Theoretical and Applied Genetics (2023) 136:65
https://doi.org/10.1007/s00122-023-04298-x

ORIGINAL ARTICLE

Fully efficient, two-stage analysis of multi-environment trials with directional dominance and multi-trait genomic selection

Jeffrey B. Endelman[1]

Received: 28 September 2022 / Accepted: 2 January 2023 / Published online: 22 March 2023

Abstract

Key message **R/StageWise enables fully efficient, two-stage analysis of multi-environment, multi-trait datasets for genomic selection, including support for dominance heterosis and polyploidy.**

Abstract Plant breeders interested in genomic selection often face challenges to fully utilizing multi-trait, multi-environment datasets. R package StageWise was developed to go beyond the capabilities of most specialized software for genomic prediction, without requiring the programming skills needed for more general-purpose software for mixed models. As the name suggests, one of the core features is a fully efficient, two-stage analysis for multiple environments, in which the full variance–covariance matrix of the Stage 1 genotype means is used in Stage 2. Another feature is directional dominance, including for polyploids, to account for inbreeding depression in outbred crops. StageWise enables selection with multi-trait indices, including restricted indices with one or more traits constrained to have zero response. For a potato dataset with 943 genotypes evaluated over 6 years, including the Stage 1 errors in Stage 2 reduced the Akaike Information Criterion (AIC) by 29, 67, and 104 for maturity, yield, and fry color, respectively. The proportion of variation explained by heterosis was largest for yield but still only 0.03, likely because of limited variation for the genomic inbreeding coefficient. Due to the large additive genetic correlation (0.57) between yield and maturity, naïve selection on an index combining yield and fry color led to an undesirable response for later maturity. The restricted index coefficients to maximize genetic merit without delaying maturity were identified. The software and three vignettes are available at https://github.com/jendelman/StageWise.

Introduction

During the first decade of the twenty-first century, the focus of genomic selection research was the development of theory and methods (e.g., Meuwissen et al. 2001; Habier et al. 2007; Daetwyler et al. 2008; Bernardo and Yu 2007; VanRaden 2008), and most researchers worked in animal rather than plant breeding. This changed in the following decade with the development of specialized software for genomic prediction, including rrBLUP (Endelman 2011), GAPIT (Lipka et al. 2012), synbreed (Wimmer et al. 2012), BGLR (Pérez and de los Campos 2014), and sommer (Covarrubias-Pazaran 2016). Over the last several years, new software development has emphasized multi-trait prediction models (Montesinos-López et al. 2019; Runcie et al. 2021; Pérez-Rodríguez and de los Campos 2022). Collectively, these software publications have been cited several thousand times, which reflects their enabling role for the adoption of genomic selection, particularly in plant breeding.

However, these packages have limitations to handle the full complexity of plant breeding data, with different experimental designs, heritabilities, and spatial models for non-genetic variation. The challenge of properly analyzing multi-environment datasets existed before genomic selection, which led to the concept of a two-stage analysis (Frensham et al. 1997). In Stage 1, genotype means are estimated as fixed effects for each environment, which become the response variable in Stage 2. The errors of the Stage 1 estimates are typically different, and failure to account for this in Stage 2 leads to sub-optimal results (Möhring and Piepho 2009). A "fully efficient" two-stage analysis uses the full variance–covariance matrix of the Stage 1 genotype means in Stage 2, rather than a diagonal approximation (Piepho et al. 2012; Damesa et al. 2017). Previous examples of a properly weighted, two-stage analysis have used one of three

Communicated by Huihui Li.

✉ Jeffrey B. Endelman
endelman@wisc.edu

[1] Department of Horticulture, University of Wisconsin-Madison, Madison, WI 53706, USA

well-established, REML-based programs for mixed models: SAS PROC MIXED (SAS Institute Inc, Cary, NC), ASReml (Gilmour et al. 2015), or ASReml-R (Butler et al. 2018). All three software allow the variance–covariance matrix of the random effect for Stage 1 errors to be specified while estimating the other, unknown variance components of the Stage 2 model. Despite this precedent, many studies continue to ignore Stage 1 errors, and I believe a major reason is the additional programming skill required.

The goal of the current research was to develop a new R package (R Core Team 2022) for genomic selection that makes fully efficient, two-stage analysis more accessible to plant breeders. The software, called StageWise, returns empirical BLUPs using variance components estimated with ASReml-R. It also works for polyploids and incorporates advanced features such as directional dominance and multi-trait selection indices.

Methods

Single trait with homogeneous GxE

The response variable for Stage 2 is the Stage 1 BLUEs for the effect of genotype in environment. The mixed model with homogeneous GxE can be written as

$$BLUE\left[g_{ij}\right] = y_{ij} = E_j + g_i + gE_{ij} + s_{ij} \quad (1)$$

where g_{ij} is the genotypic value for individual (or clone) i in environment j, E_j is the fixed effect for environment j, g_i is the random effect for individual i across environments, and the GxE effect, gE_{ij}, is actually the model residual (Damesa et al. 2017). The s_{ij} effect, which represents the Stage 1 estimation error, is multivariate normal with no free variance parameters: the variance–covariance matrix is the direct sum of the variance–covariance matrices of the Stage 1 BLUEs (Damesa et al. 2017). The gE_{ij} are independent and identically distributed (i.i.d.), which implies a single genetic correlation between all environments. Without marker data, the software assumes the g_i effects are i.i.d.

When marker data are provided, the software decomposes g_i into additive and non-additive values. The vector of additive values is multivariate normal with covariance proportional to a genomic additive matrix **G** (VanRaden 2008 Method 1, extended to arbitrary ploidy). If **W** represents the centered matrix of allele dosages (n individuals x m bi-allelic markers with frequencies $p = 1–q$), then for ploidy ϕ,

$$\mathbf{G} = \frac{\mathbf{W}\mathbf{W}^{\mathrm{T}}}{\phi \sum_k p_k q_k} \quad (2)$$

If a three-column pedigree is provided, **G** can be blended with the pedigree relationship matrix **A** (calculated using R package AGHmatrix (Amadeu et al. 2016)) to produce $\mathbf{H} = (1-\omega)\mathbf{G} + \omega\mathbf{A}$, for $0 \le \omega \le 1$ (Legarra et al. 2009; Christensen and Lund 2010). In addition to the additive polygenic effect, the user can indicate some markers should be included as additive (fixed effect) covariates in Eq. (1), to capture large effect QTL.

Directional dominance

Two models for the non-additive genetic values are available. In the genetic residual model, the non-additive values are i.i.d. The other option is a directional (digenic) dominance model, which follows the classical framework of Fisher (1941) and Kempthorne (1957) and is a refinement of recent research (Vitezica et al. 2013; Xiang et al. 2016; Endelman et al. 2018; Batista et al. 2022). For a locus with two alleles designated 0/1, there are three digenic dominance effects $\beta_{00}, \beta_{01}, \beta_{11}$, which equal the dominance deviation in diploids, but more generally for any ploidy are the coefficients for regressing the dominance deviation on diplotype dosage. (Higher order dominance effects for polyploids are not considered.) These dominance effects can be expressed in terms of a parameter that has no established name but may be called a digenic substitution effect, β, by analogy with the allele substitution effect α for additive effects. The β parameter represents the average change in dominance deviation per unit increase in dosage of the heterozygous diplotype:

$$\beta = \beta_{01} - \frac{1}{2}\left(\beta_{00} + \beta_{11}\right) \quad (3)$$

(This differs from the scaling in Endelman et al. (2018) by –2 so that β in Eq. (3) equals d in the classical diploid model of Vitezica et al. (2013).) Designating the frequency of allele 1 as $p = 1 - q$, the dominance effects can be expressed in terms of the substitution effect:

$$\begin{aligned} \beta_{00} &= -2p^2\beta \\ \beta_{01} &= 2pq\beta \\ \beta_{11} &= -2q^2\beta \end{aligned} \quad (4)$$

The dominance value of an individual is the sum of its dominance effects and can be written as $Q\beta$, where the dominance coefficient Q for ploidy ϕ and allele dosage X (of allele 1) is

$$Q = -2\binom{\phi}{2}p^2 + 2p(\phi-1)X - X(X-1) \quad (5)$$

In Eq. (5), $\binom{\phi}{2}$ is the binomial coefficient. The dominance genetic variance, V_D, is $\binom{\phi}{2}$ times the variance of the dominance effects, $4p^2q^2\beta^2$. Extending this framework to m

loci, the dominance value is $\sum_{k=1}^{m} Q_k\beta_k$, and the dominance variance is

$$V_D = \binom{\phi}{2} \sum_{k=1}^{m} 4p_k^2 q_k^2 \beta_k^2 + \sum_k \sum_{k' \neq k} \beta_k \beta_{k'} \mathrm{cov}\left[Q_k, Q_{k'}\right] \quad (6)$$

The first term in Eq. (6) is the dominance *genic* variance, which depends on allele frequencies but not LD between loci. The second term is the disequilibrium covariance, which can be positive or negative.

