\boldsymbol{u}_{m_+}$ against the E-BLUPs of $\boldsymbol{u}_{m_-}$ for the QTL-MAT model. The four points which are labelled correspond to DH lines which have large deviations from the genetic regression.

Table 3: Chromosome contributions to the overall mGBLUP for four selected DH lines.

	X67	X109	X108	X98
1A	0.186	0.066	-0.250	-0.134
1B	0.057	0.024	-0.048	0.026
1D	-0.047	0.043	-0.018	-0.059
2A	-0.092	0.112	0.054	-0.105
2B	0.069	0.063	-0.153	0.124
2D	0.017	0.018	0.010	-0.012
3A	-0.046	-0.047	0.053	-0.006
3B	0.188	0.162	-0.184	-0.213
3D	-0.004	0.007	0.007	-0.004
4A	0.125	-0.070	-0.141	-0.155
4B	0.040	0.043	0.043	-0.043
4D	-0.016	-0.013	-0.014	0.013
5A	0.222	-0.069	-0.241	-0.075
5B	0.157	0.102	0.139	0.001
5D	0.010	-0.001	0.023	0.013
6A	0.069	0.005	0.056	-0.143
6B	-0.137	0.135	-0.056	-0.134
6D	0.031	-0.031	-0.001	0.003
7A	0.139	0.120	-0.188	-0.034
7B	-0.017	0.132	-0.132	-0.069
7D	0.009	-0.009	0.006	0.009
Total	0.959	0.792	-1.035	-0.997

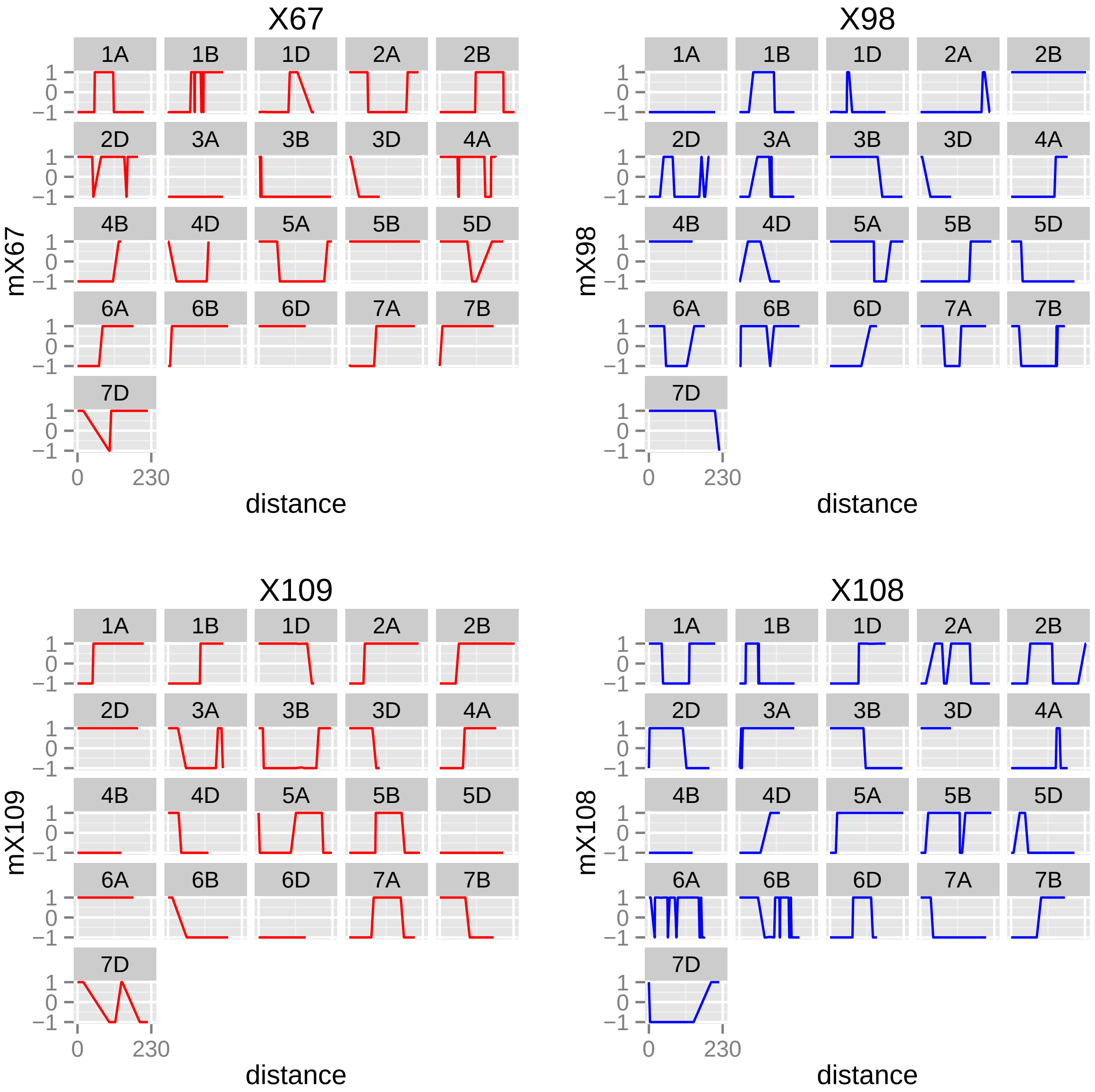

Figure 4: Trellis plot of the marker profiles across the 21 chromosomes for the four selected DH lines which have large deviations from the genetic regression.

APPENDIX

Derivations for the genotype to marker model

To simplify the following we consider the genotype model rather than the genotype joint regression model. The latter model involves a total of four random terms associated with the additive and residual genetic reduced rank, plus diagonal '+' treatment respectively. The former has two components, one associated with the additive and residual genetic terms respectively. Our derivation is based on the joint-likelihood approach used by Henderson (1975) in obtaining the mixed model equations.

The log joint-density for the genotype model is given by

$$\ell(\boldsymbol{y}, \boldsymbol{u}_m, \boldsymbol{u}_e, \boldsymbol{\alpha}) = \ell(\boldsymbol{y} \mid \boldsymbol{u}_m, \boldsymbol{u}_e) + \ell(\boldsymbol{u}_m \mid \boldsymbol{\alpha}) + \ell(\boldsymbol{u}_e) + \ell(\boldsymbol{\alpha})$$

Since $\boldsymbol{u}_m = \boldsymbol{M}_T\boldsymbol{\alpha}$ where $\boldsymbol{M}_T = \boldsymbol{I}_2 \otimes \boldsymbol{M}$ then $\ell(\boldsymbol{u}_m \mid \boldsymbol{\alpha})$ is a constant and so

$$\ell(\boldsymbol{y}, \boldsymbol{u}_m, \boldsymbol{u}_e, \boldsymbol{\alpha}) = \ell(\boldsymbol{y} \mid \boldsymbol{u}_m, \boldsymbol{u}_e) + \ell(\boldsymbol{u}_e) + \ell(\boldsymbol{\alpha}) \tag{17}$$

To maximise (17) with respect to $(\boldsymbol{u}_m^\top, \boldsymbol{\alpha}^\top, \boldsymbol{\tau}^\top, \boldsymbol{u}_e^\top)^\top$ subject to the constraint $\boldsymbol{u}_m = \boldsymbol{M}_T\boldsymbol{\alpha}$ we introduce a vector of Lagrangian multipliers, which yield the quantity, ignoring constants,

$$Q = -\frac{1}{2}[\boldsymbol{e}^\top\boldsymbol{R}^{-1}\boldsymbol{e} + \boldsymbol{u}_e^\top\boldsymbol{G}_{ee}^{-1}\boldsymbol{u}_e + \boldsymbol{\alpha}^\top\boldsymbol{G}_{\alpha\alpha}^{-1}\boldsymbol{\alpha}] - \boldsymbol{a}^\top(\boldsymbol{u}_m - \boldsymbol{M}_T\boldsymbol{\alpha}) \tag{18}$$

Differentiation of Q with respect to $(\boldsymbol{u}_m^\top, \boldsymbol{\alpha}^\top, \boldsymbol{\tau}^\top, \boldsymbol{u}_e^\top)^\top$ and setting these quantities to zero yields

$$\begin{bmatrix} \boldsymbol{Z}_g^\top\boldsymbol{R}^{-1}\boldsymbol{Z}_g & \boldsymbol{I}_m & \boldsymbol{0} & \boldsymbol{Z}_g^\top\boldsymbol{R}^{-1}\boldsymbol{Z}_g & \\ \boldsymbol{I}_m & \boldsymbol{0} & -\boldsymbol{M}_T & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{M}_T^\top & \boldsymbol{G}_{\alpha\alpha}^{-1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{X}^\top\boldsymbol{R}^{-1}\boldsymbol{Z}_g & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{X}^\top\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^\top\boldsymbol{R}^{-1}\boldsymbol{Z}_g \\ \boldsymbol{Z}_g^\top\boldsymbol{R}^{-1}\boldsymbol{Z}_g & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{Z}_g^\top\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}_g^\top\boldsymbol{R}^{-1}\boldsymbol{Z}_g + \boldsymbol{G}_{ee}^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{u}}_m \\ \tilde{\boldsymbol{a}} \\ \tilde{\boldsymbol{\alpha}} \\ \tilde{\boldsymbol{\tau}} \\ \tilde{\boldsymbol{u}}_e \end{bmatrix} \begin{bmatrix} \boldsymbol{Z}_g^\top\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{X}^\top\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}_g^\top\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix}$$

Absorption of $(\tilde{\boldsymbol{\tau}}^\top, \tilde{\boldsymbol{u}}_e^\top)^\top$ into the equations for $(\tilde{\boldsymbol{u}}_m^\top, \tilde{\boldsymbol{a}}^\top, \tilde{\boldsymbol{\alpha}}^\top)^\top$ yields

$$\begin{bmatrix} \boldsymbol{Z}_g^\top\boldsymbol{S}\boldsymbol{Z}_g & \boldsymbol{I}_m & \boldsymbol{0} \\ \boldsymbol{I}_m & \boldsymbol{0} & -\boldsymbol{M}_T \\ \boldsymbol{0} & -\boldsymbol{M}_T^\top & \boldsymbol{G}_{\alpha\alpha}^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{u}}_m \\ \tilde{\boldsymbol{a}} \\ \tilde{\boldsymbol{\alpha}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z}_g^\top\boldsymbol{S}\boldsymbol{y} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix} \tag{19}$$

where $\boldsymbol{S} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{W}_0\boldsymbol{C}_0^{-1}\boldsymbol{W}_0^\top\boldsymbol{R}^{-1}$, $\boldsymbol{W}_0 = [\boldsymbol{X}\ \boldsymbol{Z}_g]$, $\boldsymbol{C}_0 = \boldsymbol{W}_0^\top\boldsymbol{R}^{-1}\boldsymbol{W}_0 + \boldsymbol{G}_0^*$ and

$$\boldsymbol{G}_0^* = \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{G}_{ee}^{-1} \end{bmatrix}$$

From equation 19 we see that

$$\begin{aligned} \tilde{\boldsymbol{u}}_m &= \boldsymbol{M}_T\tilde{\boldsymbol{\alpha}} \\ \boldsymbol{G}_{\alpha\alpha}^{-1}\tilde{\boldsymbol{\alpha}} &= \boldsymbol{M}_T^\top\tilde{\boldsymbol{a}} \end{aligned} \tag{20}$$

Absorbing $\tilde{\boldsymbol{\alpha}}$ into the equations for $(\tilde{\boldsymbol{u}}_m^\top, \tilde{\boldsymbol{a}}^\top)^\top$ yields

$$\begin{bmatrix} \boldsymbol{Z}_g^\top \boldsymbol{S} \boldsymbol{Z}_g & \boldsymbol{I}_m \\ \boldsymbol{I}_m & -\boldsymbol{G}_{mm}^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{u}}_m \\ \tilde{\boldsymbol{a}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z}_g^\top \boldsymbol{S} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} \tag{21}$$

and hence

$$\tilde{\boldsymbol{a}} = \boldsymbol{G}_{mm}^{-1} \tilde{\boldsymbol{u}}_m \tag{22}$$

Finally absorbing the equation for $\tilde{\boldsymbol{a}}$ into the equation for $\tilde{\boldsymbol{u}}_m$ yields

$$(\boldsymbol{Z}_g^\top \boldsymbol{S} \boldsymbol{Z}_g + \boldsymbol{G}_{mm}^{-1}) \tilde{\boldsymbol{u}}_m = \boldsymbol{Z}_g^\top \boldsymbol{S} \boldsymbol{y}$$

and

$$\text{pev}\,(\tilde{\boldsymbol{u}}_m) = (\boldsymbol{Z}_g^\top \boldsymbol{S} \boldsymbol{Z}_g + \boldsymbol{G}_{mm}^{-1})^{-1}$$

as required. Using 22 and 20 we have

$$\begin{aligned} \tilde{\boldsymbol{\alpha}} &= \boldsymbol{G}_{\alpha\alpha} \boldsymbol{M}_T^\top \boldsymbol{G}_{mm}^{-1} \tilde{\boldsymbol{u}}_m \\ &= (\boldsymbol{G}_{T_\alpha} \otimes \boldsymbol{D})(\boldsymbol{I}_2 \otimes \boldsymbol{M}^\top)(\boldsymbol{G}_{T_\alpha}^{-1} \otimes \boldsymbol{K}^{-1}) \tilde{\boldsymbol{u}}_m \\ \Rightarrow \quad \begin{bmatrix} \tilde{\boldsymbol{\alpha}}_- \\ \tilde{\boldsymbol{\alpha}}_+ \end{bmatrix} &= \begin{bmatrix} \boldsymbol{D} \boldsymbol{M}^\top \boldsymbol{K}^{-1} \tilde{\boldsymbol{u}}_{m_-} \\ \boldsymbol{D} \boldsymbol{M}^\top \boldsymbol{K}^{-1} \tilde{\boldsymbol{u}}_{m_+} \end{bmatrix} \end{aligned}$$

It follows that using results on the inverse of partitioned matrices that

$$\begin{aligned} \text{pev}\,(\tilde{\boldsymbol{\alpha}}) &= \boldsymbol{G}_{\alpha\alpha} - \boldsymbol{G}_{\alpha\alpha} \boldsymbol{M}_T^\top \boldsymbol{G}_{mm}^{-1} \boldsymbol{M}_T \boldsymbol{G}_{\alpha\alpha} + \\ &= \boldsymbol{G}_{\alpha\alpha} \boldsymbol{M}_T^\top \boldsymbol{G}_{mm}^{-1} \text{pev}\,(\tilde{\boldsymbol{u}}_m)\, \boldsymbol{G}_{mm}^{-1} \boldsymbol{M}_T \boldsymbol{G}_{\alpha\alpha} \\ &= \boldsymbol{G}_{\alpha\alpha\cdot} + (\boldsymbol{I}_2 \otimes \boldsymbol{D} \boldsymbol{M}^\top \boldsymbol{K}^{-1}) \text{pev}\,(\tilde{\boldsymbol{u}}_m)\, (\boldsymbol{I}_2 \otimes \boldsymbol{K}^{-1} \boldsymbol{M} \boldsymbol{D}) \end{aligned} \tag{23}$$

where $\boldsymbol{G}_{\alpha\alpha\cdot} = \boldsymbol{G}_{\alpha\alpha} - \boldsymbol{G}_{\alpha\alpha} \boldsymbol{M}_T^\top \boldsymbol{G}_{mm}^{-1} \boldsymbol{M}_T \boldsymbol{G}_{\alpha\alpha}$. Hence from equation 23

$$\begin{aligned} \text{pev}\,(\tilde{\boldsymbol{\alpha}}_-) &= \lambda_{\alpha_-}^2 \boldsymbol{D}_{\alpha\alpha\cdot} + \boldsymbol{D} \boldsymbol{M}^\top \boldsymbol{K}^{-1}) \text{pev}\left(\tilde{\boldsymbol{u}}_{m_-}\right) \boldsymbol{K}^{-1} \boldsymbol{M} \boldsymbol{D}) \\ \text{pev}\,(\tilde{\boldsymbol{\alpha}}_+) &= (\lambda_{\alpha_+}^2 + \psi_\alpha) \boldsymbol{D}_{\alpha\alpha\cdot} + \boldsymbol{D} \boldsymbol{M}^\top \boldsymbol{K}^{-1}) \text{pev}\left(\tilde{\boldsymbol{u}}_{m_+}\right) \boldsymbol{K}^{-1} \boldsymbol{M} \boldsymbol{D}) \end{aligned}$$

References

Borg, L., Smith, A., Taylor, J., Cullis, B., & Statistics, A. (2015). Statistics for the Australian Grains Industry Technical Report Series Osmotic stress experiment for Cranbrook Halberd mapping population. Technical report, University of Wollongong, Wolllongong.

