Plant Methods

COMMENTARY Open Access

Open-source analytical pipeline for robust data analysis, visualizations and sharing in crop breeding

Waseem Hussain*, Mahender Anumalla, Margaret Catolos, Apurva Khanna, Ma. Teresa Sta. Cruz, Joie Ramos and Sankalp Bhosale

Abstract

Background: Developing a systematic phenotypic data analysis pipeline, creating enhanced visualizations, and interpreting the results is crucial to extract meaningful insights from data in making better breeding decisions. Here, we provide an overview of how the Rainfed Rice Breeding (RRB) program at IRRI has leveraged R computational power with open-source resource tools like R Markdown, *plotly*, LaTeX, and HTML to develop an open-source and end-to-end data analysis workflow and pipeline, and re-designed it to a reproducible document for better interpretations, visualizations and easy sharing with collaborators.

Results: We reported the state-of-the-art implementation of the phenotypic data analysis pipeline and workflow embedded into a well-descriptive document. The developed analytical pipeline is open-source, demonstrating how to analyze the phenotypic data in crop breeding programs with step-by-step instructions. The analysis pipeline shows how to pre-process and check the quality of phenotypic data, perform robust data analysis using modern statistical tools and approaches, and convert it into a reproducible document. Explanatory text with R codes, outputs either in text, tables, or graphics, and interpretation of results are integrated into the unified document. The analysis is highly reproducible and can be regenerated at any time. The analytical pipeline source codes and demo data are available at https://github.com/whussain2/Analysis-pipeline.

Conclusion: The analysis workflow and document presented are not limited to IRRI's RRB program but are applicable to any organization or institute with full-fledged breeding programs. We believe this is a great initiative to modernize the data analysis of IRRI's RRB program. Further, this pipeline can be easily implemented by plant breeders or researchers, helping and guiding them in analyzing the breeding trials data in the best possible way.

Keywords: Rice, Breeding analytics, Open-resource, Interactive visualizations, Reproducibility

Background

The International Rice Research Institute (IRRI), established in the 1960s, is the world's premier research organization dedicated to rice science. Rainfed rice breeding (RRB) at IRRI started since the establishment of the institute and is continuously committed to innovate and develop improved rice germplasm for improving the livelihood of farmers encountering challenging climates [1]. Currently, the ongoing rice breeding project, "Accelerated Genetic Gains in Rice Alliance" at IRRI, funded by the Bill and Melinda Gates Foundation (BMGF), is mandated to modernize breeding strategies and framework to increase the current rates of genetic gains in close collaboration with NARES network-partner's across South Asia (India, Bangladesh, and Nepal), East and Southern Africa (Kenya, Mozambique, Tanzania, and Burundi).

*Correspondence: waseem.hussain@irri.org
Rice Breeding Innovation Platform, International Rice Research Institute (IRRI), Los Banos, Laguna, Philippines

Every year RRB at IRRI shares the breeding germplasm tolerant to drought, salt, heat, and submergence with the regional partner's for phenotypic evaluation and, in return, receives raw phenotypic data from several trials at different locations. For instance, the RRB during the year 2019 received data from approximately 20 trials from the NARES partners. It is crucial to demystify data analysis for regional partner's to make better breeding decisions and present the results in an easy and understandable format. Detailed documentation will contribute to a clear interpretation and understanding of results along with promoting collaborations. Furthermore, simultaneously analyzing and documenting the results has not been possible with readily available computational tools that require a 'copy and paste' system to document or report the highly error-prone results. Thus, we believe an immediate up-gradation of data analysis workflow is crucial to be more effective and enhance reproducibility [2]. The high-end improvement is necessary for conveniently documenting and sharing the result reports.

Technology advances have made data management, analysis, interpretation, visualization, and sharing more convenient. For example, R software [3] packages viz., *ggplot2* [4], *plotly* (https://plotly.com/), *DT* (https://rstudio.github.io/DT/) has made the data mining manageable and visualizations interactive and dynamic. Similarly, with *R Markdown* [5], data analysis can be turned into high-quality, reproducible reports in which codes, text, tables, graphics, and more are embedded in one unified document. Furthermore, the reports can be generated in various formats, including MS Word, PDF, HTML (Hyper-Text Markup Language), and more for seamless sharing (https://rmarkdown.rstudio.com/).

Here, we provide an overview of how the RRB program at IRRI has leveraged in R computational power with open-source resource tools of R Markdown, plotly, LaTeX [6] (https://www.latex-project.org/get/) and HTML to develop an analysis workflow of phenotypic data analysis, and re-designing it to a reproducible document for better interpretations, visualization and easy sharing with collaborators. The developed analysis workflow demonstrates how to pre-process and check data quality and perform robust data analysis using modern statistical tools and approaches. Besides developing this analytical pipeline and workflow, we showed how this workflow could be embedded into a well-descriptive document or report. In practice, we provide an open-source analytical pipeline with comprehensive details, procedures, and end-to-end steps. It integrates the analysis workflow, explanatory text with R codes, outputs either in text, tables, or graphics, and interpretation of results into a single document or, in simpler words, 'everything is at one place. The complete and detailed description of results will act as a guide for phenotypic data analysis. It can be easily put into practice by the plant breeders and or plant researchers having a full-fledged breeding program.

Overview of analysis workflow and pipeline

Figure 1 illustrates the improved analysis workflow adopted in this study to analyze multi-environment trial data. The workflow is divided into four main

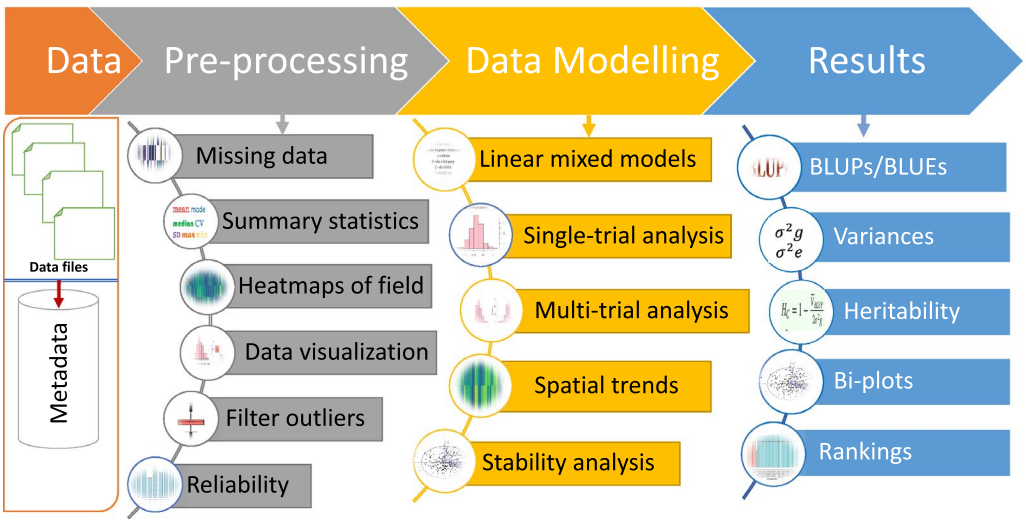

Fig. 1 Schematic representation of data analysis workflow adapted in the current report. The four main steps involved in the analysis workflow process are **a** data importing, **b** data pre-processing, **c** data modeling, and **d** results generation. The main steps are divided into individual components required to develop a comprehensive and robust analysis pipeline