In classical quantitative genetics, the substitution effects are fixed parameters, but to compute dominance values by BLUP, we switch to viewing them as random normal effects (de los Campos et al. 2015), with mean μ_β and variance σ^2_β. For a trait with no average heterosis in the population, $\mu_\beta = 0$ (Varona et al. 2018). Let $\mathbf{Q}$ denote the $n \times m$ matrix of dominance coefficients for n individuals at m loci. The vector of dominance values $\mathbf{Q\beta}$ is multivariate normal, with mean $\mathbf{Q1}\mu_\beta$ and variance–covariance matrix $\mathbf{QQ}^{\mathrm{T}}\sigma^2_\beta$. Equivalently, the dominance values can be written as

$$\mathbf{Q}\boldsymbol{\beta} = -b\mathbf{F} + \mathbf{d}_0 \quad (7)$$

where $\mathbf{F}$ is a vector of genomic inbreeding coefficients, with regression coefficient b (positive value implies heterosis), and $\mathbf{d}_0 \sim \mathrm{MVN}\left(0, \mathbf{D}\sigma^2_D\right)$ represents dominance with no average heterosis. The genomic dominance matrix $\mathbf{D}$ is defined by interpreting its variance component σ^2_D as the expected value of the classical dominance variance with respect to the substitution effects, assuming no overall heterosis. From Eq. (6) the result is

$$\sigma^2_D = E\left[V_D\right] = \sigma^2_\beta \binom{\phi}{2} \sum_k 4p_k^2 q_k^2 \quad (8)$$

which leads to

$$\mathbf{D} = \frac{\mathbf{QQ}^{\mathrm{T}}}{\binom{\phi}{2} \sum_k 4p_k^2 q_k^2} \quad (9)$$

From Eq. (7), the vector of genomic inbreeding coefficients $\mathbf{F}$ is proportional to the row sum of $\mathbf{Q}$. The correct scaling is derived by considering the expected value of Q (Eq. 5) in the classical sense (where genotypes are random and parameters are fixed), for a completely inbred population in which homozygotes of allele 1 occur with frequency p. Under these conditions, $E[X] = \phi p$ and $E\left[X^2\right] = \phi^2 p$, which leads to $E[Q] = -2pq\binom{\phi}{2}$. Extending this to multiple loci and equating the result to $F = 1$ sets the proportionality constant and leads to the following definition:

$$\mathbf{F} = \frac{-\mathbf{Q1}}{\binom{\phi}{2} \sum_k 2p_k q_k} \quad (10)$$

The vector of genomic inbreeding coefficients is included as a fixed effect covariate in the Stage 2 model. Inbreeding coefficients can also be computed from the diagonal elements of the additive relationship matrix (either A or G) according to $(G-1)/(\phi-1)$ (Henderson 1976; Gallais 2003; Endelman and Jannink 2012).

Extension to multiple locations or traits

StageWise has the option of including a random effect $g(L)$ in Stage 2 for genotype within location (or L can represent some other factor, such as management). Using the subscript k to designate location, the linear model (Eq. 1) becomes

$$\mathrm{BLUE}\left[g_{ijk}\right] = y_{ijk} = E_j + g(L)_{ik} + gE_{ijk} + s_{ijk} \quad (11)$$

The $g(L)_{\mathrm{ik}}$ effect is modeled using a separable covariance structure, $\mathbf{I} \otimes \boldsymbol{\Gamma}$ in the absence of marker data, where the genetic covariance between locations $\boldsymbol{\Gamma}$ follows a second-order factor-analytic (FA2) model. The FA2 model provides a good balance between statistical parsimony and complexity for many plant breeding applications, and *Stage2* returns the rotated and scaled factor loadings (Cullis et al. 2010). A heterogeneous variance model is used for gE_{ijk} (which is the model residual as before), with different variance parameters for each location.

When marker data are provided, genotypic value is partitioned into additive and non-additive values, and the FA2 model is still used for the additive covariance between locations. Attempts to use an FA2 model for non-additive values were unsuccessful in several datasets, and even with a compound symmetry model, the correlation parameter was always on the boundary (equal to 1). The non-additive correlation parameter was therefore fixed at 1 and accepted as a model limitation. When markers are included as fixed effect covariates, different regression coefficients are estimated for each location. Similarly, different regression coefficients for genomic inbreeding are estimated per location.

A similar framework is used for multi-trait analysis, with trait replacing location in Eq. (11), except that all trait covariance matrices are unstructured. In Stage 1, a separable covariance model is used for the residuals, and in Stage 2, the fixed effects for environment are trait-specific. When markers are used to partition additive and non-additive genetic value, separate unstructured covariance matrices are estimated for each. Multi-trait models are limited to the homogeneous GxE structure described for single trait analysis (i.e., the genetic correlation between all environments is the same, regardless of location).

Proportion of variance explained

The aim is to quantify the proportion of variance (PVE) explained by each effect in the Stage 2 model, excluding the main effect E_j (which mirrors how heritability is calculated). The core idea is to compute variances based on the method of Legarra (2016), and the PVE is the variance of each effect divided by the sum. This is not a true partitioning of variance because the Stage 2 effects are not necessarily orthogonal.

First consider effects such as gE_{ij} and s_{ij} (Eq. 1), which are indexed by both genotype i and environment j. Representing these effects by vector $\mathbf{y}$ of length t, the variance is

$$V_y = \frac{1}{t}\sum_{ij} y_{ij}^2 - \left(\frac{1}{t}\sum_{ij} y_{ij}\right)^2 = \frac{1}{t}\mathbf{y}'\mathbf{y} - \frac{1}{t^2}\left(\mathbf{1}_t'\mathbf{y}\right)^2 \tag{12}$$

The symbol $\mathbf{1}_t$ in Eq. (12) is a $t \times 1$ vector of 1's. For multivariate normal (MVN) $\mathbf{y}$ with mean $\boldsymbol{\mu}$ and variance–covariance matrix $\mathbf{K}$, the expectation of V_y can be computed using the following general formula for quadratic forms (Searle et al. 1992):

$$E\left[\mathbf{y}'\mathbf{A}\mathbf{y}\right] = \mathrm{tr}(\mathbf{A}\mathbf{K}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} \tag{13}$$

The "tr" in Eq. (13) stands for trace, which equals the sum of the diagonal elements. It follows that

$$E\left[V_y\right] = \left[\overline{\mathrm{diag}(\mathbf{K})} - \overline{\mathrm{K}_{..}}\right] + \left[\overline{\mu^2_{\cdot}} - \left(\overline{\mu}_{\cdot}\right)^2\right] \tag{14}$$

where $\overline{\mathrm{diag}(\mathbf{K})}$ is the mean of the diagonal elements of $\mathbf{K}$. Equation (14) follows the convention of using an overbar to indicate averaging with respect to dotted subscripts.

For effects indexed only by genotype, such as g_i, Eq. (14) needs to be modified to accommodate unbalanced experiments. If $\mathbf{x} \sim \mathrm{MVN}(\boldsymbol{\mu}, \mathbf{K})$, and $\mathbf{Z}$ is the incidence matrix relating $\mathbf{x}$ to the gE basis of the Stage 2 model, then $\mathbf{y} = \mathbf{Z}\mathbf{x}$ is the random vector for which we need to compute the expected variance. The result is identical to Eq. (14) provided the averages are interpreted as weighted averages:

$$\begin{aligned}
\overline{\mathrm{diag}(\mathbf{K})} &= \frac{1}{t}\sum_i w_i \mathrm{K}_{ii} \\
\overline{\mathrm{K}_{..}} &= \frac{1}{t^2}\sum_{i,j} w_i \mathrm{K}_{ij} w_j \\
\overline{\mu^b_{\cdot}} &= \frac{1}{t}\sum_i w_i \mu_i^b \text{ for exponent } b = 1, 2, \ldots
\end{aligned} \tag{15}$$

The weights w_i in Eq. (15) come from $\mathbf{w} = \mathbf{1}_t'\mathbf{Z}$ and represent the number of environments for genotype i.

