Supplementary information: method to quantify population of starch granule types

Emi Tanaka¹*, Jean-Phillippe F Ral², Sean Li², Raj Gaire², Colin R Cavanagh³, Brian R Cullis¹ and Alex Whan²

Experimental Design
The trait of interest in our experiment is the proportion of the granule types. This is estimated from the granule size distribution using the two methods described in the main paper, namely the size threshold measure and mixture measure. The distribution of the granule size is obtained from a multi-phase experiment. Specifically, the experiment involved four phases: grain samples taken from the field trial (Phase I); then milled to flour (Phase II); processed further for extraction in the laboratory (Phase III); and the run through Mastersizer (Phase IV).

The treatment for our experiment is the genotype. To distinguish the sources of variation from genetic sources of variation, we require a separate randomisation and replication for each phase and an appropriate analysis that takes into account the blocking structures for each phase [1–3]. The details of each phase are given below and factors that will be used to specify the model later are given in this font.

Phase I: Field trial
The field trial was arranged in a rectangular array of 81 rows by 20 columns, i.e. a total of 1620 plots. There were a total of 1077 genotypes sown. The trial was designed using the DiGGer software [4] for a p-replicate design [5] in which a proportion p of lines were sown with two plots each and the remaining with a single plot each. The 14 check genotypes had additional 3-7 plots each. The value of p is 0.48 (521 out of 1077 genotypes). The trial had 3 near-resolvable blocks, coinciding with the bay (in the sense that replicated genotypes, excluding checks, appeared at most once in each block), which were aligned with the rows such that the first replicated block comprised of rows 1-27, the second rows 28-54, and the third rows 55-81.

The pEUs are the 1620 physical locations within the field indexed by the rows and columns. The pOUs are the 1620 grain samples from these plots. The pRUs are the 1042 packets of seeds. Of the 1620 pOUs, 705 pOUs are used for the next phase of which 215 pOUs were relevant to the final phase of our experiment.

The pIU in the next phase consisted of a uniform mixture of grain samples from 2 plots or of a single plot. Each Phase II pIU (and thus pEU) is indexed by two factors, Leftplot and Rightplot, corresponding to the plot number (1-1620) and these two factors differ where the unit is a mixture of two different plots, otherwise, have the same level. Similar factors are defined for column (1-20), row (1-81) and bay (1-3) numbers (Leftcol, Rightcol, Leftrow, Rightrow, Leftbay, and Rightbay).

Phase II: Milling
In the milling phase, 10 grain samples are milled for each of the 62 days (MillDay) resulting in a total of 620 milled samples. These 620 milled samples are the Phase II pOUs. The pEUs are the “positions” associated with the milling run and is indexed by milling day and order. Note that MillOrder consists of 10 levels and is nested within MillDay. The grain samples were either a composite from two plots sown with the same genotype; from a single plot or split from a single plot to serve as a replicate in Phase II - we refer these grain samples as composite, single or duplicate samples, respectively, and these comprise the 620 pIUs. Specifically, the grain samples were obtained from the 705 plots from the previous phase and consisted of 131 composite samples, 397 single samples and 92 duplicate samples. Of the 397 single samples, 132 have field replicates. In total there were 508 genotypes in this phase. The milling samples are replicated such that we have two replicate blocks (MillRep) that coincides with the milling day. Specifically, milling replicate 1 is from milling day 1 to 31 inclusive and milling replicate 2 is from milling day 32 to 62 inclusive. The pRUs are the 92 grain samples. From 620 milled samples, 180 are used for the extraction phase.

Phase III: Extraction
In the extraction phase, the 180 milled samples (pIUs) consisting of 155 genotypes from Phase II, of which 44 milled samples were split to form as a replicate
(thus there were 88 pRUs), were then treated as described in the Methods in the main paper. In total, there were 224 milled samples prepared for extractions with 32 extractions (ExtractOrder) per day over 7 days (ExtractDay). Here ExtractOrder consists of 32 levels and is nested within ExtractDay. The pEUs are the positions corresponding to the extraction indexed by extraction day and order and the pOUs are the 224 extractions. About 21.1% (38) of the 180 milled samples were milling of the composite grain samples.

Phase IV: Mastersizer
All of the 224 extractions (Phase IV pIUs) from the previous phase, each with a replicate except for 16 extractions, was then run through the Mastersizer with 36 runs (MSOrder) per day for each of the 12 days (MSDay). Each run consisted of 2 technical replicates (MS TecRep) resulting in 864 particle size distribution. Thus the pOUs are these 864 particle size distribution; the OU is the technical run associated with each pOUs; and the 416 extractions are the pRU. The pEUs are the positions of the Mastersizer run indexed by day and order.