components: data import, pre-processing and quality check, data analysis, and result extractions. In the pre-processing and quality check, we demonstrated a detailed procedure and instruction on checking the quality of data and ensuring only high-quality phenotypic data points are advanced for downstream analysis to get reliable estimates or predictions of genotypes. The sample document for this available is on GitHub (https://github.com/whussain2/Analysis-pipeline). For the data analysis step, we provide a detailed overview of how to analyze the data separately or jointly using linear mixed-model (LMM) approaches. The analytical pipeline is demonstrated both in the ASReml-R package and in the lme4 R package available on GitHub (https://github.com/whussain2/Analysis-pipeline). We applied mixed models ranging from basic to higher advanced models in separate-trial analysis accounting for experimental design factors and spatial trends. Similarly, in multi-environment trial (MET) data, we showed single-stage or two-stage analysis approaches ranging from basic models to higher advanced factor analytical models. In the results step, we demonstrated selecting the best model and using it to extract different results. Results including BLUPs, heritability, correlation and covariance matrix of environments, G x E BLUPs, principal component analysis (PCA) biplot-showing stability and relationship of environments, and latent regression plots to accesses the stability of genotypes (Fig. 2) were presented. All the instructions, R source codes, examples, and the data sets are freely available in the GitHub repository at https://github.com/whussain2/Analysis-pipeline. Additional resources on analyzing the MET data and checking the stability of genotypes are given in section 1.4 of the ASReml analysis workflow.

Data importing

In this step, raw phenotypic data is imported into the R workspace, and metadata information is generated (Fig. 1). Information about the field trial, data collection, experimental design, and more are given in section 1 of the sample pre-processing HTML file available on GitHub (https://github.com/whussain2/Analysis-pipeline). The raw phenotypic data imported can also be visualized in the table format in the report. Interestingly, the table generated for raw phenotypic data is highly interactive and can be easily managed, searched, and sorted like a mini excel sheet (section 2 in the sample pre-processing HTML file). Further, the table generated can be easily exported in various formats or printed directly within the document. The raw phenotypic data used for the demo purpose in this study comes from one of the rainfed breeding program trials, which were evaluated in alpha lattice design with two replications and six blocks across multiple environments in Africa with a total of 200 unique genotypes per environment. Besides replications and blocks, row and column information is noted down to account for the spatial trends.

Data pre-processing and quality check

Data pre-processing involving the quality control of data is the most critical and complex step in data analysis. The workflow to pre-process and check the quality of data is given on a pre-processing HTML file available

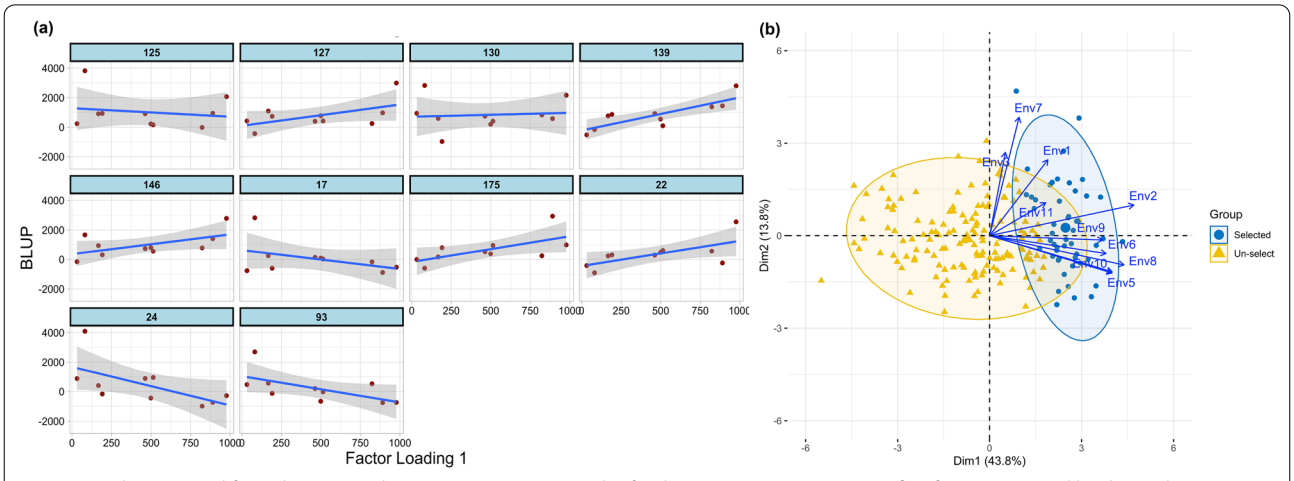

Fig. 2 Results extracted from the MET analysis. **a** Latent regression plot for the top 10 genotypes using first factor estimated loadings. The solid blue line and the gray shade correspond to the latent regression line and the confidence interval of 95%, respectively. **b** Biplot of selected genotypes (in blue color) and un-selected genotypes (in yellow triangles) based on predicted breeding values adjusted across all environments based on factor analytic covariance structure. The blue lines with arrows show the environments and their correlations

on GitHub page (https://github.com/whussain2/Analysis-pipeline). We provide a series of quality control steps (Fig. 1) to ensure better data quality and advance quality phenotypes for downstream analysis to get more reliable and accurate estimates or predictors. In general, data pre-processing steps include checking for noise, i.e., removing outliers, errors, or missing data, removing corrupt or inaccurate records, checking for normality assumptions for more reliable estimates, and looking for linearity or co-linearity for best model fit. Briefly, the steps mentioned in the document are:

