Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

03-11

Detecting QTL for photoperiod sensitivity in a Brassica napus doubled haploid population using a linear mixed model with correlated marker effects

Alison Smith, Brian Cullis and Matthew Nelson

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Detecting QTL for photoperiod sensitivity in a Brassica napus doubled haploid population using a linear mixed model with correlated marker effects

Alison Smith, Honorary Principal Fellow,
Centre for Statistical and Survey Methodology, University of Wollongong
Brian Cullis, Professor of Biometry,
Centre for Statistical and Survey Methodology, University of Wollongong
Matthew Nelson, Research Assistant Professor,
School of Plant Biology, University of Western Australia

June 3, 2011

1 Aims and description of data

Here we describe the analysis of a glasshouse experiment designed to investigate quantitative trait loci (QTL) controlling photoperiod sensitivity in a doubled haploid (DH) population of *Brassica napus*.

A total of 142 DH lines derived from two genetically diverse parents (an Australian spring line, MONTY, and a European spring line, LYNX) was grown in a glasshouse experiment. Note that it was later discovered that some of these DH lines were genetically identical (see section 1.1). Also grown were the parents (M-28DH and L-37DH), direct and reciprocal hybrids (LM F1 and ML F1) and control varieties (CB TRIBUNE, CB TELFER, TOPAS, CAMPINO and WESTAR-10DH) making a total of 151 lines in the experiment. Two treatments were investigated in the glasshouse, namely long daylength (LD) with 16hr photoperiod (enabled with lamps) and 8hr dark period and short daylength (SD) where the plants received sunlight alone (without any additional lighting).

In the glasshouse there were 4 adjacent benches with 200 pots on each arranged in a rectangular array of 20 rows by 10 ranges. The daylength treatments were allocated to benches (2 benches per treatment) with the LD treatment being allocated to the middle two benches to facilitate the additional lighting. This allocation provides one possible realisation of a randomised complete block design for daylength treatments so we regard it as such in the analysis. Within each bench 49 lines were replicated (ie. were grown in 2 pots) and 102 were unreplicated (single pot only). The randomisation was carried out as for a partially replicated design (see Cullis et al., 2006) with blocks aligned with ranges (block 1 corresponding to ranges 1-5; block 2 to ranges 6-10). The replication of lines was balanced as far as possible to ensure a fairly even distribution of total number of pots per line across the full design. This distribution is given in Table 1 for each treatment separately and overall.

Table 1: Number of lines with specified number of pots for each treatment and overall.

Treatment	2 pots	3 pots	4 pots	5 pots	6 pots	8 pots
LD	57	90	4			
SD	82	40	29			
overall			42	50	65	4

The trait of interest in this study is days to first flowering (DFL). We commence with an analysis of this trait to examine sources of variation (see Section 2) then expand the analysis to encompass the detection of QTL (see Section 3). For the latter we had 327 DArT markers classified into 19 linkage groups. For reasons of confidentiality the map is not presented here.

1.1 Final make-up of data-set

As previously stated the experiment as planned contained 151 lines with 142 DH lines. However after the conduct of the experiment it was found that 9 of the DH lines were duplicates of others and so for the purposes of the analysis have been re-named accordingly. Additionally there was no phenotypic data for two DH lines so the final set of lines evaluated consisted of 140 lines with 9 controls and 131 DH lines. Thus in the data-set there are two factors indexing the lines, namely a factor 'Id' that indexes the original 151 lines and a factor 'Id2' that indexes the final 140 lines.

Genotypic data was available for 126 DH lines and was unavailable for the DH lines LM30-013, LM30-014, LM30-052, ML31-029 and ML31-084. This is important for the subsequent analyses. Also note that marker information was fairly complete (89%) for the 126 lines that were genotyped. Missing marker information was imputed using Broman's qtl package in R. This ensures that all marker values are either 1 (corresponding to a LYNX allele) or -1 (corresponding to a MONTY allele).

