

Supplemental Material:
Robust Modelling of Additive and Non-additive Variation
with Intuitive Inclusion of Expert Knowledge

Ingeborg Gullikstad Hem^{1*}, Maria Lie Selle*, Gregor Gorjanc[†], Geir-Arne Fuglstad*
and Andrea Riebler*

* Department of Mathematical Sciences, Norwegian University of Science and
Technology, Trondheim, Norway.

[†] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of
Edinburgh, Easter Bush, Edinburgh, UK.

¹Corresponding author: ingeborg.hem@ntnu.no

Contents

1	Note S1: Detailed Method Description	3
1.1	Model description	3
1.2	Component-wise priors	3
1.3	Model-wise prior	5
1.4	Data example	6
1.5	Extension to the additive and dominance model	9
2	Note S2: Results	12
2.1	Estimating genetic values	12
2.2	Estimating variances	12
2.3	Increasing number of observations	13
3	Supplemental Figures and Tables	18

1 Note S1: Detailed Method Description

In this note we describe the two approaches A-comp* and A-tree* in detail. We focus on the additive model (Model A) since this allows us to maximize readability and is sufficient to illustrate the main ideas, but the final section provides a description on how to extend to the additive and dominance model (Model AD) with the two approaches AD-comp* and AD-tree*. We also provide examples of the resulting prior and posterior distributions for two specific datasets with 100 and 500 observations. The aim is that this note contains all details necessary to reproduce the results.

1.1 Model description

The additive genetic model is given by

$$y_i = \mu + a_i + e_i, \quad i = 1, 2, \dots, n,$$

where n is the number of individuals, μ is the intercept, $\mathbf{a} = (a_1, a_2, \dots, a_n) \sim \mathcal{N}_n(\mathbf{0}, \sigma_a^2 \mathbf{A})$ are the additive values, and $\mathbf{e} = (e_1, e_2, \dots, e_n) \sim \mathcal{N}_n(\mathbf{0}, \sigma_e^2 \mathbf{I}_n)$ is the environmental noise. The covariance matrix \mathbf{A} is singular with rank less than n due to the construction from the SNP matrix. If we collect the phenotypes in a vector $\mathbf{y} = (y_1, y_2, \dots, y_n)$, the model can be formulated as the Gaussian likelihood

$$\pi(\mathbf{y} | \mu, \sigma_a^2, \sigma_e^2) = \left(\frac{1}{2\pi} \right)^{n/2} \frac{1}{|\Sigma(\sigma_a^2, \sigma_e^2)|^{n/2}} \exp \left(-\frac{1}{2} (\mathbf{y} - \mathbf{1}\mu)^T \Sigma(\sigma_a^2, \sigma_e^2)^{-1} (\mathbf{y} - \mathbf{1}\mu) \right), \quad \mathbf{y} \in \mathbb{R}^n, \quad (1)$$

where $\Sigma(\sigma_a^2, \sigma_e^2) = \sigma_a^2 \mathbf{A} + \sigma_e^2 \mathbf{I}_n$, and $\mathbf{1} = (1, 1, \dots, 1)$ is the n -dimensional vector of ones.

In Bayesian statistics, the likelihood must be combined with a prior distribution for the parameters, $\pi(\mu, \sigma_a^2, \sigma_e^2)$. This prior should encapsulate our prior beliefs about the parameters based on expert knowledge, and stabilize inference in low-data settings. Bayes' theorem gives the posterior distribution for the parameters as

$$\pi(\mu, \sigma_a^2, \sigma_e^2 | \mathbf{y}) = \frac{\pi(\mu, \sigma_a^2, \sigma_e^2) \pi(\mathbf{y} | \mu, \sigma_a^2, \sigma_e^2)}{\pi(\mathbf{y})} \propto \pi(\mu, \sigma_a^2, \sigma_e^2) \pi(\mathbf{y} | \mu, \sigma_a^2, \sigma_e^2), \quad (2)$$

where $\pi(\mathbf{y})$ is the marginal distribution of the phenotypes, and the proportionality is with respect to everything that does not vary as functions of the parameters. The constant $\pi(\mathbf{y})$ in Equation (2) is not needed for Markov chain Monte Carlo methods.

1.2 Component-wise priors

The common approach for selecting priors is to select independent component-wise prior distributions for σ_a^2 , σ_e^2 and μ so that $\pi(\sigma_a^2, \sigma_e^2, \mu) = \pi(\sigma_a^2) \pi(\sigma_e^2) \pi(\mu)$. A recent development are the penalized complexity (PC) priors that are derived in a principled way (Simpson et al., 2017). They are constructed based on a *base model* and "distance" to a more complex model that extends the base model. In the case of a model (1.1) the base model does not have a genetic effect or equivalently genetic variance is zero. The proposed prior penalizes increased model complexity through Kullback-Leibler divergence between the base and complex model, but the details are not essential to our presentation, and we encourage interested readers to see Simpson et al. (2017) for details. The PC prior for the variance of a Gaussian distribution is an exponential distribution on the standard deviation, which leads to the following priors transformed to

variances for the model (1.1):

$$\begin{aligned}\pi(\sigma_a^2) &= \frac{\lambda_a}{2\sqrt{\sigma_a^2}} \exp\left(-\lambda_a\sqrt{\sigma_a^2}\right), \quad \sigma_a^2 > 0, \\ \pi(\sigma_e^2) &= \frac{\lambda_e}{2\sqrt{\sigma_e^2}} \exp\left(-\lambda_e\sqrt{\sigma_e^2}\right), \quad \sigma_e^2 > 0.\end{aligned}$$

These priors are combined with a weakly informative Gaussian prior $\mu \sim \mathcal{N}_1(0, \sigma_{\text{Int}}^2)$ for the intercept. This is a $\text{PC}_0(\cdot)$ prior on variance.

Simpson et al. (2017) proposes to select λ_a and λ_e by eliciting a statement about a quantile for each variance from experts. In this paper we consider the specification of the two hyperparameters by specifying the upper quartiles of the priors through V_a and V_e for σ_a^2 and σ_e^2 , respectively. In this case, one must select

$$\lambda_a = -\frac{\ln(0.25)}{\sqrt{V_a}}, \quad \text{and} \quad \lambda_e = -\frac{\ln(0.25)}{\sqrt{V_e}}.$$

Applying Bayes' theorem from Equation (2) results in

$$\begin{aligned}\pi(\sigma_a^2, \sigma_e^2, \mu | \mathbf{y}) &\propto \pi(\mu) \pi(\sigma_a^2) \pi(\sigma_e^2) \pi(\mathbf{y} | \mu, \sigma_a^2, \sigma_e^2) \\ &\propto \frac{\lambda_a \lambda_e}{4\sqrt{\sigma_a^2 \sigma_e^2} |\Sigma(\sigma_a^2, \sigma_e^2)|^{n/2}} \\ &\quad \times \exp\left(-\frac{1}{2}(\mathbf{y} - \mu \mathbf{1})^T \Sigma(\sigma_a^2, \sigma_e^2)^{-1} (\mathbf{y} - \mu \mathbf{1}) - \frac{\mu^2}{2\sigma_{\text{Int}}^2} - \lambda_a \sqrt{\sigma_a^2} - \lambda_e \sqrt{\sigma_e^2}\right),\end{aligned}\quad (3)$$

for $\mu \in \mathbb{R}$, $\sigma_a^2 > 0$ and $\sigma_e^2 > 0$.

Even though the normalizing constant is not analytically tractable in Equation (3), Markov chain Monte Carlo (MCMC) methods can be devised for performing Bayesian inference through sampling. In this paper, we used Hamiltonian Monte Carlo (HMC) through Stan (Carpenter et al., 2017), a probabilistic programming language for statistical inference. Stan takes advantage of the No U-Turn Sampler (NUTS Hoffman and Gelman, 2014), which replaces random walks with a more efficient exploration strategy based on numerical solution of differential equations. NUTS also has a reduced need for tuning compared to other MCMC algorithms and, in some cases, the algorithm can be run completely without manually setting tuning parameters. Sampling from the posterior in Equation (3) through Stan requires writing the expression for $\ln(\pi(\sigma_a^2, \sigma_e^2, \mu | \mathbf{y}))$ in a Stan description file. Clever parametrizations such as using the logarithms of the variances instead of the variances themselves will improve the efficiency, but the complexity involved in implementing MCMC is greatly reduced from writing one's own MCMC algorithms. The coding language in the description file is similar as C++.

We have used Stan through the R-package **RStan** (Stan Development Team, 2019) to implement the inference in the paper. This required us only to provide the expression for the joint posterior, and we did not need to calculate the different full conditionals such as in a standard Gibbs sampling algorithm. We rephrased the additive model as a hierarchical model, where the additive values are be sampled together with the parameters,

$$\begin{aligned}\mathbf{y} | \mu, \mathbf{a}, \sigma_e^2 &\sim \mathcal{N}_n(\mathbf{1}\mu + \mathbf{a}, \sigma_e^2 \mathbf{I}_n) \\ \mathbf{a} | \sigma_a^2 &\sim \mathcal{N}_n(\mathbf{0}, \sigma_a^2 \mathbf{A}) \\ (\mu, \sigma_a^2, \sigma_e^2) &\sim \pi(\mu, \sigma_a^2, \sigma_e^2),\end{aligned}$$

where $\mathbf{1} = (1, 1, \dots, 1)$ is the n -dimensional vector of ones. Again Bayes' theorem provides

$$\pi(\mathbf{a}, \mu, \sigma_a^2, \sigma_e^2 | \mathbf{y}) \propto \pi(\mu) \pi(\sigma_a^2) \pi(\sigma_e^2) \pi(\mathbf{a} | \sigma_a^2) \pi(\mathbf{y} | \mu, \mathbf{a}, \sigma_e^2),$$

where all terms in the product are known distributions. This joint posterior is implemented in Stan through a calculation of $\ln(\pi(\mathbf{a}, \mu, \sigma_a^2, \sigma_e^2 | \mathbf{y}))$. The details of the sampling is handled by the software. This approach is termed A-comp* in the paper, and the prior $\pi(\sigma_a)$ is denoted $\text{PC}_0(\sqrt{V_a^2}, 0.25)$ and the prior $\pi(\sigma_e)$ is denoted $\text{PC}_0(\sqrt{V_e^2}, 0.25)$.

1.3 Model-wise prior

A shortcoming of using independent component-wise priors is that it does not provide a direct way to include expert knowledge about the relative sizes of variance parameters. In the context of the additive model, this refers to the situation where expert knowledge could be available about the phenotypic variance $\sigma_P^2 = \sigma_a^2 + \sigma_e^2$ and the broad-sense heritability $h_g^2 = p_{\frac{g}{g+e}} = \sigma_a^2 / (\sigma_a^2 + \sigma_e^2)$, i.e., $\sigma_a^2 = p_{\frac{g}{g+e}} \sigma_P^2$ and $\sigma_e^2 = (1 - p_{\frac{g}{g+e}}) \sigma_P^2$. If component-wise priors are chosen for σ_a^2 and σ_e^2 , it can be extremely challenging to understand the *a priori* assumptions being imposed on σ_P^2 and $p_{\frac{g}{g+e}}$. Furthermore, we are not ensuring that we end up with reasonable families of priors for σ_P^2 and $p_{\frac{g}{g+e}}$, which have desirable properties.

A complementary framework to PC priors are the recent hierarchical decomposition (HD) priors by Fuglstad et al. (2020). Using this framework, we can directly incorporate expert knowledge about the quantities σ_P^2 and $p_{\frac{g}{g+e}}$. Assume that the expected variability in phenotypes is not known *a priori* and that we aim for a scale-invariant prior. We can address this situation with a Jeffreys' prior

$$\pi(\sigma_P^2) \propto 1/\sigma_P^2, \quad \sigma_P^2 > 0.$$

Next, we apply the approach detailed in Fuglstad et al. (2020) by assuming that the model described by Equation (1) is a flexible extension of the *base model* $\mathbf{y} | \mu, \sigma_P^2 \sim \mathcal{N}(\mathbf{1}\mu, \sigma_P^2 \mathbf{I}_n)$, where the heritability $p_{\frac{g}{g+e}} = 0$. This means that we use a $\text{PC}_0(\cdot)$ prior on $p_{\frac{g}{g+e}}$. Based on Fuglstad et al. (2020), we can then calculate the prior

$$\pi(p_{\frac{g}{g+e}}) = \lambda_h |d'(p_{\frac{g}{g+e}})| \exp\left(-\lambda_h d(p_{\frac{g}{g+e}})\right), \quad 0 < p_{\frac{g}{g+e}} < 1,$$

where

$$d(p_{\frac{g}{g+e}}) = \sqrt{p_{\frac{g}{g+e}} (\text{tr}(\mathbf{A}) - n) - \ln(\det(p_{\frac{g}{g+e}} \mathbf{A} + (1 - p_{\frac{g}{g+e}}) \mathbf{I}_n))},$$

and

$$d'(p_{\frac{g}{g+e}}) = \frac{\text{tr}(\mathbf{A}) - n - \text{tr}\left\{(p_{\frac{g}{g+e}} \mathbf{A} + (1 - p_{\frac{g}{g+e}}) \mathbf{I}_n)^{-1} (\mathbf{A} - \mathbf{I}_n)\right\}}{2\sqrt{p_{\frac{g}{g+e}} (\text{tr}(\mathbf{A}) - n) - \ln(\det(p_{\frac{g}{g+e}} \mathbf{A} + (1 - p_{\frac{g}{g+e}}) \mathbf{I}_n))}}$$

denotes the derivative of the function $d(\cdot)$. Here λ_h is a hyperparameter, $\text{tr}(\cdot)$ denotes the matrix trace, and $\det(\cdot)$ denotes the matrix determinant. The function $d(\cdot)$ is the Kullback-Leibler distance and expresses the added complexity of having a broad-sense heritability $p_{\frac{g}{g+e}} > 0$ compared to having the broad-sense heritability $p_{\frac{g}{g+e}} = 0$. We set the hyperparameter λ_h by specifying the median $R_{\frac{g}{g+e}}$ of the prior for $p_{\frac{g}{g+e}}$. This is achieved by setting

$$\lambda_h = -\frac{\ln(0.5)}{d(p_{\frac{g}{g+e}} = R_{\frac{g}{g+e}})}.$$

We choose to use independent priors for μ , σ_P^2 and $p_{\frac{g}{g+e}}$ so that $\pi(\mu, \sigma_P^2, p_{\frac{g}{g+e}}) = \pi(\mu) \pi(\sigma_P^2) \pi(p_{\frac{g}{g+e}})$.

Since the prior is formulated in terms of phenotypic variance and heritability, it is useful to reparametrize the hierarchical model as

$$\begin{aligned} \mathbf{y}|\mu, \sigma_P^2, p_{\frac{g}{g+e}} &\sim \mathcal{N}_n(\mathbf{1}\mu + \mathbf{a}, \sigma_P^2(1 - p_{\frac{g}{g+e}})\mathbf{I}_n) \\ \mathbf{a}|\sigma_P^2, p_{\frac{g}{g+e}} &\sim \mathcal{N}_n(\mathbf{0}, \sigma_P^2 p_{\frac{g}{g+e}} \mathbf{A}) \\ \mu, \sigma_P^2, p_{\frac{g}{g+e}} &\sim \pi(\mu, \sigma_P^2, p_{\frac{g}{g+e}}). \end{aligned}$$

We calculate the posterior up to proportionality through Bayes' theorem,

$$\pi(\mathbf{a}, \mu, \sigma_P^2, p_{\frac{g}{g+e}}|\mathbf{y}) \propto \pi(\mu)\pi(\sigma_P^2)\pi(p_{\frac{g}{g+e}})\pi(\mathbf{a}|\sigma_P^2, p_{\frac{g}{g+e}})\pi(\mathbf{y}|\mu, \mathbf{a}, \sigma_P^2, p_{\frac{g}{g+e}}),$$

where all terms in the product are known distributions. The sampling in Stan is automatically handled based on code that calculates the value of the joint posterior up to proportionality. We precompute $\ln(\pi(p_{\frac{g}{g+e}}))$ for a range of values and approximate the function by a spline. This greatly reduces the computational burden and provides a large speed-up under MCMC sampling. This approach is termed A-tree* in the paper, and the prior $\pi(p_{\frac{g}{g+e}})$ is denoted $\text{PC}_0(R_{\frac{g}{g+e}})$.

1.4 Data example

We simulated two datasets using the breeding program described in the main paper by reducing the 10,000 individuals at the first trial stage (headrow) to $n = 700, 600, \dots, 100$ individuals, and used one dataset of size 100 and one of size 500. The choices of hyperparameters for the priors follow the main paper: $V_a = 0.25 \cdot 1.86$ and $V_e = 0.75 \cdot 1.86$. Figures S1.1 and S1.2 clearly demonstrates that the dataset of size $n = 500$ provides more information than the data set of size $n = 100$. Figure S1.2a indicates that the inference about the phenotypic variance is not sensitive to the choice between the two priors for both $n = 100$ and $n = 500$, whereas Figure S1.2b shows that inference about heritability is influenced by the prior for $n = 100$, but not for $n = 500$. This suggests that it is important to select a plausible prior that encodes prior belief for heritability. This is an argument for a principled approach for prior construction directly targeting heritability and phenotypic variance instead of additive and environmental variances separately.