Statistical Analysis
Statistical Model
We consider the analysis of the proportion of the granule types separately in an univariate model. Let $p_A$, $p_B$ and $p_C$ be the $n \times 1$ vectors of estimates of the proportion of A-type, B-type and C-type granules with each ordered by the observational units. Note that we expect the sum of the corresponding elements of $p_A$, $p_B$ and $p_C$ to be less than or equal to 1, with most cases holding the equality. This is because $p_A$, $p_B$ and $p_C$ are estimates which may result in rounding errors and also these proportions may not account for all the granule types measured. Let $k$ denote the number of phases in the experiment.

We assume that there were $m_d$ genotypes grown in the trial with $m \leq m_d$ genotypes that were used in the final phase of the experiment.

For each granule type $T \in \{A, B, C\}$, we can write the general linear mixed model as

$$y_T = Z_g u_g + \sum_{r=1}^{k} X_r \tau_r + \sum_{r=1}^{k} Z_p u_p + e$$  \hspace{0.5cm} (1)

where $y_T$ is the $n \times 1$ vector of the logit transformation of $p_T$, the proportion estimate of granule type $T$; $u_g$ is the $n \times 1$ vector of total genetic effects with associated design matrix $Z_g$; $\tau_r$ is the vector of fixed effects associated with phase $r$ with corresponding design matrix $X_r$; $u_p$ is the vector of random peripheral effects associated with phase $r$ with corresponding design matrix $Z_p$; and $e$ is the $n \times 1$ vector of residuals. The vectors of random peripheral effects will typically contain sub-vectors that will be denoted by $u_{pr}$. We assume that these sub-vectors, $u_g$ and $e$ are jointly Gaussian distributed and mutually independent with variances given as var ($u_g$) = $G_g$, var ($u_{pr}$) = $G_{pr}$ and var ($e$) = $R$. In the simplest model, we assume that $G_{pr} = \sigma^2_{pr} I$ with the dimensions of the identity matrix commensurate with the length of the corresponding vector of effects although we can include more complex forms such as a separable auto-regressive process of order 1 for the modelling of spatial correlations in the field plot effects [6, 7]. Note here that the above design matrices are the same regardless of the granule type although the corresponding effects as well as residuals are granule type dependant, we remove the subscript $T$ for brevity.

The experimental units in the second phase include composite mixtures of the observational units in the first phase, as such the design matrix $X_1$ and $Z_1$ in model (1) have non-standard formats. We can rewrite $X_1 = C_1 \ X_1^\ast$ and $Z_1 = C_1 \ Z_1^\ast$ where $C_1$ is a matrix such that $C_1 \ X_1^\ast \tau_1$ and $C_1 \ Z_1^\ast \ u_1$ results in a vector containing averages of the associated Phase I effects over the same level of Phase II pEU. See [?] for examples of the $C_1$ matrix.

Reliability calculation
We compute the reliability as the mean of the squared accuracy of each genotype [?]. Specifically, the accuracy of the $i$-th genotype for a particular granule type is the correlation of the corresponding total genetic effect to its predicted value and given as

$$r_i = \sqrt{1 - \frac{C_{ii}^{gg}}{G_{ii}}}$$  \hspace{0.5cm} (2)

where $C_{ii}^{gg}$ is the $i$-th diagonal element of the prediction error variance matrix for $u_g$ in model (1) and $G_{ii}$ is the $i$-th diagonal element of $G_g$. Then the (mean-genotype) reliability is given as

$$r^2 = \frac{1}{m} \sum_{i=1}^{m} r_i^2 .$$

Analysis of data
All analysis in this paper were fitted using ASReml-R package [8] within the R statistical environment [9] which uses the average information algorithm [7] for the residual maximum likelihood estimates (REML) of the variance parameters. The estimates and prediction of the fixed and random effects are then obtained by
Table 1 The number of phase-related units and treatment of the experiment.

<table>
<thead>
<tr>
<th>Unit</th>
<th>Phase I</th>
<th>Phase II</th>
<th>Phase III</th>
<th>Phase IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>1077 genotypes</td>
<td>508 genotypes</td>
<td>155 genotypes</td>
<td>155 genotypes</td>
</tr>
<tr>
<td>pRU</td>
<td>521 x 2 packets of seed</td>
<td>46 x 2 grain samples</td>
<td>44 x 2 flour samples</td>
<td>208 x 2 extractions</td>
</tr>
<tr>
<td>pLU</td>
<td>1620 packets of seed</td>
<td>620 grain samples</td>
<td>224 flour samples</td>
<td>432 extractions</td>
</tr>
<tr>
<td>pEU</td>
<td>1620 plots</td>
<td>620 milling runs</td>
<td>224 extraction runs</td>
<td>432 analyser runs</td>
</tr>
<tr>
<td>pOU</td>
<td>1620 grain samples</td>
<td>620 flour samples</td>
<td>224 extractions</td>
<td>864 size distribution</td>
</tr>
</tbody>
</table>

Table 2 The REML estimation of the variance components and the mean-line reliability from the univariate models for data vector given by Method 1 (size-threshold) or 2 (mixture distribution) from the main article. Note that Phase I-IV residuals are indexed by Plot (LeftPlot and RightPlot in the model), MillDay:MillOrder, ExtractDay:ExtractOrd, and MSDay:MSOrd:MSTecRep, respectively.