2 Base-line analysis of DFL data

The aim of this analysis was to provide a suitable base-line linear mixed model for use in the QTL detection analysis. Thus it was important to investigate non-genetic sources of variation in the data. The initial linear mixed model fitted to the DFL data included terms that reflected the randomisation employed in the design. This included random effects for daylength treatment replicate blocks (fitted as the factor 'Rep' with 2 levels); random effects for benches within replicate blocks (fitted as the factor 'Bench' with 4 levels labelled as LD1, LD2, SD1 and SD2); random effects for blocks within benches (fitted as the compound term 'Bench:Block' where Block is a factor with 2 levels indexing blocks within benches). In terms of the residual effects we commenced with a separate spatial covariance model for each bench (thus

there is a separate variance, range and row correlation for each bench).

In terms of the genetic effects we regard the two daylength treatments as two separate "traits" so fit a linear mixed model that allows for a separate genetic variance for each treatment and a genetic correlation between treatments. Given the aim of this experiment we must ensure that these parameters relate only to the DH lines (and only those with genotypic data) rather than the full set of lines tested. This requires a factor called 'Idtype', say, that has 15 levels (9 levels to index the control lines, 5 levels to index the DH lines with no genetic data and a single level for all remaining DH lines). The effects for this factor for each treatment are then fitted as fixed effects in the model. All subsequent references to "genetic variance" and "genetic correlation" relate to the genotyped DH lines only.

The linear mixed model can be written as

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Z_g u_g} + \boldsymbol{Z_1 u_1} + \boldsymbol{Z_2 u_2} + \boldsymbol{Z_3 u_3} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{y}$ is the $n \times 1$ vector of days to flowering data where $n = 800$ (the number of data records); $\boldsymbol{\tau}$ is the vector of fixed effects (Idtype effects for each treatment); $\boldsymbol{u_g}$ is the 280×1 vector of line effects for each treatment (ordered as lines within treatments); $\boldsymbol{u_1}$ is the 2×1 vector of replicate effects; $\boldsymbol{u_2}$ is the 4×1 vector of bench within replicate effects; $\boldsymbol{u_3}$ is the 8×1 vector of block effects for each bench and $\boldsymbol{e}$ is the $n \times 1$ vector of residual effects. Note that the vector of data (and vector of residuals) is ordered as rows within ranges within replicates within treatments.

The matrices $\boldsymbol{X},\ \boldsymbol{Z_g},\ \boldsymbol{Z_1},\ \boldsymbol{Z_2}$ and $\boldsymbol{Z_3}$ are the associated design matrices. In terms of the genetic variance assumptions we have

$$\text{var}\left(\boldsymbol{u_g}\right) = \begin{bmatrix} \sigma_{g_{LL}} & \sigma_{g_{LS}} \\ \sigma_{g_{LS}} & \sigma_{g_{SS}} \end{bmatrix} \otimes \boldsymbol{I}_{140} \tag{2}$$

where $\sigma_{g_{LL}}$ and $\sigma_{g_{SS}}$ are the genetic variances for the long and short daylength treatments respectively and $\sigma_{g_{LS}}$ is the genetic covariance (so that the genetic correlation is given by $\sigma_{g_{LS}}/\sqrt{\sigma_{g_{LL}} \times \sigma_{g_{SS}}}$).

In terms of the other random effects we assume simple variance component structures so there is a single variance associated with each of the vectors $\boldsymbol{u_1},\ \boldsymbol{u_2}$ and $\boldsymbol{u_3}$. For the residuals we have

$$\text{var}\left(\boldsymbol{e}\right) = \text{diag}\left(\boldsymbol{R}_i\right) = \text{diag}\left(\sigma_i^2 \boldsymbol{\Sigma_{c}}_i \otimes \boldsymbol{\Sigma_{r}}_i\right) \tag{3}$$

where i indexes the 4 benches (ordered as LD1, LD2, SD1, SD2); $\boldsymbol{\Sigma_{c}}_i$ and $\boldsymbol{\Sigma_{r}}_i$ are the 10×10 and 20×20 spatial correlation matrices for the range

and row dimensions for bench i and σ_i^2 is the variance of the spatial process for the i^{th} bench. We assume auto-regressive processes of order one for each correlation matrix.

This model was fitted within ASReml-R using the following statement:

```
asreml(fixed = dfl ~ Trt * Idtype, random = ~corh(Trt):Id2 +
    Rep + Bench + Bench:Block, rcov = ~at(Bench):ar1(Range):ar1(Row),
    data = flg.df)
```

Recall that 'Id2' is the factor with 140 levels corresponding to the final set of lines phenotyped in the experiment. Once the fixed Idtype effects are fitted the effects for 'Id2' corresponding to the controls and non-genotyped DH lines are zero (effectively eliminated).