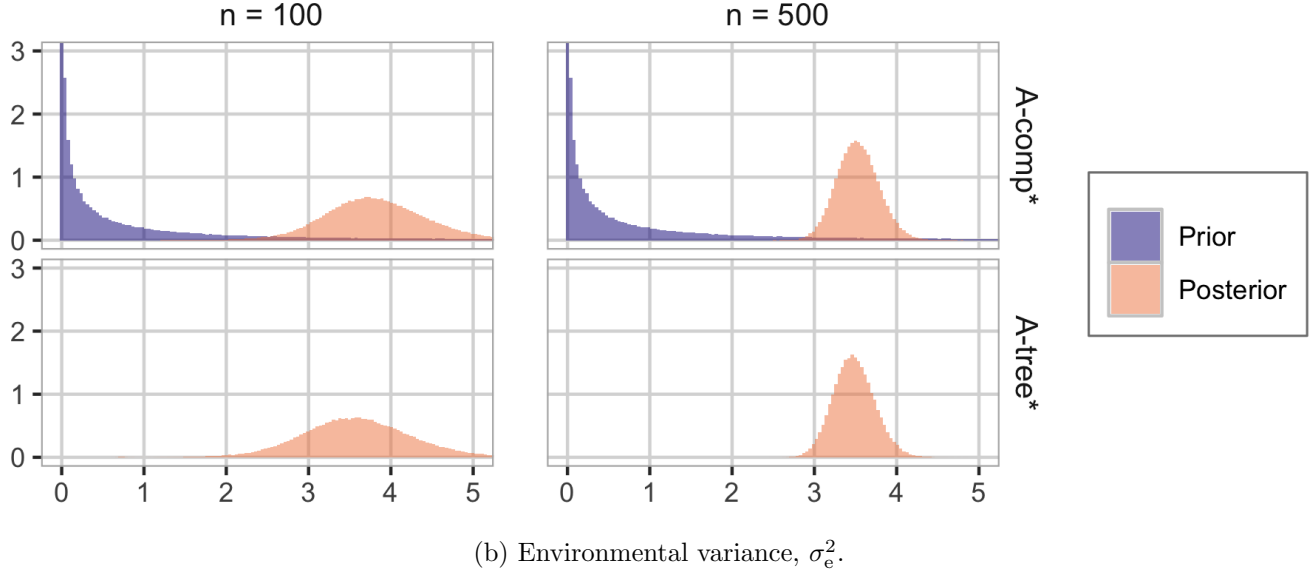
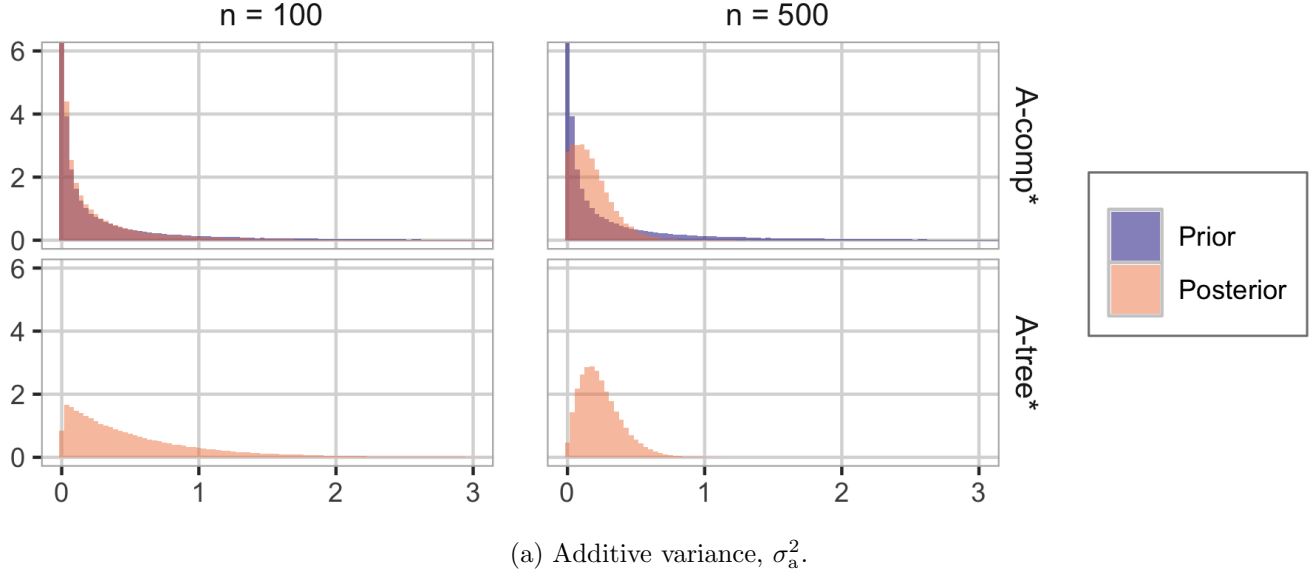
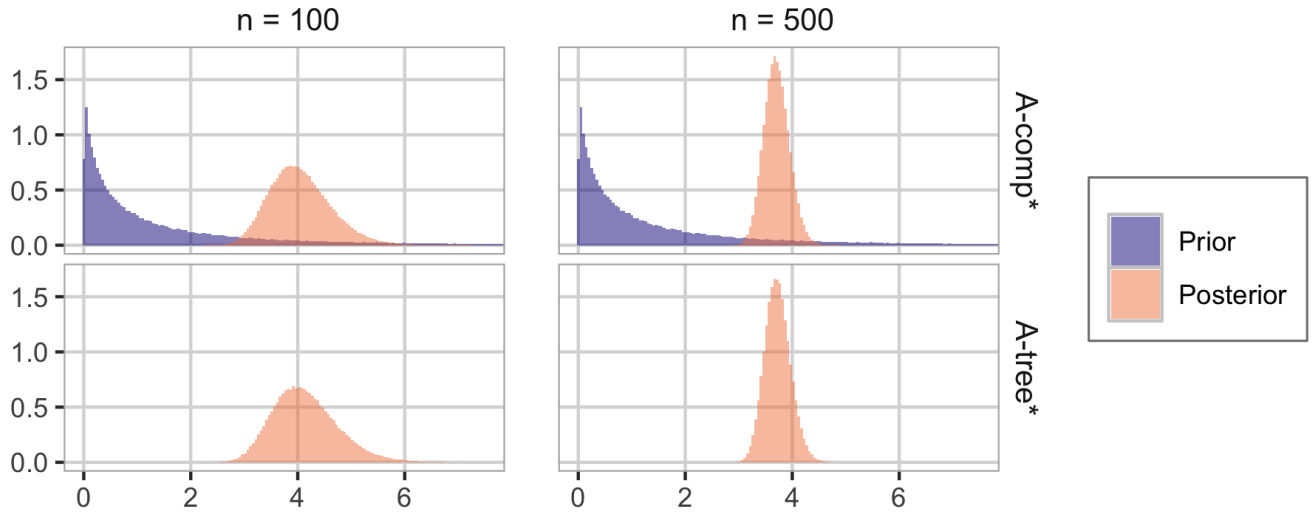
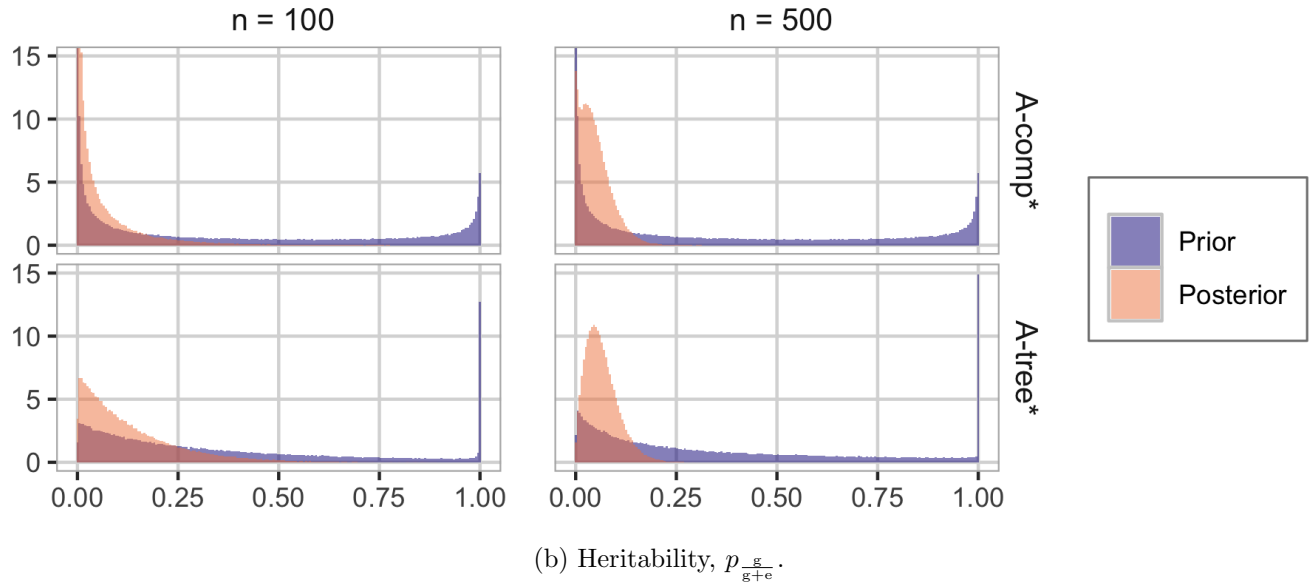


Figure S1.1: Columns indicate data sizes $n = 100$ and $n = 500$, and rows indicate priors A-comp* and A-tree*. The upper panel (a) shows priors and posteriors for additive variance σ_a^2 and the lower panel (b) shows priors and posteriors for environmental variance σ_e^2 . Priors are not plotted for A-tree* because the prior on the phenotypic variance σ_P^2 and thus also on $\sigma_a^2 = p_{\frac{g}{g+e}} \sigma_P^2$ and $\sigma_e^2 = (1 - p_{\frac{g}{g+e}}) \sigma_P^2$ are scale-invariant, and therefore improper.



(a) Phenotypic (total) variance, σ_P^2 .



(b) Heritability, $p_{\frac{g}{g+e}}$.

Figure S1.2: Columns indicate data sizes $n = 100$ and $n = 500$, and rows indicate priors A-comp* and A-tree*. The upper panel (a) shows priors and posteriors for phenotypic variance σ_P^2 and the lower panel (b) shows priors and posteriors for heritability $p_{\frac{g}{g+e}}$. Priors are not plotted for the combination A-tree* and σ_P^2 because the prior on σ_P^2 is scale-invariant, and therefore improper.

1.5 Extension to the additive and dominance model

For Model AD, dominance values are added to the model,

$$y_i = \mu + a_i + d_i + e_i, \quad i = 1, 2, \dots, n,$$

where n is the number of individuals, μ is the intercept, $\mathbf{a} = (a_1, a_2, \dots, a_n) \sim \mathcal{N}_n(\mathbf{0}, \sigma_a^2 \mathbf{A})$ are the additive values, $\mathbf{d} = (d_1, d_2, \dots, d_n) \sim \mathcal{N}_n(\mathbf{0}, \sigma_d^2 \mathbf{D})$ are dominance values, and $\mathbf{e} = (e_1, e_2, \dots, e_n) \sim \mathcal{N}_n(\mathbf{0}, \sigma_e^2 \mathbf{I}_n)$ is the environmental noise. Let $\sigma_P^2 = \sigma_a^2 + \sigma_d^2 + \sigma_e^2$ be the phenotypic variance, $p_{\frac{g}{g+e}} = (\sigma_a^2 + \sigma_d^2) / (\sigma_a^2 + \sigma_d^2 + \sigma_e^2)$ be the broad-sense heritability.

Following the main paper, we plan to describe the prior on the variances through a joint prior on σ_P^2 , $p_{\frac{g}{g+e}}$, and the proportion of additive to genetic variance, $p_{\frac{a}{g}} = \sigma_a^2 / \sigma_g^2$. The details are technical, and we therefore present the rationale behind each prior before we describe the technical details. We will apply independent priors to σ_P^2 , $p_{\frac{g}{g+e}}$, and $p_{\frac{a}{g}}$. These three priors will be described in reverse order.

We believe that $p_{\frac{a}{g}}$ should be around $R_{\frac{a}{g}}$, and desire a prior that favors this value and penalizes deviations from $R_{\frac{a}{g}}$. Therefore, we apply Fuglstad et al. (2020, Theorem 1) with the *base model* $p_{\frac{a}{g}} = R_{\frac{a}{g}}$, which yields

$$\pi(p_{\frac{a}{g}}) = \begin{cases} \frac{\lambda_1 |d'_1(p_{\frac{a}{g}})|}{2(1 - \exp(-\lambda_1 d_1(0)))} \exp(-\lambda_1 d_1(p_{\frac{a}{g}})), & 0 < p_{\frac{a}{g}} < R_{\frac{a}{g}}, \\ \frac{\lambda_1 |d'_1(p_{\frac{a}{g}})|}{2(1 - \exp(-\lambda_1 d_1(1)))} \exp(-\lambda_1 d_1(p_{\frac{a}{g}})), & R_{\frac{a}{g}} < p_{\frac{a}{g}} < 1. \end{cases}$$

This formulation guarantees that the median is $R_{\frac{a}{g}}$. This is a $\text{PC}_M(\cdot)$ prior. The distance is calculated as

$$d_1(p_{\frac{a}{g}}) = \sqrt{\text{tr}(\Sigma_0^{-1} \Sigma(p_{\frac{a}{g}})) - n - \ln(\det(\Sigma_0^{-1} \Sigma(p_{\frac{a}{g}})))},$$

where $\Sigma_0 = R_{\frac{a}{g}} \mathbf{A} + (1 - R_{\frac{a}{g}}) \mathbf{D}$ and $\Sigma(p_{\frac{a}{g}}) = p_{\frac{a}{g}} \mathbf{A} + (1 - p_{\frac{a}{g}}) \mathbf{D}$. In practice, we have found

$$\pi(p_{\frac{a}{g}}) = \frac{\lambda_1 |d'_1(p_{\frac{a}{g}})|}{2} \exp(-\lambda_1 d_1(p_{\frac{a}{g}})), \quad 0 < p_{\frac{a}{g}} < 1,$$

to be a reasonable approximation for the datasets in this paper as $d_1(0)$ and $d_1(1) \gg 1/\lambda_1$. The hyperparameter λ_1 controls the spread of the prior around the median $R_{\frac{a}{g}}$ and is selected by numerical optimization so that *a priori*

$$\text{P}(\text{logit}(R_{\frac{a}{g}}) - 1 < \text{logit}(p_{\frac{a}{g}}) < \text{logit}(R_{\frac{a}{g}}) + 1) = 0.75,$$

where $\text{logit}(p) = \ln(p/(1-p))$.

In the next step we construct a prior for $p_{\frac{g}{g+e}}$ with the *base model* $p_{\frac{g}{g+e}} = 0$. In this construction we assume the total genetic effect $\mathbf{g} | \sigma_P^2, p_{\frac{g}{g+e}} = (\mathbf{a} + \mathbf{d}) | \sigma_P^2, p_{\frac{g}{g+e}} \sim \mathcal{N}_n(\mathbf{0}, \sigma_P^2 p_{\frac{g}{g+e}} (R_{\frac{a}{g}} \mathbf{A} + (1 - R_{\frac{a}{g}}) \mathbf{D}))$. This means that we fix $p_{\frac{a}{g}} = R_{\frac{a}{g}}$ under the construction of the prior for $p_{\frac{g}{g+e}}$. Following Fuglstad et al. (2020, Theorem 1), a prior is constructed based on the distance measure

$$d_2(p_{\frac{g}{g+e}}) = \sqrt{\text{tr}(p_{\frac{g}{g+e}} \Sigma_2 + (1 - p_{\frac{g}{g+e}}) \mathbf{I}_n) - n - \ln(\det(p_{\frac{g}{g+e}} \Sigma_2 + (1 - p_{\frac{g}{g+e}}) \mathbf{I}_n))},$$

where $\Sigma_2 = R_{\frac{a}{g}} \mathbf{A} + (1 - R_{\frac{a}{g}}) \mathbf{D}$. The resulting prior is

$$\pi(p_{\frac{g}{g+e}}) = \lambda_2 |d'_2(p_{\frac{g}{g+e}})| \exp(-\lambda_2 d_2(p_{\frac{g}{g+e}})), \quad 0 < p_{\frac{g}{g+e}} < 1.$$

We set the hyperparameter λ_2 by specifying the median $R_{\frac{g}{g+e}}$ of the prior for $p_{\frac{g}{g+e}}$. This is achieved by setting

$$\lambda_2 = -\frac{\ln(0.5)}{d_2(p_{\frac{g}{g+e}} = R_{\frac{g}{g+e}})}.$$

This is a $PC_0(\cdot)$ prior

We want a scale-invariant prior for the phenotypic variance, and choose a Jeffreys' prior:

$$\pi(\sigma_P^2) \propto \frac{1}{\sigma_P^2}, \quad \sigma_P^2 > 0.$$

These three priors are combined with the previous prior on μ to form $\pi(\mu, \sigma_P^2, p_{\frac{g}{g+e}}, p_{\frac{a}{g}}) = \pi(\mu)\pi(\sigma_P^2)\pi(p_{\frac{g}{g+e}})\pi(p_{\frac{a}{g}})$. To allow sampling of both additive and dominance effects, we describe the model as a hierarchical model

$$\begin{aligned} \mathbf{y}|\mu, \mathbf{a}, \mathbf{d}, \sigma_P^2, p_{\frac{g}{g+e}} &\sim \mathcal{N}_n(\mathbf{1}\mu + \mathbf{a} + \mathbf{d}, (1 - p_{\frac{g}{g+e}})\sigma_P^2\mathbf{I}_n) \\ \mathbf{a}|\sigma_P^2, p_{\frac{g}{g+e}}, p_{\frac{a}{g}} &\sim \mathcal{N}_n(\mathbf{0}, p_{\frac{a}{g}}p_{\frac{g}{g+e}}\sigma_P^2\mathbf{A}) \\ \mathbf{d}|\sigma_P^2, p_{\frac{g}{g+e}}, p_{\frac{a}{g}} &\sim \mathcal{N}_n(\mathbf{0}, (1 - p_{\frac{a}{g}})p_{\frac{g}{g+e}}\sigma_P^2\mathbf{D}) \\ \mu, \sigma_P^2, p_{\frac{g}{g+e}}, p_{\frac{a}{g}} &\sim \pi(\mu, \sigma_P^2, p_{\frac{g}{g+e}}, p_{\frac{a}{g}}). \end{aligned}$$

Bayes' theorem results in

$$\pi(\mu, \mathbf{a}, \mathbf{d}, \sigma_P^2, p_{\frac{g}{g+e}}, p_{\frac{a}{g}}|\mathbf{y}) \propto \pi(\mu)\pi(\sigma_P^2)\pi(p_{\frac{g}{g+e}})\pi(p_{\frac{a}{g}})\pi(\mathbf{a}|\sigma_P^2, p_{\frac{g}{g+e}}, p_{\frac{a}{g}})\pi(\mathbf{d}|\sigma_P^2, p_{\frac{g}{g+e}}, p_{\frac{a}{g}})\pi(\mathbf{y}|\mu, \mathbf{a}, \mathbf{d}, \sigma_P^2, p_{\frac{g}{g+e}}, p_{\frac{a}{g}}).$$

Inference is implemented through **RStan** by writing code to calculate $\ln(\pi(\mathbf{a}, \mathbf{d}, \sigma_P^2, p_{\frac{g}{g+e}}, p_{\frac{a}{g}}|\mathbf{y}))$, and parameterization through the logarithm of the variances and the logit transformation of the proportions are needed. This is the approach termed AD-tree* in the paper, and the prior $\pi(p_{\frac{g}{g+e}})$ is denoted $PC_0(R_{\frac{g}{g+e}})$ and the prior $\pi(p_{\frac{a}{g}})$ is denoted $PC_M(R_{\frac{a}{g}})$. Plots of priors and posteriors offer little added value over those shown for Model A and are therefore omitted.