Butler, D., Cullis, B., Gilmour, A., & Gogel, B. J. (2009). ASReml-R Reference Manual, Release 3.

Cullis, B. R., Smith, A. B., & Coombes, N. E. (2006). On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological and Environmental Statistics* **11**, 381–393.

de Los Campos, G., Gianola, D., & Rosa, G. J. (2009). Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *Journal of Animal Science* **87**, 1883–1887.

Genc, Y., Oldach, K., Verbyla, A., Lott, G., Hassan, M., Tester, M., Wallwork, H., & McDonald, G. (2010). Sodium exclusion QTL associated with improved seedling growth in bread wheat under salinity stress. *Theoretical and Applied Genetics* **121**, 877–894.

Gianola, D., Perez-Enciso, M., & Toro, M. a. (2003). On marker-assisted prediction of genetic value: Beyond the ridge. *Genetics* **163**, 347–365.

Gilmour, A., Cullis, B., Welham, S., Gogel, B., & Thompson, R. (2004). An efficient computing strategy for prediction in mixed linear models. *Computational Statistics and Data Analysis* **44**, 571–586.

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). {AI}, an efficient algorithm for {REML} estimation in linear mixed models. *Biometrics* **51**, 1440–1450.

Green, P. J. & Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.

Haskard, K. A. (2005). *Anisotropic Matérn correlation and other issues in model-based geostatistics*. PhD thesis, BiometricsSA, University of Adelaide.

Haskard, K. A., Cullis, B. R., & Verbyla, A. P. (2007). Anisotropic Matérn correlation and spatial prediction using REML. *Journal of Agricultural and Biological Sciences* **12**, 147–160.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423–477.

Kammann, E. E. & Wand, M. P. (2003). Geoadditive models. *Applied Statistics* **52**, 1–18.

Kammholz, S. J., Campbell, a. W., Sutherland, M. W., Hollamby, G. J., Martin, P. J., Eastwood, R. F., Barclay, I., Wilson, R. E., Brennan, P. S., & Sheppard, J. a. (2001). Establishment and characterisation of wheat genetic mapping populations. *Australian Journal of Agricultural Research* **52**, 1079–1088.

Linsell, K. J., Rahman, M. S., Taylor, J. D., Davey, R. S., Gogel, B. J., Wallwork, H., Forrest, K. L., Hayden, M. J., Taylor, S. P., & Oldach, K. H. (2014). QTL for resistance to root lesion nematode (Pratylenchus thornei) from a synthetic hexaploid wheat source. *Theoretical and Applied Genetics* **127**, 1409–1421.

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829.

Ober, U. (2010). Kernel-Based BLUP with Genomic Data. In *9th World Congress on Genetics Applied to Livestock Production.*, pages 2–5.

Patterson, H. D. & Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* **31**, 545–554.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86**, 677–690.

R Core Team (2015). R: A Language and Environment for Statistical Computing.

Smith, A., Cullis, B. R., & Thompson, R. (2001). Analysing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* **57**, 1138–1147.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag, New York.

Stram, D. O. & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects setting. *Biometrics* **50**, 1171–1177.

Taylor, J. & Verbyla, A. (2011). R Package wgaim : QTL analysis in bi-parental populations using linear mixed models. *Journal of Statistical Software* **40**, 1–19.

Taylor, J. D. & Butler, D. (2014). ASMap: An (A)ccurate and (S)peedy linkage map construction package for inbred populations that uses the extremely efficient MSTmap algorithm.

Thompson, R. (1985). A Note on Restricted Maximum Likelihood Estimation with an Alternative Outlier Model. *Journal of the Royal Statistical Society. Series B (Methodological)* **47**, 53–55.

Thompson, R., Cullis, B., Smith, A., & Gilmour, A. (2003). A Sparse Implementation of the Average Information Algorithm for Factor Analytic and Reduced Rank Variance Models. *Australian & New Zealand Journal of Statistics* **45**, 445–459.

VanRaden, P. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* **91**, 4414–4423.

Verbyla, A. P., Cullis, B. R., & Thompson, R. (2007). The analysis of QTLs by simultaneous use of the full linkage map. *Theoretical and Applied Genetics* **116**, 95–111.

Verbyla, A. P., Taylor, J. D., & Verbyla, K. L. (2012). RWGAIM: An efficient high dimensional random whole genome average QTL interval mapping approach. *Genetics Research* **94**, 291–306.

Wahba, G. (1990). *Spline models for observational data*. SIAM:Philadelphia.

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., Maccaferri, M., Salvi, S., Milner, S. G., Cattivelli, L., Mastrangelo, A. M., Whan, A., Stephen, S., Barker, G., Wieseke, R., Plieske, J., Lillemo, M., Mather, D., Appels, R., Dolferus, R., Brown-Guedira, G., Korol, A., Akhunova, A. R., Feuillet, C., Salse, J., Morgante, M., Pozniak, C., Luo, M.-C., Dvorak, J., Morell, M., Dubcovsky, J., Ganal, M., Tuberosa, R., Lawley, C., Mikoulitch, I., Cavanagh, C., Edwards, K. J., Hayden, M., & Akhunov, E. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90000 single nucleotide polymorphism array. *Plant Biotechnology Journal* **12**, 787–796.

Webster, R. & Oliver, M. A. (2001). *Geostatistics for Environmental Scientists.* John Wiley and Sons, Chichester.

Wilkinson, G. N. & Rogers, C. E. (1973). Symbolic description of factorial models for analysis of variance. *Applied Statistics* **22**, 392–399.

Yang, W. & Tempelman, R. J. (2010). A Bayesian Antedependence Model to Account for Linkage Disequilibrium in Whole Genome Selection. In *9th World Congress on Genetics Applied to Livestock Production.*, pages 3–6.

Statistics for the Australian Grains Industry Technical Report Series

The analysis of QTL and QTL×treatment experiments using spatial models for marker effects

Brian Cullis and Alison Smith
National Institute for Applied Statistics and Research Australia
School of Mathematics and Applied Statistics
University of Wollongong

email: bcullis@uow.edu.au

January 14, 2016

1 Introduction

The continued increase in the availability of markers has led to much interest in their use in genetic improvement programs of crop species, such as wheat and maize. There is a large amount of literature on this topic and much of the focus has turned from marker assisted selection and the identification of quantitative trait loci (QTL) to genomic selection. The basic idea in marker-assisted selection is to exploit statistical dependencies (linkage disequilibrium, LD) existing in the joint distribution of marker and QTLs. Linkage disequilibrium between markers and QTL has two main objectives, and in some way these are not disjoint and not surprisingly the statistical models used in these two applications are similar. We refer to these objectives as (i) QTL analysis in which the aim is to infer genomic locations and effects (i.e. QTLs) which affect a (quantitative) trait and (ii) genomic selection in which the aim is to obtain predictions of genetic merit of individuals for selection as parents in a breeding program. Since the seminal paper of Meuwissen et al. (2001) there has been significant progress made in the second objective where there was a realisation that unravelling the genetic architecture of a trait via identification of (major) QTLs is not necessary for prediction of genetic merit. This concept built on the idea that a trait is the result of the influences of many, possibly small QTLs which would be very difficult if not impossible, to detect and hence routinely use within a breeding program.

A plethora of statistical methods have been developed for the first objective. Some of the so-called whole genome approaches are remarkably similar to the approaches used for genomic selection. Verbyla et al. (2007) presented a whole genome average interval mapping (WGAIM) approach for QTL analysis of a single trait in a single trial. They used an approach which was embedded within the framework of ridge regression or or so-called genomic BLUP (GBLUP), in which all the intervals on a linkage map are used simultaneously avoiding, in some sense, repeated genome scans to detect QTLs. Their approach uses forward selection, commencing by fitting a model similar to ridge regression (except based on intervals rather than markers) and then choosing putative QTLs using the concept of the alternative outlier model of Thompson (1985). Once an interval (or marker) is chosen it is then fitted in the model as a fixed effect and the process repeated until all significant QTLs have been identified and included in the model as fixed effects. Their method was shown to be much more powerful than composite interval mapping although there is a small increase in selecting false positives. Their approach has been implemented in the R statistical computing environment (R Core Team, 2015) in the package **wgaim** (Taylor & Verbyla, 2011).

Recently Verbyla et al. (2012) addressed two issues of WGAIM. Firstly they improved the efficiency of the analysis when the number of markers is large. Secondly, they considered the issue of (selection) bias involved in moving the selected QTL to the fixed effects. They addressed the first issue by considering a reformulation of the ridge regression model, which was similar to the approach originally proposed by VanRaden (2008), in which they fitted a model using a genomic relationship matrix to avoid inclusion of marker effects. Verbyla et al. (2012) used a variant of this idea which also avoided fitting marker effects directly and hence was found to be computationally efficient when the number of markers (r) exceeded the number of genotypes (m). They addressed the second issue of selection bias by fitting the set of selected intervals (markers) as random in the final step. We note however that this does not really fully address the issue of selection bias.

There has been an increasing interest in the use of so-called spatial models in both QTL analysis and genomic selection. Gianola et al. (2003) considered a range of alternative models for use in marker-assisted selection, which included an extension in which the model for marker effects included both chromosomal effects and within chromosomal deviations which were correlated according to a first-order autoregressive process. They extended the first-order autoregressive model to extend the implicit assumption that the markers are equally spaced (in a genetic sense), by considering the exponential model, which is the continuous-lag extension of the first-order autoregressive model. They noted that distances between markers could be based on physical units such as kilobases. They did not apply these models to real data-sets. Yang & Tempelman (2010) considered the use of the class of ante-dependence models for genomic selection. These models were popularised by (Pourahmadi, 1999) for the analysis of longitudinal data. Their approach was framed in a Bayesian context and they concluded that, on the basis of a simulation study that the models offered a "biologically reasonable and computationally tractable method to accommodate LD", and that the antedependence based model "should lead to measurably greater gains in accuracy of whole genome selection as greater levels of LD are attained between markers with newly developed SNP marker panels".

In a related approach there has also been interest in the use of spatial (and related) models for genomic selection, but rather than extending the ridge regression model for markers, various authors have considered alternative models to the genomic relationship matrix generated via the ridge regression model for markers (see for example, VanRaden (2008)). de Los Campos et al. (2009) considered the use of reproducing kernel Hilbert spaces regression (RKHS) for genomic selection and developed an approach based on the assumption that the additive genetic signal is a arbitrary function of the set of markers. The specification of the function is based on the class of semi-parametric regression models used in the smoothing splines literature and advocated by Green & Silverman (1994). Their choice of penalty function comes from the RKHS class of models (Wahba, 1990). Specifically they suggest use of the so-called gaussian kernel, which is a one parameter covariance model allowing for flexibility in the rate of decay of covariance as the "distance" between individuals increases. They do not provide a formal approach for estimation of this rate constant.

In a similar approach, Ober (2010) suggested the application of a high-dimensional kriging-extension to genomic selection. Their model is similar to the model of de Los Campos et al. (2009) but they choose the covariance model for the genomic relationship matrix to be based on the Matérn class of covariance functions (Stein, 1999). This model allows for more flexibility in capturing the functional dependency of the covariances on the (genetic) distance of individuals based on their SNP profiles. They compared their approach to GBLUP in a small simulation study. Their results suggest that there was little to choose between the spatial approach and conventional GBLUP.

There is a clear interest in the use of spatial models in genomic selection, but as yet these models have not been applied to the analysis of real data-sets. Spatial models have been proposed to model both the elements of the covariance between individuals and the covariance of marker effects within a chromosome. In this paper we will develop a general class of spatial models for the identification of the genetic architecture of complex traits in QTL mapping experiments by exploiting the existence of high LD between markers in modern marker panels. Our approach is a natural extension of the WGAIM approach, based on markers, not intervals, which uses the Matérn class of spatial covariance models. Stein (1999) advocates the use of the Matèrn model as a model which can be broadly and effectively employed to the problem of prediction in irregularly spaced spatial data. He argues that the Matérn model has much more flexibility then other models such as the class of smoothing splines or RKHS models which lead to a serious loss in efficiency. The additional flexibility of the Matérn class comes from the inclusion of the parameter which controls the so-called "smoothness" of the (gaussian) random field and he illustrates that many other covariance models are simply specific forms of a Matérn model in which the smoothness parameter is chosen a priori. Haskard et al. (2007) and Kammann & Wand (2003) have demonstrated the utility of the Matérn class for prediction in a spatial context.

The structure of the paper is as follows.