For the multi-location model, the genotype within location variance is computed using $\mathbf{K} = \mathbf{G} \otimes \boldsymbol{\Gamma}$ and weights equal to the number of times each gL combination is present. For a balanced experiment with n individuals and s locations, the result is

$$\begin{aligned}
E\left[V_{g(L)}\right] &= \frac{\mathrm{tr}(\mathbf{G}\otimes\boldsymbol{\Gamma})}{ns} - \frac{\left(\mathbf{1}_n' \otimes \mathbf{1}_s'\right)(\mathbf{G}\otimes\boldsymbol{\Gamma})\left(\mathbf{1}_n \otimes \mathbf{1}_s\right)}{n^2 s^2} \\
&= \left[\overline{\mathrm{diag}(\mathbf{G})}\right]\left[\overline{\mathrm{diag}(\boldsymbol{\Gamma})}\right] - \left(\overline{\mathrm{G}_{..}}\right)\left(\overline{\Gamma_{..}}\right)
\end{aligned} \tag{16}$$

Following Rogers et al. (2021), Eq. (16) is partitioned into a main effect V_g plus genotype x loc interaction V_{gL}. The main effect is based on the average of the $\frac{s(s-1)}{2}$ off-diagonal elements of $\boldsymbol{\Gamma}$:

$$E\left[V_g\right] = \left[\overline{\mathrm{diag}(\mathbf{G})} - \overline{\mathrm{G}_{..}}\right]\left[\frac{2}{s(s-1)}\sum_i\sum_{j>i}\Gamma_{ij}\right] \tag{17}$$

Equation (17) is extended to the unbalanced case by using weighted averages for $\mathbf{G}$.

BLUP

Empirical BLUPs are calculated conditional on the variance components estimated in Stage 2. All Stage 2 models described above can be written in the following standard form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\delta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \tag{18}$$

where $\boldsymbol{\delta}$ is a vector of fixed effects (for environments, markers, and inbreeding), $\mathbf{u}$ is a vector of multivariate normal genetic effects, and $\boldsymbol{\varepsilon}$ is the "residual" vector (for the g x env and Stage 1 error effects). Let $\hat{\mathbf{u}}$ denote BLUP[$\mathbf{u}$], which is calculated one of two ways for numerical efficiency. If the length of $\mathbf{y}$ exceeds the length of $\mathbf{u}$, then $\hat{\mathbf{u}}$ is calculating by inverting the coefficient matrix of the mixed model equations (MME; Henderson 1975). Otherwise, $\hat{\mathbf{u}}$ is calculated by inverting $\mathbf{V} = var(\mathbf{y})$ and using the following result (Searle et al. 1992):

$$\begin{aligned}
&\hat{\mathbf{u}} = cov(\mathbf{u}, \mathbf{y})\mathbf{P}\mathbf{y} = var(\mathbf{u})\mathbf{Z}'\mathbf{P}\mathbf{y} \\
&\text{where } \mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}
\end{aligned} \tag{19}$$

Genetic merit is a linear combination of random and fixed effects. For random effects, the structure of $\mathbf{u}$ is trait nested within individual, nested within additive vs. non-additive values. For fixed effects (ignoring the environment effects), $\boldsymbol{\delta}$ contains trait nested within marker effects, followed by trait nested within the regression coefficient for heterosis. If $\mathbf{W}$ represents the centered matrix of allele dosages for the fixed effect markers (n individuals x m markers), $\mathbf{F}$ is the vector of genomic inbreeding coefficients, and $\mathbf{c}$ is the vector of economic weights for multiple traits or locations, then the genetic merit vector for the population is

$$\boldsymbol{\theta} = \left(\left[\mathbf{I}_n \ \gamma\mathbf{I}_n\right] \otimes \mathbf{c}'\right)\mathbf{u} + \left(\left[\mathbf{W} \ \gamma\mathbf{F}\right] \otimes \mathbf{c}'\right)\boldsymbol{\delta} \tag{20}$$

The value of γ depends on which genetic value is predicted: 0 for additive value, 1 for total value, and $\left(\frac{\phi}{2} - 1\right)/(\phi - 1)$ for breeding value and ploidy ϕ (Gallais 2003). Because BLUP is a linear operator, $\hat{\boldsymbol{\theta}} = \text{BLUP}[\boldsymbol{\theta}]$ (i.e., the selection index) is given by Eq. (20) with $\mathbf{u}$ and $\boldsymbol{\delta}$ replaced by their predicted values.

Index coefficients entered by the user are interpreted as relative weights for standardized traits (or locations). To generate the vector $\mathbf{c}$, the software divides the user-supplied weights by the standard deviations of the breeding values (estimated in Stage 2); it also applies an overall scaling such that $\|\mathbf{c}\| = 1$, which ensures predictions are commensurate with the original trait scale in multi-location models.

The reliability r_i^2 of the predicted merit $\hat{\theta}_i$ for individual i is the squared correlation with its true value θ_i, which depends only on the random effects. If $\mathbf{u}_i$ represents the vector of random genetic effects for individual i, and $\boldsymbol{\lambda}$ denotes $\left[\begin{array}{cc} 1 & \gamma \end{array}\right]' \otimes \mathbf{c}$, then the random effects component of θ_i is $\boldsymbol{\lambda}'\mathbf{u}_i$, and the reliability is

$$r_i^2 = \frac{cov^2(\theta_i, \hat{\theta}_i)}{var(\theta_i)var(\hat{\theta}_i)} = \frac{\left[\boldsymbol{\lambda}' cov(\mathbf{u}_i, \hat{\mathbf{u}}_i)\boldsymbol{\lambda}\right]^2}{\left[\boldsymbol{\lambda}' var(\mathbf{u}_i)\boldsymbol{\lambda}\right]\left[\boldsymbol{\lambda}' var(\hat{\mathbf{u}}_i)\boldsymbol{\lambda}\right]} = \frac{\boldsymbol{\lambda}' var(\hat{\mathbf{u}}_i)\boldsymbol{\lambda}}{\boldsymbol{\lambda}' var(\mathbf{u}_i)\boldsymbol{\lambda}} \tag{21}$$

The final equality in Eq. (21) relies on the following property of BLUP: $cov\left(\mathbf{u}, \hat{\mathbf{u}}\right) = var\left(\hat{\mathbf{u}}\right)$. For the MME solution method, the $var\left(\hat{\mathbf{u}}\right)$ matrix is computed as $var(\mathbf{u}) - \mathbf{C}_{22}$, where $\mathbf{C}_{22}$ is from the partitioned inverse coefficient matrix (Henderson 1975). For the $\mathbf{V}$ inversion method, $var\left(\hat{\mathbf{u}}\right) = var(\mathbf{u})\left(\mathbf{Z}'\mathbf{P}\mathbf{Z}\right)var(\mathbf{u})$ (Searle et al. 1992).

Selection response

The breeder's equation provides the expected response to truncation selection on predicted merit $\hat{\theta}$. If $\mathbf{b}$ denotes the multi-trait vector of breeding values for an individual, then its predicted merit is $\hat{\theta} = \mathbf{c}'\hat{\mathbf{b}}$ (see Eq. 20), and the multi-trait response $\mathbf{x}$ under selection intensity i is

$$\mathbf{x} = \left[i\sigma_{\hat{\theta}}\right]\left[\frac{cov_n\left(\mathbf{b}, \hat{\theta}\right)}{\sigma_{\hat{\theta}}^2}\right] = i\sigma_{\hat{\theta}}^{-1} cov_n\left(\mathbf{b}, \hat{\mathbf{b}}\right)\mathbf{c} \tag{22}$$

(To connect Eq. (22) with a familiar form of the breeder's equation, the first bracketed term is the selection differential, and the second bracketed term represents heritability.) The subscript n on cov_n indicates it is the covariance with respect to the n individuals in the population, which differs slightly from the covariance of a vector with respect to its MVN distribution (see "Appendix"). As mentioned earlier, under BLUP, the latter covariance satisfies $cov\left(\mathbf{u}, \hat{\mathbf{u}}\right) = var\left(\hat{\mathbf{u}}\right)$. Combining this result with "Appendix" Eq. (35), it follows that $cov_n\left(\mathbf{b}, \hat{\mathbf{b}}\right) = var_n(\hat{\mathbf{b}})$, which is denoted $\mathbf{B}$. The formula for traits j and k is

$$\begin{aligned} B_{jk} &= \overline{\text{diag}(\mathbf{L})} - \overline{L_{..}} + \overline{\mu_{j\cdot}\mu_{k\cdot}} - \left(\overline{\mu_{j\cdot}}\right)\left(\overline{\mu_{k\cdot}}\right) \\ \mathbf{L} &= \left[\mathbf{I}_n \ \gamma\mathbf{I}_n\right] cov\left(\hat{\mathbf{u}}_j, \hat{\mathbf{u}}_k\right)\left[\mathbf{I}_t \ \gamma\mathbf{I}_t\right]' \\ \boldsymbol{\mu}_j &= \left[\mathbf{W} \ \gamma\mathbf{F}\right]\boldsymbol{\delta}_j \end{aligned} \tag{23}$$