The approach given in this section extends further to the model also including epistasis (Model ADX). One additional step is required, but the same approach is taken for all steps as described above for Model AD.

Full details of the implementation in Stan for the approaches used in the paper are found in Files S3 and S4. File S3 contains the full Stan-code used for model fitting in the simulated case study, with necessary R functions and scripts for constructing the prior distributions and running inference, and File S4 contains the full Stan-code used in the real case study, with necessary R functions and scripts.

References

- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2017), 'Stan: a probabilistic programming language', *Journal of Statistical Software* **76**(1).
- Fuglstad, G.-A., Hem, I. G., Knight, A., Rue, H. and Riebler, A. (2020), 'Intuitive joint priors for variance parameters', *Bayesian Analysis*. Advance publication.
URL: <https://doi.org/10.1214/19-BA1185>

- Hoffman, M. D. and Gelman, A. (2014), ‘The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo’, *J. Mach. Learn. Res.* **15**(1), 1593–1623.
- Simpson, D., Rue, H., Riebler, A., Martins, T. G. and Sørbye, S. H. (2017), ‘Penalising model component complexity: A principled, practical approach to constructing priors’, *Statistical Science* **32**(1), 1–28.
- Stan Development Team (2019), ‘RStan: the R interface to Stan’. R package version 2.19.2. <http://mc-stan.org/>.
- URL:** <http://mc-stan.org/>

2 Note S2: Results

Here we give a detailed description of the results for the the additive model with model-wise default prior (A-tree) and the maximum likelihood approach (A-ML), the additive and dominance model and the non-additive model with model-wise expert knowledge prior (AD-tree* and ADX-tree*), in addition to phenotype selection. We show the results of the remaining settings in the Figures S11-S16.

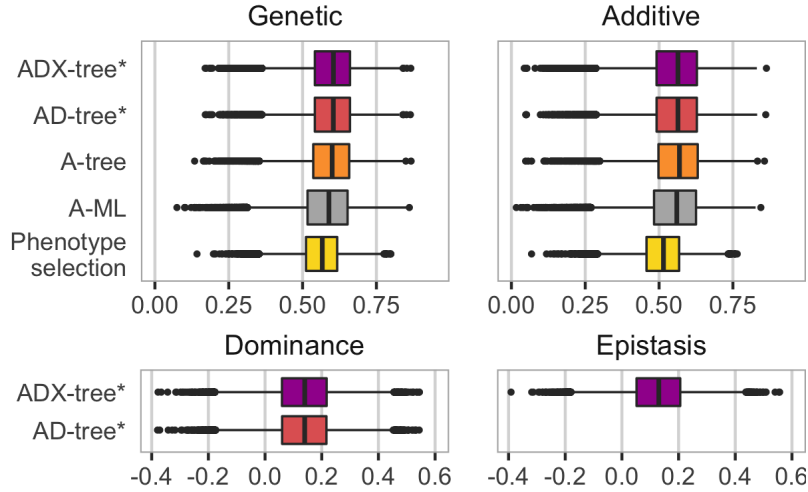


Figure S2.1: Accuracy of estimating the different genetic values for all individuals by model and prior setting - correlation (high value is desired, boxplots show variation over replicates). Genetic (upper left) means the estimated additive values for Model A, and the sum of estimated additive, dominance and epistasis values for Model ADX.

2.1 Estimating genetic values

While using the model-wise priors and expert knowledge significantly improved the selection of the genetically best individuals compared to the maximum-likelihood approach (see the main paper), it did not significantly improve the accuracy of estimating different genetic values across all individuals. We show this in Figure S2.1 with the correlation between the true (simulated) values and corresponding posterior means and in Figure S2.2 with the continuous rank probability score (CRPS) between the true values and corresponding posterior distributions. We show this for the genetic, additive, dominance and epistasis values separately. While there was a tendency of more favourable correlation and CRPS for certain model and prior settings, the variation between replicates was much larger than variation between the model and prior settings. The model-wise prior tended to perform better than the component-wise prior, expert knowledge tended to perform better than the default non-informative prior knowledge and use of prior knowledge via the Bayesian approach tended to perform better than the maximum likelihood. All models performed better in estimating the genetic and additive values, especially in the terms of CRPS, than the phenotype selection where we treat the phenotype as a point estimate of the genetic value.

2.2 Estimating variances

Variance component estimates varied considerably around the true values for all models and prior settings, but the estimates from the Bayesian inference showed slightly larger biases and smaller variance estimates

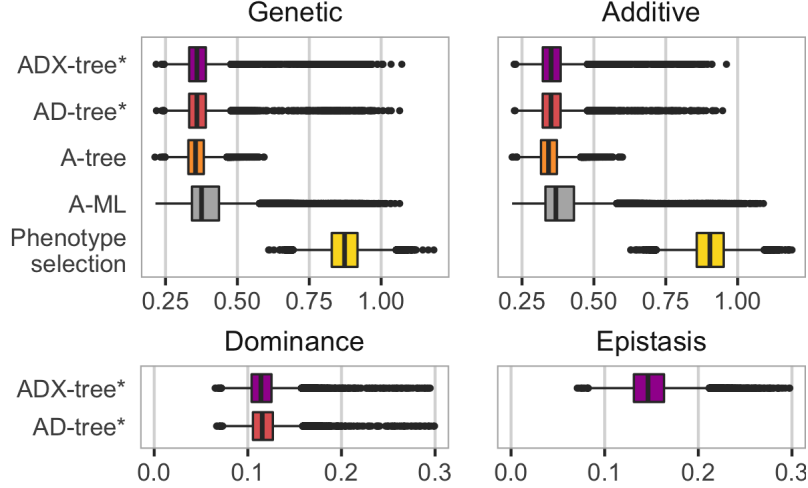


Figure S2.2: Accuracy of estimating the different genetic values for all individuals by model and prior setting - continuous rank probability score (CRPS; low value is desired, boxplots show variation over replicates). Genetic (upper left) means the estimated additive values for Model A, and the sum of estimated additive, dominance and epistasis values for Model ADX.

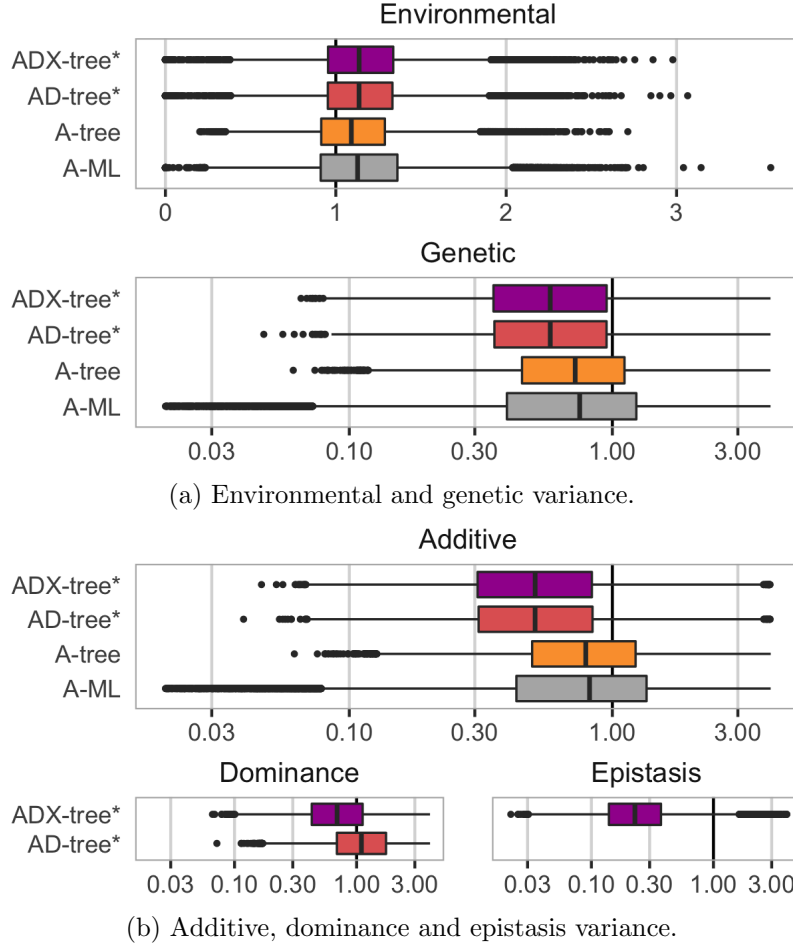
than the maximum likelihood approach. We show this in Figure S2.3 with the ratio of estimated to true variances (value close to 1 is desired and values below/above 1 denote underestimation/overestimation). Of the model and prior settings in Figure S2.3, A-ML was the closest to the true value on average in estimating the genetic variance, but also had the largest variation between replicates. Bayesian analysis with A-tree reduced variance between replicates, but did not improve bias. AD-tree* and ADX-tree* further increased the bias (underestimation) compared to A-ML and A-tree. When estimating dominance variance, AD-tree* performed better than ADX-tree*, but does not give estimates of the epistasis variance. Estimates for epistasis variance were considerably more underestimated than for the dominance variance.

In Figure S2.5 we show the posterior distributions of the environmental, additive, dominance and epistasis variances from one year in one simulated breeding program for the ADX-tree* setting (model-wise expert knowledge priors for the additive and non-additive model). We see what we would expect: The environmental variance is larger than the variances of the genetic components, the additive effect stands for most of the genetic variation, and the dominance and epistasis variances are small.

We show the prior and posterior distributions of the phenotypic variance, the proportion of genetic to phenotypic variance, proportion of additive to genetic variance, proportion of dominance to non-additive variance, also for the ADX-tree* setting in Figure S2.6. We see that the data informs about the phenotypic variance and the broad-sense heritability, but only weakly informs about the two other splits.

2.3 Increasing number of observations

Figure S2.4 shows the ability of estimating the model variances for increasing number of individuals for the additive (A) model and the additive and non-additive (ADX) model. The plot shows the posterior median from each model fit divided by the true variance from the simulated breeding program. Figure S2.4 shows the variance estimates from Models A and ADX, and datasets of size 100, 300, 500 and 700 here, and include the full results with variance estimates, correlation and continuous rank probability score (CRPS) for all models and number of individuals in the Figures S14-S16. From the environmental



(a) Environmental and genetic variance.

(b) Additive, dominance and epistasis variance.

Figure S2.3: Accuracy of estimating (a) environmental and genetic variance and (b) additive, dominance and epistasis variance by model and prior setting - expressed as the estimated posterior median divided by the true value (a value close to 1 is desired, boxplots show variation over replicates, x -axes have a log-scale (except for environmental variance) and is focused on area around 1 with some outliers excluded).

variance estimates we see that the variation between replicates decreases for all models for increasing number of observations. The maximum likelihood approach underestimated the additive, dominance and epistasis variances to a larger extent than the Bayesian approach did, and this underestimation decreased when the number of individuals increased. However, 700 observations is not enough for the maximum likelihood approach to obtain a bias in dominance and epistasis variance estimates as low as the Bayesian approach, indicating that the need for good priors decrease with increasing number of observations, but suitable priors are still necessary also for 700 observations. The inference stability did not increase with increasing number of observations for any of the models fitted with the maximum likelihood approach. The Bayesian models had the same high inference stability as in Table 2 in the main paper.

The correlation did not differ significantly between the models and approaches, and increased with increasing number of observations for all settings (Figure S15). The CRPS of the genetic and additive effects was significantly lower (better) for the models fitted with the Bayesian approach for a low number of individuals, but the maximum likelihood approach improved quickly when the dataset size increased

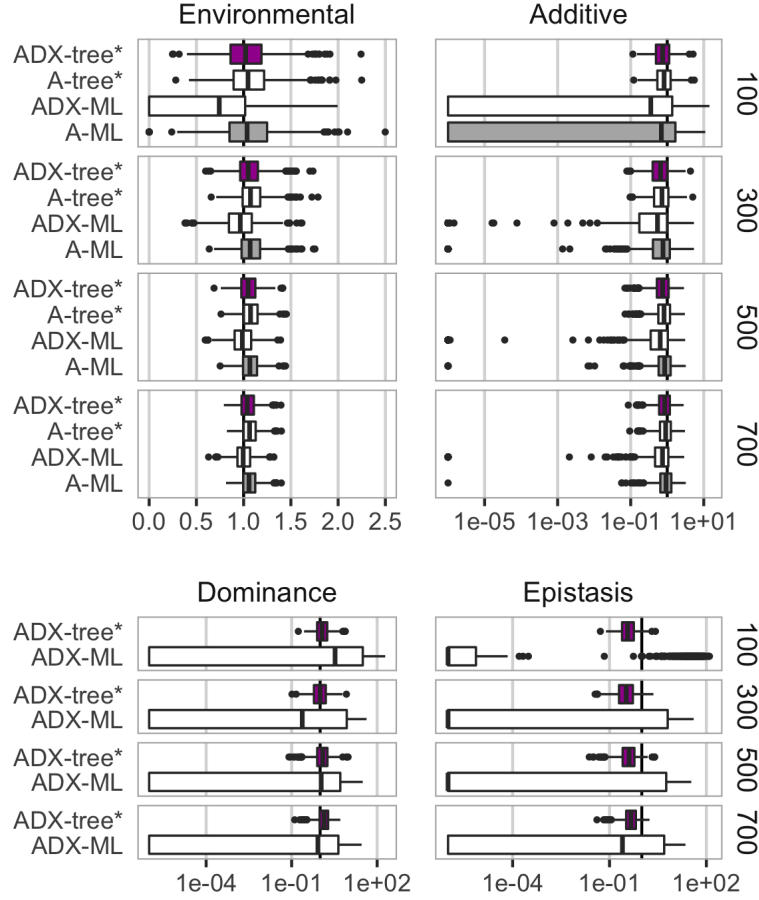


Figure S2.4: Accuracy of estimating environmental, additive, dominance and epistasis variance expressed as the estimated posterior median divided by the true value (a value close to 1 is desired). The dataset size is indicated for each box, and the y-axis shows the model and prior settings. x -axes have a log-scale (except for environmental variance), and all values smaller than 10^{-6} are set to 10^{-6} as those values are essentially zero.

(Figure S14). The Bayesian models had a significantly lower CRPS of the dominance and especially epistasis effects than the maximum likelihood approach for all dataset sizes. The results from the additive and dominance (AD) model did, with an exception of slightly more overestimation of the dominance variation for the maximum likelihood approach, not differ from the results from the non-additive model (ADX).