2 Statistical model for a simple QTL mapping experiment

We commence by considering the analysis of a simple QTL mapping experiment. By this we mean an experiment with only a single treatment factor, namely the genotypes from the mapping population, including parental and check varieties. Let $\boldsymbol{y}$ denote the $n \times 1$ vector of phenotypic data, where n is the number of observations in the experiment. We can write a model for the data vector as

$$\boldsymbol{y} = \boldsymbol{X}_p\boldsymbol{\tau}_p + \boldsymbol{Z}_g^*\boldsymbol{u}_g^* + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{\tau}$ is a vector of (incidental) fixed effects with associated design matrix $\boldsymbol{X}$; $\boldsymbol{u}_g^*$ is the $(m + m_0) \times 1$ vector of random total genetic effects corresponding to all genotypes, both those with marker data (m) and those without (m_0) marker data. The latter genotypes may include both parental and check varieties but also those DH lines which were genotyped but were discarded from the marker set during construction of the linkage map on the basis of either too many cross-overs or too much missing data. We consider the partition of both $\boldsymbol{u}_g^*$ and $\boldsymbol{Z}_g^*$ which is conformal with this classification. That is,

$$\boldsymbol{u}_g^* = (\boldsymbol{u}_{g_0}^\top \; \boldsymbol{u}_g^\top)^\top \qquad \text{and} \qquad \boldsymbol{Z}_g^* = [\boldsymbol{Z}_{g_0} \; \boldsymbol{Z}_g]$$

Equation 1 can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{u}_g + \boldsymbol{e} \tag{2}$$

where $\boldsymbol{X} = [\boldsymbol{X}_p \; \boldsymbol{Z}_{g0}]$ and $\boldsymbol{\tau} = (\boldsymbol{\tau}_p^\top \; \boldsymbol{u}_{g0}^\top)^\top$. This notation and form of the model is necessary as we fit the genetic effects for those genotypes without marker data as fixed effects to exclude their influence on the genetic analysis. This model is now extended to consider the partitioning of the total genetic effects for those genotypes with marker data into additive and residual genetic effects. The extension commences with the decomposition:

$$\boldsymbol{u}_g = \boldsymbol{u}_a + \boldsymbol{u}_e \tag{3}$$

where the two terms are the additive and residual genetic effects respectively. Given $\boldsymbol{M}$, the matrix of (SNP) marker data (assumed known, without missing values and columns ordered according to linkage groups and in map order within linkage groups) of size $m \times r$, then we consider a model for the additive genetic effects given by

$$\boldsymbol{u}_a = \boldsymbol{u}_m + \boldsymbol{u}_\epsilon, \quad \text{and} \quad \boldsymbol{u}_m = \boldsymbol{M}\boldsymbol{\alpha} \tag{4}$$

where $\boldsymbol{\alpha}$ is the vector of marker regression coefficients and $\boldsymbol{u}_\epsilon$ is the vector of lack of fit additive genetic effects. For simple mapping populations such as doubled haploid populations and recombinant inbred lines, genotypes are usually derived from a bi-parental cross of in-bred lines and so the lack of fit term can be assumed to be (effectively) zero. We note that the non-imputed values in $\boldsymbol{M}$ are coded as -1 and 1 and r is much larger than m for most of our applications. Extensions to non-inbred populations is possible.

Thus the model in equation 2 can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{M}\boldsymbol{\alpha} + \boldsymbol{Z}_g\boldsymbol{u}_e + \boldsymbol{e} \tag{5}$$

This is referred to as the marker model. The genotype model is given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{u}_m + \boldsymbol{Z}_g\boldsymbol{u}_e + \boldsymbol{e} \tag{6}$$

These models can be extended to include peripheral random effects which often are associated with blocking factors arising in the analysis of designed comparative experiments. We omit this extension for pedalogical reasons.

2.1 Variance models for random effects: simple QTL experiments

We assume that the variance matrices of the residual genetic effects and the residuals are given by

$$\begin{aligned} \text{var}(\boldsymbol{u}_e) &= \sigma_e^2\boldsymbol{I}_m \\ \text{var}(\boldsymbol{e}) &= \boldsymbol{R} \end{aligned}$$

where typically the matrix $\boldsymbol{R}$ is a function of an (R-level) variance parameter vector.

The variance matrix for the random marker effects is given by

$$\text{var}(\boldsymbol{\alpha}) = \sigma_\alpha^2\boldsymbol{D} = \oplus_{i=1}^c \boldsymbol{D}_i = \boldsymbol{G}_{\alpha\alpha}, \quad \text{say} \tag{7}$$

where c is the number of chromosomes and $\boldsymbol{D}_i$ is an $r_i \times r_i$ variance matrix of within-chromosome marker effects for chromosome i, and r_i is the number of markers in chromosome i. Note that $r = \sum_{i=1}^{c} r_i$. The block diagonality of $\boldsymbol{D}$ in equation 7 assumes that the within chromosome marker effects are independent across chromosomes. We assume further that the elements of $\boldsymbol{D}_i$ are given by

$$d_{i;j,j'} = (\phi|t_{i;j,j'}|)^{\nu}\mathcal{K}_{\nu}(\phi|t_{i;j,j'}|) \tag{8}$$

where $\mathcal{K}_\nu$ is a modified Bessel function, ϕ is the range parameter of the process and $|t_{i;j,j'}|$ is the absolute value of the distance between markers j and j' on chromosome i. Markers can be assumed equidistant within a chromosome, or more often, in the absence of physical distances we use the map distance in centimorgans. Underlying this variance model is the assumption that $\boldsymbol{\alpha}$ is a realisation of a Gaussian random process at specific locations (along the genome).

For a given ν, the range parameter ϕ affects the rate of decay of the correlation function with increasing $|t_{i;j,j'}|$. The parameter $\nu > 0$ controls the analytic smoothness of the underlying (genetic) process, the process being $\lceil \nu \rceil - 1$ times mean-square differentiable, where $\lceil \nu \rceil$ is the smallest integer greater than or equal to ν (Stein, 1999, page 31). Larger ν correspond to smoother processes. We note that $\nu = \frac{1}{2}$ yields the exponential correlation function,

$$d_{i;j,j'} = \exp(-\phi|t_{i;j,j'}|), \tag{9}$$

while $\nu = 1$ yields Whittles elementary correlation function, (Webster & Oliver, 2001, page 119). When ν is of the form $h + \frac{1}{2}$, with h a nonnegative integer, the correlation function in equation 8 is of the form $\exp(-\phi|t_{i;j,j'}|)$ times a polynomial in $|t_{i;j,j'}|$ of degree h. Kammann & Wand (2003) use the model where $\nu = \frac{3}{2}$, in which case

$$d_{i;j,j'} = \exp(-\phi|t_{i;j,j'}|)(1 + \phi|t_{i;j,j'}|) \tag{10}$$

and they term this model a “geo-additive” model. It has the advantage of being computationally simple to differentiate and gives rise to a process which is once differentiable.

It follows from equation 4 that the variance matrix for $\boldsymbol{u}_m$ is given by $\sigma_\alpha^2 \boldsymbol{MDM}^\top = \sigma_\alpha^2 \boldsymbol{K} = \boldsymbol{G}_{mm}$, say, where we call the matrix $\boldsymbol{K}$ or order $m \times m$, matern-genomic relationship. This matrix is dense, but is relatively cheap to compute given the block diagonal form for $\boldsymbol{D}$.

3 Statistical model for a factorial QTL mapping experiment

Here we extend the models for the analysis of a simple QTL mapping experiment to the analysis of a QTL mapping experiment with a factorial treatment structure. These experiments are often conducted to determine the genetic architecture of the tolerance or resistance of crops to a range of abiotic and biotic stresses. Recent examples include Linsell et al. (2014) and Genc et al. (2010), and the treatment structure of these experiments usually involves the factorial combination of the genotypes with a treatment with two levels, namely a control (‘-’) and a stress (‘+’). The most common approach to the analysis of these experiments is to undertake a two-stage approach, forming differences or ratios of the '+' and the '-' treatment means for each genotype, thence subjecting these to a QTL analysis. This approach is piecemeal and results in a loss of information. Our approach is to extend the approach outlined in the previous section by jointly modelling the variance of the treatment × genotype effects using a particular form of the factor analytic models suggested by Smith et al. (2001) for the analysis of multi-environment trials.

Our model is of the same form as equations 5 and 6 except additional fixed effects are included in $\boldsymbol{\tau}$. These effects represent the saturated factorial structure between the factor associated with those genotypes without marker data and the treatment factor. This includes, by default the main effect of the treatment factor and hence the vector $\boldsymbol{u}_g$ represents the total genetic effect nested within treatments. For brevity we refer to the latter as the genotype by treatment total genetic effects. The vector $\boldsymbol{u}_g$ is $2m \times 1$ with the elements ordered genotypes within treatments, and hence $\boldsymbol{u}_g = (\boldsymbol{u}_{g_-}^\top \; \boldsymbol{u}_{g_+}^\top)^\top$ where the two sub-vectors are the effects for the '-' and '+' respectively. As before we consider the decomposition of $\boldsymbol{u}_g$ given by

$$\boldsymbol{u}_g = \begin{bmatrix} \boldsymbol{u}_{g_-} \\ \boldsymbol{u}_{g_+} \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_{a_-} \\ \boldsymbol{u}_{a_+} \end{bmatrix} + \begin{bmatrix} \boldsymbol{u}_{e_-} \\ \boldsymbol{u}_{e_+} \end{bmatrix}$$

and further

$$\begin{aligned} \boldsymbol{u}_a = \begin{bmatrix} \boldsymbol{u}_{a_-} \\ \boldsymbol{u}_{a_+} \end{bmatrix} &= \begin{bmatrix} \boldsymbol{M} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_- \\ \boldsymbol{\alpha}_+ \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{u}_{m_-} \\ \boldsymbol{u}_{m_+} \end{bmatrix} = \boldsymbol{u}_m \end{aligned} \tag{11}$$

3.1 Variance models for random effects: factorial QTL experiments

We model the vector of marker by treatment effects and residual genetic by treatment effects using a constrained factor analytic variance model of order 1. This model has three parameters and therefore has the same number of parameters as an unstructured variance matrix for two "traits" (i.e. treatments), but it permits a natural and biologically meaningful interpretation. It has the added advantage of dealing with non-positive definite variance matrices. The so-called extended factor analytic models were introduced by Thompson et al. (2003) as a computationally efficient alternative to the approach presented by Smith et al. (2001). The model we consider here is a sub-class of these models, in which one of the specific variances is set to zero. The regression form of the model for marker by treatment effects is given by

$$\begin{aligned} \boldsymbol{\alpha} &= \begin{bmatrix} \lambda_{\alpha_-} \boldsymbol{f}_\alpha \\ \lambda_{\alpha_+} \boldsymbol{f}_\alpha \end{bmatrix} + \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{\delta}_\alpha \end{bmatrix} \\ &= (\boldsymbol{\lambda}_\alpha \otimes \boldsymbol{I}_r) \boldsymbol{f}_\alpha + \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{\delta}_\alpha \end{bmatrix} \end{aligned}$$

where $\boldsymbol{\lambda}_\alpha = (\lambda_{\alpha-}\ \lambda_{\alpha+})^\top$ and we assume that

$$\text{var}\begin{pmatrix} \boldsymbol{f}_\alpha \\ \boldsymbol{\delta}_\alpha \end{pmatrix} = \begin{bmatrix} \boldsymbol{G}_{f_\alpha f_\alpha} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}_{\delta_\alpha \delta_\alpha} \end{bmatrix}$$

where $\boldsymbol{G}_{f_\alpha f_\alpha} = \boldsymbol{D}$ and $\boldsymbol{G}_{\delta_\alpha \delta_\alpha} = \psi_\alpha \boldsymbol{D}$. Hence it follows that

$$\text{var}(\boldsymbol{\alpha}) = \boldsymbol{G}_{T_\alpha} \otimes \boldsymbol{D} = \boldsymbol{G}_{\alpha\alpha} \tag{12}$$

where

$$\boldsymbol{G}_{T_\alpha} = \boldsymbol{\lambda}_\alpha \boldsymbol{\lambda}_\alpha^\top + \text{diag}(0, \psi_\alpha)$$

This regression representation of the model admits a simple interpretation. The vector $\boldsymbol{f}_\alpha$ represents the pleiotropic marker effects across treatments. The effects for the `'-'` and `'+'` treatments are scaled by $\lambda_{\alpha-}$ and $\lambda_{\alpha+}$ respectively. The vector $\boldsymbol{\delta}_\alpha$ represents the deviations from the additive genetic regression of $\boldsymbol{\alpha}_+$ on $\boldsymbol{\alpha}_-$. Hence this term captures the non-pleiotropic effects associated with the `'+'` treatment which are independent of the `'-'` treatment. Specifically this additive genetic regression is given by

$$\boldsymbol{\alpha}_+ = \frac{\lambda_{\alpha+}}{\lambda_{\alpha-}} \boldsymbol{\alpha}_- + \boldsymbol{\delta}_\alpha \tag{13}$$

Consistent with how this model is fitted, for either $\boldsymbol{\alpha}$ (or $\boldsymbol{u}_m$ or $\boldsymbol{u}_e$), we refer to it as a `rr(trt) + diag(+)` model which is an abbreviation for a reduced rank model of order one, which is a factor model of order 1 for two traits where both specific variances are set to zero, plus a scaled identity or default variance matrix for the marker effects for the stress or '+' treatment. This parameterisation is computationally efficient, numerically stable and defaults to a positive semi-definite variance matrix (i.e a 2 × 2 matrix of order one) when $\psi_\alpha = 0$. When $\psi_\alpha = 0$ then $\boldsymbol{u}_{m-}$ and $\boldsymbol{u}_{m+}$ are perfectly correlated.

It is straightforward to extend the marker regression model to the genotype regression model

$$\boldsymbol{u}_m = (\boldsymbol{\lambda}_\alpha \otimes \boldsymbol{I}_m)\boldsymbol{f}_m + \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{\delta}_m \end{bmatrix} \tag{14}$$

where

$$\boldsymbol{f}_m = \boldsymbol{M}\boldsymbol{f}_\alpha \qquad \text{and} \qquad \boldsymbol{\delta}_m = \boldsymbol{M}\boldsymbol{\delta}_\alpha \tag{15}$$

Hence

$$\begin{aligned} \text{var}(\boldsymbol{f}_m) &= \boldsymbol{K} \\ \text{var}(\boldsymbol{\delta}_m) &= \psi_\alpha \boldsymbol{K} \\ \text{var}(\boldsymbol{u}_m) &= \boldsymbol{G}_{T_\alpha} \otimes \boldsymbol{K} \\ &= \boldsymbol{G}_{mm}, \quad \text{say} \end{aligned}$$

and this is termed the `rr(Trt) + diag(+)` variance model for (additive) genetic effects (i.e. $\boldsymbol{u}_m$).

Similarly applying the same model to the residual genetic by treatments effects, $\boldsymbol{u}_e$, the residual genetic regression model is

$$\boldsymbol{u}_e = (\boldsymbol{\lambda}_e \otimes \boldsymbol{I}_m)\boldsymbol{f}_e + \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{\delta}_e \end{bmatrix}$$

where

$$\begin{aligned} \text{var}\,(\boldsymbol{f}_e) &= \boldsymbol{I}_m \\ \text{var}\,(\boldsymbol{\delta}_e) &= \psi_e \\ \text{var}\,(\boldsymbol{u}_e) &= \boldsymbol{G}_{T_e} \otimes \boldsymbol{I}_m \\ &= \boldsymbol{G}_{ee}, \quad \text{say} \end{aligned}$$

where

$$\boldsymbol{G}_{T_e} = \boldsymbol{\lambda}_e \boldsymbol{\lambda}_e^\top + \text{diag}\,(0, \psi_e)$$

and this is termed the `rr(Trt) + diag(+)` variance model for the (residual) genetic effects (i.e. $\boldsymbol{u}_e$).