a) *Missing data*: Here, the raw phenotypic data is visualized, and the proportion of missing data is visualized (section 3.1 in the sample pre-processing HTML file). Data can be filtered based on a certain proportion of missing data. For example, we dropped the trials having more than 20% of the data in this demonstration.
b) *Descriptive statistics*: In this step, phenotypic data points are summarized as mean, mode, coefficient of variation, the standard deviation for a given variable. Descriptive statistics helps to understand data much better and is the initial step to draw conclusions from the data and plan the model fitting. In addition, the descriptive statistics components may give a clue about the possible errors in the data. For example, the coefficient of variation (CV) calculated in this section can be used to measure variability for a given trait, determine the best plot size in uniformity trials, measure the stability of phenotypes, or measure variation in other individuals or populations attributes [7].
c) *Generate heatmaps of field*: In this step, interactive heatmaps of experimental field design are plotted to check for the field's spatial trends and the trend in the missing data. The presence of spatial effects in the data means that advanced models accounting for the spatial effects is required to get better estimates or predictions (section 3.3 in the sample pre-processing HTML file)
d) *Data visualization*: In this step, data is visualized using box plots, histograms, and QQ plots. Histograms show the data distribution and ideas about normality assumptions, and QQ plots depict quantiles of the datasets and assess correlated errors among data points for a given variable (section 3.4 in the sample pre-processing HTML file). Similarly, box plots shown in the data are an excellent technique to visualize the data distribution, dispersion, outlier detection, and trait variation. We also interactively presented the boxplots so that more information is obtained that is hidden in static boxplots.
e) *Filter for outliers*: In this step, outliers are identified and filtered using the Bonferroni-Holm test [8, 9]. Bonferroni-Holm test is more powerful and reliable when dealing with either small or large data sets. It can identify outliers based on the significance of residuals (section 3.5 in a sample pre-processing HTML file).
f) *Reliability of trial*: Based on yield data, we also look at the reliability of each trial or environment as a quality criterion. Any experimental trials having reliability lower than 0.2 are dropped from the analysis (section 3.6 in the sample pre-processing HTML file). More details on reliability and how to calculate it are given in the sample pre-processing HTML file.

Data analysis

The data analysis demonstrated here is divided into single or separate-trial analysis and multi-environment trial analysis. We demonstrated data analysis of MET both in the ASReml-R package [10] and lme4 R package [11]. We also demonstrated the analysis using marker data and extracting the genomic estimated breeding values (GEBVs) using the gBLUP model.

Data analysis in ASReml-R package

a) Single-trial analysis

In single-trial analysis, each trial or environment is analyzed separately. Data for a given variable is analyzed using a mixed model approach in the ASReml-R package. Data analysis includes basic mixed models accounting only for experimental design factors (blocks and replications here) and advanced mixed models accounting for experimental design factors and spatial effects or trends [12–14]. In total, five mixed models were implemented to analyze the data and correct for the experimental design factors and spatial trends (correlated residuals across the field dimensions). More advanced models can be used in phenotypic data analysis to account for the spatial trends [13, 15]. However, in this demo, we just showed examples of five mixed models. The best model selected based on AIC values (lower the AIC value better the model) and residual plot information [16] was used to extract the Best Linear Unbiased Predictors (BLUPs). In the analysis, we used genotypes as a random effect to extract the BLUPs, which are good for the phenotypic selection and ranking of the lines in the breeding programs [17, 18]. However, suppose we are going to use genomic selection or predictions in the breeding program. In that case, it is better to use the genotypes as a fixed effect and extract the BLUEs, which can be used as a response variable in the genomic prediction model to

extract the BLUPs or breeding values. The reason to use lines as fixed effects is to avoid double shrinkage if we use genotypes as random effects in both cases [17]. We also demonstrated how to use marker data and genomic relationship matrix to model the phenotypic data and extract the genomic estimated breeding values for each line using single-step genomic selection approach [19, 20].

The details of the five models used in the demonstration are given below and in the sample document of ASReml-R workflow HTML file (section 1.1) available on GitHub.

Model 1: In this model, we account for just experimental design factors, blocks and replications and no spatial trends, i.e., correlated residuals across the trial's dimensions (rows and columns). Here in this model, blocks and genotypes are used as random effects. The description of model 1 is as:

$$y_{ijk} = \mu + g_i + r_j + b_{jk} + \epsilon_{ijk}$$

where, y_{ijk} is the effect of *i-th* genotype in *j*-th replication and *k*-th block nested within *j*-th replication; μ is the overall mean; g_i is the random effect of the *i*-th genotype; r_j is the fixed effect of *j*-th replication; b_{jk} is the random effect of *k*-th block nested within *j*-th replication; ϵ_{ijk} is the residual error.

Here we assume residuals are independent and identically distributed as $\epsilon \sim iidN(0, \sigma_\epsilon^2)$.

Model 2: In this model, we account for experimental design factor blocks, replications, rows and columns, and no spatial trends. Blocks, rows, and columns, and genotypes were used as random effects. The description of model 2 is as:

$$y_{ijklm} = \mu + g_i + r_j + b_{jk} + c_l + ro_m + \epsilon_{ijklm}$$

where, y_{ijklm} is the effect of *i*-th genotype in the *j*-th replication, *k*-th block nested within *j*-th replication, *l*-th column and *m*-th row; μ is the overall mean; g_i is the random effect of the *i*-th genotype; r_j is the fixed effect of *j*-th replication; b_{jk} is the random effect of *k*-th block nested within *j*-th replication; c_l is the random effect of the *l*-th column; ro_m is the random effect of the *m*-th row; ϵ_{ijklm} is the residual error.

Here we assume residuals are independent and identically distributed as $\epsilon \sim iidN(0, \sigma_\epsilon^2)$.

Model 3: In this model, we account for experimental design factors, replications and blocks, and spatial trends, i.e., correlated residuals across rows and columns. Blocks and genotypes were used as random effects. The description of model 3 is as:

$$y_{ijk} = \mu + g_i + r_j + b_{jk} + \epsilon_{ijk}$$

where, y_{ijk} is the effect of *i*-th genotype in *j*-th replication and *k*-th block within *j*-th replication; μ is the overall mean; g_i is the random effect of the *i*-th genotype; r_j is the fixed effect of *j*-th replication; b_{jk} is the random effect of *k*-th block nested within *j*-th replication; ϵ_{ijk} is the residual error.

Here, we assume ϵ is a random effect representing correlated residuals based on the distance between plots along the rows and columns, where $\epsilon \sim N(0, \mathbf{R})$ and $\mathbf{R}$ is the covariance matrix of ϵ. The difference between this model and model 1 and model 2 described above is the structure of the covariance residuals $\mathbf{R} = \sigma_\epsilon^2 \mathbf{\Sigma_c}(\rho_c) \otimes \mathbf{\Sigma_r}(\rho_r)$. σ_ϵ^2 is the variance of spatially dependent residual, $(\rho_c)and$ $\mathbf{\Sigma_r}(\rho_r)$ represents the first-order autoregressive correlation matrices and ρ_c and ρ_r are the autocorrelation parameters for the columns and rows; $\otimes$ represents the Kronecker product between separable auto-regressive processes of the first order in the row-column dimensions [21–24].