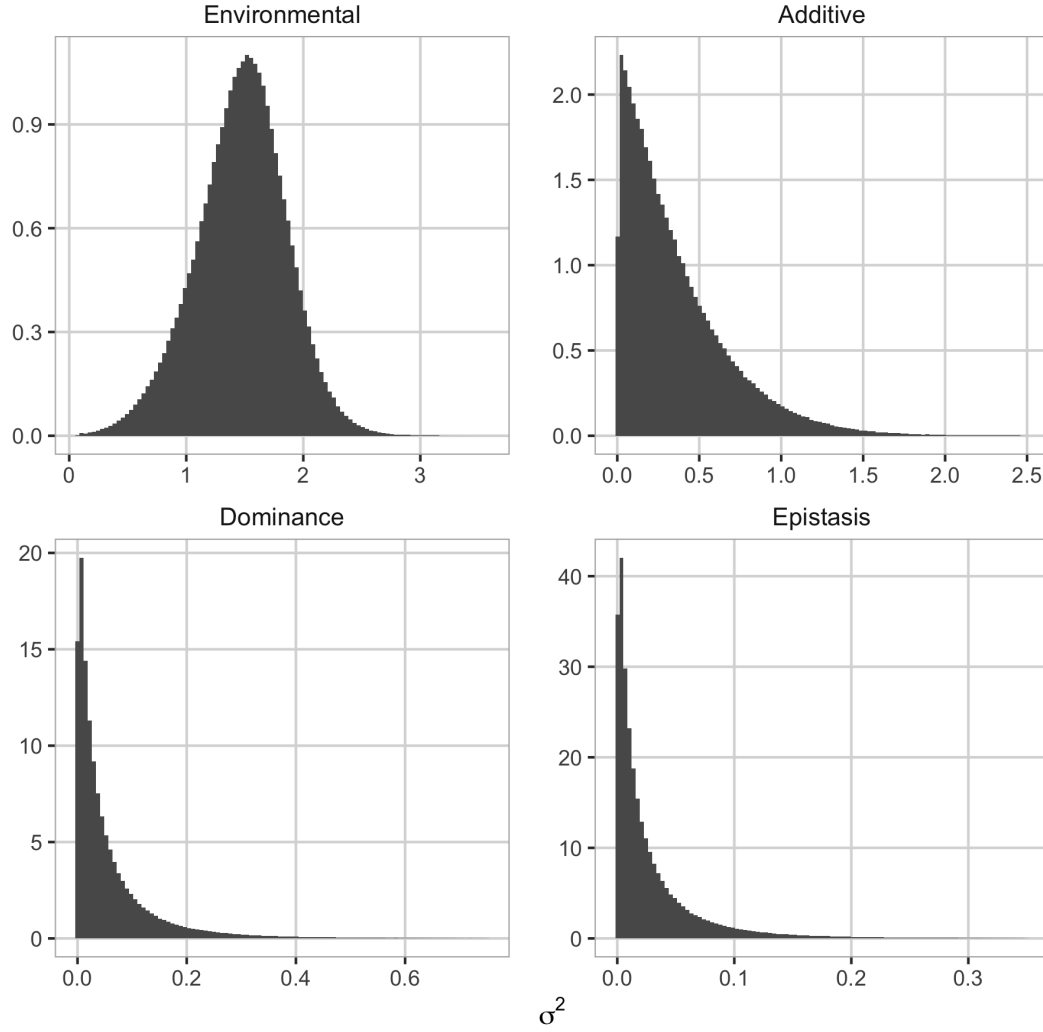


Figure S2.5: Posterior distribution of the environmental, additive, dominance and epistasis variance from the ADX-tree* setting, from one year in one simulated breeding program. Priors are not plotted because the prior on the phenotypic variance σ_P^2 and thus also on the variance parameters σ_e^2 , σ_a^2 , σ_d^2 and σ_x^2 , are scale-invariant, and therefore improper.

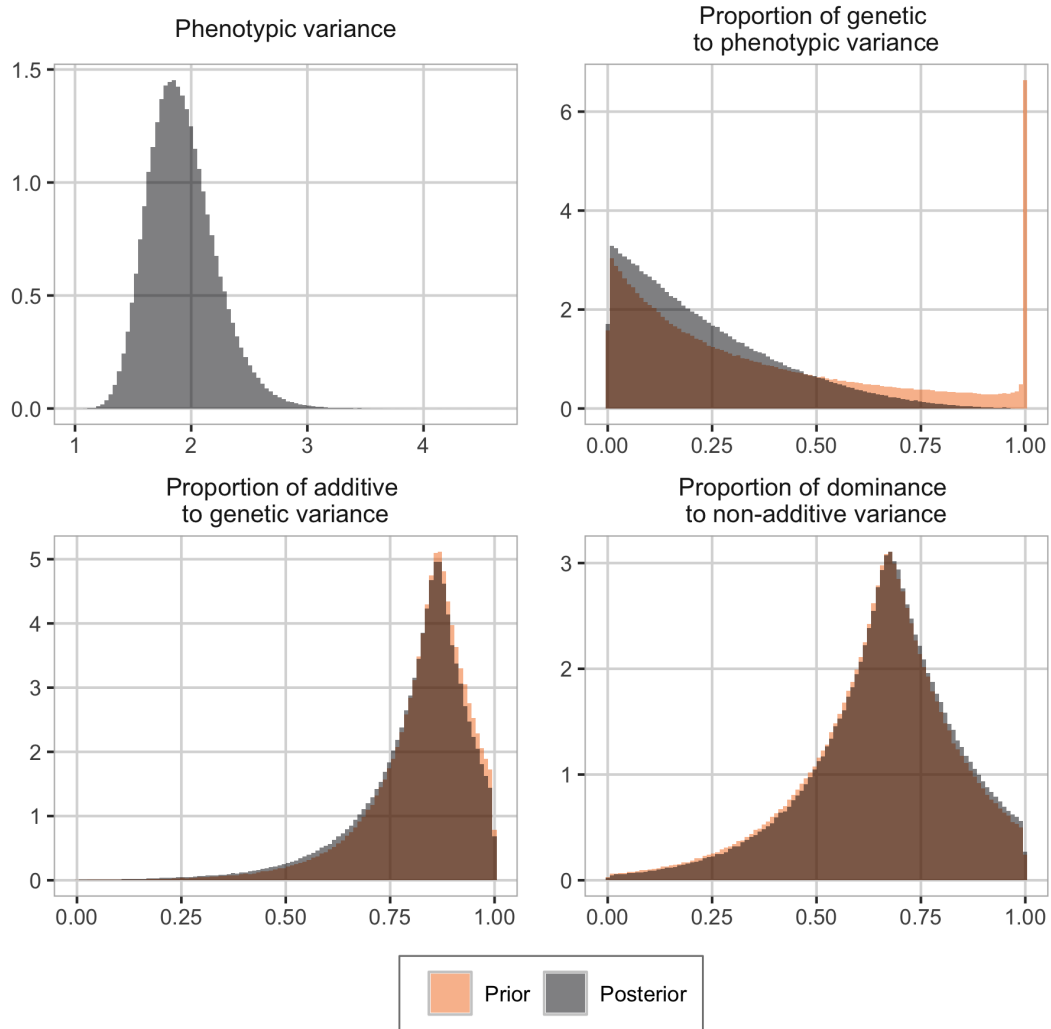


Figure S2.6: Prior and distribution of the phenotypic variance and variance proportions from the ADX-tree* setting, from one year in one simulated breeding program. The prior is not plotted for phenotypic variance σ_P^2 because it is scale-invariant, and therefore improper.

3 Supplemental Figures and Tables

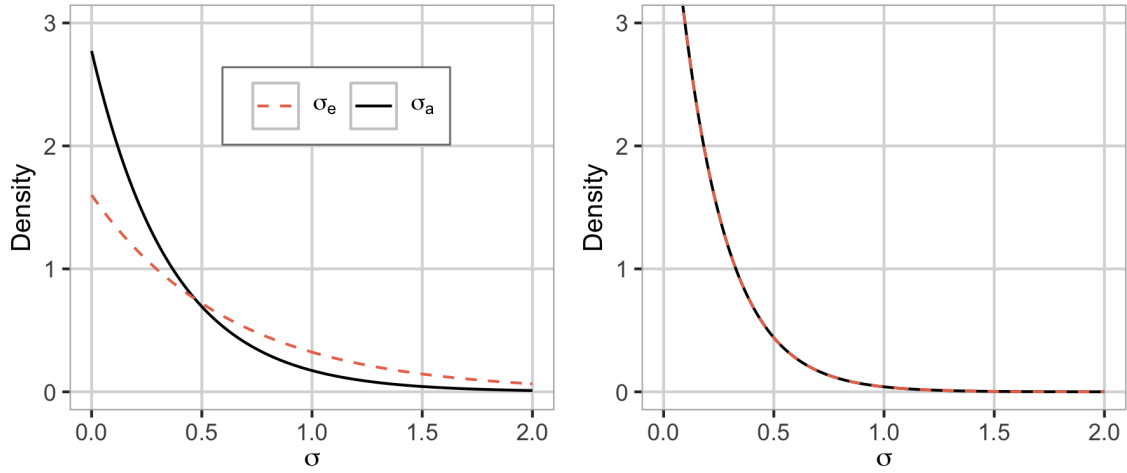


Figure S1: The prior used for the A-comp* (additive model with component-wise expert knowledge prior) (left) and A-comp (additive model with component-wise default prior) (right) settings. For A-comp*, we use $h_g^2 = 0.25$. We have plotted the priors for $V_P = 1$. For A-comp, additive and environmental variances have the same prior.

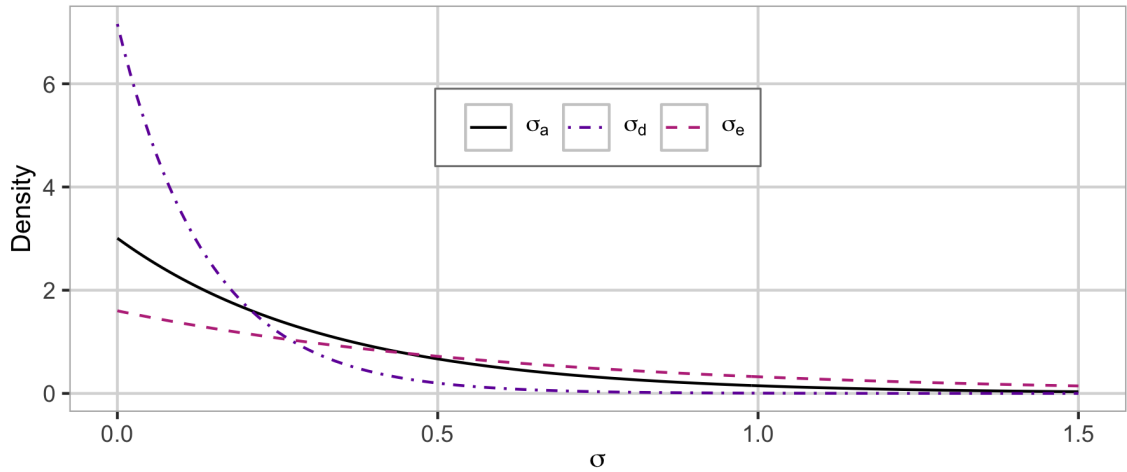


Figure S2: The prior used for the AD-comp* (additive and dominance model with component-wise expert knowledge prior) setting. We use $R_{\frac{g}{g+e}} = 0.25$ and $R_{\frac{a}{g}} = 0.85$. We have plotted the priors for $V_P = 1$.

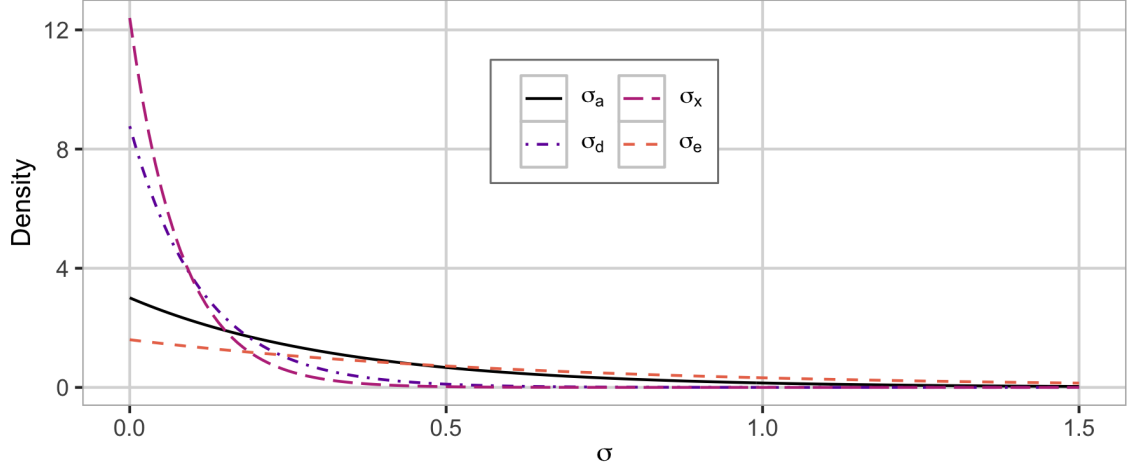


Figure S3: The prior used for the ADX-comp* (additive and non-additive model with component-wise expert knowledge prior) setting. We use $R_{\frac{g}{g+e}} = 0.25$, $R_{\frac{a}{g}} = 0.85$ and $R_{\frac{d}{d+x}} \approx 0.67$. We have plotted the priors for $V_p = 1$.

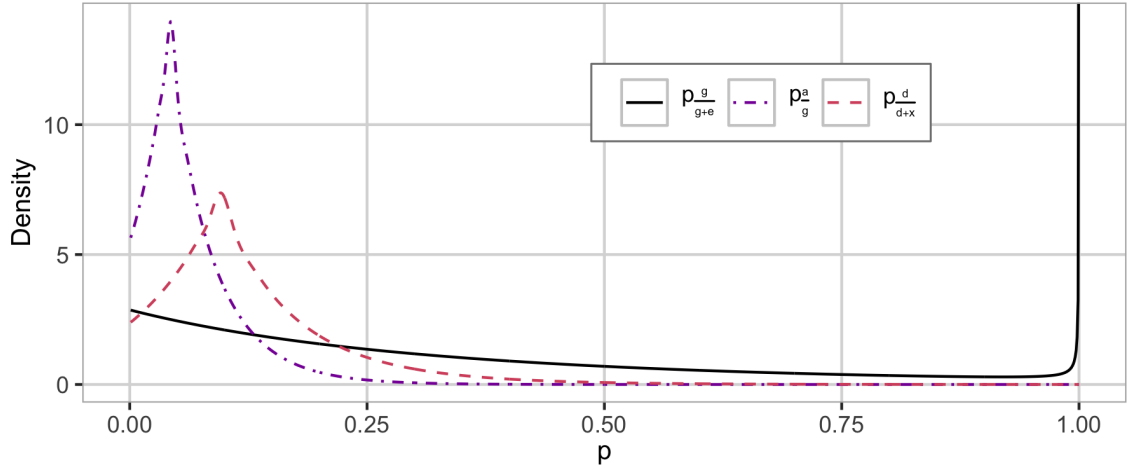
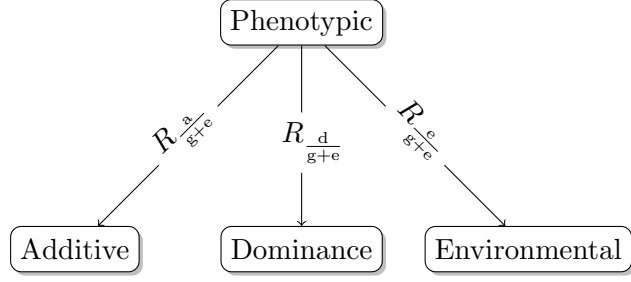
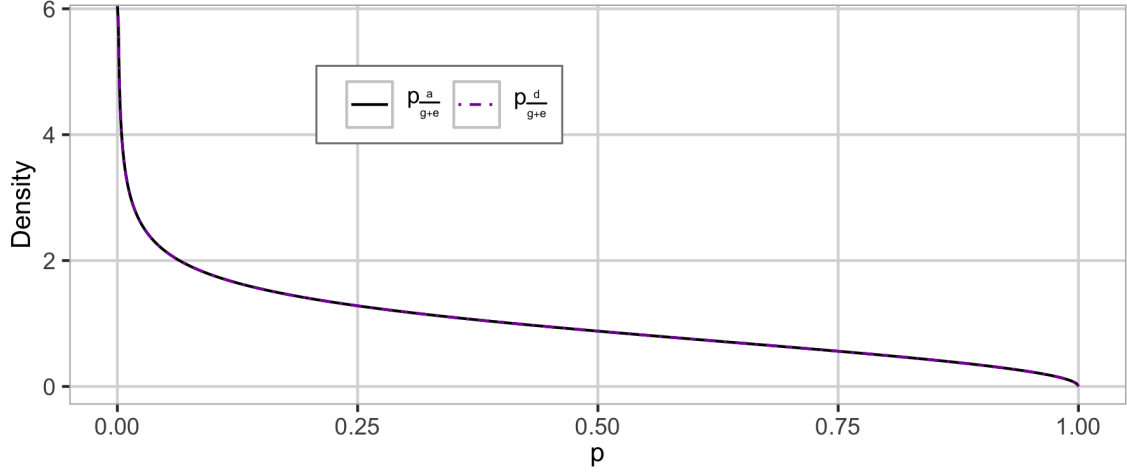


Figure S4: The HD prior used for the ADX-tree-opp* (additive and non-additive model with model-wise opposite expert knowledge prior) setting with the proportion of genetic to phenotypic variance $p_{\frac{g}{g+e}}$, additive to genetic variance $p_{\frac{a}{g}}^2$, and dominance to non-additive variance $p_{\frac{d}{d+x}}^d$. We use $R_{\frac{g}{g+e}} = 0.25$, $R_{\frac{a}{g}} = 0.05$ and $R_{\frac{d}{d+x}} \approx 0.11$. This is a dataset specific prior.