To complete this section, we now form the full model based on the two genetic regression models for $\boldsymbol{u}_m$ and $\boldsymbol{u_e}$. This formulation is the most computationally efficient and numerically stable (Thompson et al., 2003). Substitution of equations 14 and 15 for $\boldsymbol{u}_m$ and $\boldsymbol{u_e}$ respectively into equation 6 gives

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X\tau} + \boldsymbol{Z}_{f_\alpha}\boldsymbol{f}_m + \boldsymbol{Z}_{g+}\boldsymbol{\delta}_m + \boldsymbol{Z}_{f_e}\boldsymbol{f}_e + \boldsymbol{Z}_{g+}\boldsymbol{\delta}_e + \boldsymbol{e} \\ &= \boldsymbol{W\beta} + \boldsymbol{e} \end{aligned} \tag{16}$$

where $\boldsymbol{Z}_g = [\boldsymbol{Z}_{g-} \;\; \boldsymbol{Z}_{g+}]$, $\boldsymbol{\beta} = (\boldsymbol{\tau}^\top, \boldsymbol{f}_m^\top, \boldsymbol{\delta}_m^\top, \boldsymbol{f}_e^\top, \boldsymbol{\delta}_e^\top)^\top$, $\boldsymbol{W} = [\boldsymbol{X} \;\; \boldsymbol{Z}_{f_\alpha} \;\; \boldsymbol{Z}_{g+} \;\; \boldsymbol{Z}_{f_e} \;\; \boldsymbol{Z}_{g+}]$, $\boldsymbol{Z}_{f_\alpha} = \boldsymbol{Z}_g(\boldsymbol{\lambda}_\alpha \otimes \boldsymbol{I}_m)$, $\boldsymbol{Z}_{f_e} = \boldsymbol{Z}_g(\boldsymbol{\lambda}_e \otimes \boldsymbol{I}_m)$ and

$$\text{var}\begin{pmatrix} \boldsymbol{f}_m \\ \boldsymbol{\delta}_m \\ \boldsymbol{f}_e \\ \boldsymbol{\delta}_e \end{pmatrix} = \text{diag}\,(\boldsymbol{K}, \psi_\alpha \boldsymbol{K}, \boldsymbol{I}_m, \psi_e \boldsymbol{I}_m) = \boldsymbol{G}_{rr}$$

This model is similar to the model in equation (4) of Thompson et al. (2003).

4 Estimation and Prediction

Since r is usually large and much greater than m, we prefer to fit the genotype joint regression model and thence obtain predictions and (model-based) prediction error variances of the set of genetic effects in the marker joint regression model from those obtained in fitting the genotype joint regression model as a post processing step. Details are only provided for the factorial QTL experiment, as the results for the simple QTL experiment can be inferred from these.

The first step in fitting the genotype joint regression model is the estimation of variance parameters, the most common method being Residual Maximum Likelihood (REML, Patterson & Thompson (1971)). This usually involves an iterative process. In this paper we use the Average Information (AI) algorithm (Gilmour et al., 1995) as implemented in the R package **ASReml-R** (Butler et al., 2009). Haskard (2005) presents details for the fitting the Matérn model and this has been implemented in **ASReml-R**. This required evaluation of the derivatives of the residual likelihood with respect to ϕ and ν. The differential for ϕ is relatively straightforward to compute analytically, though more complex for ν, unless ν is of the form $h + \frac{1}{2}$, with h a nonnegative integer. Following Haskard (2005) we avoid occasional numerical problems, and use numerical methods to obtain the differentials with respect to ν. This issue is mostly avoided through our modelling strategy (see section 6). We have found that, given sensible starting values and sequential model building, the AI algorithm to obtain REML estimates of the variance parameters performs well for most cases, again this is aided by our approach to modelling.

However, one potential obstacle to routine use of REML is the burden in computing the likelihood, score and AI matrix for large numbers of markers. Unlike general spatial problems, however, we can exploit the block diagonality of $\boldsymbol{D}$, significantly reducing the computational load to the inversion of $r_i \times r_i$ matrices.

Given estimates of the variance parameters we obtain Empirical Best Linear Unbiassed Estimates (E-BLUEs) of the fixed effects and Empirical Best Linear Unbiassed Predictions (E-BLUPs) of the random effects, E- denoting that we replace all variance parameters with their REML estimates. In particular, our interest centres on the E-BLUPs of the genetic for both genotypes and markers. The E-BLUPs for the terms in the genotype regressions can be obtained from the solutions to the mixed model equations. The mixed model equations for the genotype joint regression model given by equation 16 are

$$\boldsymbol{C}\tilde{\boldsymbol{\beta}} = \boldsymbol{W}^\top \boldsymbol{R}^{-1}\boldsymbol{y}$$

where $\tilde{\boldsymbol{\beta}} = (\hat{\boldsymbol{\tau}}^\top, \tilde{\boldsymbol{f}}_m^\top, \tilde{\boldsymbol{\delta}}_m^\top, \tilde{\boldsymbol{f}}_e^\top, \tilde{\boldsymbol{\delta}}_e^\top)^\top$, $\boldsymbol{C} = \boldsymbol{W}^\top \boldsymbol{R}^{-1}\boldsymbol{W} + \boldsymbol{G}^*$ and

$$\boldsymbol{G}^* = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{G}_{rr}^{-1} \end{bmatrix}$$

The (model-based) prediction error variances of the $\tilde{\boldsymbol{f}}_m$ and $\tilde{\boldsymbol{\delta}}_m$ are given by pev $\left(\tilde{\boldsymbol{f}}_m\right)$ and pev $\left(\tilde{\boldsymbol{\delta}}_m\right)$, where pev () is a matrix function which extracts the block diagonal matrix of $\boldsymbol{C}^{-1}$ relating to its vector argument. It is then straightforward to obtain the E-BLUP of $\tilde{\boldsymbol{u}}_m$ and its associated prediction error variance matrix from these quantities (including the prediction error covariance between $\tilde{\boldsymbol{f}}_m$ and $\tilde{\boldsymbol{\delta}}_m$). This is achieved in **ASReml-R** as a post-processing procedure using the `predict.asreml` method. This implements the strategies outlined in Gilmour et al. (2004).

We now present a summary of the results required to obtain the E-BLUPS and prediction error variances of the effects of interest in the marker joint regression model as functions of the effects of interest in the genotype joint regression model. A sketch of the proof of these results is given in the appendix. In the following if we define

$$\boldsymbol{D}_{mm\cdot} = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{M}^\top\boldsymbol{K}^{-1}\boldsymbol{M}\boldsymbol{D}$$

then

$$\begin{aligned}
\tilde{\boldsymbol{f}}_\alpha &= \boldsymbol{D}\boldsymbol{M}^\top\boldsymbol{K}^{-1}\tilde{\boldsymbol{f}}_m \\
\text{pev}\left(\tilde{\boldsymbol{f}}_\alpha\right) &= \boldsymbol{D}_{mm\cdot} + \boldsymbol{D}\boldsymbol{M}^\top\boldsymbol{K}^{-1}\text{pev}\left(\tilde{\boldsymbol{f}}_m\right)\boldsymbol{K}^{-1}\boldsymbol{M}\boldsymbol{D} \\
\tilde{\boldsymbol{\delta}}_\alpha &= \boldsymbol{D}\boldsymbol{M}^\top\boldsymbol{K}^{-1}\tilde{\boldsymbol{\delta}}_m \\
\text{pev}\left(\tilde{\boldsymbol{\delta}}_\alpha\right) &= \psi_\alpha\boldsymbol{D}_{mm\cdot} + \boldsymbol{D}\boldsymbol{M}^\top\boldsymbol{K}^{-1}\text{pev}\left(\tilde{\boldsymbol{\delta}}_m\right)\boldsymbol{K}^{-1}\boldsymbol{M}\boldsymbol{D} \\
\tilde{\boldsymbol{\alpha}}_- &= \boldsymbol{D}\boldsymbol{M}^\top\boldsymbol{K}^{-1}\tilde{\boldsymbol{u}}_{m_-} \\
\text{pev}\left(\tilde{\boldsymbol{\alpha}}_-\right) &= \lambda^2_{\alpha_-}\boldsymbol{D}_{mm\cdot} + \boldsymbol{D}\boldsymbol{M}^\top\boldsymbol{K}^{-1}\text{pev}\left(\tilde{\boldsymbol{u}}_{m_-}\right)\boldsymbol{K}^{-1}\boldsymbol{M}\boldsymbol{D} \\
\tilde{\boldsymbol{\alpha}}_+ &= \boldsymbol{D}\boldsymbol{M}^\top\boldsymbol{K}^{-1}\tilde{\boldsymbol{u}}_{m_+} \\
\text{pev}\left(\tilde{\boldsymbol{\alpha}}_+\right) &= (\lambda^2_{\alpha_+} + \psi_\alpha)\boldsymbol{D}_{mm\cdot} + \boldsymbol{D}\boldsymbol{M}^\top\boldsymbol{K}^{-1}\text{pev}\left(\tilde{\boldsymbol{u}}_{m_+}\right)\boldsymbol{K}^{-1}\boldsymbol{M}\boldsymbol{D}
\end{aligned}$$

5 Inference and examination of genomic marker profiles

Need to add in the chromosome score tests and then the way to look at the turning points within those chromosomes selected from scores.

6 Example

6.1 Phenotypic data and experimental design

Borg et al. (2015) describe the analysis of a QTL factorial experiment on wheat. We present a brief overview here. The aim of the experiment was to investigate whether phenotyping for osmotic stress tolerance could be a valid surrogate screening method for drought tolerance in wheat. A total of 168 genotypes (166 doubled haploid (DH) lines and the two parents; Cranbrook and Halberd) were grown in two successive runs within a glasshouse. Full details concerning the DH population can be found in Kammholz et al. (2001). The glasshouse contained four hydroponic tubs which were placed on benches. Each hydroponic tub could hold a maximum of 144 pots arranged in a 16 row by 9 column rectangular array. The key practical constraint was the assignment of genotypes to maturity groups, which were placed to be located within in so-called maturity blocks (`Matblks`) and these maturity blocks were systematically located across the tubs (and benches). Each genotype was classified into one of five maturity groups; very quick, quick, moderate, slow and very slow. This classification was based on days until flowering observed in field trials conducted at Yanco and Narrabri experimental stations in New South Wales. Whenever maturity blocks were adjacent within a tub at least one row of plots was left vacant between them. The genotypes were allocated (or randomised) to a set of three row-adjacent pots within their respective maturity block. The osmotic treatment was applied to two randomly chosen pots in each set and the remaining pot was the control. Tillers at the right stage within the pots assigned to the '+' treatment were tagged and these pots were then placed into tanks located in the same glasshouse which contained varying concentrations of $NaCl^-$. The pots were exposed to increasing concentrations over a 6 day period after which time they were returned to their original position within the maturity group. A sample of spikes from the control pots were also tagged for identification and measurement at the completion of the experiment.

Spike grain number (SGN) was recorded on tagged tillers from each pot. The number of tillers tagged varied per pot depending on the number of tillers at the appropriate developmental stage during tagging. At the time of measurement, the identity of the pot within each set of three pots was not recorded. The tiller was identified (non-uniquely) by the genotype, the treatment and the run. There was only one set of three pots for each genotype per run.

The experiment design is non-standard and is complicated by the following issues. Firstly, the same randomisation of genotypes to sets of three row-adjacent pots was used for each run. Secondly, the maturity blocks were located in the same position for each run and lastly, we do not know the allocation of treatments to pots within each set of three row-adjacent pots. The design is strictly unreplicated for genotypes as the experimental unit (EU) for genotypes is the set of three row-adjacent pots, which is invariant across runs and within maturity blocks. The EU for the treatment by genotype combination is a pot within a set of three row-adjacent pots, however we do not know if the allocation of treatments changed between runs and we cannot identify the allocation of treatments to pots. Fortuitously as a result of the genetic analysis, genetic clones were identified among the DH lines. (A genetic clone is defined to be a pair (or in some cases up to 4) of DH lines which have identical matching of alleles for greater than 99.5%. This meant that there was (minimal) p-artial replication of the genotypes (Cullis et al., 2006) for each run.

6.2 Genotypic data and map construction

A total of 165 lines were genotyped using a 90K SNP chip containing gene-associated SNPs that provided dense coverage of the wheat genome (Wang et al., 2014). These markers were combined with three phenological markers which were available for this population. There were a total of 16231 markers available for linkage map construction. The consensus map of Wang et al. (2014) was used as a reference map during construction. We used the R package **ASMap** (Taylor & Butler, 2014) for map construction.

The final linkage map had a total of 15601 which was thinned to 1383 non-redundant markers for the QTL analysis. The number of markers per chromosome ranged from 14 to 111 with a median number of 74. The overall length of the map (using the Kosambi distance measure) was 3867, individual lengths ranging from 95 to 234. There were a total of 9 genotypes excluded from the QTL analysis, which included the two parents. A total of 143 genotypes were included and some of these genotypes being groups of genetic clones.

6.3 Linear mixed model for the base-line model

The complex and multi-strata nature of the experimental design, necessitated use of a base-line mixed model which contained many non-genetic terms. See Borg et al. (2015) for details of the derivation of this base-line model. Using the extended model formulae syntax of Butler et al. (2009) after Wilkinson & Rogers (1973), this model is given by

```
fixed    = ~ Run*Gdrop*Trt
random   = ~ rr(Trt):Gkeep + at(Trt,‘+’):Gkeep + Matblk + Tub + Mplot +
             Run:Matblk + Run:Tub + Gkeep:Run + Gkeep:Trt:Run +
             Run:Mplot
residual = ~ idv(units)
```

The terms of interest are `rr(Trt):Gkeep + at(Trt,‘+’):Gkeep`, which represent the total genetic effects for the genotype × treatment effects. The variance matrix for these effects is the model described in section 3.1, as the `rr(trt) + diag(+)` model. We exclude the additive genetic effects from this model, as this model serves as the reference model for the subsequent QTL analysis. The set of models presented in section 6.4 include all of the terms which were fitted in the base-line model. A shifted power transformation was used to improve the normality of the SGN data.

6.4 Model fitting

Table 1 presents a summary of the four models fitted to the SGN data for the CxH factorial QTL experiment. The first model we fit is the base-line model which does not consider the decomposition of the total genetic effects into additive and residual genetic effects. The model denoted ’QTL-ide’ includes a pair of terms to model the additive genotype × treatment effects. This model assumes that the marker effects are independent, that is, the standard WGAIM or GBLUP variance model. The REMLLRT for testing H_0 that $\boldsymbol{\lambda}_a = \mathbf{0}$ and $\psi_a = 0$ using this model, was 21.73, which, when compared to the reference distribution of a mixture of χ^2 variates (Stram & Lee, 1994) gave a p-value <0.001. Thus there is strong evidence for additive genetic variance due to the markers in the genotype × treatment effects.

