VARIATION IN AUSTRALIAN DURUM WHEAT GERMPLASM FOR PRODUCTIVITY TRAITS UNDER IRRIGATED AND RAINFED CONDITIONS. I. GENOTYPE PERFORMANCE FOR AGRONOMIC TRAITS AND BENCHMARKING

Supplementary material: statistical models and methods

1 Analysis of individual trials

We consider the analysis of DTM for the 2013 trial and assume the data-frame includes factors for the genotypes and irrigation blocks. These are labelled as Genotype (with $v = 36$ levels) and Iblock (with $b = 2$ levels) respectively. The total number of plots (and thence data records) is $n = 144$. The baseline LMM includes fixed effects for Iblock and random effects for Genotype main effects and Genotype by Iblock interactions. These random effects will be denoted by $\boldsymbol{u}_1$ (a $v \times 1$ vector) and $\boldsymbol{u}_2$ (a $bv \times 1$ vector) and are assumed to have simple variance component structures, namely $\mathrm{var}(\boldsymbol{u}_1) = \sigma_1^2 \boldsymbol{I}_v$ and $\mathrm{var}(\boldsymbol{u}_2) = \sigma_2^2 \boldsymbol{I}_{bv}$. The residuals are assumed to have a variance matrix of $\sigma^2 \boldsymbol{I}_n$. To fit this model in ASReml-R we use the following call:

```
DTM2013.asr <- asreml(DTM ~ Iblock,
random = ~ Genotype + Iblock:Genotype, data=data2013.df)
```

Additional terms relating to extraneous variation associated with field rows and columns were added to this baseline model as required on a trial and trait basis (see Table 1 below). They were included if deemed statistically significant ($p < 0.05$) using residual maximum likelihood ratio tests. A spatial correlation model (separable autoregressive process of order 1) was included for the grain yield trait for all 3 trials.

It is instructive to note that the sum of the Genotype main effects and Genotype by Iblock interactions gives the genotype effects nested within irrigation blocks. We denote the latter by the $bv \times 1$ vector $\boldsymbol{u}$ so we have

$$\boldsymbol{u} = (\boldsymbol{I}_v \otimes \boldsymbol{1}_b)\, \boldsymbol{u}_1 + \boldsymbol{u}_2 \tag{1}$$

and the use of the variance component structures leads to a separable variance matrix for $\boldsymbol{u}$ given by

$$\begin{aligned} \mathrm{var}(\boldsymbol{u}) &= \boldsymbol{G_V} \otimes \boldsymbol{G_B} \\ \text{where } \boldsymbol{G_V} = \boldsymbol{I}_v \quad &\text{and} \quad \boldsymbol{G_B} = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 + \sigma_2^2 \end{bmatrix} \end{aligned} \tag{2}$$

Table 1: Individual trial analysis for each trait: trials in which random field row and column effects were fitted.

Trait	Rows	Columns
PYLD	2014	
TGW	2012, 2014	2012, 2013, 2014
GRM2		2013
HI	2012, 2014	2012, 2014
PLHT	2012, 2013	2012, 2013
SCR	2014	2014
DTM	2012	2012, 2013, 2014

The form of $\boldsymbol{G_B}$ is known as a compound symmetric form. The parameters have the interpretation that $\sigma_1^2 + \sigma_2^2$ is the variance of the genotype effects within an irrigation block and σ_1^2 is the covariance between the genotype effects in the two blocks and thence $\rho = \sigma_1^2/(\sigma_1^2 + \sigma_2^2)$ is the correlation. REML estimates of ρ for individual trials and traits are given in Table 6 in the main document.

We also note that the separable form of the variance matrix for $\boldsymbol{u}$ could be fitted directly in ASReml-R using the model call:

```
DTM2013.asr <- asreml(DTM ~ Iblock,
random = ~ corv(Iblock):Genotype, data=data2013.df)
```

In this model there are also two variance parameters associated with the random effects, namely a correlation and a variance. The former is ρ and the latter is $\sigma_v^2 = \sigma_1^2 + \sigma_2^2$. Thus the two model calls differ only in their parameterisation of the variance structures for the random effects. The latter form is useful in that it shows how we may generalise to a three-way separable form for the multi-environment trial analysis.