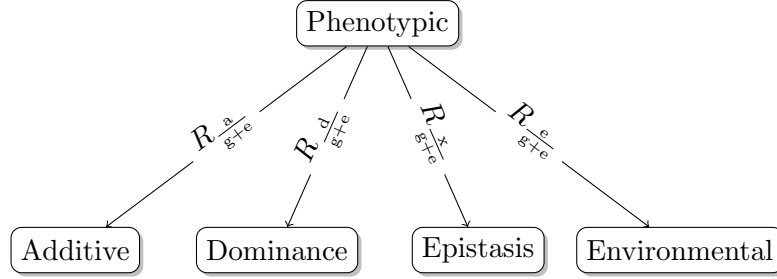


(a) Tree structure.

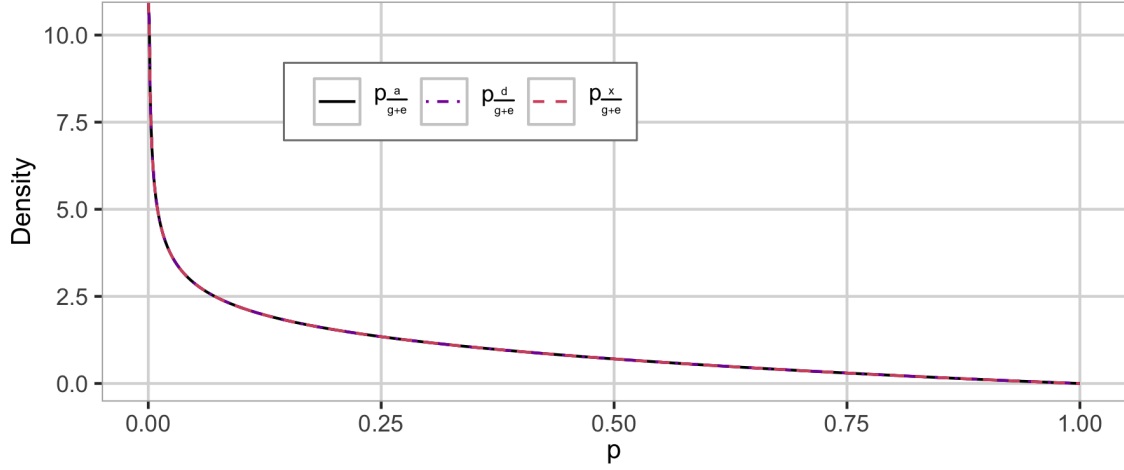


(b) Prior.

Figure S5: The (a) tree structure and (b) HD prior for the AD-tree (additive and dominance model with model-wise default prior) setting with equal magnitude for the four sources of variation without using expert knowledge - the proportion of additive to phenotypic variance $p_{\frac{a}{g+e}}$, and dominance to phenotypic variance $p_{\frac{d}{g+e}}$. This corresponds to a Dirichlet (3) prior on the variance proportions.



(a) Tree structure.



(b) Prior.

Figure S6: The (a) tree structure and (b) HD prior for the ADX-tree (additive and non-additive model with model-wise default prior) setting with equal magnitude for the four sources of variation without using expert knowledge - the proportion of additive to phenotypic variance $p_{\frac{a}{g+e}}$, dominance to phenotypic variance $p_{\frac{d}{g+e}}$, and epistasis to phenotypic variance $p_{\frac{x}{g+e}}$. This corresponds to a Dirichlet (4) prior on the variance proportions.

Trial	No. of obs.
Adenstedt (Ade13)	1,729
Böhnshausen (Boh12)	1,101
Böhnshausen (Boh13)	1,692
Hadmersleben (Had12)	1,738
Hadmersleben (Had13)	1,669
Harzhof (Hhof12)	1,736
Harzhof (Hhof13)	1,738
Hohenheim (Hoh12)	1,720
Hohenheim (Hoh13)	1,703
Seligenstadt (Sel12)	834
Seligenstadt (Sel13)	1,739

Table S1: The number of observed phenotypes for each of the 11 trials in the 6 locations in Germany for the Central European wheat dataset. The total number of individuals in the dataset is 1,739, where 15 are male parents, 120 are female parents and 1,604 are hybrids. Names in parentheses are the abbreviations used.

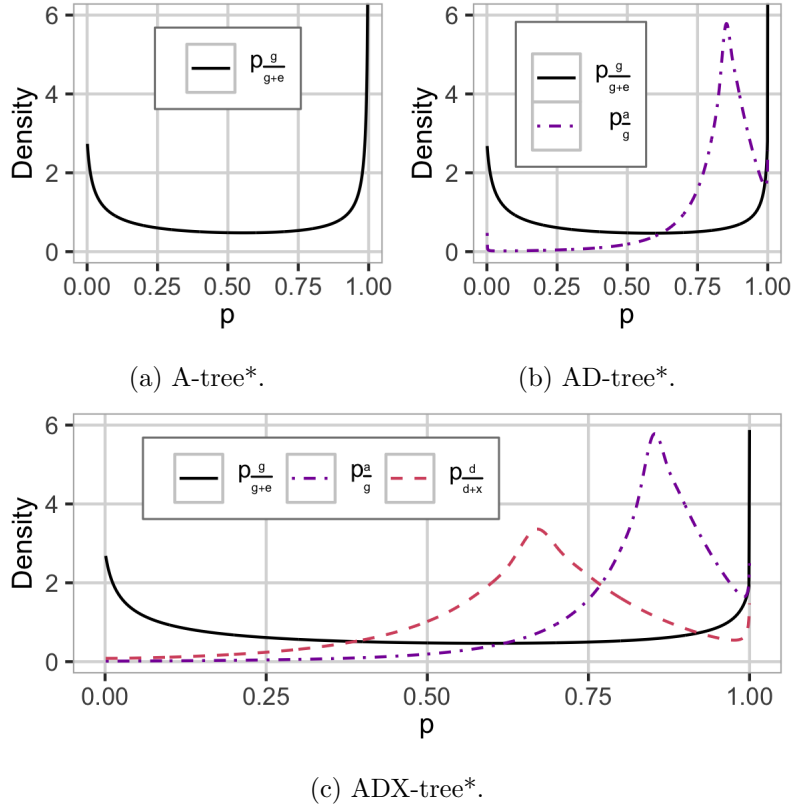


Figure S7: The model-wise expert knowledge HD prior used in (a) A-tree*, (b) AD-tree* and (c) ADX-tree* settings in the analysis of the Central European winter wheat data. $R_{\frac{g}{g+e}} = 0.75$, $R_{\frac{a}{g}} = 0.85$ and $R_{\frac{d}{d+x}} \approx 0.67$.

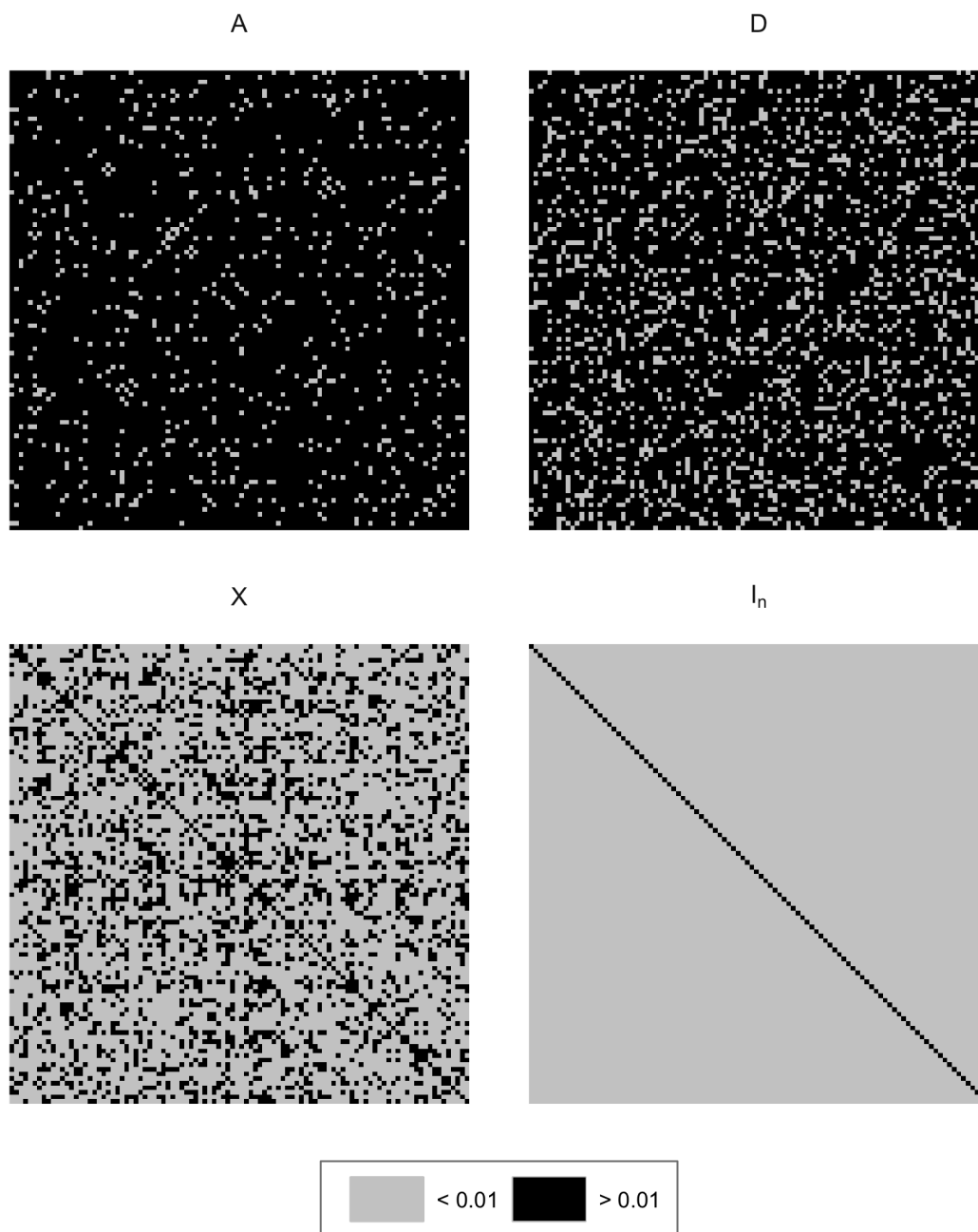


Figure S8: Covariance matrices for the additive (**A**), dominant (**D**), epistasis (**X**) and environmental (**I_n**) sources of variation for one year in one simulated breeding program.

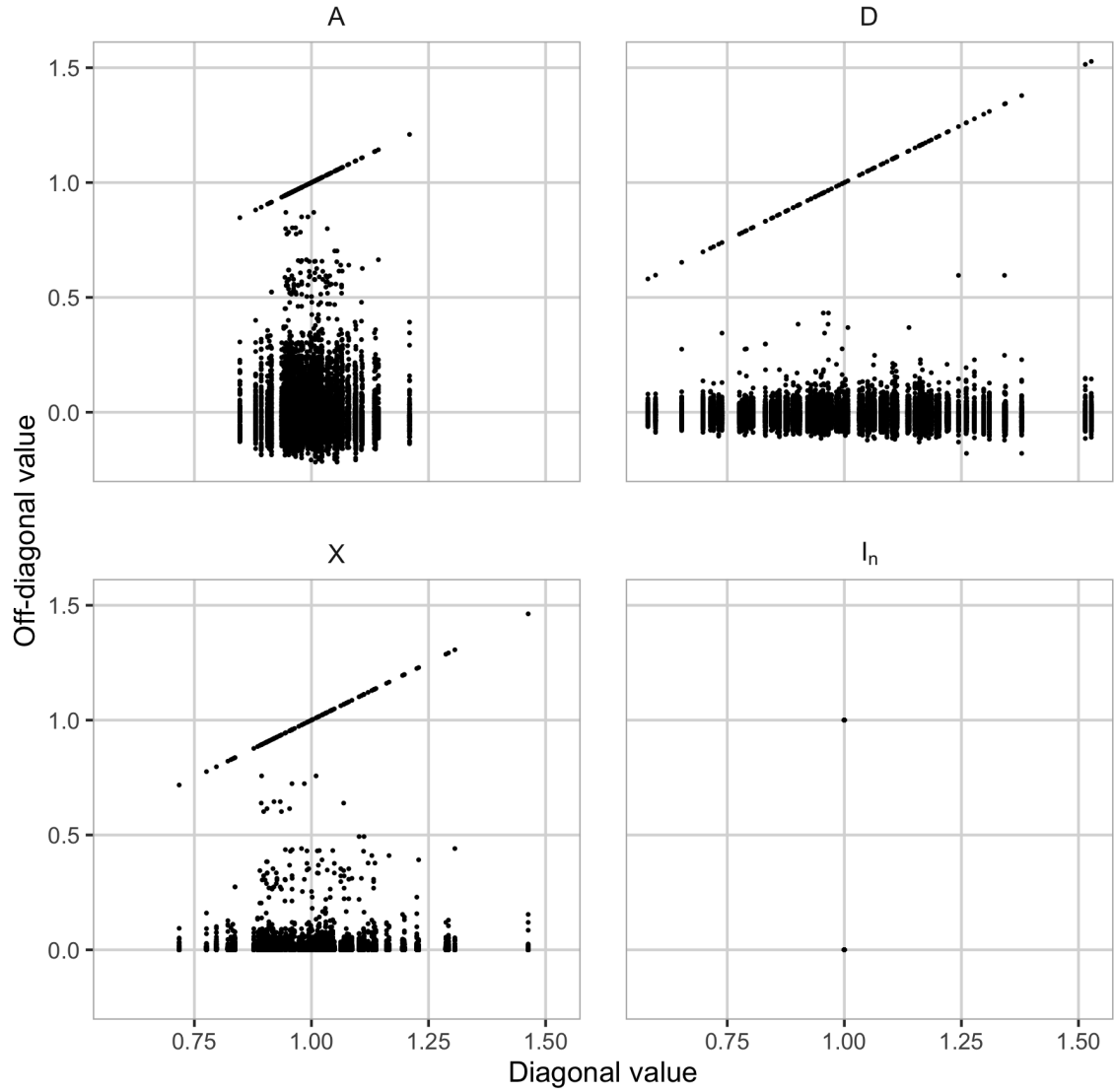


Figure S9: Scatterplot of entries of the covariance matrices for the additive (**A**), dominant (**D**), epistasis (**X**) and environmental (I_n) sources of variation for one year in one simulated breeding program. The off-diagonal values for each row is plotted pairwise against the diagonal value on the same row.

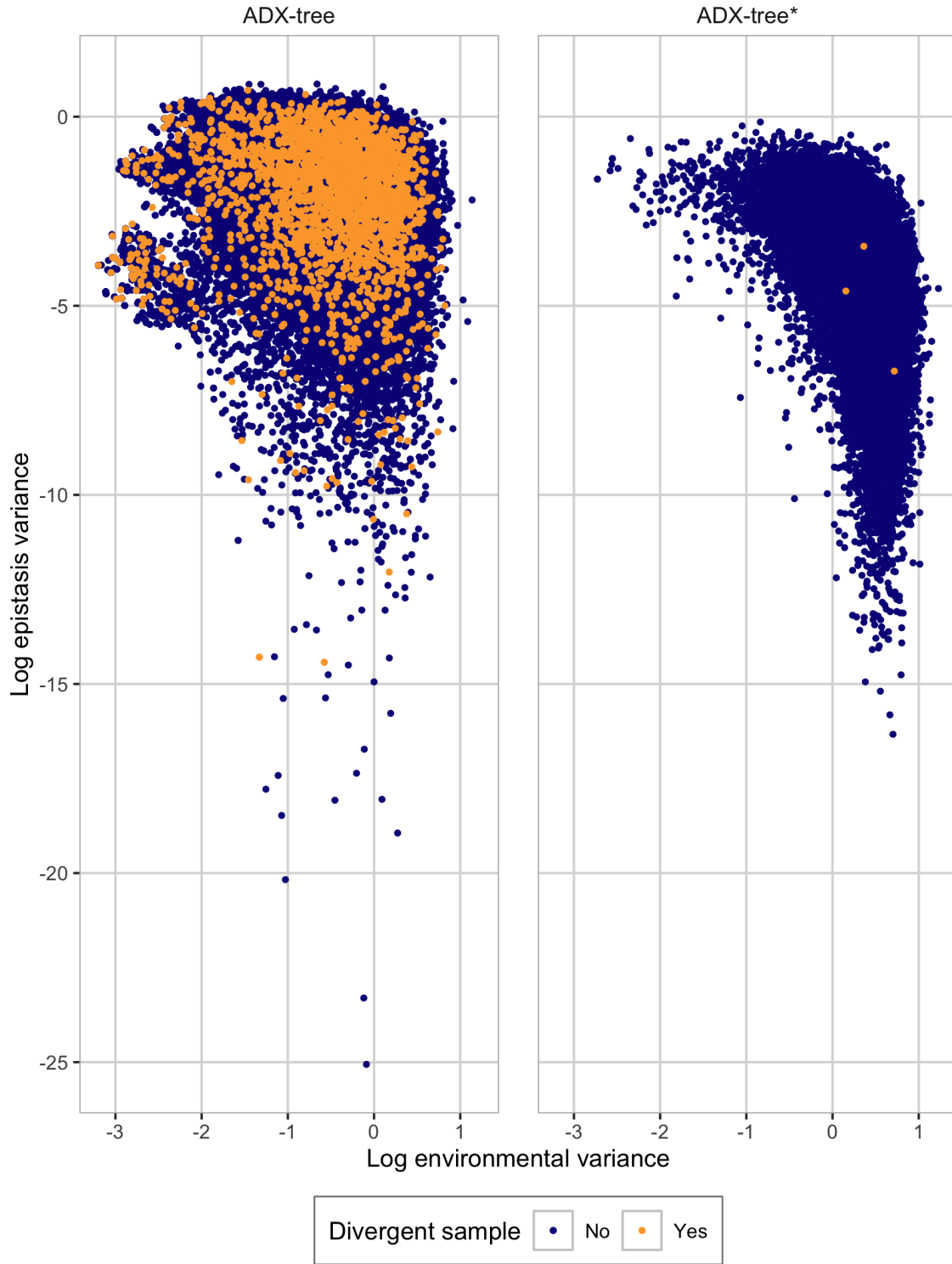


Figure S10: The joint posterior of the environmental and epistasis variances (log-scale) for one year in one simulated breeding program with the ADX-tree (left) and ADX-tree* (right) settings. By a divergent sample we mean a sample where the MCMC sampler had a divergent transition.