The next model we fit is the correlated marker (effects) QTL model, which considers the vector $\boldsymbol{\alpha}$ as a realisation of a Gaussian process with correlation function given by equation 9. This model has an additional two variance parameters. Our approach is to examine the fit of the Matérn model for a range of values of ν. It is likely that direct REML estimation of ν will be problematic (see for example Stein (1999) and Haskard (2005)). We chose $\nu = 0.5$ and $\nu = 1.5$, the former being the exponential form for the correlation function and the latter is the geo-additive model used by Kammann & Wand (2003). The Matérn model with $\nu = 1.5$ resulted in a marginally better fit and hence we choose $\nu = 1.5$. This correlated marker effects model results in a marker additive genetic process being once differentiable, which has practical and biological advantages which we will exploit in the identification of the putative QTLs.

Table 1 indicates that this model gave only a modest improvement in fit compared to the fit from the QTL-ide model. This is not surprising, though given the substantial amount of non-genetic variation in the data, and the moderate sample size. The mean accuracies of the key genetic effects from the base-line model demonstrates this point. The mean accuracies of the E-BLUPs of the total genetic effects for the '-' and '+' genotype × treatment effects were 0.439 and 0.536 respectively. The mean accuracy for the E-BLUPs of the deviations from the (total) genetic regression of $\boldsymbol{u}_{e_+}$ on $\boldsymbol{u}_{e_-}$ (i.e. of $\tilde{\boldsymbol{\delta}}_e$ was only 0.163.

Table 1: Summary of residual maximum likelihood (REML) and REML estimates of the key variance parameters for the four models fitted to the CxH factorial QTL experiment. Note that ϕ and ψ_α are fixed at zero for models QTL-ide and QTL-MATf respectively.

	logl	$\hat{\lambda}_{\alpha_-}$	$\hat{\lambda}_{\alpha_+}$	$\hat{\psi}_\alpha$	$\hat{\phi}$	$\hat{\lambda}_{e_-}$	$\hat{\lambda}_{e_+}$	$\hat{\psi}_e$
Base	-2282.25					0.542	1.198	0.341
QTL-ide	-2271.38	0.805	0.703	0.510	0.000	0.231	1.004	0.000
QTL-MAT	-2269.73	0.777	0.637	0.449	0.055	0.287	1.089	0.000
QTL-MATf	-2272.33	0.783	0.633	0.000	0.035	0.251	1.251	0.000

The REMLLRT for testing H$_0$ that $\phi = 0$ (as ν is redundant when $\phi = 0$) was 3.299. An approximate p-value for this statistic is 0.0693. There is only mild evidence to reject H$_0$ using this test.

The final model we fit is the model which constrains the variance for the residuals from the (additive) genetic regression of $\boldsymbol{\alpha}_+$ on $\boldsymbol{\alpha}_-$ (or equivalently $\boldsymbol{u}_{m_+}$ on $\boldsymbol{u}_{m_-}$) to zero. This model is referred to as model 'QTL-MATf' model in table 1. The REMLLRT for testing H$_0$ that $\psi = 0$ was 5.203 with an approximate p-value <0.001. Hence we strongly reject H$_0$, concluding that there is significant deviation from the additive genetic regression of $\boldsymbol{\alpha}_+$ on $\boldsymbol{\alpha}_-$, supporting further investigation of the additive genetic (marker) profiles to identify genomic regions of interest.

Another interesting feature of this modelling approach can be seen by closer examination of the REML estimates of the variance parameters associated with the residual genetic effects. The REML estimates of $\lambda^2_{e_-}$ were 0.293 and 0.082 for the Base and QTL-MAT models respectively. On the other hand the REML estimates of ψ_{e_-} were 0.341 and 0 for the Base and QTL-MAT models respectively. That is, inclusion of the additive effects in the model(s) resulted in some reduction in the residual (polygenic and pleiotropic) genetic variance for the overall genotype effects, but a substantial reduction in the residual (polygenic and specific) genetic variance for the deviations from the residual genetic regression by inclusion of the markers.

Table 2 presents a summary of the model based accuracies for the E-BLUPs of genetic effects for the QTL-ide and QTL-MAT models respectively. There is a large increase in accuracy for the additive genetic effects associated with the most important trait, namely $\boldsymbol{\delta}_{\boldsymbol{\alpha}}$.

Table 2: Summary of model based accuracies for the E-BLUPs of key genetic effects for two models fitted to the CxH factorial QTL experiment

	QTL-ide	QTL-MAT
acc($\tilde{\boldsymbol{\delta}}_m$)	0.384	0.494
acc($\tilde{\boldsymbol{u}}_{m_-}$)	0.495	0.609
acc($\tilde{\boldsymbol{u}}_{m_+}$)	0.501	0.593
acc($\tilde{\boldsymbol{\delta}}_a$)	0.016	0.264
acc($\tilde{\boldsymbol{\alpha}}_-$)	0.023	0.345
acc($\tilde{\boldsymbol{\alpha}}_+$)	0.023	0.333

6.5 Identification of putative QTLs

A key step we consider before proceeding with formal identification of putative QTLs, based on the fit of the QTL-MAT model, is to examine the set of additive genetic effects in more detail across the genome. We propose a graphical exploration of the genome-wide additive genetic profiles of the effects associated with the additive genetic regression for both the marker × treatment and genotype × treatment effects. This is analogous to the genome-wide scan conducted in CIM and we have found that it can provide useful informal information in determining regions of interest for a particular trait.

Another useful graphical tool is to plot the E-BLUPs of $\boldsymbol{u}_{m_+}$ against the E-BLUPs of $\boldsymbol{u}_{m_-}$. This plot allows identification of genotypes which may provide information to support the genomic regions identified in the previous plot, and those identified using the formal approaches we present later in this section.

Figure 1 presents the genome-wide trellis plot of the E-BLUPs of $\boldsymbol{\alpha}_-$ and $\boldsymbol{\delta}_\alpha$ for each of the 21 chromosomes from the QTL-MAT model. This plot provides a useful overview of the contrasting genetic profiles for the pleiotropic and specific effects associated with both treatments and the deviations from the genetic regression respectively. Our focus is on the latter. There appears to be little evidence of putative QTLs on the D genome, while there appears to be regions of interest on chromosomes 1A, 3B and 5A. The trough in the profile on 5A is close to a peak for $\boldsymbol{\alpha}_-$. There appears to be some separation between these, although this warrants further formal examination.

Figure 2 presents the genome-wide trellis plot of the E-BLUPs of $\boldsymbol{\alpha}_-$ and $\boldsymbol{\delta}_\alpha$ for each of the 21 chromosomes from the QTL-ide model. There is good agreement between this plot and the plot for the QTL-MAT model apart from the (assumed) smoothness of the process for the QTL-MAT model.

Figure 3 presents the scatter plot of the E-BLUPs of $\boldsymbol{u}_{m_+}$ against the E-BLUPs of $\boldsymbol{u}_{m_-}$. We refer to these E-BLUPs as mGBLUPs, as they are analogous to the GBLUPs obtained from the QTL-ide model. The genetic regression accounts for 47.5 of the total variation in $\boldsymbol{u}_{m_+}$. Four DH lines are labelled on the plot. These correspond to DH lines which exhibit large (absolute) deviations from the genetic regression.

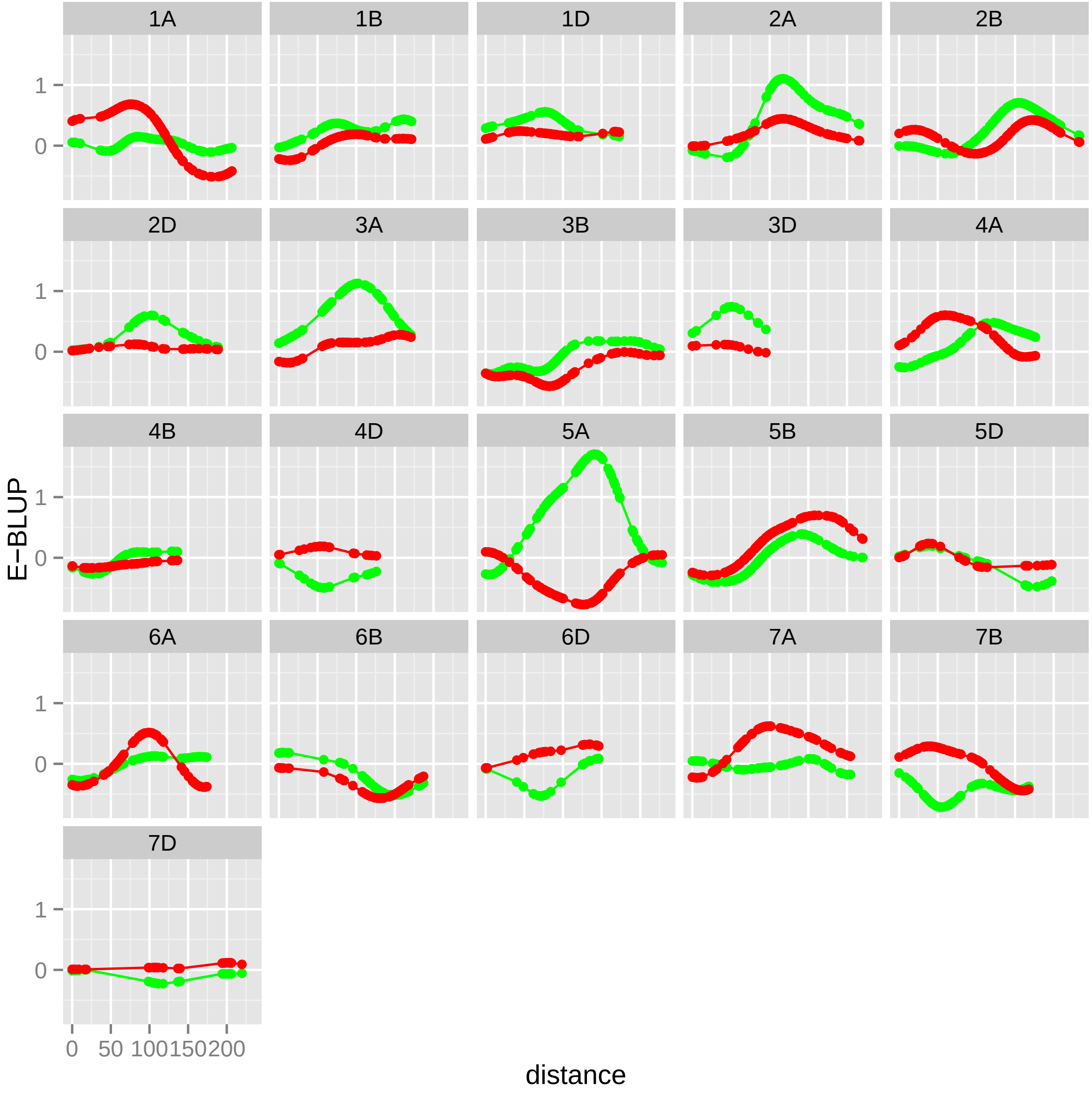

Figure 1: Genome-wide trellis plot of the E-BLUPs of $\boldsymbol{\alpha}_{-}$ (green) and $\boldsymbol{\delta}_{\alpha}$ (red) for each of the 21 chromosomes from the QTL-MAT model.

Table 3 presents the chromosome contributions to the mGBLUP for the four DH lines which were labelled in figure 3, while figure 4 presents the marker profiles for these DH lines in a trellis plot across the 21 chromosomes as ancillary information to assist in the interpretation of these contributions. Although there is some agreement in the relative magnitude of the absolute contributions for each of the four lines across the 21 chromosomes, it is clear that there may be (minor) QTLs present on chromosomes other than 1A, 3B and 5A.

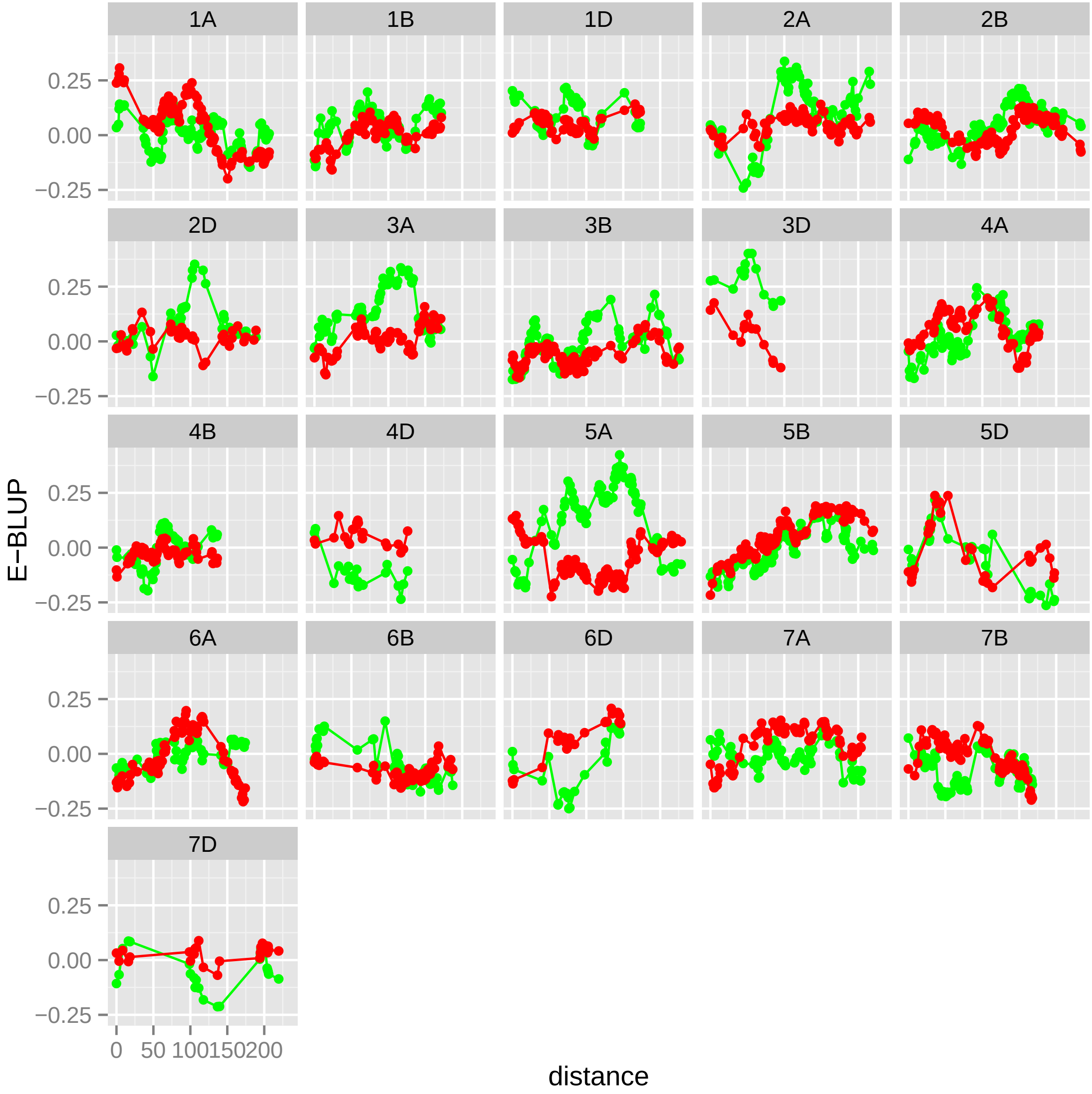

Figure 2: Genome-wide trellis plot of the E-BLUPs of $\boldsymbol{\alpha}_{-}$ (green) and $\boldsymbol{\delta}_{\alpha}$ (red) for each of the 21 chromosomes from the QTL-ide model.