The vector $\hat{\mathbf{u}}_j$ is a $2n \times 1$ stacked vector of the predicted additive and non-additive values for a population of size n. The calculation of $cov(\hat{\mathbf{u}}_j, \hat{\mathbf{u}}_k)$ follows the same procedure described above (see Eq. 19), and the contribution from $\boldsymbol{\delta}$ is calculated using the fixed effect estimates. Since the overall scaling of the index coefficients is arbitrary, we can impose $\sigma_{\hat{\theta}}^2 = 1$. Inverting Eq. (22) under this constraint leads to an expression for the index coefficients:

$$\mathbf{c} = i^{-1}\mathbf{B}^{-1}\mathbf{x} \tag{24}$$

Substituting this result into $1 = \sigma_{\hat{\theta}}^2 = \mathbf{c}'\mathbf{B}\mathbf{c}$ leads to an implicit equation for the response:

$$\mathbf{x}'\mathbf{B}^{-1}\mathbf{x} - i^2 = 0 \tag{25}$$

Equation (25) is the matrix representation of an ellipsoid in t dimensions, which is used by StageWise to provide a geometric visualization of selection tradeoffs. (The software DESIRE (Kinghorn 2013) is an earlier example of plotting the elliptical multi-trait response.) If the response is expressed in units of genetic standard deviation, a diagonal matrix $\boldsymbol{\Delta}$ with elements $\sigma_b = \sqrt{\sigma_{\text{A}}^2 + \gamma^2\sigma_{\text{D}}^2}$ is used to rescale the matrix of the quadratic form as $\boldsymbol{\Delta}\mathbf{B}^{-1}\boldsymbol{\Delta}$. The principal axes of the ellipse are given by the eigenvectors of this matrix, and the lengths of the semi-axes equal the inverse square-root of the eigenvalues.

This geometric model provides a convenient method for implementing a restricted selection index, in which the response for some traits is constrained to be zero (Kempthorne and Nordskog 1959). From above, the change in genetic merit associated with response $\mathbf{x}$ is $\mathbf{c}'\mathbf{x}$, which is the projection of $\mathbf{x}$ onto $\mathbf{c}$ times the magnitude of $\mathbf{c}$. For the unrestricted index, the response that maximizes genetic gain is therefore the solution of the following convex optimization problem:

$$\begin{aligned} &\max_{\mathbf{x}} \ \mathbf{c}'\mathbf{x} \\ &\mathbf{x}'\mathbf{B}^{-1}\mathbf{x} \le 1 \end{aligned} \tag{26}$$

The linear inequality constraint in Eq. (26), which is convex, replaces the linear equality constraint of Eq. (25), which is not convex. This substitution is valid because the linear objective ensures the optimum is on the boundary (Boyd and Vandenberghe 2004). For the restricted index,

the restricted traits are not included in the objective $\mathbf{c}'\mathbf{x}$, and equality or inequality constraints on the genetic gain x_i for restricted trait i are added to Eq. (26). Convex optimization is performed using CVXR (Fu et al. 2020), and the index coefficients are computed from the optimal $\mathbf{x}$ via Eq. (24) with intensity $i = 1$.

Marker effects and GWAS

Marker effects and GWAS scores are also calculated by BLUP. Let $\boldsymbol{\alpha}$ represent the $mt \times 1$ vector of additive (substitution) effects for t traits/locations nested within m markers, with variance–covariance matrix $\mathbf{I}_m \otimes \boldsymbol{\Gamma}\left(\phi \sum_k p_k q_k\right)^{-1}$ for ploidy ϕ (Endelman et al. 2018). From the linearity of BLUP, the predicted multi-trait index of marker effects is $\left(\mathbf{I}_m \otimes \mathbf{c}'\right)\hat{\boldsymbol{\alpha}}$, and from Eq. (19), $\hat{\boldsymbol{\alpha}}$ can be written in terms of the predicted additive values $\hat{\mathbf{a}}$ as

$$\begin{aligned}\hat{\boldsymbol{\alpha}} &= cov(\boldsymbol{\alpha}, \mathbf{y})\mathbf{P}\mathbf{y} = var(\boldsymbol{\alpha})\left[\mathbf{W}' \otimes \mathbf{I}_t\right]\left[\mathbf{G}^{-1} \otimes \boldsymbol{\Gamma}^{-1}\right]\hat{\mathbf{a}} \\ &= \frac{\left(\mathbf{W}'\mathbf{G}^{-1} \otimes \mathbf{I}_t\right)\hat{\mathbf{a}}}{\phi \sum_k p_k q_k} \\ \Rightarrow \left(\mathbf{I}_m \otimes \mathbf{c}'\right)\hat{\boldsymbol{\alpha}} &= \frac{\left(\mathbf{W}'\mathbf{G}^{-1} \otimes \mathbf{c}'\right)\hat{\mathbf{a}}}{\phi \sum_k p_k q_k}\end{aligned} \tag{27}$$

The $\mathbf{W}$ matrix in Eq. (27) is the centered matrix of allele dosages (individuals x markers). A similar result holds for relating the multi-trait index of digenic substitution effects $\boldsymbol{\beta}$ to the predicted dominance values $\hat{\mathbf{d}}$ (Eq. 7):

$$\left(\mathbf{I}_\mathrm{m} \otimes \mathbf{c}'\right)\hat{\boldsymbol{\beta}} = \frac{\left(\mathbf{Q}'\mathbf{D}^{-1} \otimes \mathbf{c}'\right)\hat{\mathbf{d}}}{\binom{\phi}{2} \sum_k 4p_k^2 q_k^2} \tag{28}$$

The fixed effect for inbreeding is included in $\hat{\mathbf{d}}$ and therefore represented in the predicted marker effects.

GWAS p-values are computed from the standardized BLUPs of the marker effects, which are asymptotically standard normal (Gualdrón Duarte et al. 2014). If $\mathbf{w}_k$ denotes the kth column of the $\mathbf{W}$ matrix, then the standard error of the predicted additive effect for marker k is

$$\frac{\left[\left(\mathbf{w}_k'\mathbf{G}^{-1} \otimes \mathbf{c}'\right)var(\hat{\mathbf{a}})\left(\mathbf{G}^{-1}\mathbf{w}_k \otimes \mathbf{c}\right)\right]^{1/2}}{\phi \sum_k p_k q_k} \tag{29}$$

The formula for dominance effects is analogous, based on Eq. (28). StageWise provides the option to parallelize this computation across multiple cores. To control for multiple testing, the desired significance level specified by the user is divided by the effective number of markers (Moskvina and Schmidt 2008) to set the p value discovery threshold.

Potato data analysis

The potato dataset is an updated version of the data from Endelman et al. (2018), which spanned 2012–2017 at one location (Hancock, WI) and contained 571 clones from both preliminary and advanced yield trials. The current version spans 2015–2020 and contains 943 clones. Fixed effects for block or trial, as well as stand count, were used in Stage 1. Three traits were analyzed: total yield (Mg ha^{-1}), vine maturity (1 [early] to 9 [late] visual scale at 100 days after planting), and potato chip fry color (Hunter L) after 6 months of storage. The G matrix was used for multi-trait analysis, instead of H, due to convergence problems with the latter.

Marker data files contain the estimated allele dosage (0–4) from genotyping with potato SNP array v2 or v3 (which contains most of v2) (Felcher et al. 2012; Vos et al. 2015). Genotype calls were made with R package fitPoly (Zych et al. 2019). Data from the two array versions were combined with the command *merge_impute* from R package polyBreedR (https://github.com/jendelman/polyBreedR). This command performs one iteration of the EM algorithm described in Poland et al. (2012) (only one iteration is needed for complete datasets at low and high density), followed by shift and scaling (if necessary) to ensure all data are in the interval [0, ploidy].

Results

The workflow to analyze data with StageWise is illustrated in Fig. 1. Any software can be used to compute genotype BLUEs and their variance–covariance matrix in Stage 1. For convenience, the package has a command named *Stage1*, which can accommodate any number of fixed or i.i.d. random covariates, as well as spatial analysis using SpATS (Rodríguez-Álvarez et al. 2018). To partition genetic value into additive and non-additive components, genome-wide marker data is processed with the command *read_geno*, and the output is then included in the call to *Stage2*. After estimating the variance components with *Stage2*, the *blup_prep* command inverts either the coefficient matrix of the mixed model equations or the variance–covariance matrix of the Stage 2 response variable, whichever is smaller. This allows for rapid, iterative use of the *blup* command to obtain different types of predictions and standard errors, which are used in the calculation of reliability (i.e., squared accuracy) for individuals and GWAS scores for markers. Three vignettes, or tutorials, come with the software to give detailed examples of using the commands. The following results represent a condensed version of this information.