Model 4: In this model, we account for experimental design factors, replications and blocks, and spatial trends, i.e., correlated residuals across rows only.

$$y_{ijk} = \mu + g_i + r_j + b_{jk} + \epsilon_{ijk}$$

where, y_{ijk} is the effect of *i*-th genotype in *j*-th replication and *k*-th block within *j*-th replication; μ is the overall mean; g_i is the random effect of the *i*-th genotype; r_j is the fixed effect of *j*-th replication; b_{jk} is the random effect of *k*-th block nested within *j*-th replication; ϵ_{ijk} is the residual error.

Here, we assume ϵ is a random effect representing correlated residual based on the distance between plots along the rows only, where $\epsilon \sim N(0, \mathbf{R})$ and $\mathbf{R}$ is the covariance matrix of ϵ. Here, $\mathbf{R} = \mathbf{I}_c$. $\sigma_\epsilon^2 \otimes \mathbf{\Sigma_{ro}}(\rho_{ro})$. σ_ϵ^2 is the variance of spatially dependent residual; $\mathbf{\Sigma_{ro}}(\rho_r)$ represents the first-order auto-regressive correlation matrices and ρ_{ro} is the auto-correlation parameters for the rows; $\otimes$ represents the Kronecker product between separable auto-regressive processes of the first order in the row dimensions. $\mathbf{I}_c$ represents independently and identically distributed variance structure for columns.

Model 5: In this model, we account for experimental design factors replications and blocks, and spatial trends i.e., correlated residuals across columns only.

$$y_{ijk} = \mu + g_i + r_j + b_{jk} + \epsilon_{ijk}$$

where, y_{ijk} is the effect of *i*-th genotype in *j*-th replication and *k*-th block within *j*-th replication; μ is the overall mean; g_i is the random effect of the *i*-th genotype; r_j is the fixed effect of *j*-th replication; b_{jk} is the fixed effect of *k*-th block within *j*-th replication; ϵ_{ijk} is the residual error.

Here, we assume ϵ is a random effect that represents correlated residual across columns only, where, $\epsilon \sim N(0, \mathbf{R})$ and $\mathbf{R}$ is the covariance matrix ofϵ, and

$\mathbf{R} = \sigma_{\epsilon}^2 \mathbf{\Sigma_c}(\rho_c) \otimes \mathbf{I}_r$. σ_{ϵ}^2 is the variance of spatially dependent residual; $\Sigma_c(\rho_c)$ represents the first-order autoregressive correlation matrices and ρ_c the autocorrelation parameters for the columns only; $\mathbf{I}_r$ represents independently and identically distributed variance structure for rows.

b) Multi-environment trial (MET) analysis

Depending upon the number of environments, MET analysis can be performed using single-stage or stage-wise approaches for analysis [12, 22, 25, 26]. The single-stage analysis is the golden standard to analyze the MET data. However, stage-wise analysis is more appropriate in the experiments or trials with unbalanced data sets, different experimental design factors across trials, and to avoid computational challenges of analyzing a huge number of trials. In a stage-wise or two-step approach, adjusted means are estimated per trial or environment, and weighted adjusted means (associated variance–covariance matrix) are fitted in the second step to get the predicted means for each genotype. This demonstration showed how to perform MET analysis using both the single-stage and two-stage or step analysis. The details on the MET analysis using the ASReml-R package is given on sample ASReml-R workflow HTML file (section 1.2) available on GitHub (https://github.com/whussain2/Analysis-pipeline).

i) Single-stage approach

In single-stage analysis, all the trials are analyzed jointly. Here, a joint analysis of MET is performed using a linear mixed model (LMM). The mixed model used is defined as:

$$y_{ijkl} = \mu + g_i + e_j + (ge)_{ij} + r_{jk} + b_{jkl} + \epsilon_{ijkl}$$

where, y_{ijkl} is the effect of i-th genotype is j-th environment, k-th replication nested within j-th environment and l-th block nested within k-th replication and j-th environment; μ is overall mean; g_i is the random effect of i-th genotype; e_j is the random effect of j-th environment; ge_{ij} is the interaction effect of i-th genotype with j-th environment; r_{jk} is the fixed effect of k-th replication nested within j-th environment, b_{jkl} is the random effect of l-th block nested within k-th replication and j-th environment, ϵ_{ijkl} is the residual.

In the matrix notation the mixed model can be represented as:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z_g u}_1 + \mathbf{Z_b u}_2 + \boldsymbol{\epsilon}$$

where, $\mathbf{y}$ is a vector of phenotypic trait values in all the genotypes; $\mathbf{X}$ is the design matrix of fixed effects of replications; $\mathbf{Z_g}$ is the design matrix of genotypes within environments that combine the main effects of genotypes, environments and genotype by environment interactions; $\mathbf{Z_b}$ is the random effect of blocks nested within the replications. $\boldsymbol{\beta}$ is the vector of fixed effects estimates; $\mathbf{u}_1$, $\mathbf{u}_2$, $\boldsymbol{\epsilon}$ are the vector of random effects of genotypes, blocks nested within replications, and residuals within environments, respectively.

Random effects are assumed to be random and normally distributed with zero mean vectors and variance–covariance matrices **B**, **G**, **R**, respectively, such that the joint distribution of these three terms is given as:

$$\begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \boldsymbol{\epsilon} \end{bmatrix} \sim \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R} \end{bmatrix} \right)$$

G is a variance–covariance (VCOV) matrix for the effect of genotypes within environments. For the **G** matrix, different VCOV structures were tested, such as compound symmetry (CS), diagonal (Diag), common genetic correlations (corgh), and FA of order k, in which k is the number of multiplicative components (FAk). For the **R** matrix, identity and diagonal and spatial trends VCOV structures were tested [27–33].

Here in this demo, we applied 10 models depending upon the variance–covariance structure of random and residual effects. The brief description of these models is given below:

Model 1 and model 2: Models 1 and 2 were basic models in which we assume the variance for residuals and random effects are independent and normally distributed, and implying genotypes have the same variance over the environments. Genotypic variance and covariance between pairs of environments are homogeneous, which corresponds to compound symmetry (CS) variance–covariance structure in the mixed model.

Model 3: In this model we assume different variances across environments, i.e., heterogeneous error variances across environments.

Model 4: In this model, we assume different variances across environments with spatial variance structure same for all the environments. It is assumed that each environment comprises of a rectangular array of rows(r) and columns (c) with $\mathbf{R} = \sigma_{\epsilon}^2 \mathbf{\Sigma_c}(\rho_c) \otimes \mathbf{\Sigma_r}(\rho_r)$. σ_{ϵ}^2 is the variance of spatially dependent residual; $\mathbf{\Sigma_c}(\rho_c)$ *and* $\mathbf{\Sigma_r}(\rho_r)$ represents the first-order autoregressive correlation matrices and ρ_c and ρ_r are the autocorrelation parameters for the columns and rows; $\otimes$ represents the Kronecker product between separable auto-regressive processes of the first order in the row-column dimension.