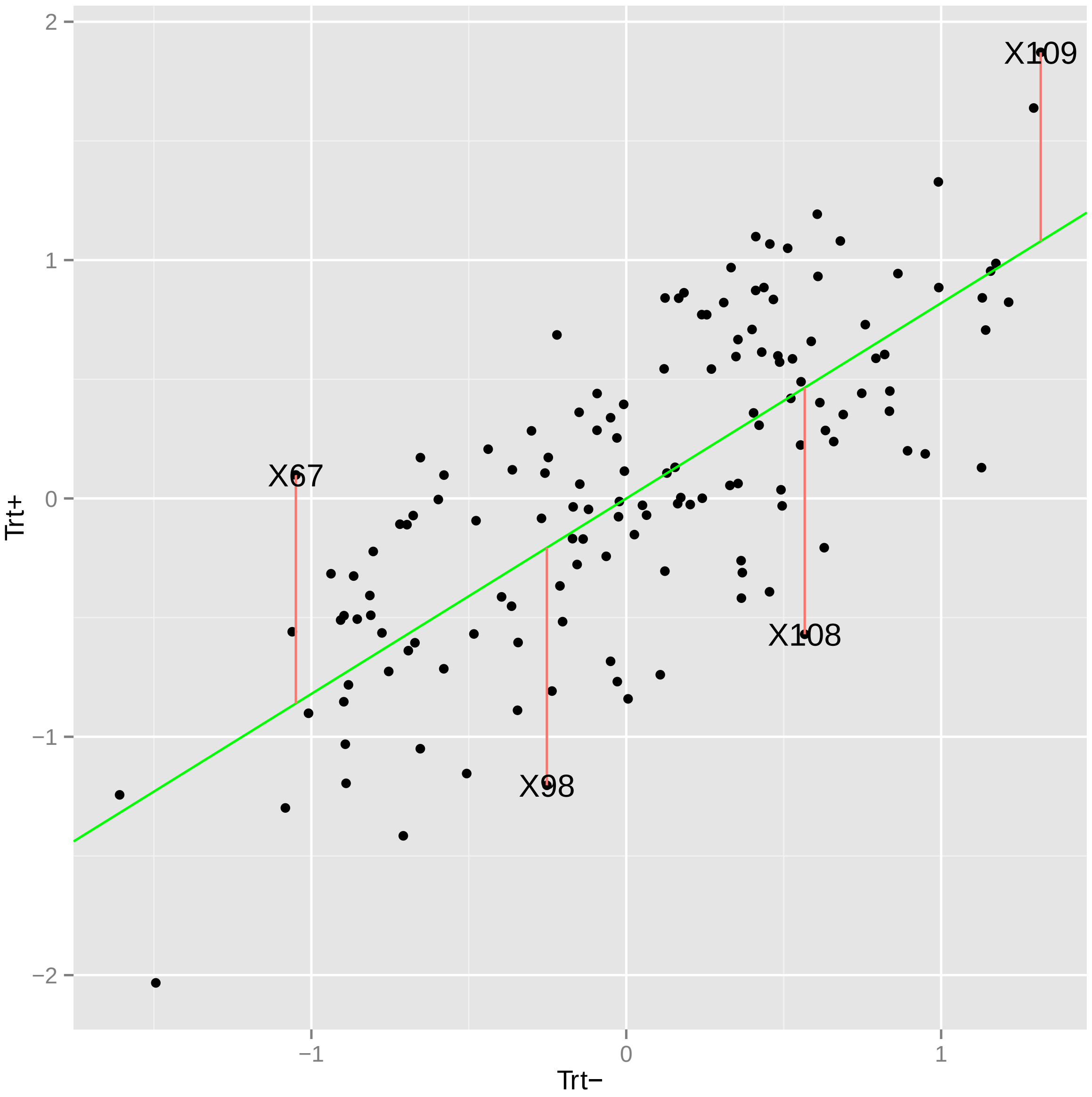

Figure 3: Scatter plot of the E-BLUPs of $\boldsymbol{u}_{m_+}$ against the E-BLUPs of $\boldsymbol{u}_{m_-}$ for the QTL-MAT model. The four points which are labelled correspond to DH lines which have large deviations from the genetic regression.

Table 3: Chromosome contributions to the overall mGBLUP for four selected DH lines.

	X67	X109	X108	X98
1A	0.186	0.066	-0.250	-0.134
1B	0.057	0.024	-0.048	0.026
1D	-0.047	0.043	-0.018	-0.059
2A	-0.092	0.112	0.054	-0.105
2B	0.069	0.063	-0.153	0.124
2D	0.017	0.018	0.010	-0.012
3A	-0.046	-0.047	0.053	-0.006
3B	0.188	0.162	-0.184	-0.213
3D	-0.004	0.007	0.007	-0.004
4A	0.125	-0.070	-0.141	-0.155
4B	0.040	0.043	0.043	-0.043
4D	-0.016	-0.013	-0.014	0.013
5A	0.222	-0.069	-0.241	-0.075
5B	0.157	0.102	0.139	0.001
5D	0.010	-0.001	0.023	0.013
6A	0.069	0.005	0.056	-0.143
6B	-0.137	0.135	-0.056	-0.134
6D	0.031	-0.031	-0.001	0.003
7A	0.139	0.120	-0.188	-0.034
7B	-0.017	0.132	-0.132	-0.069
7D	0.009	-0.009	0.006	0.009
Total	0.959	0.792	-1.035	-0.997

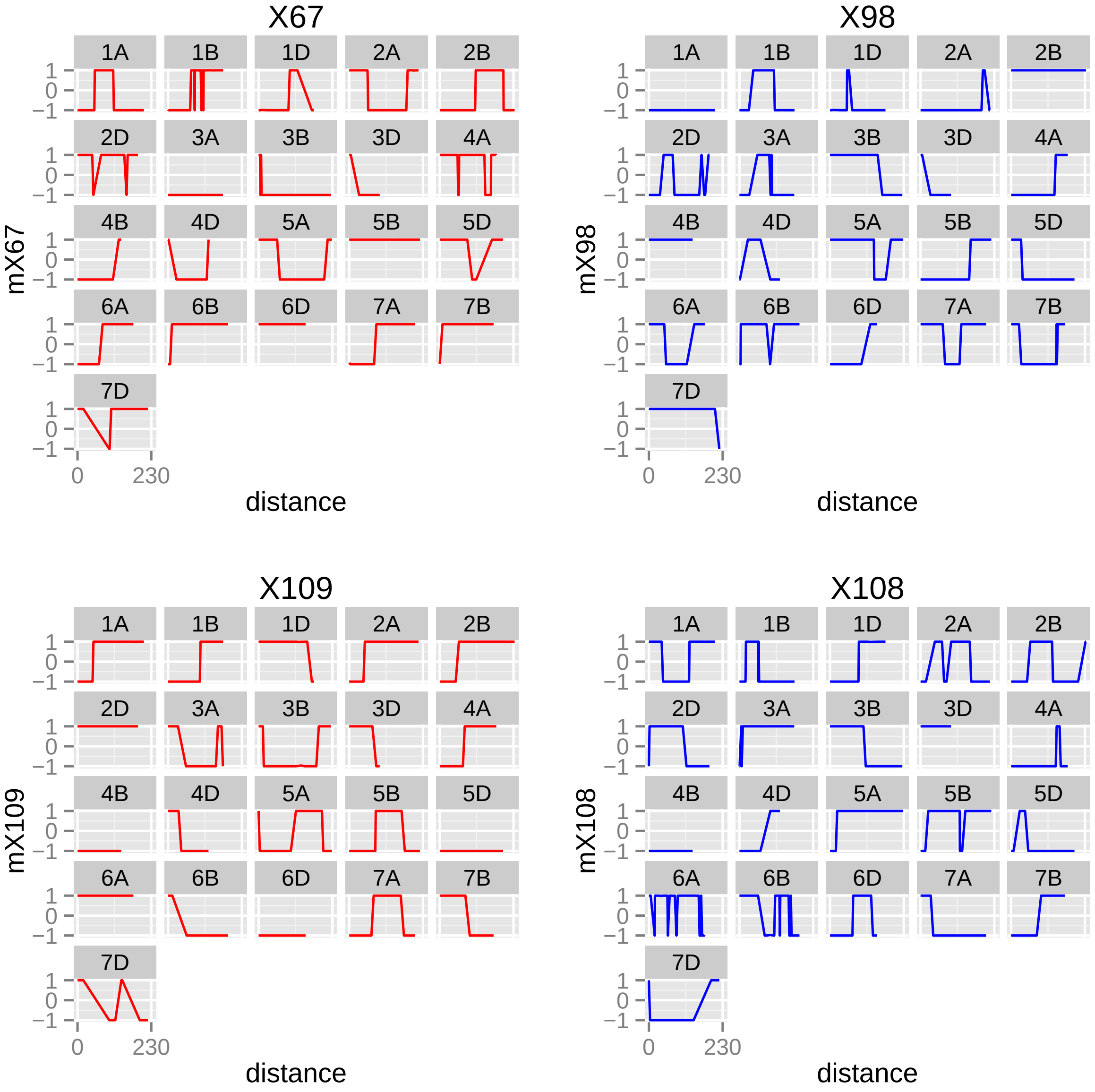

Figure 4: Trellis plot of the marker profiles across the 21 chromosomes for the four selected DH lines which have large deviations from the genetic regression.

APPENDIX

Derivations for the genotype to marker model

To simplify the following we consider the genotype model rather than the genotype joint regression model. The latter model involves a total of four random terms associated with the additive and residual genetic reduced rank, plus diagonal '+' treatment respectively. The former has two components, one associated with the additive and residual genetic terms respectively. Our derivation is based on the joint-likelihood approach used by Henderson (1975) in obtaining the mixed model equations.

The log joint-density for the genotype model is given by

$$\ell(\boldsymbol{y}, \boldsymbol{u}_m, \boldsymbol{u}_e, \boldsymbol{\alpha}) = \ell(\boldsymbol{y} \mid \boldsymbol{u}_m, \boldsymbol{u}_e) + \ell(\boldsymbol{u}_m \mid \boldsymbol{\alpha}) + \ell(\boldsymbol{u}_e) + \ell(\boldsymbol{\alpha})$$

Since $\boldsymbol{u}_m = \boldsymbol{M}_T\boldsymbol{\alpha}$ where $\boldsymbol{M}_T = \boldsymbol{I}_2 \otimes \boldsymbol{M}$ then $\ell(\boldsymbol{u}_m \mid \boldsymbol{\alpha})$ is a constant and so

$$\ell(\boldsymbol{y}, \boldsymbol{u}_m, \boldsymbol{u}_e, \boldsymbol{\alpha}) = \ell(\boldsymbol{y} \mid \boldsymbol{u}_m, \boldsymbol{u}_e) + \ell(\boldsymbol{u}_e) + \ell(\boldsymbol{\alpha}) \tag{17}$$

To maximise (17) with respect to $(\boldsymbol{u}_m^\top, \boldsymbol{\alpha}^\top, \boldsymbol{\tau}^\top, \boldsymbol{u}_e^\top)^\top$ subject to the constraint $\boldsymbol{u}_m = \boldsymbol{M}_T\boldsymbol{\alpha}$ we introduce a vector of Lagrangian multipliers, which yield the quantity, ignoring constants,

$$Q = -\frac{1}{2}[\boldsymbol{e}^\top\boldsymbol{R}^{-1}\boldsymbol{e} + \boldsymbol{u}_e^\top\boldsymbol{G}_{ee}^{-1}\boldsymbol{u}_e + \boldsymbol{\alpha}^\top\boldsymbol{G}_{\alpha\alpha}^{-1}\boldsymbol{\alpha}] - \boldsymbol{a}^\top(\boldsymbol{u}_m - \boldsymbol{M}_T\boldsymbol{\alpha}) \tag{18}$$

Differentiation of Q with respect to $(\boldsymbol{u}_m^\top, \boldsymbol{\alpha}^\top, \boldsymbol{\tau}^\top, \boldsymbol{u}_e^\top)^\top$ and setting these quantities to zero yields

$$\begin{bmatrix} \boldsymbol{Z}_g^\top\boldsymbol{R}^{-1}\boldsymbol{Z}_g & \boldsymbol{I}_m & \boldsymbol{0} & \boldsymbol{Z}_g^\top\boldsymbol{R}^{-1}\boldsymbol{Z}_g & \\ \boldsymbol{I}_m & \boldsymbol{0} & -\boldsymbol{M}_T & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{M}_T^\top & \boldsymbol{G}_{\alpha\alpha}^{-1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{X}^\top\boldsymbol{R}^{-1}\boldsymbol{Z}_g & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{X}^\top\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^\top\boldsymbol{R}^{-1}\boldsymbol{Z}_g \\ \boldsymbol{Z}_g^\top\boldsymbol{R}^{-1}\boldsymbol{Z}_g & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{Z}_g^\top\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}_g^\top\boldsymbol{R}^{-1}\boldsymbol{Z}_g + \boldsymbol{G}_{ee}^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{u}}_m \\ \tilde{\boldsymbol{a}} \\ \tilde{\boldsymbol{\alpha}} \\ \tilde{\boldsymbol{\tau}} \\ \tilde{\boldsymbol{u}}_e \end{bmatrix} \begin{bmatrix} \boldsymbol{Z}_g^\top\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{X}^\top\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}_g^\top\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix}$$

Absorption of $(\tilde{\boldsymbol{\tau}}^\top, \tilde{\boldsymbol{u}}_e^\top)^\top$ into the equations for $(\tilde{\boldsymbol{u}}_m^\top, \tilde{\boldsymbol{a}}^\top, \tilde{\boldsymbol{\alpha}}^\top)^\top$ yields

$$\begin{bmatrix} \boldsymbol{Z}_g^\top\boldsymbol{S}\boldsymbol{Z}_g & \boldsymbol{I}_m & \boldsymbol{0} \\ \boldsymbol{I}_m & \boldsymbol{0} & -\boldsymbol{M}_T \\ \boldsymbol{0} & -\boldsymbol{M}_T^\top & \boldsymbol{G}_{\alpha\alpha}^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{u}}_m \\ \tilde{\boldsymbol{a}} \\ \tilde{\boldsymbol{\alpha}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z}_g^\top\boldsymbol{S}\boldsymbol{y} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix} \tag{19}$$

where $\boldsymbol{S} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{W}_0\boldsymbol{C}_0^{-1}\boldsymbol{W}_0^\top\boldsymbol{R}^{-1}$, $\boldsymbol{W}_0 = [\boldsymbol{X}\ \boldsymbol{Z}_g]$, $\boldsymbol{C}_0 = \boldsymbol{W}_0^\top\boldsymbol{R}^{-1}\boldsymbol{W}_0 + \boldsymbol{G}_0^*$ and

$$\boldsymbol{G}_0^* = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}_{ee}^{-1} \end{bmatrix}$$