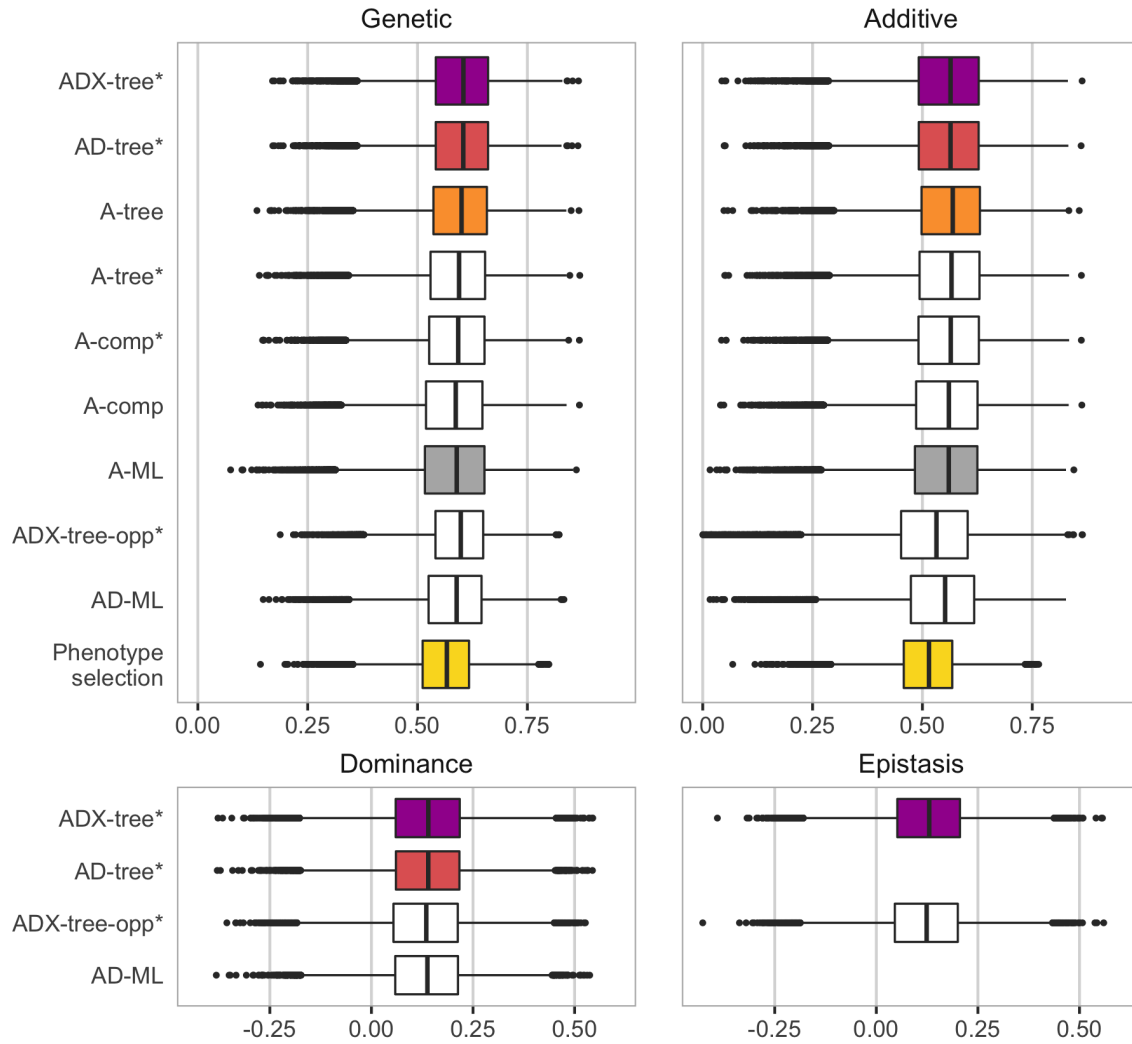


Figure S11: The ability to estimate the different genetic values for all individuals by the model and prior setting - correlation (high value is desired, boxplots show variation over replicates). Genetic (upper left) means the estimated additive values for Model A, the sum of the estimated additive and dominance values for Model AD, and the sum of estimated additive, dominance and epistasis values for Model ADX.

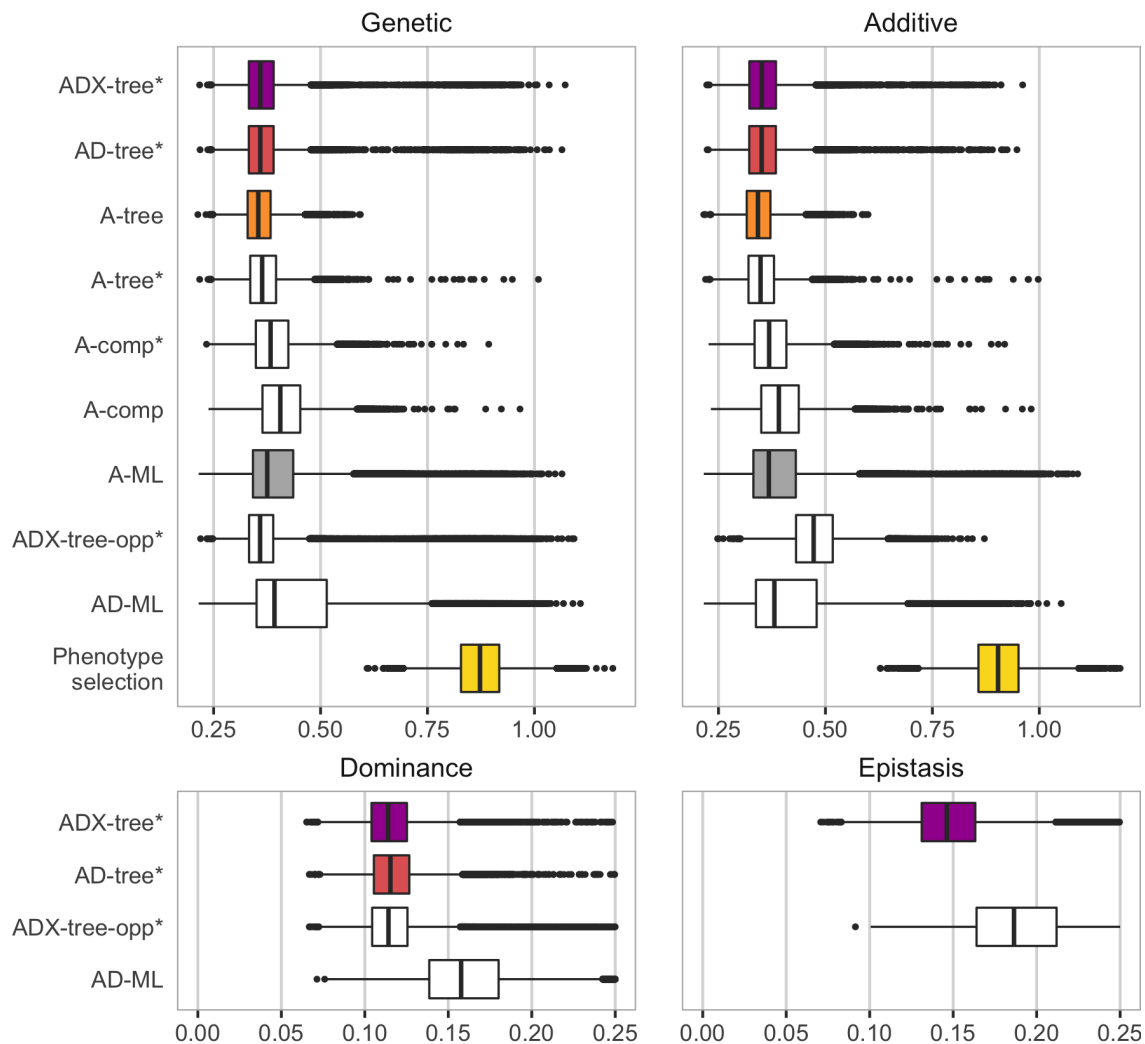
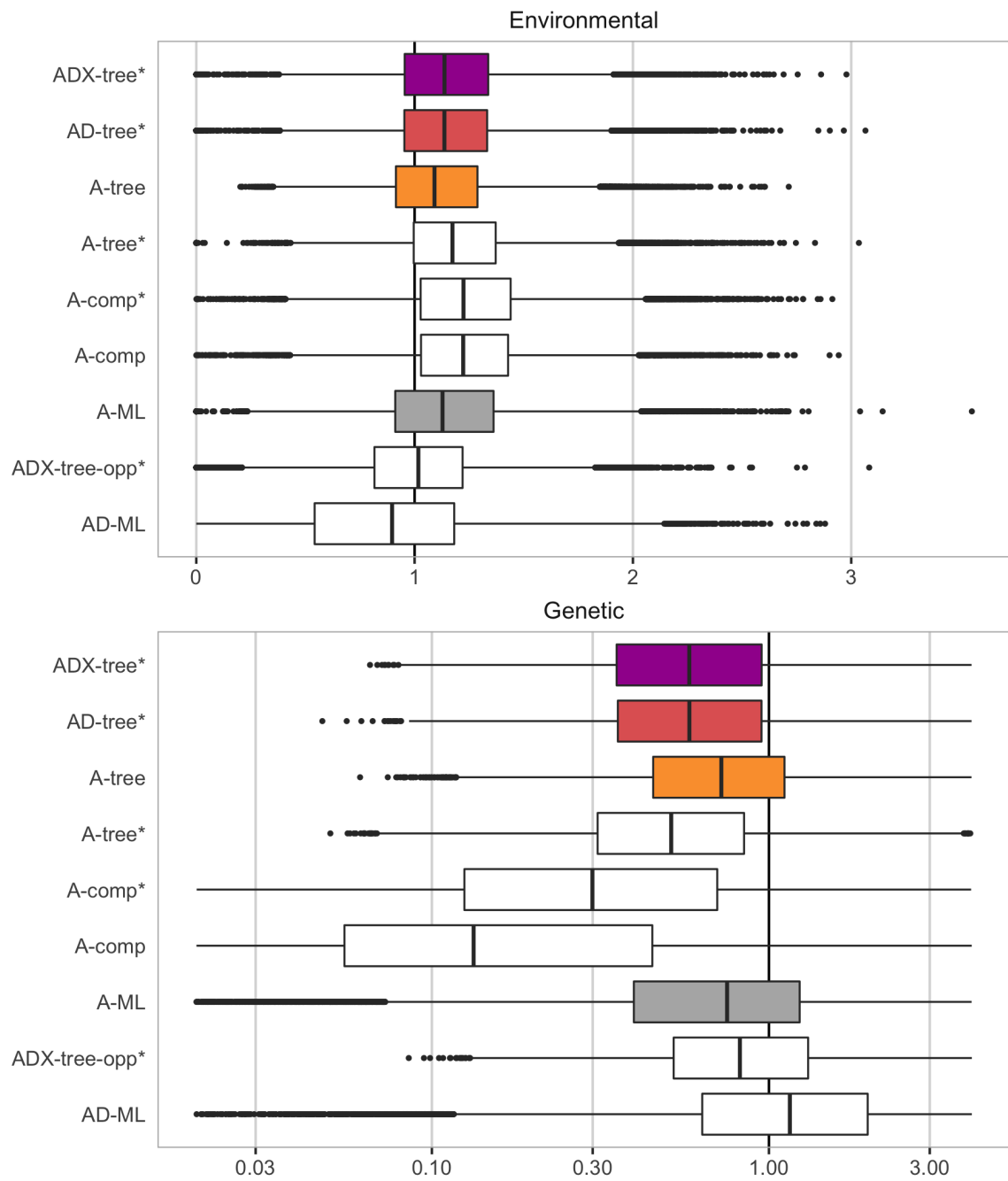
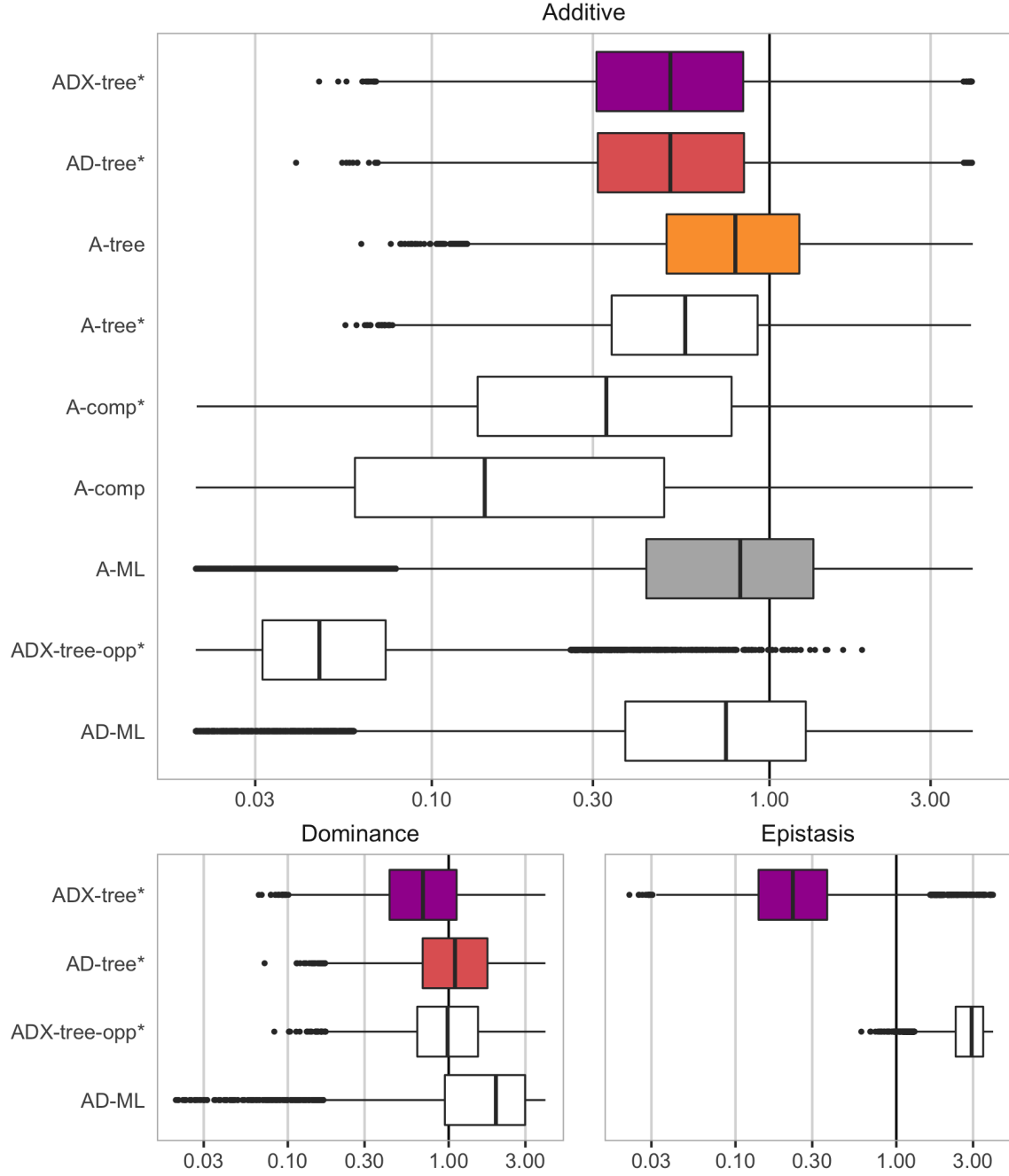


Figure S12: The ability to estimate the different genetic values for all individuals by the model and prior setting - continuous rank probability score (CRPS; low value is desired, boxplots show variation over replicates). Genetic (upper left) means the estimated additive values for Model A, the sum of the estimated additive and dominance values for Model AD, and the sum of estimated additive, dominance and epistasis values for Model ADX.