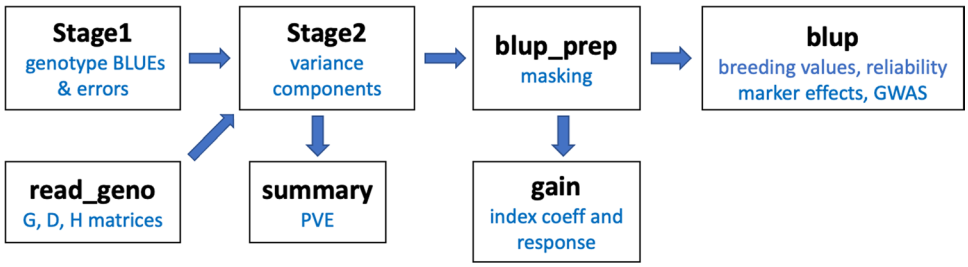

Fig. 1 Overview of the commands and workflow in R/StageWise

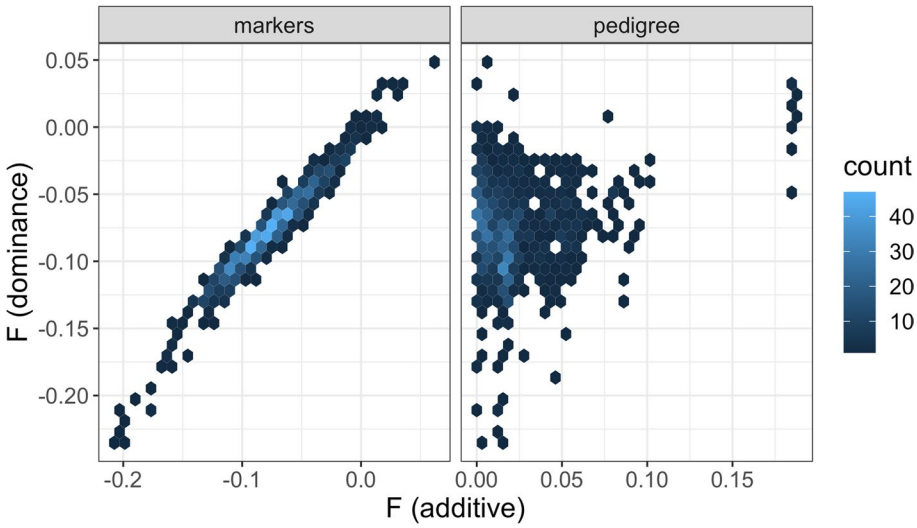

Fig. 2 Comparison of inbreeding coefficients (F) for a population of 943 potato breeding lines. The vertical axis is computed from the dominance coefficients, and the horizontal axis is computed from the additive relationship matrix

The primary dataset comes from six years of potato yield trials at a single location and includes 943 genotyped clones. The genotypic values of heterozygous clones have both additive and non-additive components. Non-additive values can be modeled in StageWise either as genetic residuals (no covariance) or as dominance values. In the context of genomic prediction, directional dominance models use inbreeding coefficients to estimate heterosis. Figure 2 compares three types of inbreeding coefficients for this population: (1) F_D, from the directional dominance model, (2) F_G, from the diagonal elements of the additive genomic relationship matrix, and (3) F_A, from the diagonal elements of the pedigree relationship matrix. The F_G and F_D coefficients from the genomic models were highly correlated ($r = 0.98$) and have the same population mean, -0.08, which indicates a slight excess of heterozygosity compared to panmictic (Hardy–Weinberg) equilibrium. Although there was some concordance between the genomic and pedigree coefficients for the most inbred individuals, there was little agreement at small values of F_A (Fig. 2).

Table 1 Akaike Information Criterion (AIC) for the Stage 2 model with vs. without inclusion of the Stage 1 errors

	Yield	Fry color	Vine maturity
Without	7007	4662	2018
With	6940	4558	1989
Change	−67	−104	−29

Single trait analysis

Initially, the three traits in the potato dataset–total yield, chip fry color, and vine maturity–were analyzed independently. In Stage 1, broad-sense heritability on a plot basis was highest for yield (0.70–0.83), with similar results for fry color (0.25–0.74) and maturity (0.38–0.74) (Figure S1, ESM1). The benefit of including Stage 1 errors in the Stage 2 model was assessed based on the change in AIC, which ranged from −29 for maturity to −104 for fry color (Table 1). Applying the *summary* command to the output from *Stage2* generates a table with the proportion of variation explained (PVE). The PVE for additive effects, which can be called genomic heritability, ranged from 0.34 (yield) to 0.43 (maturity) (Table 2). The PVE for dominance effects has two parts: one due to the variance of the dominance

Table 2 Proportion of variation explained for the multi-year potato dataset

	Yield	Fry color	Vine maturity
Additive	0.34	0.38	0.43
Dominance	0.12	0.04	0.02
Heterosis	0.03	0.00	0.00
Genotype x year	0.30	0.24	0.16
Stage 1 error	0.21	0.34	0.39

Both "Dominance" and "Heterosis" come from the directional dominance model

effects ("Dominance" in Table 2), and the other from variation in the genomic inbreeding coefficient ("Heterosis" in Table 2). Of the three traits, yield had the largest influence of dominance, with a combined PVE of 0.15.

StageWise has the ability for genomic prediction with the H matrix, which is a weighted average of G and A that was originally developed to use ungenotyped individuals in the training population (Legarra et al. 2009; Christensen and Lund 2010). Even when all individuals are genotyped, H may still outperform G due to the sparsity of A (Fig. 3). For the potato dataset, the change in AIC with H ranged from −6 (fry color) to −13 (yield). The optimum weight for A was 0.3 for vine maturity and fry color and 0.5 for yield. As the weight for A increased, the estimate for genomic heritability (solid line in Fig. 3) also increased, at the expense of dominance (dashed line).

The *blup_prep* command has an option to mask Stage 1 BLUEs, which can be used to estimate the accuracy of predicting new individuals or new environments. Figure 4 compares the reliability of genome-wide marker-assisted selection (MAS) vs. marker-based selection (MBS) for the last breeding cohort in the potato dataset. The distinction between MAS and MBS is that the selection candidates are part of the training set with MAS but not with MBS (Bernardo 2010). The reliability of MAS (r_A^2) was 0.14–0.21 higher than MBS (r_B^2) across traits. From index theory (Lande and Thompson 1990; Riedelsheimer and Melchinger 2013), the two quantities are related by

$$r_A^2 = r_B^2 + \frac{h^2\left(1 - r_B^2\right)^2}{\left(1 - h^2 r_B^2\right)} \tag{30}$$

When used with the genomic heritability estimates from *Stage2*, this formula closely matched the data for all three traits (Fig. 4).

Although GWAS is not the emphasis of StageWise, the software can perform a fully efficient, two-stage GWAS. For the potato dataset, there was a major QTL for vine maturity on chr05 (Figure S2, ESM1), in the vicinity of the well-known regulator of potato maturity *StCDF1* (Kloosterman et al. 2013). *Stage2* has an optional argument to include markers as fixed effects for major QTL. In this case, the PVE for the marker was 0.10, which represents 21% of the total additive variance.

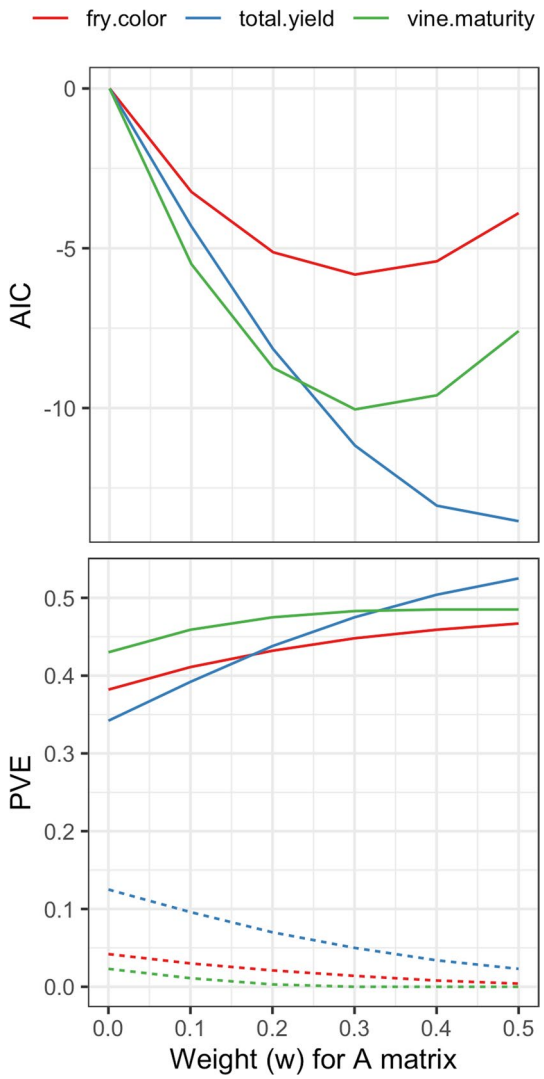

Fig. 3 Minimizing the Akaike Information Criterion (AIC) to select the optimal weighting of pedigree (A) and marker (G) additive relationship matrices: $\mathbf{H} = w\mathbf{A} + (1-w)\mathbf{G}$. The optimal weight varied by trait in a potato dataset of 943 clones. The proportion of variation explained (PVE) by the additive effects (solid line) increased with w, while the PVE for the dominance effects (dashed line) decreased