The fit of the described model resulted in the identification of 5 outliers that were subsequently removed from the analysis (ie. the DFL values were replaced by the missing value indicator) and the model re-fitted. The corresponding residual maximum likelihood (REML) estimates of the variance parameters are given in Table 2. In terms of the genetic effects we note that the genetic variances for each treatment (836.4 and 850.4 for LD and SD respectively) are very similar and large relative to residual variances. Also there is a very strong genetic correlation (0.89) between the two treatments. In terms of non-genetic effects the design effects are relatively small. The residual variances for the LD benches are similar and larger than those for the SD benches. The spatial trend within each bench are not strong and are similar for all benches.

Formal tests of siginificance were conducted for the hypotheses regarding residual parameters as discussed. This was done by fitted reduced models and using likelihood ratio tests. The final model therefore comprised a separate residual variance for each daylength treatment and common spatial correlation parameters for all benches.

Thus the model is as given in equation (1) but with the residual variance now of the form

$$\text{var}\,(\boldsymbol{e}) = \begin{bmatrix} \sigma_L^2 & 0 \\ 0 & \sigma_S^2 \end{bmatrix} \otimes \boldsymbol{I}_2 \otimes \boldsymbol{\Sigma_c} \otimes \boldsymbol{\Sigma_r} \tag{4}$$

instead of as in equation (3). Here σ_L^2 and σ_S^2 are the residual variances for long and short daylength treatments respectively.

This model was fitted within ASReml-R using the following statement:

Table 2: REML estimates of variance parameters from initial model fitted to DFL data. Parameter estimates are listed for each treatment (LD, SD) as appropriate. The column labelled 'common' relates either to parameters that are common across treatments or parameters that represent correlations between treatments.

Source	LD		common	SD	
Genetic	836.4		0.89	850.4	
Rep			16.9		
Bench			0		
Bench:Block			0		
Residual	LD1	LD2		SD1	SD2
spatial variance	293.4	393.3		162.2	210.8
spatial correlation (range)	0.22	0.20		0.24	0.16
spatial correlation (row)	0.36	0.26		0.17	0.19

```
asreml(fixed = dfl ~ Trt * Idtype, random = ~corh(Trt):Id2 +
    Rep + Bench + Bench:Block, rcov = ~diag(Trt):id(Rep):ar1(Range):ar1(Row),
    G.param = flg.asr3$G.param, R.param = flg.asr3$R.param, data = flg.df)
```

The resultant REML estimates of variance parameters are given in Table 3. It is interesting to note the very high estimated genetic correlation (0.89) between the two treatments indicating strong agreement between the genetic effects for LD and SD. This suggests that any QTL that are detected are likely to occur in similar regions for both treatments.

3 QTL detection in DFL data

We now turn to the detection of QTL for the DFL data. We use a one-stage approach in which the final model fitted (to individual pot data) in Section 2 is expanded to include marker covariate information. The marker covariates are added as random effects to the (final) base-line mixed model. In the case of a univariate problem this is similar to the first step in the Verbyla et al. (2007) approach for QTL detection except that the latter uses covariates for marker intervals rather than the markers themselves. In the Verbyla et al. (2007) approach the interval effects (across all chromosomes) are assumed to be independent with common variance. Our approach differs in that we assume the marker effects to be correlated, with the correlation between

Table 3: REML estimates of variance parameters from final model fitted to DFL data. Parameter estimates are listed for each treatment (LD, SD) as appropriate. The column labelled 'common' relates either to parameters that are common across treatments or parameters that represent correlations between treatments.

Source	LD	common	SD
Genetic	825.9	0.89	849.7
Rep		15.9	
Bench		0	
Bench:Block		0.7	
Residual	LD1&LD2		SD1&SD2
spatial variance	336.9		190.7
spatial correlation (range)		0.20	
spatial correlation (row)		0.24	

two markers being a function of the genetic distance (in cM) between them. Similar ideas were proposed by Gianola et al. (2003) who suggested the use of spatial associations between markers in their pursuit of models for predicting genetic merit. In this report we extend some of the propositions in Gianola et al. (2003) to facilitate QTL detection within a linear mixed model that accommodates sources of non-genetic variation and also accommodates the bivariate aspect of the data.