Model 5: This model assumes different variances across environments with spatial variation structure specific to each environment. The best spatial model defined for each environment structure was obtained from the separate analysis done for each environment, as shown in the above section.

Model 6: In this model, we assume a uniform correlation and heterogeneous genetic variance. Each environment has a unique genetic variance, but there were no correlations between environments.

Model 7: In this model, we assume unique genetic variance in each environment with uniform correlations between environments.

Models 8, 9, and 10: Here, we applied factor analytical models FA of order k, in which k is the number of multiplicative components (FA_k). In factor analytical model 10, we also assume the spatial variance structures are the same across each environment. More details on factor analytical models can be found in these papers [30–32, 34–38].

ii) Two-stage approach

Here in this section, the joint analysis of MET data was performed in two-stages. In the first-stage adjusted means as BLUEs and residuals in each environment were obtained by considering the genotypes as fixed effect. At this step, the adjusted means of genotypes were corrected for the experimental design factors, including blocks and replications and spatial trends in each environment. The model used follows as:

$$y_{ijk} = \mu + g_i + r_j + b_{jk} + \epsilon_{ijk}$$

where, y_{ijk} represents adjusted means for i-th observation in j-th replication and k-th block nested within j-th replication; μ is the overall mean; g_i is the effect of i-th genotype; r_j is the effect of j-th replications; b_{jk} is the effect of k-th block nested within j-th replication; ϵ_{ijk} is the residual error.

Here, we assume $\epsilon \sim N(0, \mathbf{R})$ and $\mathbf{R}$ is the covariance matrix of ϵ and $\mathbf{R} = \sigma_{\epsilon}^2 \mathbf{\Sigma_c}(\rho_c) \otimes \mathbf{\Sigma_r}(\rho_r)$. σ_{ϵ}^2 is the variance of spatially dependent residual; $\mathbf{\Sigma_c}(\rho_c)$ *and* $\mathbf{\Sigma_r}(\rho_r)$ represents the first-order autoregressive correlation matrices and ρ_c and ρ_r are the autocorrelation parameters for the columns and rows; $\otimes$ represents the Kronecker product between separable auto-regressive processes of the first order in the row-column dimensions.

In the second-stage, a mixed model was fitted across each environment using the BLUEs obtained from the first-stage as response variable. The model used follows as:

$$y_{ij} = \mu + g_i + e_j + (ge)_{ij} + \epsilon_{ij}$$

where, y_{ij} is the BLUE value for i-th observation in j-th environment; μ is the overall mean; g_i is the random effect of i-th genotype; e_j is the random effect of j-th environment; ge_{ij} is the interaction effect of i-th genotype with j-th environment; ϵ_{ij} is the residual error.

Here, we assume the error is known from the first stage. To account for the errors in the second stage, reciprocal of squared standard errors (equal to diagonal of variance–covariance matrix) as absolute weights were used, thus constraining the residual variance to one. This procedure of obtaining weights is thoroughly described in Method 2 given in [39].

Data analysis in lme4 R package

Phenotypic data modeling is also demonstrated in the lme4 R package, an open-source R package for users who don't have access to the commercial ASReml-R package.

In the lme4 R package, data analysis is again divided into two methods of separate analysis and MET analysis. The details on the analysis in lme4 are available in the lme4 R workflow HTML file available on GitHub (https://github.com/whussain2/Analysis-pipeline). Unfortunately, we cannot reproduce all the analysis depicted in ASReml analysis due to the limitation lme4 has in performing mixed-model analysis.

The description of models used in lme4 for separate and MET analysis is given below:

Model 1. lme4: For the separate analysis following mixed model was used. This is equivalent to the basic model 1 used in ASReml analysis. The model followed as:

$$y_{ijk} = \mu + g_i + r_j + b_{jk} + \epsilon_{ijk}$$

where, y_{ijk} is the effect of i-th genotype in j-th replication and k-th block within j-th replication; μ is the overall mean; g_i is the random effect of the i-th genotype; r_j is the fixed effect of j-th replication; b_{jk} is the random effect of k-th block within j-th replication; ϵ_{ijk} is the residual error.

Here, we assume residuals are independent and identically distributed as $\epsilon \sim iidN(0, \sigma_\epsilon^2)$

Model 2. lme4: For the combined analysis following mixed model was used in lme4:

$$y_{ijkl} = \mu + g_i + e_j + (ge)_{ij} + r_{jk} + b_{jkl} + \epsilon_{ijkl}$$

where, y_{ijk} is the effect of i-th genotype in j-th environment, k-th replication within j-th environment and l-th block within k-th replication within j-th environment; μ is the overall mean; g_i is the random effect of the i-th genotype; e_j is the fixed effect of j-th environment; ge_{ij} is an interaction effect of i-th genotype in j-th environment; r_{jk} is fixed effect of k-th replication within genotype j-th environment; b_{jkl} is the random effect of l-th block nested within k-th replication and j-th environment; ϵ_{ijkl} is the residual error.

The models described above are equivalent to the MET model used in ASReml-R, except we are not modeling any spatial variation here in lme4 as was done in the ASReml-R package.

Analysis by incorporating marker data

In the crop breeding programs, it is now a routine to integrate the marker data with phenotypic data to predict the genetic merit of individuals in the framework of mixed-model equations by incorporating a genomic relationship matrix (GRM) constructed by using marker data. This section demonstrated how to extend the phenotypic data analysis to marker-based analysis using a relationship matrix. Here, we show an example of how to fit the gBLUP model to get the genomic estimated breeding values (GEBV). More details on other predictive-based models using marker data can be found in these articles [17, 40, 41, 41–44]. In gBLUP genomic relationship matrix (GRM) based on marker, data is used, and GRM defines similarity or the covariance between genotypes or individuals at the genomic level. More details on how to fit the gBLUP model are given in the ASReml-R workflow HTML file (section 1.3). Briefly, here we are providing general model details and how to construct GRM.

In the matrix notation, the gBLUP model is described as:

$$\boldsymbol{y} = \mathbf{X\beta} + \mathbf{Z_g u_g} + \mathbf{Z_b u_{rb}} + \mathbf{Z_e u_e} + \boldsymbol{\varepsilon}$$

where, $\mathbf{y}$ is a vector of individual phenotypes; $\mathbf{X}$ is a design matrix of replications; $\boldsymbol{\beta}$ is a vector of fixed effects of replications; $\mathbf{Z_g}$ is a design matrix of marker effects; $\mathbf{u_g}$ is a vector of random marker effects; $\mathbf{Z_{rb}}$ is a design matrix of non-genetic block effects nested within replications; $\mathbf{u_{rb}}$ is a vector of random block effects; $\mathbf{Z_e}$ is a design matrix of non-genetic random effect of environments and genotype x environment interactions; $\mathbf{u_e}$ is a vector of main environment and interaction effects; $\boldsymbol{\varepsilon}$ is the vector of residual errors.