Multi-trait analysis

Multi-trait analysis follows the same general workflow as a single trait. In addition to the PVE, the *summary* command returns the additive correlation matrix for the traits. For the potato dataset, late maturity was correlated with higher yield ($r = 0.57$) and slightly with lighter fry color ($r = 0.23$). There

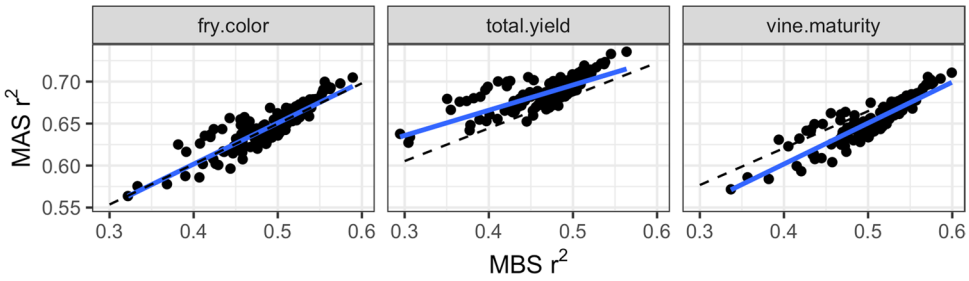

Fig. 4 Comparing the reliability (r^2) of marker-assisted (MAS) vs. marker-based (MBS) genomic selection in the potato dataset. Each point represents a clone from one breeding cohort, and the blue line is a linear trendline. The increased accuracy from having phenotypes for the selection candidates (MAS) was closely predicted by selection index theory (dashed line)

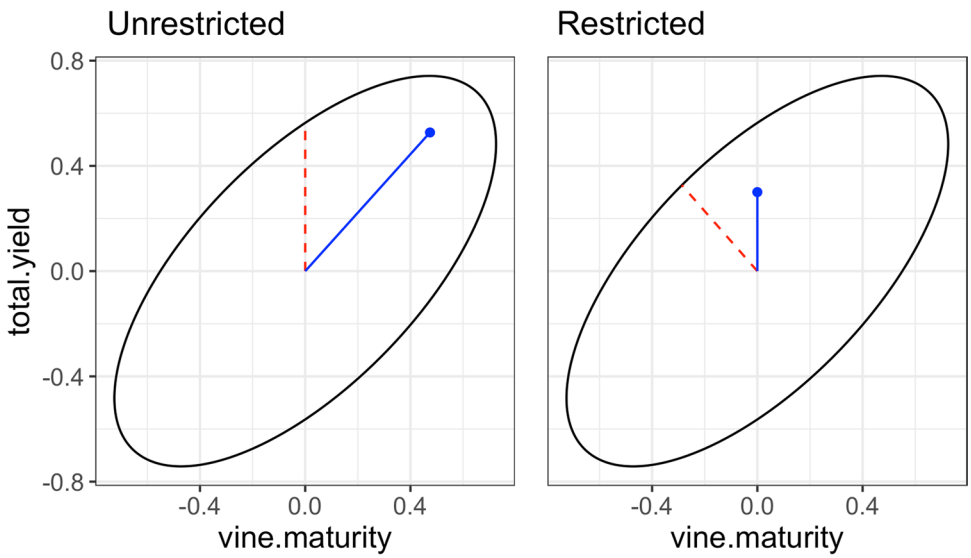

Fig. 5 Selection response tradeoffs in the potato dataset for three traits: yield, maturity, and fry color. The response surface is three-dimensional, but only the yield-maturity plane is shown to highlight the tradeoff between these two traits. The dashed red line segment is the projection of the index vector, and the solid blue line segment is the projection of the optimal response (color figure online)

was no genetic correlation ($r=0.00$) between yield and fry color.

The "index.coeff" argument for *blup* is used to specify the selection index coefficients, which determine the relative weights of the traits (after standardization to unit variance) for genetic merit. (Because StageWise uses a multi-trait BLUP, the optimal index coefficients equal the coefficients of genetic merit.) For the potato chip market, it is reasonable to give equal weight to yield and fry color. However, naïve selection on these traits alone will generate offspring with later maturity, which is undesirable. One way to avoid this is by using vine maturity as a covariate in the analysis.

Alternatively, the *gain* command in StageWise can be used to compute the coefficients of a restricted selection index, in which the response for some traits is constrained to be zero (Kempthorne and Nordskog 1959). For a given selection intensity and t traits, the set of all possible responses is a t-dimensional ellipsoid, and *gain* shows 2D slices of it. Figure 5 shows the breeding value response for yield and maturity, as well as two line segments. The dashed red line is the projection of the index vector, and the solid blue line is the projection of the optimal response. The restricted index requires negative weight for maturity to produce zero response, which reduces the yield response compared to the unrestricted index by $0.23i\sigma$ (i is selection intensity and σ is the genetic standard deviation of the breeding values; Table 3).

Table 3 Multi-trait response for potato under truncation selection, assuming yield and fry color contribute equally to genetic merit

Trait	Unrestricted index		Restricted index	
	Coefficients	Response	Coefficients	Response
Total yield	0.707	0.53	0.601	0.30
Fry color	0.707	0.52	0.601	0.51
Vine maturity	0.000	0.47	-0.527	0.00

Index coefficients are for standardized traits and scaled to have unit norm. Response is for intensity $i=1$, in units of genetic standard deviation

Discussion

StageWise was designed to enhance the use of genomic prediction in plant breeding, but there are some limitations. At present, each phenotype is associated with a single genotype identifier, which is inadequate for hybrid prediction. The options for modeling GxE are somewhat limited, particularly for multiple traits, which assume a uniform genetic correlation between environments. For single trait analysis, a more complex GxE model is possible to allow for heterogenous genetic correlation between locations. The genetic covariance between locations is based on a second- order factor-analytic (FA2) model (Smith et al. 2001), which offers enough statistical complexity for many applications. To assess model adequacy, the factor loadings returned by *Stage2* can visualized with the command *uniplot*, which generates a circular plot in which the squared radius for each location equals the proportion of genetic variance explained by the latent factors (Cullis et al. 2010). This functionality is illustrated in Vignette 2 using national trial data for potato (Schmitz Carley et al. 2019). At present, StageWise does not have functionality for genomic prediction with environmental covariates.

This is the first study to formulate and apply a model for directional dominance in polyploids. Although heterosis explained less than 5% of the variance (PVE) for yield, we should expect small PVE when there is limited variation for inbreeding. The standard deviation of F_D was only 0.03 for the population of 943 potato clones (Fig. 3).

From the theory of directional dominance, the average dominance coefficient is the covariate for estimating heterosis. Xiang et al. (2016) used average heterozygosity for the covariate because under a genotypic parameterization of dominance in diploids, this is equivalent to the average dominance coefficient. However, studies employing orthogonal parameterizations of dominance have also used this covariate (Aliloo et al. 2017; Yadav et al. 2021), even though heterozygosity is no longer equivalent to the dominance coefficient because the relative contribution of the genotypes to inbreeding depends on allele frequency (see Eq. 5). For example, the minor allele homozygote contributes more to inbreeding than the major allele homozygote, and the difference is $\phi(\phi-1)(q-p)$ for ploidy ϕ and minor allele frequency $p=1-q$ at panmictic equilibrium. To give another example, simplex dosage of the minor allele in a tetraploid contributes more to inbreeding than duplex dosage only for $p>1/3$; for $p<1/3$, duplex dosage contributes more.

A more general approach to restricted selection indices was developed in StageWise by investigating the geometry of the problem (Eq. 26). Until now, only equality constraints have been included (i.e., specifying a certain value for genetic gain), which are amenable to solution by the method of Lagrange multipliers. StageWise uses convex optimization software to allow for both equality and inequality constraints. In many situations, inequality constraints are more appropriate than equality constraints. For example, when selecting for yield, we might accept earlier but not later maturity, which is represented by response ≤ 0. With only one constrained trait, the optimal solution corresponds to zero response, so the inequality offers no advantage. But with two or more constraints, higher genetic gains are possible with inequalities (ESM2).

The "mask" argument for *blup_prep* makes it easy to investigate the potential benefit of using a correlated, secondary trait to improve genomic selection. Many plant breeding programs are exploring the use of spectral measurements from high-throughput phenotyping platforms to improve selection for yield. For example, Rutkoski et al. (2016) demonstrated that aerial measurements of canopy temperature during grain fill could be used to predict wheat grain yield. Vignette 3 shows how to recreate this result with StageWise.