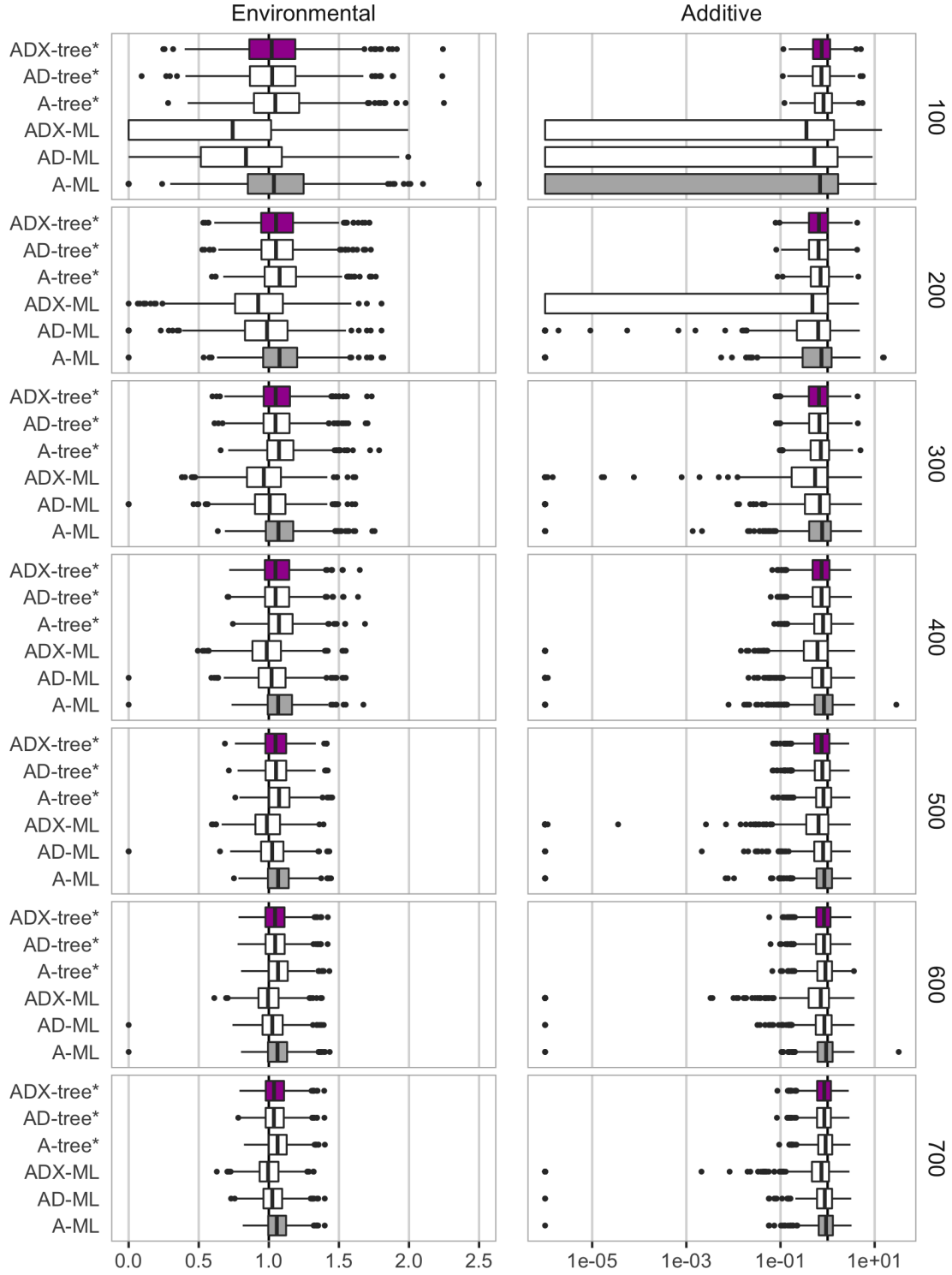


(a) The environmental and genetic variance.

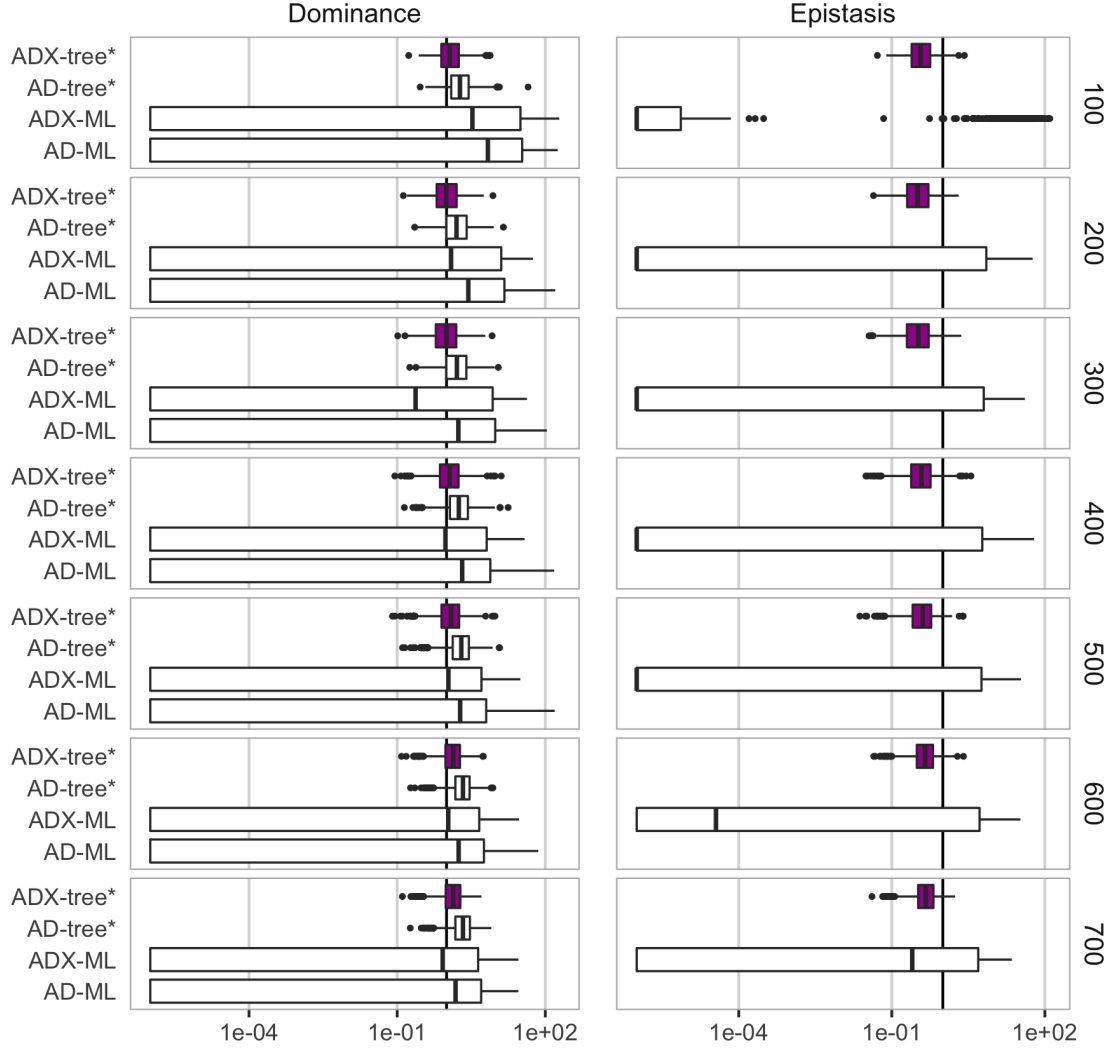


(b) The additive, dominance and epistasis variance.

Figure S13: The ability to estimate (a) environmental and genetic variance and (b) additive, dominance and epistasis variance by the model and prior setting - expressed as the estimated posterior median divided by the true value (a value close to 1 is desired, boxplots show variation over replicates, x -axes have a log-scale (except for environmental variance) and is focused on area around 1 with some outliers excluded).

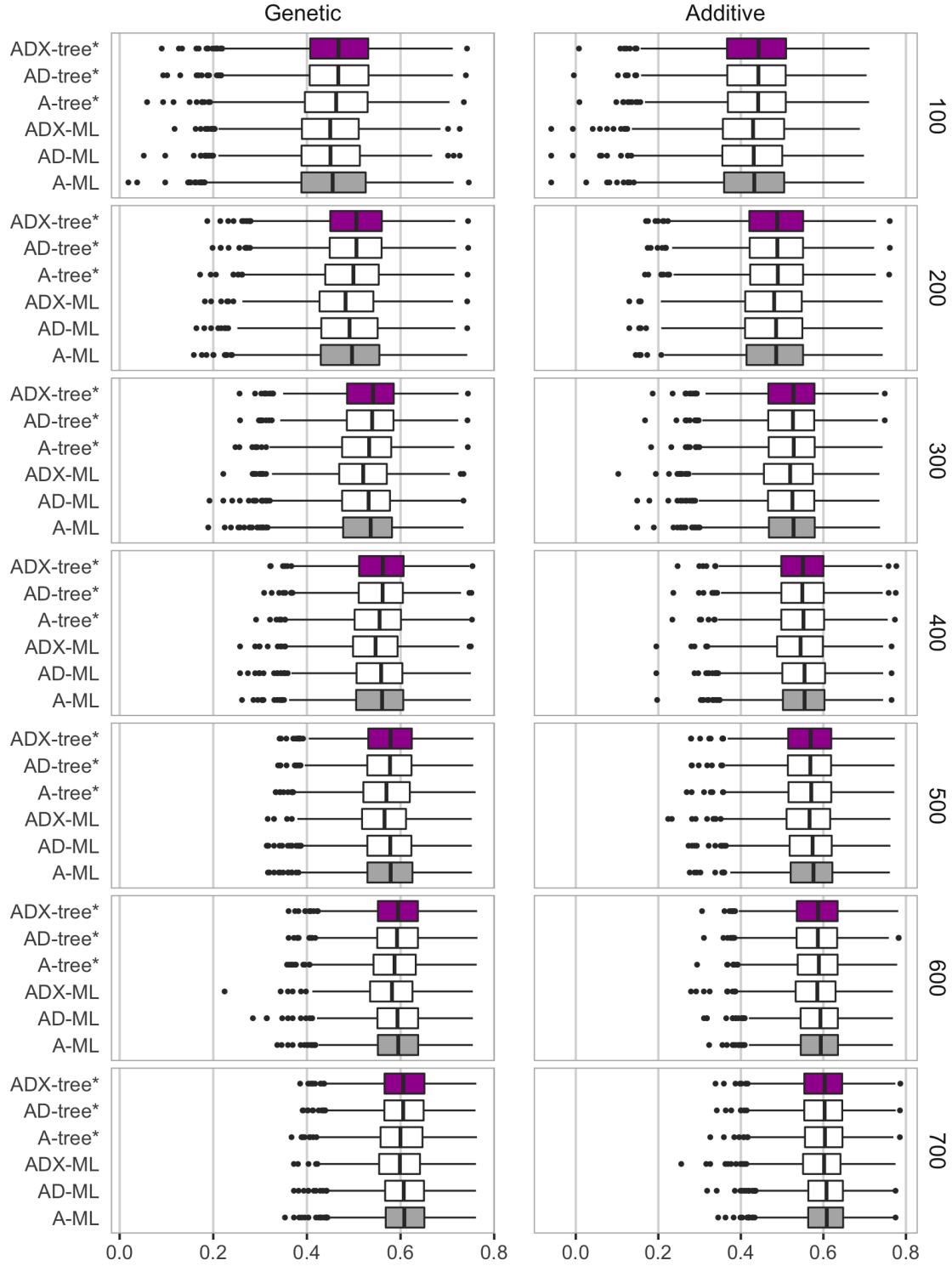


(a) The environmental and additive variance.

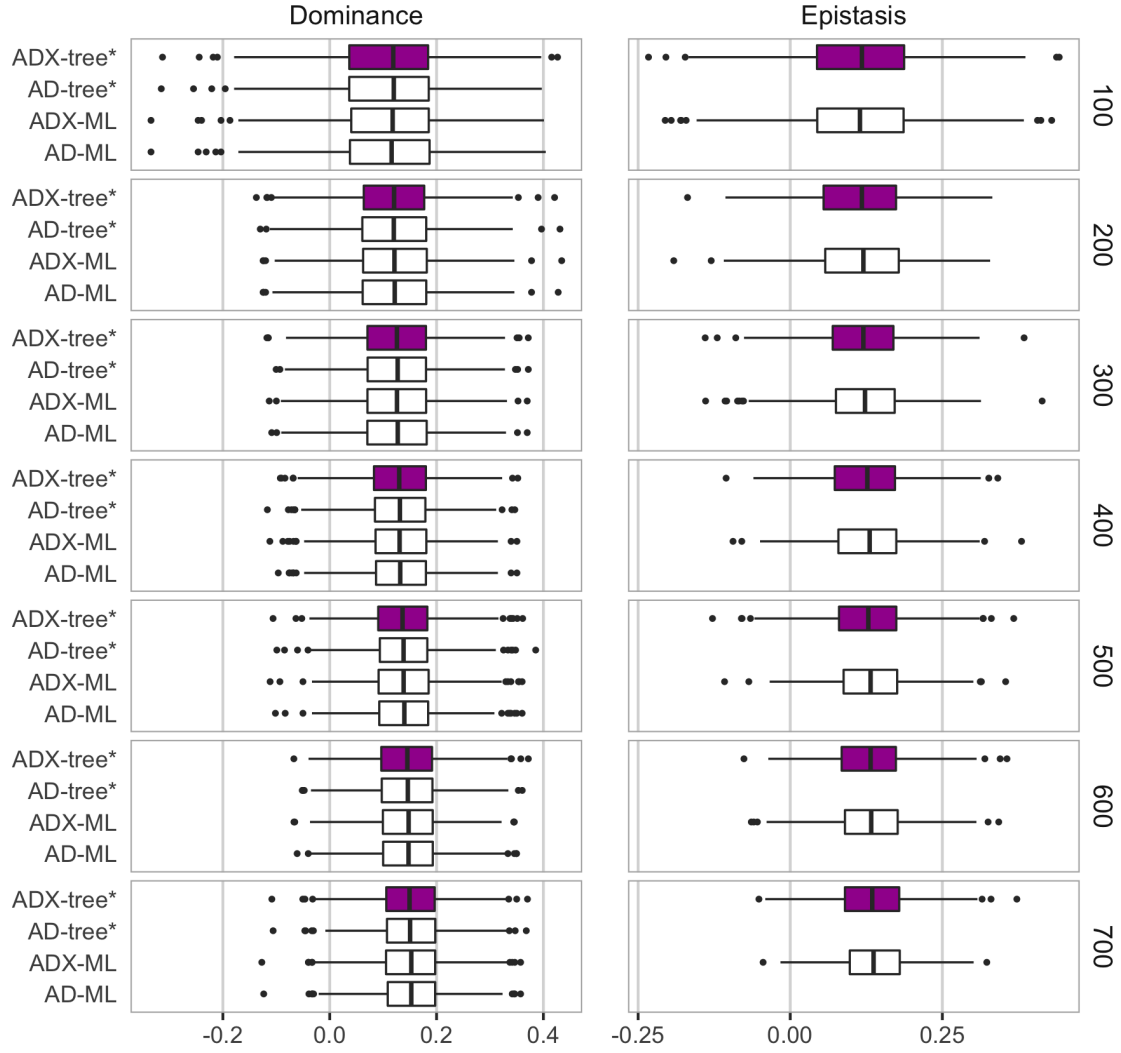


(b) The dominance and epistasis variance.

Figure S14: The ability to estimate (a) environmental and additive variance and (b) dominance and epistasis variance by model, prior setting and size - expressed as the estimated posterior median divided by the true value (a value close to 1 is desired).

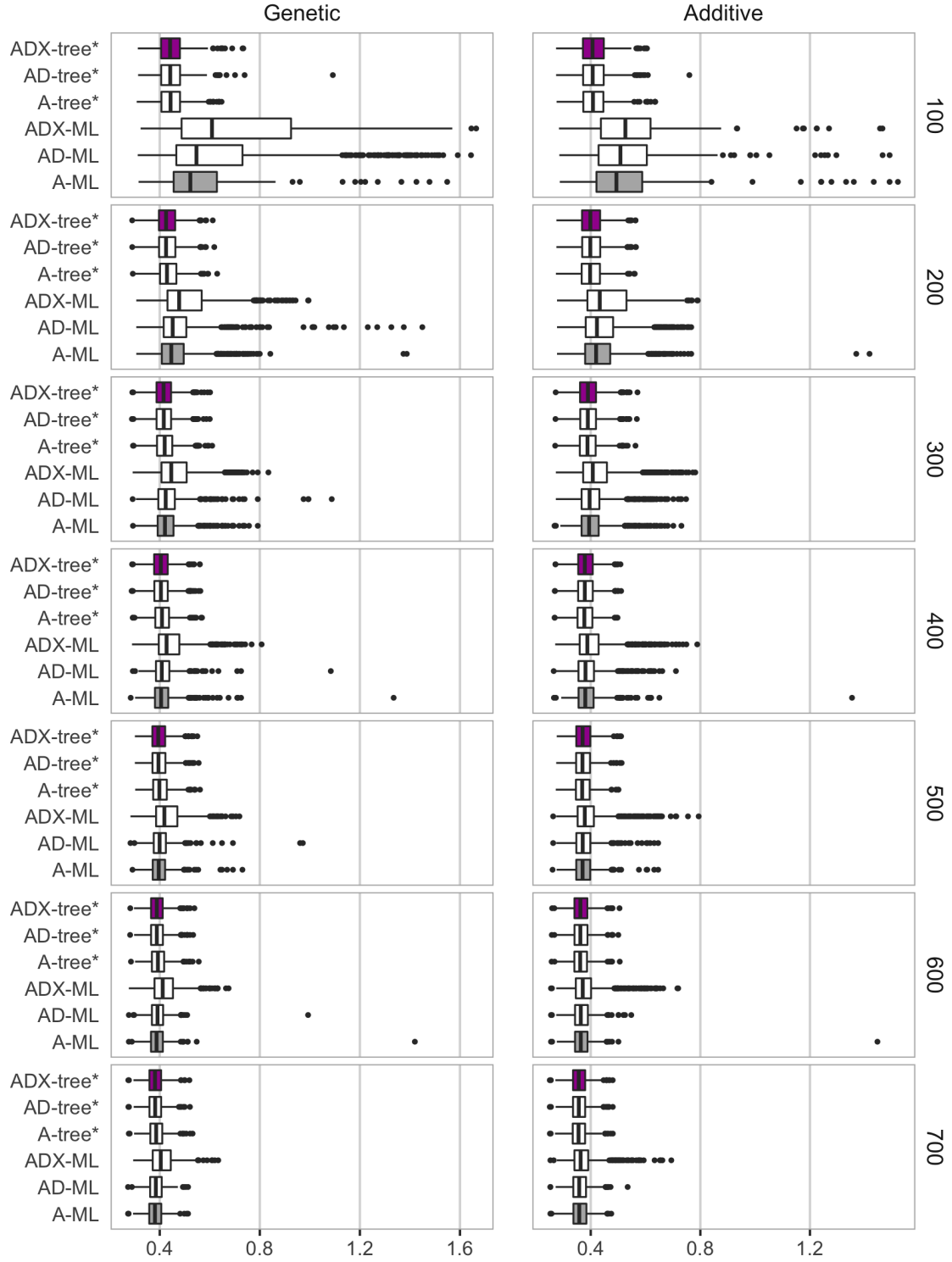


(a) The environmental and additive effect.