Further, we assume random effects are normally distributed with zero mean vectors and variance–covariance matrices **G**, **B**, **R** as described in single-stage approach of MET analysis above. Here, the expected variance of markers is given as $\text{Var}(\mathbf{u_g}) = \sigma_g^2 \mathbf{G}$, where **G** is genomic or kinship covariance matrix of n x m dimensions (n is no. of markers and m is no. of individuals) representing the genomic similarity of individuals.

Genomic (G) matrix or GRM [37, 39] is constructed using the following equation:

$$\boldsymbol{G} = \frac{\mathbf{X\acute{X}}}{\text{n}}$$

Here $\mathbf{X}$ is a scaled and centered matrix of marker data, $\acute{\mathbf{X}}$ transpose of $\mathbf{X}$ matrix, and n is the number of total markers or columns in marker data.

Results

This section demonstrated how to extract the results from the separate analysis or MET analysis using either the ASReml-R package or lme4 R packages. Users can extract the specific results depending upon the objectives. In a separate analysis, we showed how to pull BLUPs, variance components, heritability, ANOVA, and variogram to check for spatial trends for the trait using the best model. In MET analysis, besides these results mentioned above, we used the ASExtras4 R package (https://mmade.org/asextras4/) to extract additional results, including correlation and covariance matrix, G x E BLUPs, PCA biplot, and latent regression figures to

check the stability of genotypes (**Fig. 2**). For more details on these results, check the HTML workflows of ASReml and lme4 analysis available on GitHub.

Additionally, we demonstrated how to extract the heritability and generalized heritability [45] using different approaches. For example, we are dealing with spatial or complex models in this data, so calculating heritability based on the method described by [17, 45] is used to estimate heritability. Briefly, for complex residual structures and unbalanced experimental designs, heritability estimation is given by equation $H_c = 1 - \frac{\overline{V}_{BLUP}}{2\sigma_g^2}$, where $\overline{V}_{BLUP}$ is a mean–variance difference of two BLUPs and σ_g^2 is a variance of genotypes. Note that this definition of heritability is related to the reliability of breeding value predictions.

Further BLUPs extracted here are used to rank the genotypes for making selections in breeding decisions. In lme4 R package analysis, ANOVA, variance components, fixed effect as BLUEs, random effect as BLUPs, and heritability were extracted. More details on this are given in the sample lme4 R workflow HTML document available on GitHub (https://github.com/whussain2/Analysis-pipeline).

Converting analysis workflow into a document

One of the biggest challenges in data analysis is reporting it and presenting it in a well-documented format for better understanding and making breeding decisions. In data analysis, the 'copy and paste' system is mostly used to report the results, which is time-consuming and highly error-prone. Thus, the situation demands a unique technique that converts the analysis workflow explicitly into a report or document for easy interpretations, understanding, and sharing. Here, we not only reported the analysis workflow as described above but also demonstrated how this workflow could be re-designed into a reproducible document for better interpretation, visualization, and seamless sharing with partners. The generated report is the state-of-the-art implementation of an analysis workflow with a description of R scripts and results with interpretations embedded as one unified document. A sample document is available in the GitHub repository at https://github.com/whussain2/Analysis-pipeline. The main features of the document are:

1) The analysis pipeline described and given in Fig. 1 is converted into a highly reproducible document, and the same report and analysis pipeline can be generated anytime when required. The sample source codes of the analytical pipeline and demo data set can be directly downloaded from the GitHub repository (https://github.com/whussain2/Analysis-pipeline). The instructions on how to run the analysis pipeline on a local computer are given on the GitHub repository page.
2) Any new data and editing/corrections to the existing pipeline can be done by simply re-knitting the R markdown '*.Rmd*' document (https://rmarkdown.rstudio.com/articles_intro.html). This analytical pipeline avoids manually updating or generating reports or PowerPoint slides, which are otherwise highly prone to errors and time-consuming.
3) The document includes metadata (information about the field trial design, data collection, experimental design, and more) at the beginning for quick identification, location, and association of data and analysis at any given time (Fig. 3a).
4) The document is well structured and organized. For example, the document is divided into sections with headings and subheadings to increase accessibility and cognition. The table of contents is always visible in the document making it faster and easier to navigate within a document (Fig. 3a). Additionally, readers have the flexibility to hide the sections for better readability and accessibility.
5) The document is currently generated in HTML, which upon download can be easily opened in any browser without requiring any access to the internet. Further, HTML files can be shared easily and hosted on websites for easy sharing and future use.
6. The graphics in the document are highly dynamic and interactive. Simply hovering a cursor on the plot will display the additional hidden information, which is impossible in static pictures. For example, the box plots and heatmaps of experimental field design to visualize spatial trends are highly dynamic and interactive (Fig. 3b and c). Additionally, graphics can be easily exported to the local drive.
7) The output generated in the form of tables is highly dynamic and interactive. The tables generated can be easily managed, searched, and sorted like a mini excel sheet (Fig. 3d). Interestingly, tables can also be exported in various formats or printed directly within the document. The tables and result outputs being in the same file completely avoids the option of saving the files on computers and digging into them to extract useful information in making presentations or in undertaking breeding decisions.
8) Complete description and details of scripts, procedures, and methods used for analysis are elaborated in the same document. Results generated in the document in the form of figures and or tables have been thoroughly described to aid in the interpretation and better understanding. Hyperlinks have been embedded in the required sections to help in understand-