From equation 19 we see that

$$\begin{aligned} \tilde{\boldsymbol{u}}_m &= \boldsymbol{M}_T\tilde{\boldsymbol{\alpha}} \\ \boldsymbol{G}_{\alpha\alpha}^{-1}\tilde{\boldsymbol{\alpha}} &= \boldsymbol{M}_T^\top\tilde{\boldsymbol{a}} \end{aligned} \tag{20}$$

Absorbing $\tilde{\boldsymbol{\alpha}}$ into the equations for $(\tilde{\boldsymbol{u}}_m^\top, \tilde{\boldsymbol{a}}^\top)^\top$ yields

$$\begin{bmatrix} \boldsymbol{Z}_g^\top \boldsymbol{S} \boldsymbol{Z}_g & \boldsymbol{I}_m \\ \boldsymbol{I}_m & -\boldsymbol{G}_{mm}^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{u}}_m \\ \tilde{\boldsymbol{a}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z}_g^\top \boldsymbol{S} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} \tag{21}$$

and hence

$$\tilde{\boldsymbol{a}} = \boldsymbol{G}_{mm}^{-1} \tilde{\boldsymbol{u}}_m \tag{22}$$

Finally absorbing the equation for $\tilde{\boldsymbol{a}}$ into the equation for $\tilde{\boldsymbol{u}}_m$ yields

$$(\boldsymbol{Z}_g^\top \boldsymbol{S} \boldsymbol{Z}_g + \boldsymbol{G}_{mm}^{-1}) \tilde{\boldsymbol{u}}_m = \boldsymbol{Z}_g^\top \boldsymbol{S} \boldsymbol{y}$$

and

$$\text{pev}\,(\tilde{\boldsymbol{u}}_m) = (\boldsymbol{Z}_g^\top \boldsymbol{S} \boldsymbol{Z}_g + \boldsymbol{G}_{mm}^{-1})^{-1}$$

as required. Using 22 and 20 we have

$$\begin{aligned} \tilde{\boldsymbol{\alpha}} &= \boldsymbol{G}_{\alpha\alpha} \boldsymbol{M}_T^\top \boldsymbol{G}_{mm}^{-1} \tilde{\boldsymbol{u}}_m \\ &= (\boldsymbol{G}_{T_\alpha} \otimes \boldsymbol{D})(\boldsymbol{I}_2 \otimes \boldsymbol{M}^\top)(\boldsymbol{G}_{T_\alpha}^{-1} \otimes \boldsymbol{K}^{-1}) \tilde{\boldsymbol{u}}_m \\ \Rightarrow \quad \begin{bmatrix} \tilde{\boldsymbol{\alpha}}_- \\ \tilde{\boldsymbol{\alpha}}_+ \end{bmatrix} &= \begin{bmatrix} \boldsymbol{D} \boldsymbol{M}^\top \boldsymbol{K}^{-1} \tilde{\boldsymbol{u}}_{m_-} \\ \boldsymbol{D} \boldsymbol{M}^\top \boldsymbol{K}^{-1} \tilde{\boldsymbol{u}}_{m_+} \end{bmatrix} \end{aligned}$$

It follows that using results on the inverse of partitioned matrices that

$$\begin{aligned} \text{pev}\,(\tilde{\boldsymbol{\alpha}}) &= \boldsymbol{G}_{\alpha\alpha} - \boldsymbol{G}_{\alpha\alpha} \boldsymbol{M}_T^\top \boldsymbol{G}_{mm}^{-1} \boldsymbol{M}_T \boldsymbol{G}_{\alpha\alpha} + \\ &= \boldsymbol{G}_{\alpha\alpha} \boldsymbol{M}_T^\top \boldsymbol{G}_{mm}^{-1} \text{pev}\,(\tilde{\boldsymbol{u}}_m)\, \boldsymbol{G}_{mm}^{-1} \boldsymbol{M}_T \boldsymbol{G}_{\alpha\alpha} \\ &= \boldsymbol{G}_{\alpha\alpha\cdot} + (\boldsymbol{I}_2 \otimes \boldsymbol{D} \boldsymbol{M}^\top \boldsymbol{K}^{-1}) \text{pev}\,(\tilde{\boldsymbol{u}}_m)\, (\boldsymbol{I}_2 \otimes \boldsymbol{K}^{-1} \boldsymbol{M} \boldsymbol{D}) \end{aligned} \tag{23}$$

where $\boldsymbol{G}_{\alpha\alpha\cdot} = \boldsymbol{G}_{\alpha\alpha} - \boldsymbol{G}_{\alpha\alpha} \boldsymbol{M}_T^\top \boldsymbol{G}_{mm}^{-1} \boldsymbol{M}_T \boldsymbol{G}_{\alpha\alpha}$. Hence from equation 23

$$\begin{aligned} \text{pev}\,(\tilde{\boldsymbol{\alpha}}_-) &= \lambda_{\alpha_-}^2 \boldsymbol{D}_{\alpha\alpha\cdot} + \boldsymbol{D} \boldsymbol{M}^\top \boldsymbol{K}^{-1}) \text{pev}\left(\tilde{\boldsymbol{u}}_{m_-}\right) \boldsymbol{K}^{-1} \boldsymbol{M} \boldsymbol{D}) \\ \text{pev}\,(\tilde{\boldsymbol{\alpha}}_+) &= (\lambda_{\alpha_+}^2 + \psi_\alpha) \boldsymbol{D}_{\alpha\alpha\cdot} + \boldsymbol{D} \boldsymbol{M}^\top \boldsymbol{K}^{-1}) \text{pev}\left(\tilde{\boldsymbol{u}}_{m_+}\right) \boldsymbol{K}^{-1} \boldsymbol{M} \boldsymbol{D}) \end{aligned}$$

References

Borg, L., Smith, A., Taylor, J., Cullis, B., & Statistics, A. (2015). Statistics for the Australian Grains Industry Technical Report Series Osmotic stress experiment for Cranbrook Halberd mapping population. Technical report, University of Wollongong, Wolllongong.

Butler, D., Cullis, B., Gilmour, A., & Gogel, B. J. (2009). ASReml-R Reference Manual, Release 3.

Cullis, B. R., Smith, A. B., & Coombes, N. E. (2006). On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological and Environmental Statistics* **11**, 381–393.

de Los Campos, G., Gianola, D., & Rosa, G. J. (2009). Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *Journal of Animal Science* **87**, 1883–1887.

Genc, Y., Oldach, K., Verbyla, A., Lott, G., Hassan, M., Tester, M., Wallwork, H., & McDonald, G. (2010). Sodium exclusion QTL associated with improved seedling growth in bread wheat under salinity stress. *Theoretical and Applied Genetics* **121**, 877–894.

Gianola, D., Perez-Enciso, M., & Toro, M. a. (2003). On marker-assisted prediction of genetic value: Beyond the ridge. *Genetics* **163**, 347–365.

Gilmour, A., Cullis, B., Welham, S., Gogel, B., & Thompson, R. (2004). An efficient computing strategy for prediction in mixed linear models. *Computational Statistics and Data Analysis* **44**, 571–586.

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). {AI}, an efficient algorithm for {REML} estimation in linear mixed models. *Biometrics* **51**, 1440–1450.

Green, P. J. & Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.

Haskard, K. A. (2005). *Anisotropic Matérn correlation and other issues in model-based geostatistics*. PhD thesis, BiometricsSA, University of Adelaide.

Haskard, K. A., Cullis, B. R., & Verbyla, A. P. (2007). Anisotropic Matérn correlation and spatial prediction using REML. *Journal of Agricultural and Biological Sciences* **12**, 147–160.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423–477.

Kammann, E. E. & Wand, M. P. (2003). Geoadditive models. *Applied Statistics* **52**, 1–18.

Kammholz, S. J., Campbell, a. W., Sutherland, M. W., Hollamby, G. J., Martin, P. J., Eastwood, R. F., Barclay, I., Wilson, R. E., Brennan, P. S., & Sheppard, J. a. (2001). Establishment and characterisation of wheat genetic mapping populations. *Australian Journal of Agricultural Research* **52**, 1079–1088.

Linsell, K. J., Rahman, M. S., Taylor, J. D., Davey, R. S., Gogel, B. J., Wallwork, H., Forrest, K. L., Hayden, M. J., Taylor, S. P., & Oldach, K. H. (2014). QTL for resistance to root lesion nematode (Pratylenchus thornei) from a synthetic hexaploid wheat source. *Theoretical and Applied Genetics* **127**, 1409–1421.

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829.

Ober, U. (2010). Kernel-Based BLUP with Genomic Data. In *9th World Congress on Genetics Applied to Livestock Production.*, pages 2–5.

Patterson, H. D. & Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* **31**, 545–554.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86**, 677–690.

R Core Team (2015). R: A Language and Environment for Statistical Computing.

Smith, A., Cullis, B. R., & Thompson, R. (2001). Analysing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* **57**, 1138–1147.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag, New York.

Stram, D. O. & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects setting. *Biometrics* **50**, 1171–1177.

Taylor, J. & Verbyla, A. (2011). R Package wgaim : QTL analysis in bi-parental populations using linear mixed models. *Journal of Statistical Software* **40**, 1–19.

Taylor, J. D. & Butler, D. (2014). ASMap: An (A)ccurate and (S)peedy linkage map construction package for inbred populations that uses the extremely efficient MSTmap algorithm.

Thompson, R. (1985). A Note on Restricted Maximum Likelihood Estimation with an Alternative Outlier Model. *Journal of the Royal Statistical Society. Series B (Methodological)* **47**, 53–55.

Thompson, R., Cullis, B., Smith, A., & Gilmour, A. (2003). A Sparse Implementation of the Average Information Algorithm for Factor Analytic and Reduced Rank Variance Models. *Australian & New Zealand Journal of Statistics* **45**, 445–459.

VanRaden, P. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* **91**, 4414–4423.

Verbyla, A. P., Cullis, B. R., & Thompson, R. (2007). The analysis of QTLs by simultaneous use of the full linkage map. *Theoretical and Applied Genetics* **116**, 95–111.

Verbyla, A. P., Taylor, J. D., & Verbyla, K. L. (2012). RWGAIM: An efficient high dimensional random whole genome average QTL interval mapping approach. *Genetics Research* **94**, 291–306.

Wahba, G. (1990). *Spline models for observational data.* SIAM:Philadelphia.

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., Maccaferri, M., Salvi, S., Milner, S. G., Cattivelli, L., Mastrangelo, A. M., Whan, A., Stephen, S., Barker, G., Wieseke, R., Plieske, J., Lillemo, M., Mather, D., Appels, R., Dolferus, R., Brown-Guedira, G., Korol, A., Akhunova, A. R., Feuillet, C., Salse, J., Morgante, M., Pozniak, C., Luo, M.-C., Dvorak, J., Morell, M., Dubcovsky, J., Ganal, M., Tuberosa, R., Lawley, C., Mikoulitch, I., Cavanagh, C., Edwards, K. J., Hayden, M., & Akhunov, E. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90000 single nucleotide polymorphism array. *Plant Biotechnology Journal* **12**, 787–796.

Webster, R. & Oliver, M. A. (2001). *Geostatistics for Environmental Scientists.* John Wiley and Sons, Chichester.

Wilkinson, G. N. & Rogers, C. E. (1973). Symbolic description of factorial models for analysis of variance. *Applied Statistics* **22**, 392–399.

Yang, W. & Tempelman, R. J. (2010). A Bayesian Antedependence Model to Account for Linkage Disequilibrium in Whole Genome Selection. In *9th World Congress on Genetics Applied to Livestock Production.*, pages 3–6.

Local Genomic Selection

Alison Smith, Lauren Borg and Brian Cullis
National Institute for Applied Statistics and Research Australia
School of Mathematics and Applied Statistics
University of Wollongong

email: alismith@uow.edu.au

November 14, 2016

Given that the markers explained nearly all of the genetic variance and this variance has a significant impact on the trait, we seek to identify the most influential regions. This is where the emphasis will shift away from prviding LOD scores and $p-$values for single markers being linked to QTL and more towards the concept of "Local Genomic Selection". The key tools for explanation are the marker profiles as already included in the paper (Figures 11 and 10) and included here as Figures 1 and 2. Note that I have changed the scale for the y-axis on these graphs to correspond to unscaled marker scores of ±1 whereas previously these effects were scaled for the purposes of analysis. The use of the original scale makes the total effect easier to visualise since we compute the total effect for each "region" as the sum of all marker effects in that region. Regions were defined to be runs of all positive or all negative marker effects. Thus for example, for LG 1A on Figure 1 there are two regions, whereas for LG 1D there is one region (the entire LG). A 90% confidence interval was computed for the total effect for each region.

The results for the regions for SGNTol are summarised in Table 1. Only those regions containing at least 30 markers are included. The region with the largest (absolute) total effect for tolerance was on 5A and comprised the 76 markers located between 31 and 204 cM. The confidence interval does not contain zero so there is evidence to suggest the existence of a putative QTL within this region. The peak effect occurs at 126 cM (marker IWA5668(C)) as previously noted (and as given in Table 5 of submitted paper). The total effect for this region is -0.238 which means that a hypothetical variety with Cranbrook type for all markers in this region would have a loss of 0.238 "units" (ie. on the transformed SGN scale) over the population mean tolerance value. To aid with interpretation, consider Brian's original report. Figure 4 in that report shows that DH X108 had Cranbrook type for all markers in this region, so it exhibits the behaviour of the proposed hypothetical variety. The contribution to the GBLUP for X108 for tolerance for the entire LG was -0.241 (see Table 3 of Brian's original report) which is consistent with the value of -0.238 in Table 1 (which relates only to the region between 31 and 204cM).

The results for the regions for SGNCon are summarised in Table 2.

To make the interpretation even clearer, we can back-transform the total effect (and the confidence limits) to the original untransformed scale of the data (ie. actual numbers of grains per spikelet). We can then compute the predicted difference in grain number between two hypothetical varieties that differ only in that one has the Cranbrook type for all markers in the region, and the other has the Halberd type. This information is given in Tables 3 and 4. As an example, consider the region on 5A between 39 and 204 cM for the control treatment trait. Table 2 suggests that a hypothetical variety with Cranbrook type for all markers in this region would have a benefit of 0.516 "units" (ie. on the transformed SGN scale) over the population mean control treatment value. When back-transformed (see Table 4), this means that a variety which has the Cranbrook type for all markers on 5A in this region is expected, on average, to have 4.7 more grains per spikelet than a variety which has the Halberd type for all markers in this region, all other things being equal. The confidence interval tells us that the expected difference is likely to lie between 2.6 and 6.8.

Table 1: Total effects, together with lower and upper limits for a 90% confidence interval for SGNTol for all regions containing more than 30 markers. Regions are identified by the linkage group and the starting and ending marker position (cM) within the group. The number of markers in each region is given. Final two columns identify the position (marker name and distance) of the maximum absolute effect within the region. The regions are given in decreasing order of the absolute magnitude of the total effects.