2 Combined analysis across trials

We now consider the analysis of individual plot DTM data combined across all trials and assume the data-frame includes factors for the trials, genotypes and irrigation blocks. These are labelled as Trial (with $t = 3$ levels), Genotype (with $v = 61$ levels) and Iblock (with $b = 2$ levels) respectively. The total number of plots (and thence data records) is $n = 582$. The LMM includes fixed effects for Trial, Iblock and Trial by Iblock interactions. The random effects of interest comprise the Trial by Genotype by Iblock (TGB) effects and are denoted by $\boldsymbol{u}$ which is a vector of length $tvb = 366$. This is an extension of the (nested) Genotype by Iblock effects in the individual trial analysis. The variance matrix is similarly an extension of equation (2) and as per Smith et al. (2019), namely

$$\text{var}\,(\boldsymbol{u}) = \boldsymbol{G} = \boldsymbol{G_T} \otimes \boldsymbol{G_V} \otimes \boldsymbol{G_B} \tag{3}$$

where $\boldsymbol{G_T}$ is the 3×3 matrix for the trial dimension, $\boldsymbol{G_V}$ is the 61×61 matrix for the genotype dimension and $\boldsymbol{G_B}$ is the 2×2 matrix for the irrigation block

dimension. The forms for the three component matrices were discussed in the main document. Formally we assumed

- an FA1 model for the Trial dimension so that

$$\boldsymbol{G_T} = \begin{bmatrix} \lambda_1^2 + \psi_1 & \lambda_1\lambda_2 & \lambda_1\lambda_3 \\ \lambda_2\lambda_1 & \lambda_2^2 + \psi_2 & \lambda_2\lambda_3 \\ \lambda_3\lambda_1 & \lambda_3\lambda_2 & \lambda_3^2 + \psi_3 \end{bmatrix}$$

  where λ_j and ψ_j are the loading and specific variance for the j^{th} trial ($j = 1 \ldots 3$). The genetic correlation between trials i and j can then be obtained as

$$\rho_{ij} = \frac{\lambda_i\lambda_j}{\sqrt{(\lambda_i^2 + \psi_i)(\lambda_j^2 + \psi_j)}}$$

  Note that for the case of $t = 3$ trials the FA1 model involves the same number of parameters as an unstructured variance matrix. We prefer the use of the FA1 as it can accommodate cases where $\boldsymbol{G_T}$ may be singular and importantly it allows application of the selection tools in Smith & Cullis (2018) so we can obtain measures of overall genotype performance across trials (also see section 2.1).

- an identity matrix for the Genotype dimension so that $\boldsymbol{G_V} = \boldsymbol{I}_v$
- a correlation matrix for the Iblock dimension so that

$$\boldsymbol{G_B} = \begin{bmatrix} 1 & \rho_B \\ \rho_B & 1 \end{bmatrix}$$

  Note that this is a special compound symmetric form with known variance (as needed to ensure identifiability) fixed at unity.

To fit this model in ASReml-R we use the following call:

```
DTMfa.asr <- asreml(DTM ~ Trial*Iblock,
random = ~ fa(Trial):Genotype:cor(Iblock),
residual =~ dsum(~units|Trial), data=all.df)
```

Note that this model call fits a separate residual variance for each trial. Also in the final model we included random range and row effects as identified in the individual trial analyses, that is, as per Table 1. They have been excluded here for simplicity. Fitting this model provides REML estimates of the FA loadings and specific variances (and thence of the correlations, ρ_{ij}, between trials) and the correlation, ρ_B, between irrigation blocks. These are given for each trait in Table 7 of the main document.

2.1 Summarising the TGB effects

The use of the FA1 model for the Trial dimension means we can write the TGB effects as

$$\boldsymbol{u} = (\boldsymbol{\Lambda} \otimes \boldsymbol{I}_m)\, \boldsymbol{f} + \boldsymbol{\delta} \tag{4}$$

where $\boldsymbol{\Lambda} = (\lambda_1, \lambda_2, \lambda_3)^\top$ is the 3×1 matrix of trial loadings; $\boldsymbol{f}$ is the $m \times 1$ vector of scores for Genotype by Iblock combinations (so $m = 61 \times 2 = 122$) and $\boldsymbol{\delta}$ is the vector of lack of fit effects (of length 366). The variance assumptions for the scores and lack of fit effects in equation (4) are

$$\text{var} \begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{\delta} \end{pmatrix} = \begin{bmatrix} \boldsymbol{G_V} \otimes \boldsymbol{G_B} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Psi} \otimes \boldsymbol{G_V} \otimes \boldsymbol{G_B} \end{bmatrix}$$

where $\boldsymbol{\Psi}$ is the diagonal matrix of specific variances and $\boldsymbol{G_V}$ and $\boldsymbol{G_B}$ are as previously defined. These assumptions lead to the form of the variance matrix for $\boldsymbol{u}$ as given in equation (3).