Typically, the number of traits a breeder must consider for selection is too large to analyze jointly in StageWise based on the current implementation with ASReml-R. New algorithms may alleviate this limitation in the future (Runcie et al. 2021), but in the meantime, a practical approach is to split the traits into groups for multivariate analysis based on phenotypic correlations. In the final step, multiple outputs from *blup_prep* can be combined in one call to *blup*, using an index that covers all traits (example in Vignette 3).

We should acknowledge that truncation selection on breeding value is not optimal for long-term genetic gain. The design of selection methods that conserve and exploit genetic diversity more efficiently is an exciting area of research (e.g., Toro and Varona 2010; Akdemir and Sánchez 2016; Goiffon et al. 2017). Although such methods are not currently available in StageWise, the additive and dominance marker effects returned by the software can be used to implement them.

Appendix

The objective is an expression for the expected covariance between two quantities of a population of size n, represented by multivariate normal vectors $\mathbf{x}_1 \sim \mathrm{MVN}(\boldsymbol{\mu}_1, \mathbf{K}_1)$ and $\mathbf{x}_2 \sim \mathrm{MVN}(\boldsymbol{\mu}_2, \mathbf{K}_2)$, with covariance $\mathbf{L}$:

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \sim MVN\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \mathbf{K}_1 & \mathbf{L} \\ \mathbf{L} & \mathbf{K}_2 \end{bmatrix} \right) \tag{31}$$

Noting that $\mathbf{x}_1 = \left[\mathbf{I}_n \ \mathbf{0}\right]\mathbf{x}$ and $\mathbf{x}_2 = \left[\mathbf{0} \ \mathbf{I}_n\right]\mathbf{x}$, the population covariance is (cf. Eq. (12))

$$\begin{aligned}\mathrm{cov}_{12} &= n^{-1}\mathbf{x}_1'\mathbf{x}_2 - n^{-2}\left(\mathbf{1}_n'\mathbf{x}_1\right)\left(\mathbf{1}_n'\mathbf{x}_2\right) \\ &= \frac{1}{2n}\mathbf{x}'\begin{bmatrix}\mathbf{0} & \mathbf{I}_n \\ \mathbf{I}_n & \mathbf{0}\end{bmatrix}\mathbf{x} + \frac{1}{2n^2}\mathbf{x}'\begin{bmatrix}\mathbf{0} & \mathbf{J}_n \\ \mathbf{J}_n & \mathbf{0}\end{bmatrix}\mathbf{x}\end{aligned} \tag{32}$$

where $\mathbf{J}_n = \mathbf{1}_n\mathbf{1}_n'$ is a n x n matrix of ones. Using Eq. (13), the expectation of the first quadratic form in Eq. (32) is

$$\begin{aligned}&\frac{1}{2n}\mathrm{tr}\left(\begin{bmatrix}\mathbf{0} & \mathbf{I}_n \\ \mathbf{I}_n & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{K}_1 & \mathbf{L} \\ \mathbf{L} & \mathbf{K}_2\end{bmatrix}\right) + \frac{1}{2n}\left(\begin{bmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{bmatrix}'\begin{bmatrix}\mathbf{0} & \mathbf{I}_n \\ \mathbf{I}_n & \mathbf{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{bmatrix}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(L_{ii} + \mu_{1i}\mu_{2i}\right)\end{aligned} \tag{33}$$

The expectation of the second quadratic form in Eq. (32) is

$$\begin{aligned}&\frac{1}{2n^2}\mathrm{tr}\left(\begin{bmatrix}\mathbf{0} & \mathbf{J}_n \\ \mathbf{J}_n & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{K}_1 & \mathbf{L} \\ \mathbf{L} & \mathbf{K}_2\end{bmatrix}\right) + \frac{1}{2n^2}\left(\begin{bmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{bmatrix}'\begin{bmatrix}\mathbf{0} & \mathbf{J}_n \\ \mathbf{J}_n & \mathbf{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{bmatrix}\right) \\ &= \overline{L_{\cdot\cdot}} + \left(\overline{\mu_{1\cdot}}\right)\left(\overline{\mu_{2\cdot}}\right)\end{aligned} \tag{34}$$

Putting Eq. (33and34) together, the expected covariance is

$$E\left[\mathrm{cov}_{12}\right] = \overline{\mathrm{diag}(\mathbf{L})} - \overline{L_{\cdot\cdot}} + \overline{\mu_{1\cdot}\mu_{2\cdot}} - \left(\overline{\mu_{1\cdot}}\right)\left(\overline{\mu_{2\cdot}}\right) \tag{35}$$

As in the Methods, for partitioning covariance on a gE basis, the unbalanced nature of the experiment is accounted for by computing the covariance between vectors $\mathbf{y}_1 = \mathbf{Z}\mathbf{x}_1$ and $\mathbf{y}_2 = \mathbf{Z}\mathbf{x}_2$, where incidence matrix $\mathbf{Z}$ maps n individuals to gE instances. If $\mathbf{y}$ denotes the stacked vector $\left[\mathbf{y}_1 \ \mathbf{y}_2\right]'$, then

$$\mathbf{y} = \begin{bmatrix}\mathbf{y}_1 \\ \mathbf{y}_2\end{bmatrix} \sim MVN\left(\begin{bmatrix}\mathbf{Z}\boldsymbol{\mu}_1 \\ \mathbf{Z}\boldsymbol{\mu}_2\end{bmatrix}, \begin{bmatrix}\mathbf{Z}\mathbf{K}_1\mathbf{Z}' & \mathbf{Z}\mathbf{L}\mathbf{Z}' \\ \mathbf{Z}\mathbf{L}\mathbf{Z}' & \mathbf{Z}\mathbf{K}_2\mathbf{Z}'\end{bmatrix}\right) \tag{36}$$

Replacing $\mathbf{x}$ with $\mathbf{y}$ in Eq. (32), the result for expected covariance follows Eq. (35) but using averages weighted by the number of environments per genotype.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00122-023-04298-x.

Acknowledgements I would like to thank potato breeding colleagues across the US for contributing germplasm used in this study, Grace Christensen for assistance with genotyping, and the UW-Madison Hancock and Rhinelander Agricultural Research Stations.

Funding Software development has been supported by USDA Hatch Project 1013047 and the USDA National Institute of Food and Agriculture (NIFA) Award 2020–51181-32156. The potato datasets were generated with support from NIFA Awards 2016–34141-25707 and 2019–34141-30284, Potatoes USA, the Wisconsin Potato and Vegetable Growers Association, and the University of Wisconsin-Madison.

Data Availability The potato datasets and vignettes are distributed with the StageWise software, which is available at https://github.com/jendelman/StageWise under the GNU General Public License v3. The current versions of the software and vignettes at the time of publication have been archived as ESM3.

Declarations

Conflict of Interests The author has no relevant financial or non-financial interests to disclose.

References

Akdemir D, Sánchez JI (2016) Efficient breeding by genomic mating. Front Genet 7:210. https://doi.org/10.3389/fgene.2016.00210

Aliloo H, Pryce JE, González-Recio O, Cocks BG, Goddard ME, Hayes BJ (2017) Including nonadditive genetic effects in mating programs to maximize dairy farm profitability. J Dairy Sci 100:1203–1222. https://doi.org/10.3168/jds.2016-11261

Amadeu RR, Cellon C, Olmstead JW, Garcia AA, Resende MF, Muñoz PR (2016) AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: a blueberry example. Plant Genome. https://doi.org/10.3835/plantgenome2016.01.0009

Batista LG, Mello VH, Souza AP, Margarido GRA (2022) Genomic prediction with allele dosage information in highly polyploid species. Theor Appl Genet 135:723–739. https://doi.org/10.1007/s00122-021-03994-w

Bernardo R (2010) Breeding for quantitative traits in plants, 2nd edn. Stemma Press, Woodbury, MN

Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in maize. Crop Sci 47:1082–1090. https://doi.org/10.2135/cropsci2006.11.0690

Boyd S, Vandenberghe L (2004) Convex Optimization. Cambridge University Press

Butler D, Cullis B, Gilmour A, Gogel B, Thompson R (2018) ASReml-R Reference Manual Version 4. VSN International Ltd, Hemel Hempstead, UK

Christensen OF, Lund MS (2010) Genomic prediction when some animals are not genotyped. Gen Sel Evol 42:2. https://doi.org/10.1186/1297-9686-42-2

Covarrubias-Pazaran G (2016) Genome-assisted prediction of quantitative traits using the R package sommer. PloS ONE 11(6):e0156744. https://doi.org/10.1371/journal.pone.0156744

Cullis BR, Smith AB, Beeck CP, Cowling WA (2010) Analysis of yield and oil from a series of canola breeding trials. Part II. Exploring variety by environment interaction using factor analysis. Genome 53:1002–1016. https://doi.org/10.1139/G10-080

Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE 3(10):e3395. https://doi.org/10.1371/journal.pone.0003395