LG	Effect	Lower	Upper	Start	End	#markers	Name	cM
5A	-0.238	-0.467	-0.009	31	204	76	IWA5668(C)	126
3B	-0.220	-0.474	0.033	0	226	97	IWB59720(C)	83
1A	0.220	0.018	0.422	0	128	69	IWB47804(C)	76
5B	0.195	-0.015	0.406	71	221	71	IWB31506(C)	163
7A	0.171	-0.037	0.379	40	205	65	IWA3557	99
4A	0.148	-0.045	0.341	0	142	58	IWA7058(C)	60
6B	-0.143	-0.328	0.042	0	187	60	IWB35399(C)	131
2A	0.113	-0.092	0.318	15	216	67	IWB48486(C)	116
2B	0.109	-0.084	0.302	128	234	59	IWB39236(C)	172
1A	-0.086	-0.230	0.058	130	207	37	IWB34474	181
7B	-0.077	-0.231	0.077	117	168	36	IWB73668(C)	161
6A	0.075	-0.065	0.214	57	118	33	IWB23877(C)	99
1D	0.065	-0.115	0.245	0	173	53	IWB14612	44
7B	0.062	-0.124	0.247	0	108	51	IWB12630	40
3A	0.060	-0.129	0.249	56	171	53	IWB67246(C)	156
1B	0.051	-0.143	0.245	55	172	60	IWB7953(C)	99
4B	-0.043	-0.240	0.153	0	136	58	IWB23112(C)	24
5B	-0.039	-0.173	0.096	0	68	32	IWB43086(C)	24
2D	0.018	-0.128	0.164	0	189	41	IWA6520(C)	81

Table 2: Total effects, together with lower and upper limits for a 90% confidence interval for SGNCon for all regions containing more than 30 markers. Regions are identified by the linkage group and the starting and ending marker position (cM) within the group. The number of markers in each region is given. Final two columns identify the position (marker name and distance) of the maximum absolute effect within the region. The regions are given in decreasing order of the absolute magnitude of the total effects.

LG	Effect	Lower	Upper	Start	End	#markers	Name	cM
5A	0.516	0.287	0.745	39	204	76	IWB55564(C)	141
3A	0.286	0.070	0.502	0	171	67	IWB30485(C)	102
2A	0.271	0.078	0.464	67	216	58	IWB48486(C)	116
7B	-0.254	-0.497	-0.012	0	168	87	IWB40092(C)	54
2B	0.228	0.016	0.441	91	234	81	IWB26048(C)	154
1B	0.131	-0.081	0.342	14	172	74	IWB66475(C)	161
1D	0.126	-0.061	0.313	0	173	53	IWB7914(C)	77
6B	-0.116	-0.284	0.053	84	187	48	IWB21973(C)	149
3B	-0.113	-0.325	0.099	0	102	69	IWB23456	7
4A	0.089	-0.086	0.264	65	176	47	IWB41760(C)	122
5B	-0.085	-0.258	0.089	0	90	49	IWA6947(C)	32
2D	0.078	-0.085	0.240	0	189	41	IWA7332	102
5B	0.069	-0.118	0.257	91	221	54	IWB36364(C)	141
4B	-0.037	-0.182	0.107	0	62	31	IWB45064	27
1A	0.033	-0.144	0.210	62	143	57	IWA5568(C)	85
6A	-0.032	-0.175	0.112	0	66	32	IWB64917	8
6A	0.029	-0.148	0.206	67	174	44	IWB33661(C)	107
7A	-0.014	-0.179	0.151	30	121	39	IWB42745(C)	65

Table 3: Differences in number of grains per spikelet between Cranbrook and Halberd types, together with lower and upper limits for a 90% confidence interval for SGNToI for all regions containing more than 30 markers. Regions are identified by the linkage group and the starting and ending marker position (cM) within the group. The number of markers in each region is given. Final two columns identify the position (marker name and distance) of the maximum absolute effect within the region. The regions are given in decreasing order of the absolute magnitude of the total effects.

LG	Difference	Lower	Upper	Start	End	#markers	Name	cM
5A	-2.2	-4.3	-0.1	31	204	76	IWA5668(C)	126
3B	-2.0	-4.3	0.3	0	226	97	IWB59720(C)	83
1A	2.0	0.2	3.9	0	128	69	IWB47804(C)	76
5B	1.8	-0.1	3.7	71	221	71	IWB31506(C)	163
7A	1.6	-0.3	3.5	40	205	65	IWA3557	99
4A	1.4	-0.4	3.1	0	142	58	IWA7058(C)	60
6B	-1.3	-3.0	0.4	0	187	60	IWB35399(C)	131
2A	1.0	-0.8	2.9	15	216	67	IWB48486(C)	116
2B	1.0	-0.8	2.8	128	234	59	IWB39236(C)	172
1A	-0.8	-2.1	0.5	130	207	37	IWB34474	181
7B	-0.7	-2.1	0.7	117	168	36	IWB73668(C)	161
6A	0.7	-0.6	2.0	57	118	33	IWB23877(C)	99
1D	0.6	-1.1	2.2	0	173	53	IWB14612	44
7B	0.6	-1.1	2.3	0	108	51	IWB12630	40
3A	0.6	-1.2	2.3	56	171	53	IWB67246(C)	156
1B	0.5	-1.3	2.2	55	172	60	IWB7953(C)	99
4B	-0.4	-2.2	1.4	0	136	58	IWB23112(C)	24
5B	-0.4	-1.6	0.9	0	68	32	IWB43086(C)	24
2D	0.2	-1.2	1.5	0	189	41	IWA6520(C)	81

Table 4: Differences in number of grains per spikelet between Cranbrook and Halberd types, together with lower and upper limits for a 90% confidence interval for SGNCon for all regions containing more than 30 markers. Regions are identified by the linkage group and the starting and ending marker position (cM) within the group. The number of markers in each region is given. Final two columns identify the position (marker name and distance) of the maximum absolute effect within the region. The regions are given in decreasing order of the absolute magnitude of the total effects.

LG	Difference	Lower	Upper	Start	End	#markers	Name	cM
5A	4.7	2.6	6.8	39	204	76	IWB55564(C)	141
3A	2.6	0.6	4.6	0	171	67	IWB30485(C)	102
2A	2.5	0.7	4.3	67	216	58	IWB48486(C)	116
7B	-2.3	-4.6	-0.1	0	168	87	IWB40092(C)	54
2B	2.1	0.1	4.0	91	234	81	IWB26048(C)	154
1B	1.2	-0.7	3.1	14	172	74	IWB66475(C)	161
1D	1.2	-0.6	2.9	0	173	53	IWB7914(C)	77
6B	-1.1	-2.6	0.5	84	187	48	IWB21973(C)	149
3B	-1.0	-3.0	0.9	0	102	69	IWB23456	7
4A	0.8	-0.8	2.4	65	176	47	IWB41760(C)	122
5B	-0.8	-2.4	0.8	0	90	49	IWA6947(C)	32
2D	0.7	-0.8	2.2	0	189	41	IWA7332	102
5B	0.6	-1.1	2.4	91	221	54	IWB36364(C)	141
4B	-0.3	-1.7	1.0	0	62	31	IWB45064	27
1A	0.3	-1.3	1.9	62	143	57	IWA5568(C)	85
6A	-0.3	-1.6	1.0	0	66	32	IWB64917	8
6A	0.3	-1.4	1.9	67	174	44	IWB33661(C)	107
7A	-0.1	-1.6	1.4	30	121	39	IWB42745(C)	65

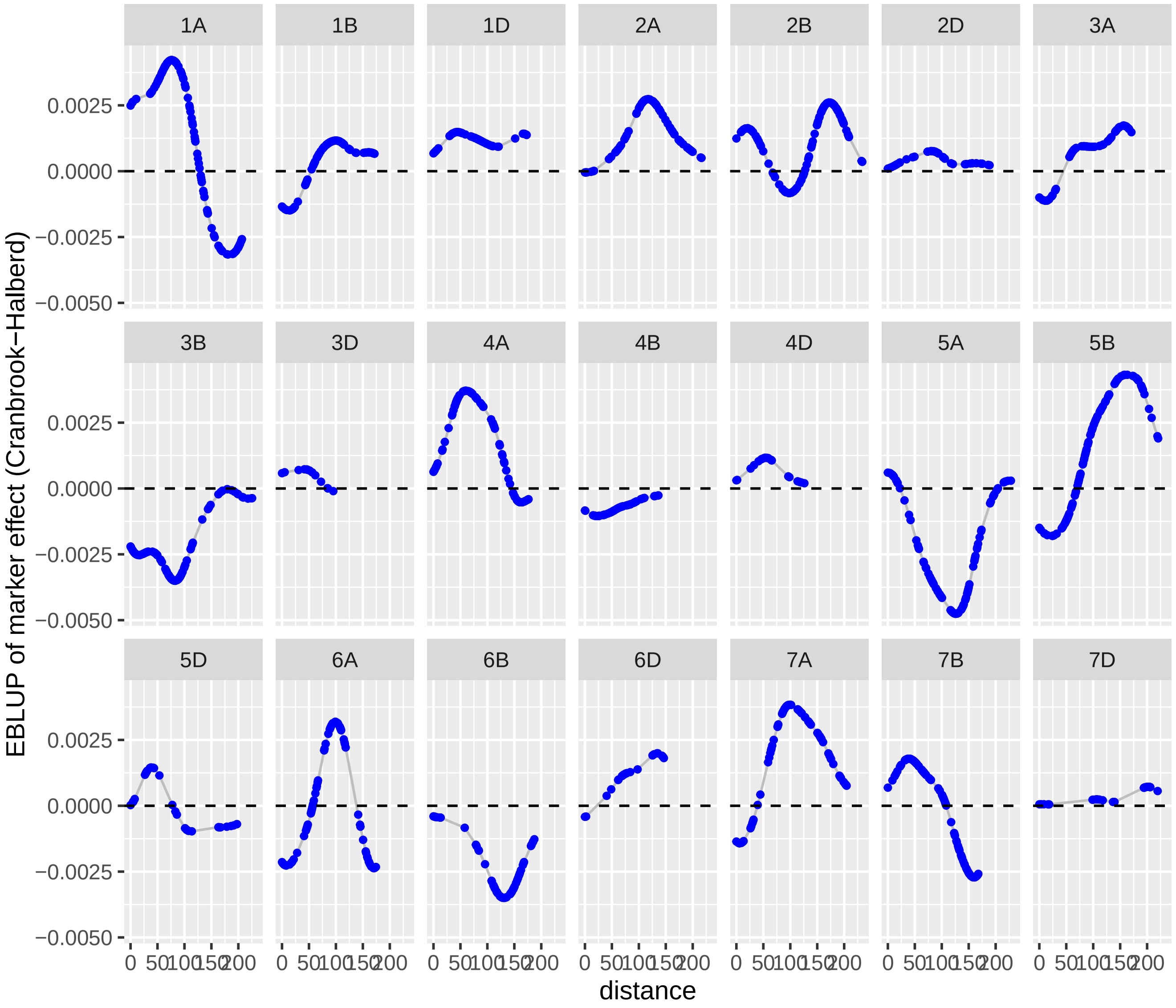

Figure 1: EBLUPs of marker effects for tolerance.

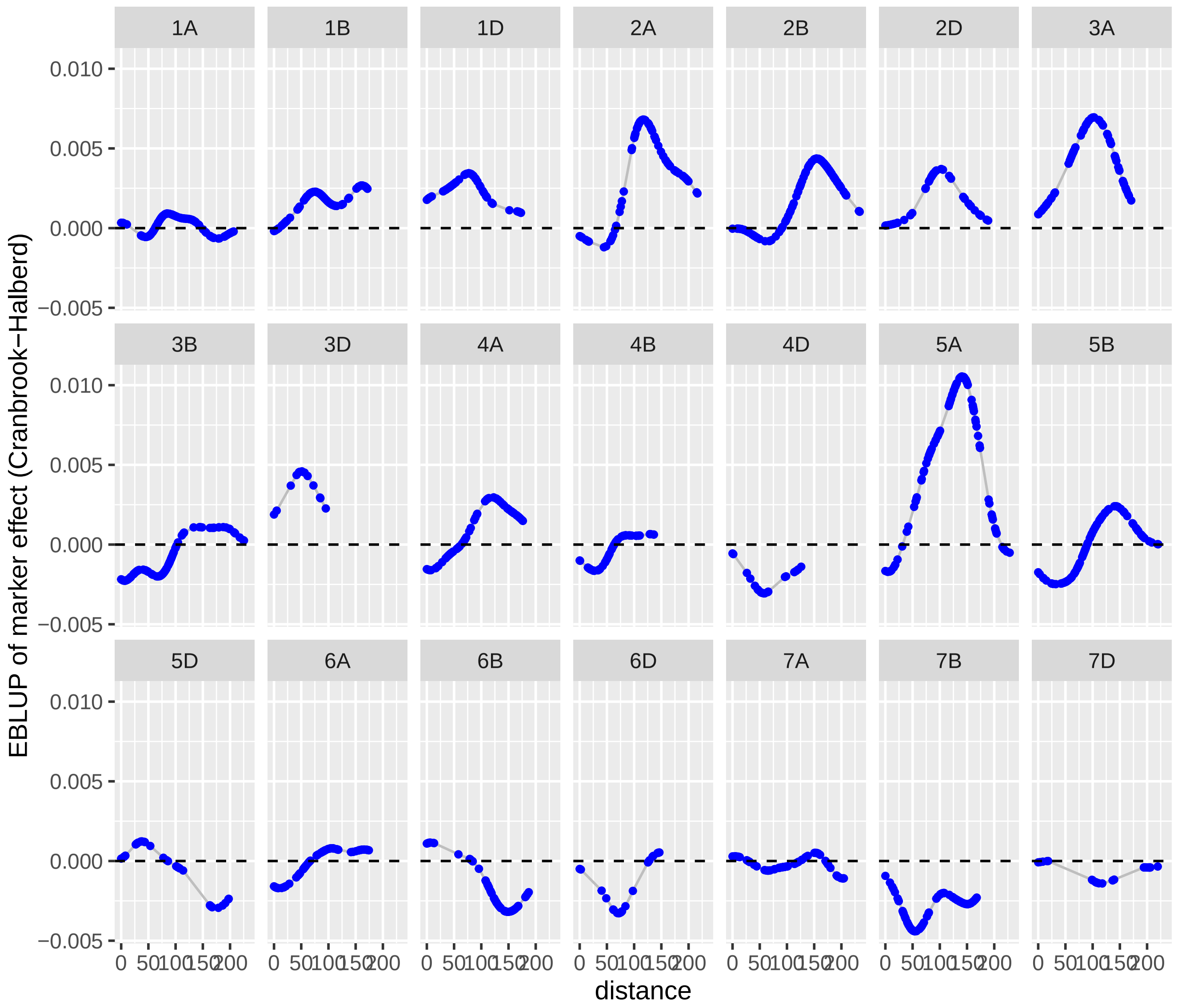

Figure 2: EBLUPs of marker effects for - treatment.