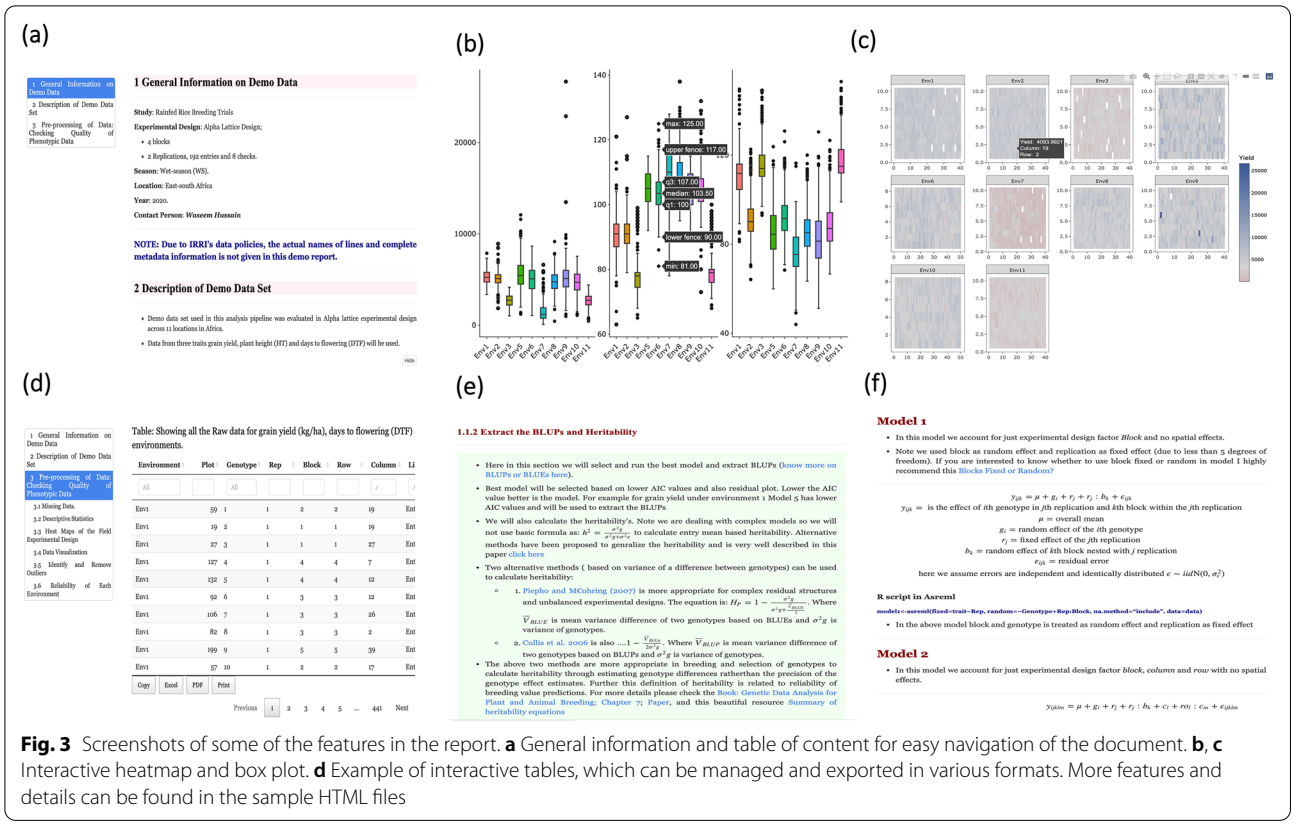

Fig. 3 Screenshots of some of the features in the report. **a** General information and table of content for easy navigation of the document. **b**, **c** Interactive heatmap and box plot. **d** Example of interactive tables, which can be managed and exported in various formats. More features and details can be found in the sample HTML files

ing the concepts and add knowledge to the users. For example, web sources on how to interpret the box plots; methods used to calculate heritability with complex models; spatial analysis modeling, and much more have been hyperlinked in the document

Conclusions

Crop breeding trial analysis and procedures are well established in the literature; however, putting them into an end-to-end analysis workflow with detailed descriptions and explanations is not available. A helpful guide and tutorial to thoroughly understand the phenotypic data analysis is a crucial requirement in breeding programs. Here, we took an initiative to modernize the data analysis of IRRI's RRB program, which can be easily put into practice and will be of great use to the crop breeding communities having full-fledged breeding programs. We believe this will serve a helpful guide specifically for researchers or plant breeders who have little knowledge about phenotypic data analysis. We reported the workflow and analytical pipeline and gave step-by-step instructions and explanations on how to analyze the phenotypic data in the best possible way for making the right breeding decisions. Conclusively, we reported end-to-end implementation of phenotypic data analyses of plant breeding field trials and re-design it into a reproducible document for easy sharing, understanding, and interpretation. In the future, we look forward to incorporating predictive analytics based on higher advanced statistical modeling and big data.

Abbreviations

BLUPs: Best linear unbiased predictions; BLUEs: Best linear unbiased estimation; HTML: Hypertext markup language; MET: Multi-environment trial; NARES: National agricultural research and extension systems.

Acknowledgements

The rainfed breeding team expresses gratitude to the IRRI's irrigated breeding team led by Joshua N. Cobb to initiate the idea of the data analysis workflow. The authors are thankful to the IRRI's internal reviewers Shalabh Dixit and Jerome Bartholome, for constructive feedback and comments for improving the manuscript. The authors are also grateful to Hans Bhardwaj (Head, Rice Breeding Innovations, IRRI) for his support. Authors are also thankful to John Damien Platten (Senior Scientist I and Head of Breeding Innovations) for the initial feedback on the analysis pipeline. The authors are also grateful to the anonymous reviewers for the substantial improvement of this manuscript.

Authors' contributions

The phenotypic data analytical pipeline workflow, different mixed model approaches, and methodology concept was designed by WH. Directions and supervision of the study by SB and WH. MA, MC, and AP involved in data gathering, compilation, and analysis. SMT and JR helped in the data collections and preparations. All authors read and approved the manuscript.

Funding
The authors would like to thank and acknowledge the Bill & Melinda Gates Foundation (BMGF) for providing research on the Accelerated Genetic Gains in Rice Alliance (AGGRI) project under ID A-2017-129.

Availability of data and materials
The datasets and R scripts used to run all the analysis demonstrated in this study are available on the GitHub page at https://github.com/whussain2/Analysis-pipeline.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 21 May 2021 Accepted: 17 January 2022
Published online: 05 February 2022