Damesa TM, Möhring K, Worku M, Piepho HP (2017) One step at a time: Stage-wise analysis of a series of experiments. Agron J 109:845–857. https://doi.org/10.2134/agronj2016.07.0395

de los Campos G, Sorensen D, Gianola D (2015) Genomic heritability: what is it? PLoS Genet 11(5):e1005048. https://doi.org/10.1371/journal.pgen.1005048

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4:50–255. https://doi.org/10.3835/plantgenome2011.08.0024

Endelman JB, Jannink JL (2012) Shrinkage estimation of the realized relationship matrix. G3 Bethesda 2:1405–1413. https://doi.org/10.1534/g3.112.004259

Endelman JB, Schmitz Carley CA, Bethke PC et al (2018) Genetic variance partitioning and genome-wide prediction with allele dosage information in autotetraploid potato. Genetics 209:77–87. https://doi.org/10.1534/genetics.118.300685

Felcher KJ, Coombs JJ, Massa AN, Hansey CN, Hamilton JP, Veilleux RE, Buell CB, Douches DS (2012) Integration of two diploid potato linkage maps with the potato genome sequence. Plos ONE 7(4):e36347. https://doi.org/10.1371/journal.pone.00363474

Fisher RA (1941) Average excess and average effect of a gene substitution. Ann Eugen 11:53–63. https://doi.org/10.1111/j.1469-1809.1941.tb02272.x

Frensham A, Cullis B, Verbyla A (1997) Genotype by environment variance heterogeneity in a two-stage analysis. Biometrics 53:1373–1383. https://doi.org/10.2307/2533504

Fu A, Narasimhan B, Boyd S (2020) CVXR: An R package for disciplined convex optimization. J Stat Software. 94:1–34. https://doi.org/10.18637/jss.v094.i14

Gallais A (2003) Quantitative genetics and breeding methods in autopolyploid plants. INRA, Paris

Gilmour AR, Gogel BJ, Cullis BR, Welham SJ, Thompson R (2015) ASReml User guide release 4.1 Structural specification. VSN International Ltd, Hemel Hempstead, UK

Goiffon M, Kusmec A, Wang L, Hu G, Schnable PS (2017) Improving response in genomic selection with a population-based selection strategy: Optimal population value selection. Genetics 206:1675–1682. https://doi.org/10.1534/genetics.116.197103

Gualdrón Duarte JL, Cantet RJC, Bates RO, Ernst CW, Raney NE, Steibel JP (2014) Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. BMC Bioinform 15:246. https://doi.org/10.1186/1471-2105-15-246

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397. https://doi.org/10.1534/genetics.107.081190

Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423–447. https://doi.org/10.2307/2529430

Henderson CR (1976) A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69–83. https://doi.org/10.2307/2529339

Kempthorne O (1957) An introduction to genetic statistics. John Wiley & Sons, New York

Kempthorne O, Nordskog AW (1959) Restricted selection indices. Biometrics 15:10–19. https://doi.org/10.2307/2527598

Kinghorn B (2013) DESIRE: Target your genetic gains. https://bkinghor.une.edu.au/desire.htm. Accessed 4 Sep. 2022.

Kloosterman B, Abelenda JA, Carretero Gomez MM et al (2013) Naturally occurring allele diversity allows potato cultivation in northern latitudes. Nature 495:246–250. https://doi.org/10.1038/nature11912

Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756. https://doi.org/10.1093/genetics/124.3.743

Legarra A (2016) Comparing estimates of genetic variance across different relationship models. Theor Pop Biol 107:26–30. https://doi.org/10.1016/j.tpb.2015.08.005

Legarra A, Aguilar I, Misztal I (2009) A relationship matrix including full pedigree and genomic information. J Dairy Sci 92:4656–4663. https://doi.org/10.3168/jds.2009-2061

Lipka AE, Tian F, Wang Q, Peiffer J et al (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28:2397–2399. https://doi.org/10.1093/bioinformatics/bts444

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829. https://doi.org/10.1093/genetics/157.4.1819

Möhring J, Piepho HP (2009) Comparison of weighting in two-stage analysis of plant breeding trials. Crop Sci 49:1977–1988. https://doi.org/10.2135/cropsci2009.02.0083

Montesinos-López OA, Montesinos-López A, Luna-Vázquez FJ, Toledo FH, Pérez-Rodríguez P, Lillemo M, Crossa J (2019) A R package for Bayesian analysis of multi-environment and multi-trait multi-environment data for genome-based prediction. G3 Bethesda 9:1355–1367. https://doi.org/10.1534/g3.119.400126

Moskvina V, Schmidt KM (2008) On multiple-testing correction in genome-wide association studies. Genet Epidemiol 32:567–573. https://doi.org/10.1002/gepi.20331

Pérez P, de los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics 198:483–495. https://doi.org/10.1534/genetics.114.164442

Pérez-Rodríguez P, de los Campos G (2022) Multitrait Bayesian shrinkage and variable selection models with the BGLR-R package. Genetics 222(1):12. https://doi.org/10.1093/genetics/iyac112

Piepho HP, Möhring J, Schulz-Streeck T, Ogutu JO (2012) A stage-wise approach for analysis of multi-environment trials. Biometrics 54:844–860. https://doi.org/10.1002/bimj.201100219

Poland J, Endelman J, Dawson J et al (2012) Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome 5:103–113. https://doi.org/10.3835/plantgenome2012.06.0006

R Core Team (2022) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Austria

Riedelsheimer C, Melchinger AE (2013) Optimizing the allocation of resources for genomic selection in one breeding cycle. Theor Appl Genet 126:2835–2848. https://doi.org/10.1007/s00122-013-2175-9

Rodríguez-Álvarez MX, Boer MP, Eeuwijk FA, Eilers PHC (2018) Correcting for spatial heterogeneity in plant breeding experiments with P-splines. Spatial Statistics 23:52–71. https://doi.org/10.1016/j.spasta.2017.10.003

Rogers AR, Dunne JC, Romay C et al (2021) The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize experiment. G3 Bethesda 11:jkaa050. https://doi.org/10.1093/g3journal/jkaa050

Runcie DE, Qu J, Cheng H, Crawford L (2021) MegaLMM: Mega-scale linear mixed models for genomic predictions with thousands of traits. Genome Biol 22:213. https://doi.org/10.1186/s13059-021-02416-w

Rutkoski J, Poland J, Mondal S, Autrique E, González Pérez L, Crossa J, Reynolds M, Singh R (2016) Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. G3 Bethesa 6:2799–2808. https://doi.org/10.1534/g3.116.032888

Schmitz Carley CA, Coombs JJ, Clough ME, De Jong WS et al (2019) Genetic covariance of environments in the potato National Chip Processing Trial. Crop Sci 58:107–114. https://doi.org/10.2135/cropsci2018.05.0314

Searle SR, Casella G, McCulloch CE (1992) Variance components. John Wiley & Sons, Hoboken, NJ

Smith A, Cullis B, Thompson R (2001) Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. Biometrics 57:1138–1147. https://doi.org/10.1111/j.0006-341X.2001.01138.x

Toro MA, Varona L (2010) A note on mate allocation for dominance handling in genomic selection. Gen Sel Evol 42:33. https://doi.org/10.1186/1297-9686-42-33

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423. https://doi.org/10.3168/jds.2007-0980

Varona L, Legarra A, Toro MA, Vitezica ZG (2018) Non-additive effects in genomic selection. Front Genet 9:78. https://doi.org/10.3389/fgene.2018.00078

Vitezica ZG, Varona L, Legarra A (2013) On the additive and dominance variance and covariance of individuals within the genomic selection scope. Genetics 195:1223–1230. https://doi.org/10.1534/genetics.113.155176

Vos PG, Uitdewilligen JGAML, Voorrips RE, Visser RGF, van Eck HJ (2015) Development and analysis of a 20K SNP array for potato (Solanum tuberosum): an insight into the breeding history. Theor Appl Genet 128:2387–2401. https://doi.org/10.1007/s00122-015-2593-y

Wimmer V, Albrecht T, Auinger HJ, Schön CC (2012) synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics 28:2086–2087. https://doi.org/10.1093/bioinformatics/bts335

Xiang T, Christensen OF, Vitezica ZG, Legarra A (2016) Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. Gen Sel Evol 48:92. https://doi.org/10.1186/s12711-016-0271-4

Yadav S, Wei X, Joyce P et al (2021) Improved genomic prediction of clonal performance in sugarcane by exploiting non-additive genetic effects. Theor Appl Genet 134:2235–2252. https://doi.org/10.1007/s00122-021-03822-1

Zych K, Gort G, Maliepaard CA, Jansen RC, Voorrips RE (2019) FitTetra 2.0: improved genotype calling for tetraploids with multiple population and parental data support. BMC Bioinformatics 20:148. https://doi.org/10.1186/s12859-019-2703-y