The correlation model we have chosen for the marker effects is from the Matern class of models. The correlation model involves two parameters, namely a range parameter (ϕ) that affects the rate of decay of correlation and a smoothness parameter (ν). We choose to fix ν at the value of 1.5, a choice recently supported by Kammann and Wand (2003). The correlation between two marker effects for markers that are separated by d cM is then given by

$$\rho = \exp(-d/\phi)(1 + d/\phi) \tag{5}$$

and the covariance by $\sigma_m^2 \rho$ where σ_m^2 is the variance of the process.

In our application we have two traits (LD and SD). We assume that the marker correlation model is the same for both traits (ie. a common ϕ parameter and $\nu = 1.5$ for both) but allow different marker variances for each trait and a covariance between the marker effects for each trait.

The linear mixed model can be written as

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Z_m}(\boldsymbol{u_m} + \boldsymbol{u_n}) + \boldsymbol{Z_g u_g} + \boldsymbol{Z_1 u_1} + \boldsymbol{Z_2 u_2} + \boldsymbol{Z_3 u_3} + \boldsymbol{e} \quad (6)$$

where $\boldsymbol{u_m}$ is the 654×1 vector of (correlated) marker effects for each treatment (ordered as markers within treatments) with associated design matrix $\boldsymbol{Z_m}$; the vector $\boldsymbol{u_n}$ is also of length 654×1 and represents independent or nugget marker effects (ie. additional noise about the correlated effects) and all other terms are as defined for equation (1). With the inclusion of marker effects the effects $\boldsymbol{u_g}$ now represent *residual* genetic effects (ie. not explained by the markers). The variance structures for the marker effects are given by

$$\begin{aligned} \text{var}(\boldsymbol{u_m}) &= \begin{bmatrix} \sigma_{m_{LL}} & \sigma_{m_{LS}} \\ \sigma_{m_{LS}} & \sigma_{m_{SS}} \end{bmatrix} \otimes \boldsymbol{\Sigma_m} \\ \text{var}(\boldsymbol{u_n}) &= \begin{bmatrix} \sigma_{n_{LL}} & \sigma_{n_{LS}} \\ \sigma_{n_{LS}} & \sigma_{n_{SS}} \end{bmatrix} \otimes \boldsymbol{I}_{327} \end{aligned}$$

where the matrix $\boldsymbol{\Sigma_m}$ is a 327×327 matrix of correlations derived from the formula in equation (5) and is a function of the two parameters ϕ (to be estimated) and ν (fixed at 1.5). The parameters $\sigma_{m_{LL}}$ and $\sigma_{m_{SS}}$ are the variances associated with this process (for the long and short daylength treatments respectively) and $\sigma_{m_{LS}}$ is the covariance. The parameters $\sigma_{n_{LL}}$, $\sigma_{n_{SS}}$ and $\sigma_{n_{LS}}$ are the nugget variances and covariance.

In order to conduct this analysis in ASReml-R we first create a data-frame 'flg.dfm' that comprises the original data-frame with 327 additional columns corresponding to the marker information. The columns of the data-frame are ordered so that the first 327 columns correspond to the marker covariates (in map order).

The full model was fitted within ASReml-R using the following statement:

```
asreml(fixed = dfl ~ Trt * Idtype, random = ~corh(Trt):mtrn(grp("marker"),
    0, phi = 1, nu = "1.5F", delta = "1F", alpha = "0F", lambda = "2F") +
    diag(Trt):grp("resmarker") + corh(Trt):Id2 + Rep + Bench +
    Bench:Block, rcov = ~diag(Trt):id(Rep):ar1(Range):ar1(Row),
    G.param = dflqtl.asr2$G.param, R.param = dflqtl.asr2$R.param,
    data = flg.dfm, na.method.X = "include", group = list(marker = 1:327,
        resmarker = 1:327), pwrpoints = list(marker = mdist),
    maxit = 10, stepsize = 1e-04)
```

The 'pwrpoints' argument is required to supply the marker distances (in a vector called 'mdist'). Note that 'mdist' was formed from the marker distances (in cM) within each linkage group but with a 100cM gap placed between linkage groups in order to ensure zero correlation between groups. The term 'diag(Trt):grp("resmarker")' represents the nugget effects. Note that we have not fitted a covariance between traits since the nugget variances were estimated as zero.