<table>
<thead>
<tr>
<th>Source</th>
<th>Method 1</th>
<th>Method 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Geno</td>
<td>0.0138</td>
<td>0.0139</td>
</tr>
<tr>
<td>Phase IV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSDay</td>
<td>0.00301</td>
<td>0.00292</td>
</tr>
<tr>
<td>MSDay:MSOrd</td>
<td>0.0060</td>
<td>0.0588</td>
</tr>
<tr>
<td>Phase III</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ExtractDay</td>
<td>0.0209</td>
<td>0.0213</td>
</tr>
<tr>
<td>Phase II</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MillRep</td>
<td>1.74e-10</td>
<td>1.24e-10</td>
</tr>
<tr>
<td>MillDay</td>
<td>4.49e-11</td>
<td>4.76e-11</td>
</tr>
<tr>
<td>Phase I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bay</td>
<td>0.00133</td>
<td>0.00142</td>
</tr>
<tr>
<td>Column</td>
<td>6.88e-09</td>
<td>9.13e-09</td>
</tr>
<tr>
<td>Row</td>
<td>0.00945</td>
<td>0.00921</td>
</tr>
<tr>
<td>Phase I Residual</td>
<td>1.24e-08</td>
<td>1.29e-08</td>
</tr>
<tr>
<td>Phase II Residual</td>
<td>8.61e-11</td>
<td>8.07e-11</td>
</tr>
<tr>
<td>Phase III Residual</td>
<td>0.0501</td>
<td>0.0492</td>
</tr>
<tr>
<td>Reliability</td>
<td>0.171</td>
<td>0.174</td>
</tr>
</tbody>
</table>

Using these variance parameter estimates in a solution of the mixed model equations [10].

Following [1] and [2], a linear mixed model was fitted to each proportion of the granule type estimated by either the size threshold or mixing probabilities as outlined in the main paper. The fixed effects incorporated the overall mean and the check genotypes and the random effects incorporated blocking structures employed in the multiphase experiment and the genotype effect. We assume that $R = \sigma^2 I_n$ and $G_g = \sigma^2_g I_m$. All variance-covariance matrices of the blocking terms are also assumed to be scaled identity. The syntax in ASReml-R for the random part of the model is given as

```
Random=~ Geno +
Leftbay:zero + and(Leftbay, 0.5) + and(Rightbay, 0.5) +
Leftcol:zero + and(Leftcol, 0.5) + and(Rightcol, 0.5) +
Leftrow:zero + and(Leftrow, 0.5) + and(Rightrow, 0.5) +
Leftplot:zero + and(Leftplot, 0.5) + and(Rightplot, 0.5) +
MillRep + MillDay + MillDay:MillOrder +
ExtractDay + ExtractDay:ExtractOrd +
MS.Day + MS.Day:MSOrd
```

where the colon operator indicates an interaction term, the factor Geno represents the genotype effects and other factors are as defined before. Here zero is a vector of zeroes used for initialising the corresponding structure for computational purposes and the collective syntax and(Leftxxx, 0.5) + and(Rightxxx, 0.5), where xxx is row, col, bay and plot, builds the non-standard design matrix $C_1Z_1^\dagger$.

The estimates of the variance components and the mean-genotype reliability are shown in Table 2. It is worth noting that the REML estimate of the genetic variance for C-type granule estimated by Method 2 is very low (0.00229) - further discussion on this matter is elaborated in the main paper.

It would be reasonable to assume the proportions of each granule type of the same sample are correlated traits. The accuracy of genotypic effects would then benefit from using a multivariate linear mixed model approach over a univariate model as above (1) [11]. In this case however, $p_C \approx 1_n - p_A - p_B$ (with inequality resulting mostly due to rounding errors or in the case of mixing weights, particle size distribution fitted with four or five Gaussian mixture distribution), as such the third proportion is almost perfectly explained by the linear combination of the other two proportions. This would necessitate only fitting a bivariate model to two
sets of proportions of granule type however as shown in Figure 1, the genetic effects for the proportions of A-type and B-type granules are almost perfectly negatively correlated. We thus did not proceed with fitting a bivariate model.

![Figure 1](image)

**Figure 1** The plot of E-BLUPs of genetic effects in model (1) for proportion of type-A vs. type-B as estimated by Method 1 (size-threshold) or Method 2 (mixture-measure). The diagonal black line corresponds to $y = -x$.

### Author details

1 School of Mathematics and Applied Statistics, University of Wollongong, Northfields Ave, 2522 Wollongong, Australia. 2 CSIRO Agriculture and Food, 1600 Clunies Ross Street, 2601 Canberra, Australia. 3 Bayer CropScience Pty Ltd, Seeds, , Australia.

### References