Fitting the model in ASReml-R provides empirical best linear unbiased predictions (EBLUPs) of the TGB effects, which we denote by $\tilde{\boldsymbol{u}}$. As discussed in the main document we are interested in summarising these in three ways, namely (a) across trials to obtain genotype EBLUPs for each irrigation block, (b) across irrigation blocks to obtain genotype EBLUPs for each trial and (c) across trials and irrigation blocks to obtain a single set of genotype EBLUPs that measure overall performance (OP) for each trait. We discuss each of these in the following sections.

2.1.1 Genotype EBLUPs for each irrigation block (across trials)

Fitting the model in ASReml-R also provides EBLUPs of the scores, denoted $\tilde{\boldsymbol{f}}$. If we re-order these as genotypes within irrigation blocks we can partition as $\tilde{\boldsymbol{f}} = \left(\tilde{\boldsymbol{f}}_W^\top, \tilde{\boldsymbol{f}}_R^\top\right)^\top$ where $\tilde{\boldsymbol{f}}_W$ is the 61×1 vector of genotype scores for the watered block and $\tilde{\boldsymbol{f}}_R$ is the 61×1 vector of genotype scores for the rainfed block. We can generalise Smith & Cullis (2018) to obtain summary measures of genotype performance across trials for each irrigation block. For the watered block we use $\bar{\lambda}\tilde{\boldsymbol{f}}_W$ where $\bar{\lambda}$ is the mean of the REML estimates of the trial loadings. Similarly for the rainfed block we use $\bar{\lambda}\tilde{\boldsymbol{f}}_R$.

These measures have been used in the main document in Figures 2 - 8 as panel (a).

2.1.2 Genotype EBLUPs for each trial (across irrigation blocks)

To obtain this summary we use the framework in section 1 in which we linked a Genotype main effect plus Genotype by Iblock interaction model to a two-way separable model for the genotype effects nested within blocks. Here we commence with the three-way separable model and show it can be expressed as the sum of Trial by Genotype effects and Trial by Genotype by Iblock interactions.

The key is the compound symmetric form for $\boldsymbol{G_B}$ in both the individual trial analyses and the MET analysis.

In an analogous manner to equation (1) we define $\boldsymbol{u}_1$ to be the $tv \times 1$ vector of Trial by Genotype effects and $\boldsymbol{u}_2$ to be the $btv \times 1$ vector of Trial by Genotype by Iblock interaction effects. We can then write the TGB effects as

$$\boldsymbol{u} = (\boldsymbol{I}_{tv} \otimes \boldsymbol{1}_b)\, \boldsymbol{u}_1 + \boldsymbol{u}_2 \tag{5}$$

We then use the following assumptions for the variance matrices:

$$\begin{aligned} \text{var}\,(\boldsymbol{u}_1) &= \rho_B \boldsymbol{G_T} \otimes \boldsymbol{G_V} \\ \text{var}\,(\boldsymbol{u}_2) &= (1-\rho_B) \boldsymbol{G_T} \otimes \boldsymbol{G_V} \otimes \boldsymbol{I}_b \end{aligned}$$

These assumptions lead to the variance matrix of $\boldsymbol{u}$ being given by equation (3). It can then be shown that EBLUPs for $\boldsymbol{u}_1$ can be obtained from the EBLUPs of $\boldsymbol{u}$. Specifically, for genotype i in trial j we have that

$$\tilde{u}_{1_{ij}} = \frac{\rho_B}{1+\rho_B} \left(\tilde{u}_{ijW} + \tilde{u}_{ijR}\right)$$

where $\tilde{u}_{ijW}$ and $\tilde{u}_{ijR}$ are the EBLUPs of the TGB effects for genotype i and trial j for the watered and rainfed blocks respectively.

These measures have been used in the main document in Figures 2 - 8 as panels (b) - (d).

2.1.3 Genotype EBLUPs that measure overall performance across trials and blocks

Combining the concepts in the previous two sections leads to a form for overall genotype performance (across trials and blocks) given by

$$\frac{\rho_B}{1+\rho_B} \bar{\lambda} \left(\tilde{\boldsymbol{f}}_W + \tilde{\boldsymbol{f}}_R\right)$$

These measures have been used in the main document in Figures 9 and 10.

References

Smith, A., Borg, L. M., Gogel, B., & Cullis, B. (2019). Estimation of factor analytic mixed models for the analysis of multi-treatment multi-environment trial data. *Journal of Agricultural, Biological and Environmental Statistics* **24**, 573588.

Smith, A. & Cullis, B. (2018). Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. *Euphytica* pages doi.org/10.1007/s10681–018–2220–5.