The resultant REML estimates of variance parameters are given in Table 4. The first point to note is the large impact of including the marker information. The (total) genetic variance prior to inclusion of the marker data was 825.9 and 849.7 for LD and SD respectively (see Table 3) whereas the residual genetic variance (after modelling the marker effects) was only 34.6 and 152.9 (see Table 4).

Table 4: REML estimates of variance parameters from the QTL model fitted to DFL data. Parameter estimates are listed for each treatment (LD, SD) as appropriate. The column labelled 'common' relates either to parameters that are common across treatments or parameters that represent correlations between treatments.

Source	LD	common	SD
Marker			
Matern variance	0.963	0.93	0.841
Matern range		1.558	
nugget variance	0		0
Residual genetic	34.6	0.72	152.9
Rep		13.6	
Bench		0	
Bench:Block		2.9	
Residual	LD1&LD2		SD1&SD2
spatial variance	336.4		191.8
spatial correlation (range)		0.22	
spatial correlation (row)		0.23	

The key output from the analysis are the best linear unbiassed predictions (BLUPs) of the marker effects from the Matern correlation model (ie. BLUP of $\boldsymbol{u_m}$). An individual effect represents the slope for the regression of the genetic effects on the marker covariate concerned. The covariates may only take two possible values, namely 1 (LYNX) and -1 (MONTY) so that the

difference in genetic effects between the two types of allele (ie. LYNX minus MONTY) is given by twice the value of the slope. We will refer to this as the "size" of the marker effect. Marker size can be graphed against genetic distance for each linkage group in order to examine the marker profile. This has been done separately for each treatment in Figures 1 and 2. Also shown on these graphs are coverage intervals that were obtained by adding and substracting twice the prediction standard error for each size estimate.

References

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

Gianola, D., M. Perez-Encisco, and M. Toro (2003). On marker-assisted prediction of genetic value: beyond the ridge. *Genetics 163*, 347–365.

Kammann, E. and M. Wand (2003). Geoadditive models. *Journal of the Royal Statistical Society, Series C 52*, 1–18.