References

1. Dar MH, Waza SA, Shukla S, Zaidi NW, Nayak S, Hossain M, et al. Drought tolerant rice for ensuring food security in Eastern India. Sustainability. 2020;12:2214.
2. Beaulieu-Jones BK, Greene CS. Reproducibility of computational workflows is automated using continuous analysis. Nat Biotechnol. 2017;35:342–6.
3. R Core Team 2018. R: A language and environment for statistical computing. e. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
4. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. 2nd ed. Springer International Publishing; 2016. https://www.springer.com/gp/book/9783319242750. Accessed 20 Jul 2020.
5. Baumer B, Udwin D. R Markdown. WIREs Computational Statistics. 2015;7:167–77.
6. Triantafyllidis CP, Papageorgiou LG. An integrated platform for intuitive mathematical programming modeling using LaTeX. PeerJ Comput Sci. 2018;4:e161.
7. Bowman DT. Common use of the CV: a statistical aberration in crop performance trials. J Cotton Sci. 2001;5:5.
8. Philipp N, Weise S, Oppermann M, Börner A, Keilwagen J, Kilian B, et al. Historical phenotypic data from seven decades of seed regeneration in a wheat ex situ collection. Sci Data. 2019;6:137.
9. Bernal-Vasquez A-M, Utz H-F, Piepho H-P. Outlier detection methods for generalized lattices: a case study on the transition from ANOVA to REML. Theor Appl Genet. 2016;129:787–804.
10. Butler DG, Cullis BR, Gilmour AR, Gogel BJ, Thompson R. ASReml estimates variance components under a general linear. 2018;188.
11. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models using lme4. [stat] 2014. arXiv:1406.5823. Accessed 21 Mar 2021.
12. Smith AB, Cullis BR, Thompson R. The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. J Agric Sci. 2005;143:449–62.
13. Isik F, Holland J, Maltecca C. Spatial analysis. In: Isik F, Holland J, Maltecca C, editors. Genetic data analysis for plant and animal breeding. Cham: Springer; 2017. p. 203–26. https://doi.org/10.1007/978-3-319-55177-7_7.
14. Giri K, Chia K, Chandra S, Smith KF, Leddin CM, Ho CKM, et al. Modelling and prediction of dry matter yield of perennial ryegrass cultivars sown in multi-environment multi-harvest trials in south-eastern Australia. Field Crops Res. 2019;243:107614.
15. Hoefler R, González-Barrios P, Bhatta M, Nunes JAR, Berro I, Nalin RS, et al. Do spatial designs outperform classic experimental designs? JABES. 2020;25:523–52.
16. Piepho HP, Williams ER. Linear variance models for plant breeding trials. Plant Breed. 2010;129:1–8.
17. Piepho HP, Möhring J, Melchinger AE, Büchse A. BLUP for phenotypic selection in plant breeding and variety testing. Euphytica. 2008;161:209–28.
18. Bernardo R. Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE. Heredity. 2020;125:375–85.
19. Oakey H, Cullis B, Thompson R, Comadran J, Halpin C, Waugh R. Genomic selection in multi-environment Crop trials. G3 Genes Genomes Genetics. 2016;6:1313–26.
20. Ovenden B, Milgate A, Wade LJ, Rebetzke GJ, Holland JB. Accounting for genotype-by-environment interactions and residual genetic variation in genomic selection for water-soluble carbohydrate concentration in wheat. G3 Genes Genomes Genetics. 2018;8:1909–19.
21. Gilmour AR, Cullis BR, Verbyla AP. Accounting for natural and extraneous variation in the analysis of field experiments. J Agric Biol Environ Stat. 1997;2:269–93.
22. Gogel B, Smith A, Cullis B. Comparison of a one- and two-stage mixed model analysis of Australia's National Variety Trial Southern Region wheat data. Euphytica. 2018;214:44.
23. Andrade MHML, Filho CCF, Fernandes MO, Bastos AJR, Guedes ML, de Marçal TS, et al. Accounting for spatial trends to increase the selection efficiency in potato breeding. Crop Sci. 2020;60:2354–72.
24. Bernardeli A, de Rocha JR, Borém A, Lorenzoni R, Aguiar R, Silva JNB, et al. Modeling spatial trends and enhancing genetic selection: an approach to soybean seed composition breeding. Crop Sci. 2020. https://doi.org/10.1002/csc2.20364.
25. Piepho H-P, Möhring J, Schulz-Streeck T, Ogutu JO. A stage-wise approach for the analysis of multi-environment trials. Biom J. 2012;54:844–60.
26. Damesa TM, Möhring J, Worku M, Piepho H-P. One step at a time: stage-wise analysis of a series of experiments. Agron J. 2017;109:845–57.
27. Malosetti M, Ribaut J-M, van Eeuwijk FA. The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. Front Physiol. 2013;4:44.
28. van Eeuwijk FA, Bustos-Korts DV, Malosetti M. What should students in plant breeding know about the statistical aspects of genotype × environment interactions? Crop Sci. 2016;56:2119–40.
29. Isik F, Holland J, Maltecca C. Multi environmental trials. In: Isik F, Holland J, Maltecca C, editors. Genetic data analysis for plant and animal breeding. Cham: Springer; 2017. p. 227–62. https://doi.org/10.1007/978-3-319-55177-7_8.
30. Jia G, Booker HM. Optimal models in the yield analysis of new flax cultivars. Can J Plant Sci. 2018;98:897–907.
31. Hernández MV, Ortiz-Monasterio I, Pérez-Rodríguez P, Montesinos-López OA, Montesinos-López A, Burgueño J, et al. Modeling genotype × environment interaction using a factor analytic model of on-farm wheat trials in the Yaqui Valley of Mexico. Agron J. 2019;111:2647–57.
32. de Souza VF, de Ribeiro PC, Júnior ICV, Oliveira ICM, Damasceno CMB, Schaffert RE, et al. Exploring genotype × environment interaction in sweet sorghum under tropical environments. Agron J. 2021;113:3005–18.
33. Piepho H-P. Analyzing genotype-environment data by mixed models with multiplicative terms. Biometrics. 1997;53:761–6.
34. Kelly AM, Smith AB, Eccleston JA, Cullis BR. The accuracy of varietal selection using factor analytic models for multi-environment plant breeding trials. Crop Sci. 2007;47:1063–70.
35. Burgueño J, Crossa J, Cornelius PL, Yang R-C. Using factor analytic models for joining environments and genotypes without crossover genotype × environment interaction. Crop Sci. 2008;48:1291–305.
36. Cullis BR, Smith AB, Beeck CP, Cowling WA. Analysis of yield and oil from a series of canola breeding trials. Part II. Exploring variety by environment interaction using factor analysis. This article is one of a selection of papers from the conference "Exploiting Genome-wide Association in Oilseed Brassicas: a model for genetic improvement of major OECD crops for sustainable farming." Genome. 2010;53:1002–16.
37. Smith AB, Ganesalingam A, Kuchel H, Cullis BR. Factor analytic mixed models for the provision of grower information from national crop variety testing programs. Theor Appl Genet. 2015;128:55–72.
38. Sjoberg SM, Carter AH, Steber CM, Campbell KAG. Application of the factor analytic model to assess wheat falling number performance and stability in multienvironment trials. Crop Sci. 2021;61:372–82.

39. Möhring J, Piepho H-P. Comparison of weighting in two-stage analysis of plant breeding trials. Crop Sci. 2009;49:1977–88.
40. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
41. Jannink J-L, Lorenz AJ, Iwata H. Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics. 2010;9:166–77.
42. Zapata-Valenzuela J, Whetten RW, Neale D, McKeand S, Isik F. Genomic estimated breeding values using genomic relationship matrices in a cloned population of Loblolly Pine. G3 Genes, Genomes, Genetics. 2013;3:909–16.
43. Wang X, Xu Y, Hu Z, Xu C. Genomic selection methods for crop improvement: current status and prospects. Crop J. 2018;6:330–40.
44. Wang J, Zhou Z, Zhang Z, Li H, Liu D, Zhang Q, et al. Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. Heredity. 2018;121:648–62.
45. Piepho H-P, Möhring J. Computing heritability and selection response from unbalanced plant breeding trials. Genetics. 2007;177:1881–8.

Publisher's Note