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

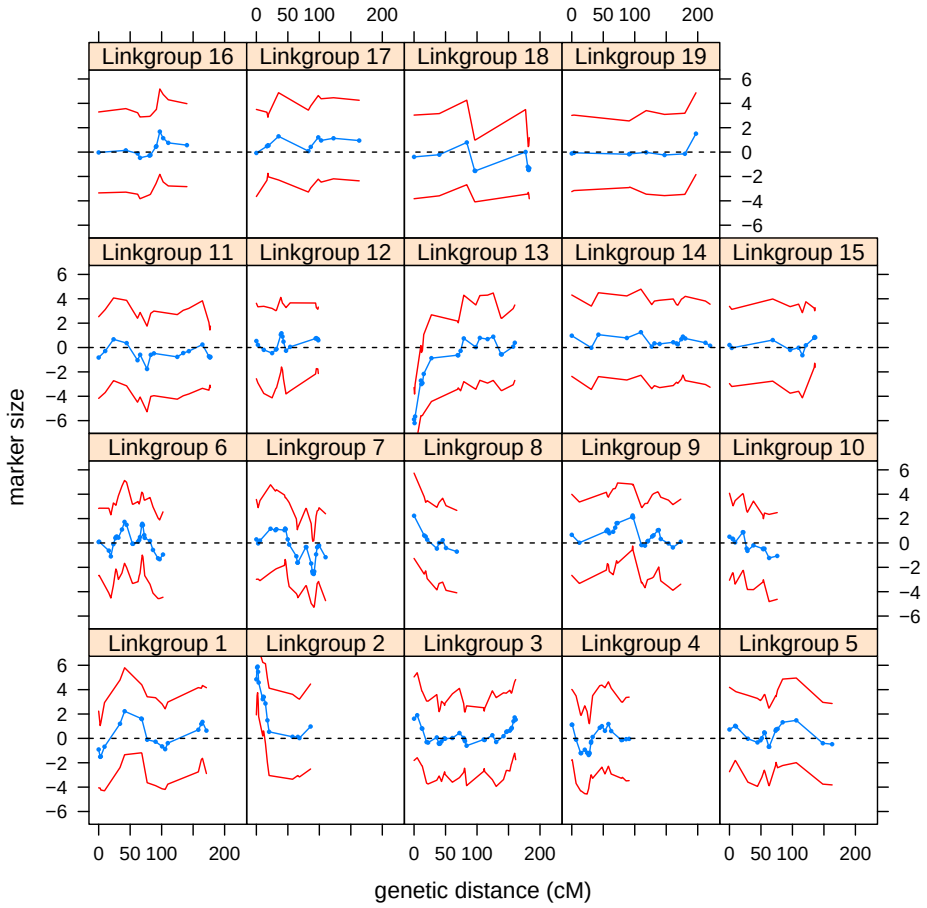

Figure 1: Marker size for LD treatment plotted against genetic distance for each linkage group. Sizes are shown as blue points that have been joined by (blue) lines. Coverage intervals are shown as red lines.

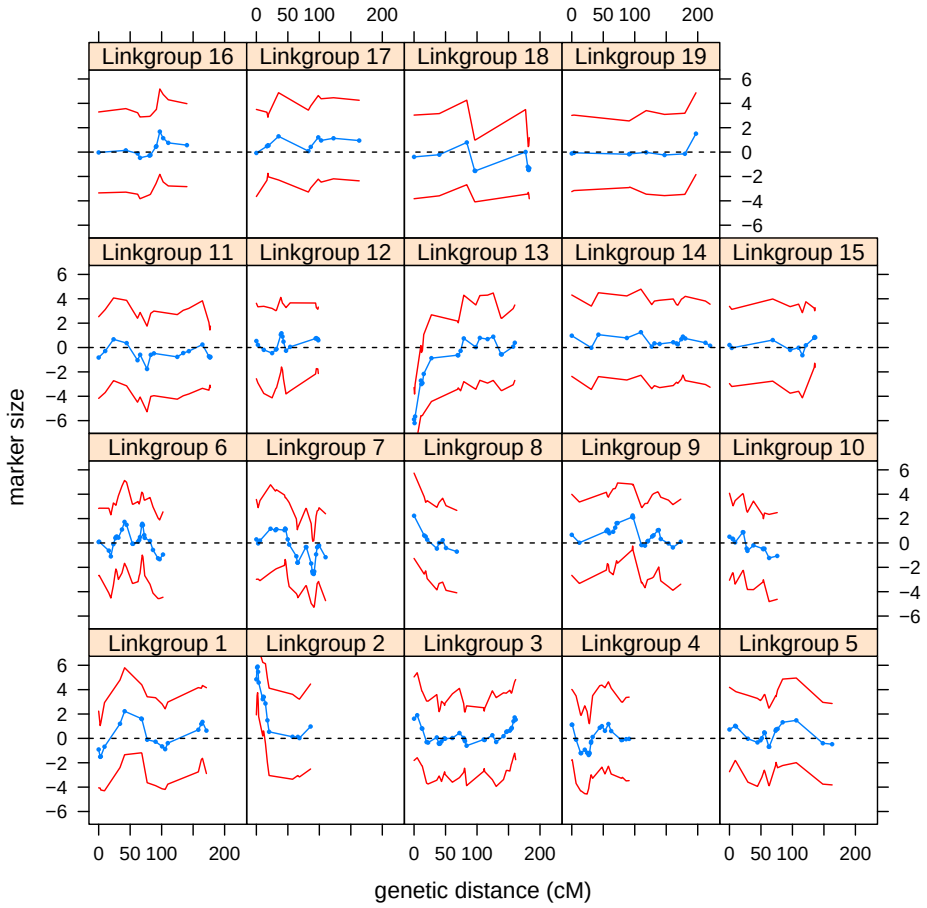

Figure 2: Marker size for SD treatment plotted against genetic distance for each linkage group. Sizes are shown as blue points that have been joined by (blue) lines. Coverage intervals are shown as red lines.