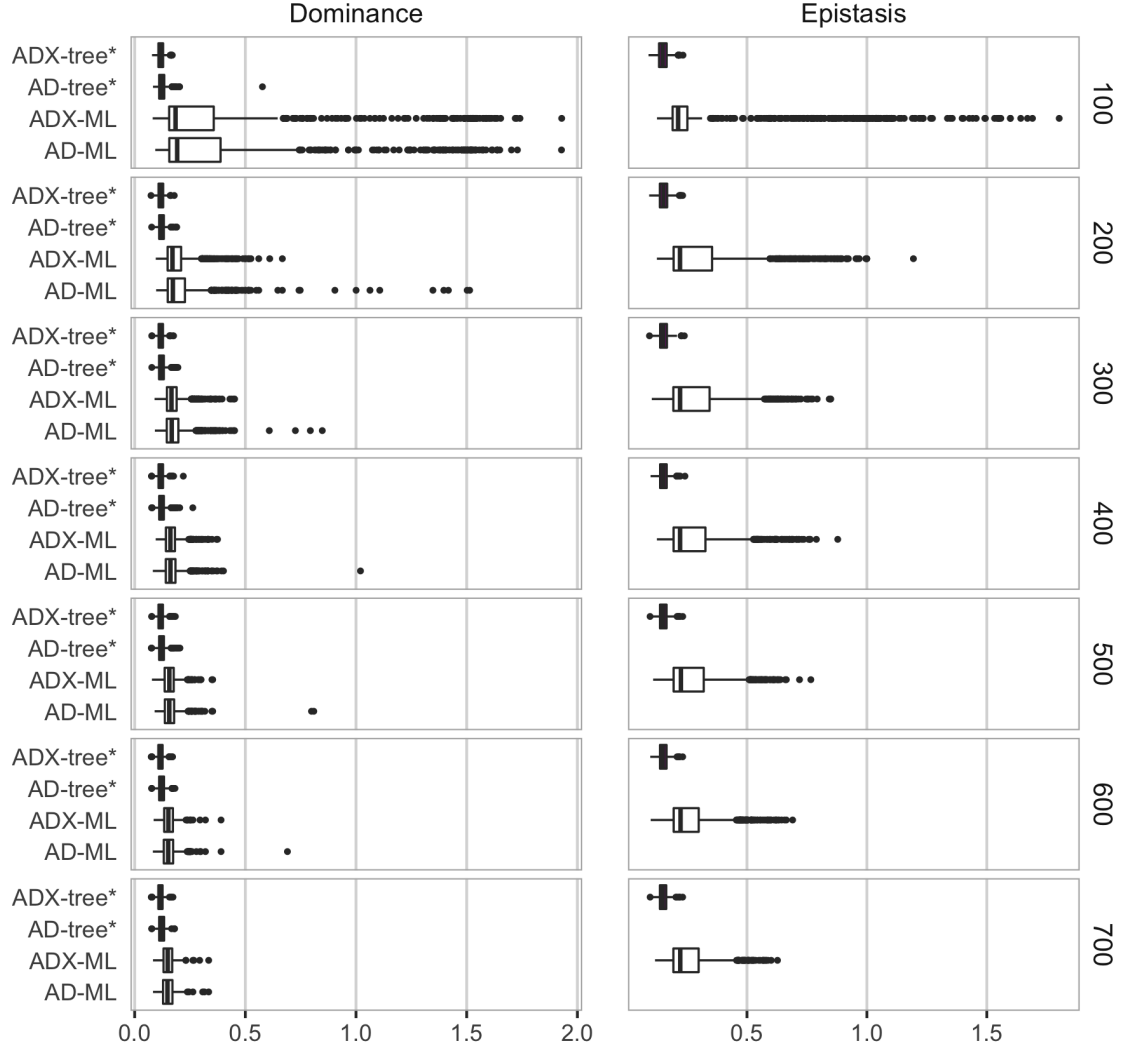


(b) The dominance and epistasis effect.

Figure S15: The ability to estimate (a) environmental and additive effect and (b) dominance and epistasis effect by the model, prior setting and size - correlation (high value is desired, boxplots show variation over replicates).



(a) The environmental and additive variance.



(b) The dominance and epistasis variance.

Figure S16: The ability to estimate (a) environmental and additive effect and (b) dominance and epistasis effect by the model, prior setting and size - continuous rank probability score (CRPS; low value is desired, boxplots show variation over replicates).

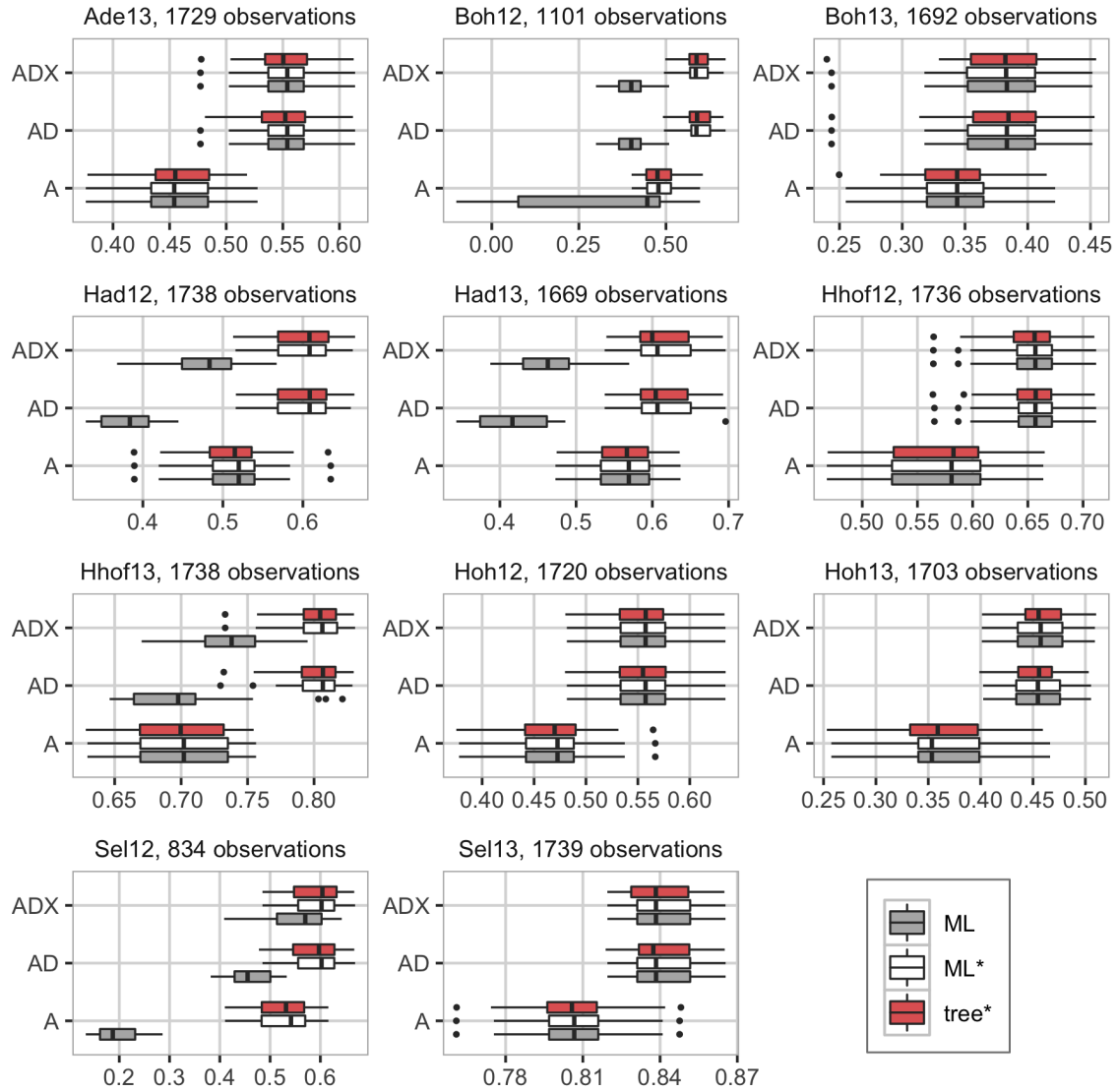


Figure S17: The ability of phenotype prediction in the real case study for all 11 trials, measured using correlation (high value is desired, boxplots show variation over cross-validations and folds). The number of observations in each dataset is indicated for each trial. The total number of parents and hybrids is 1,739.

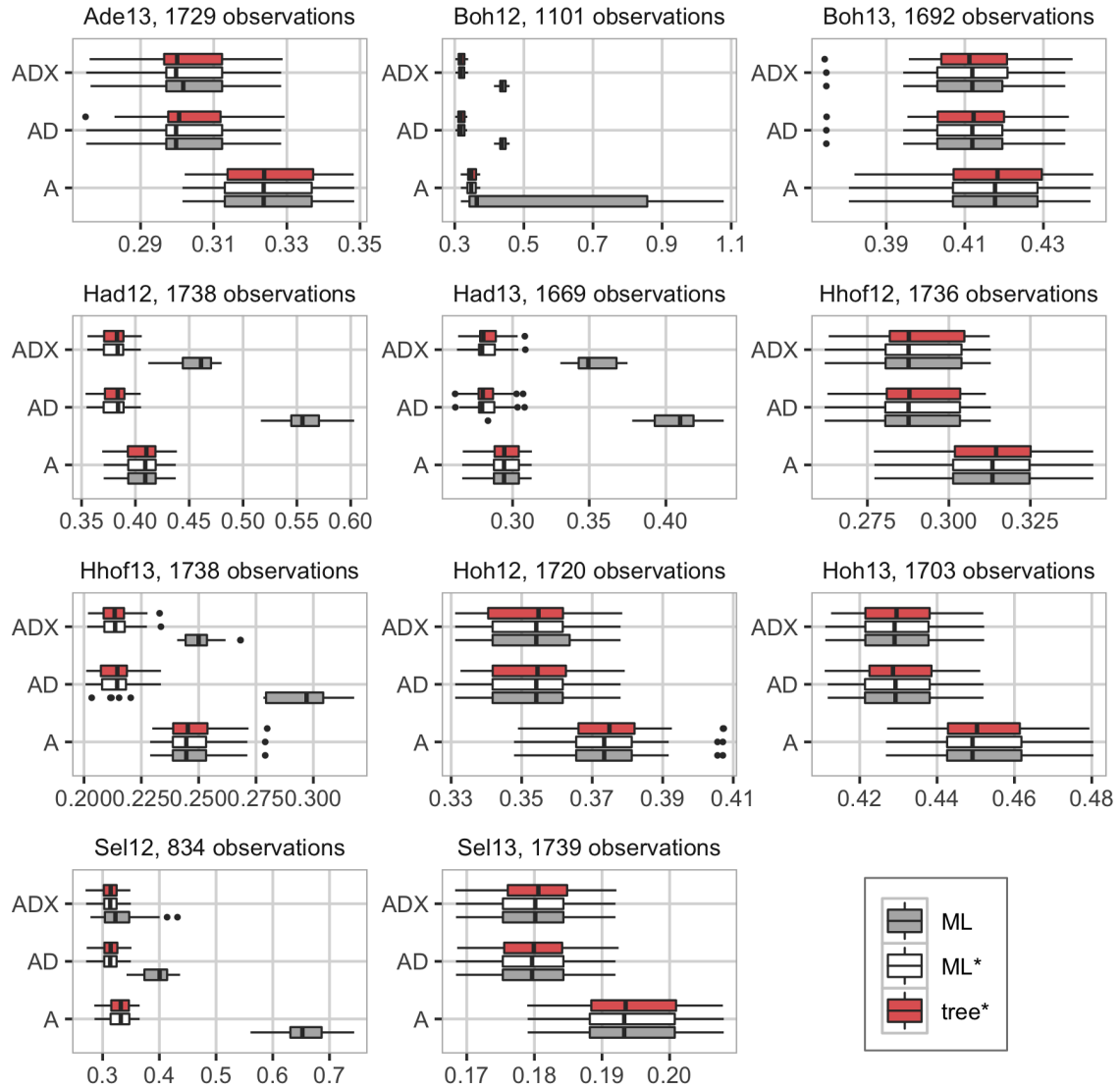
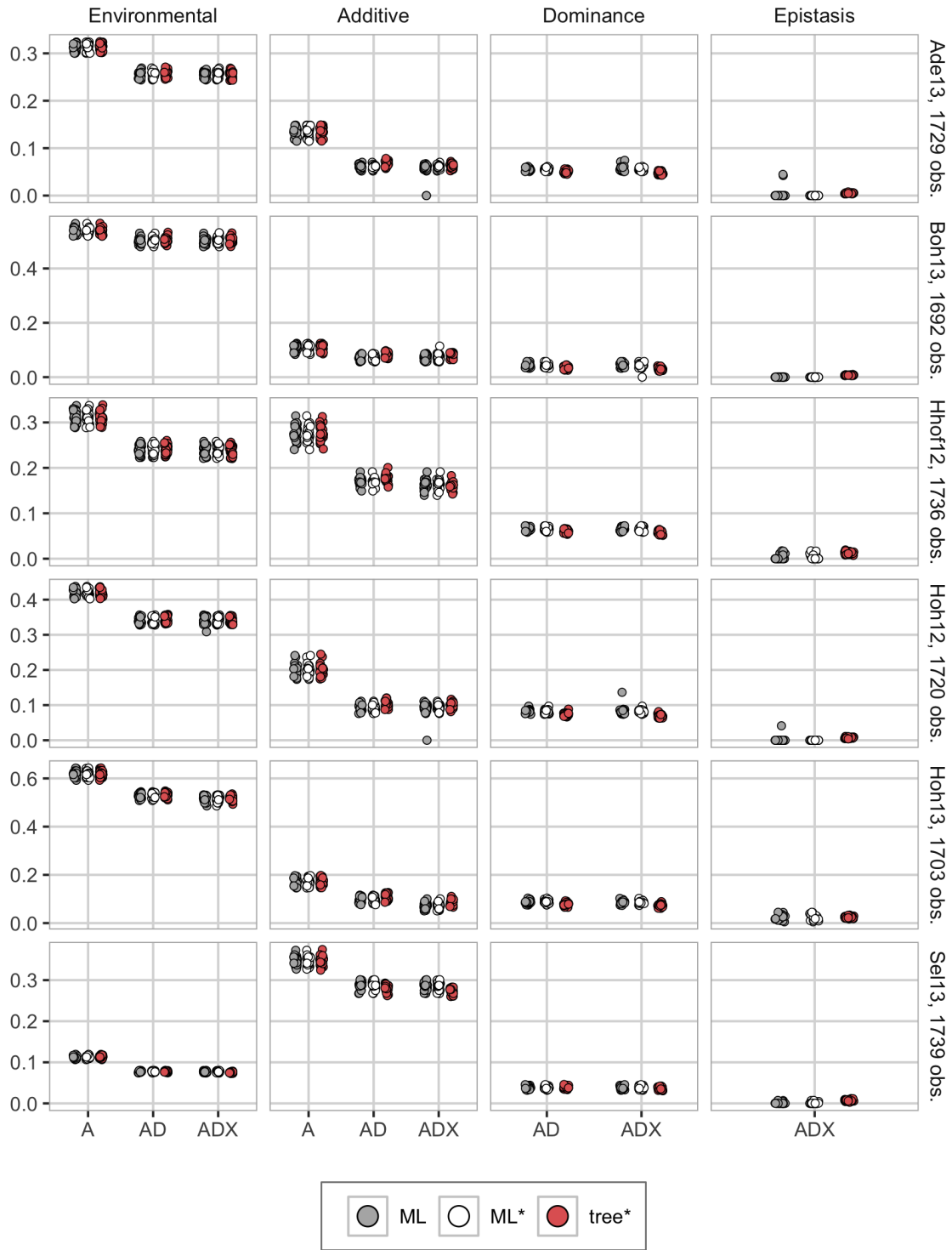


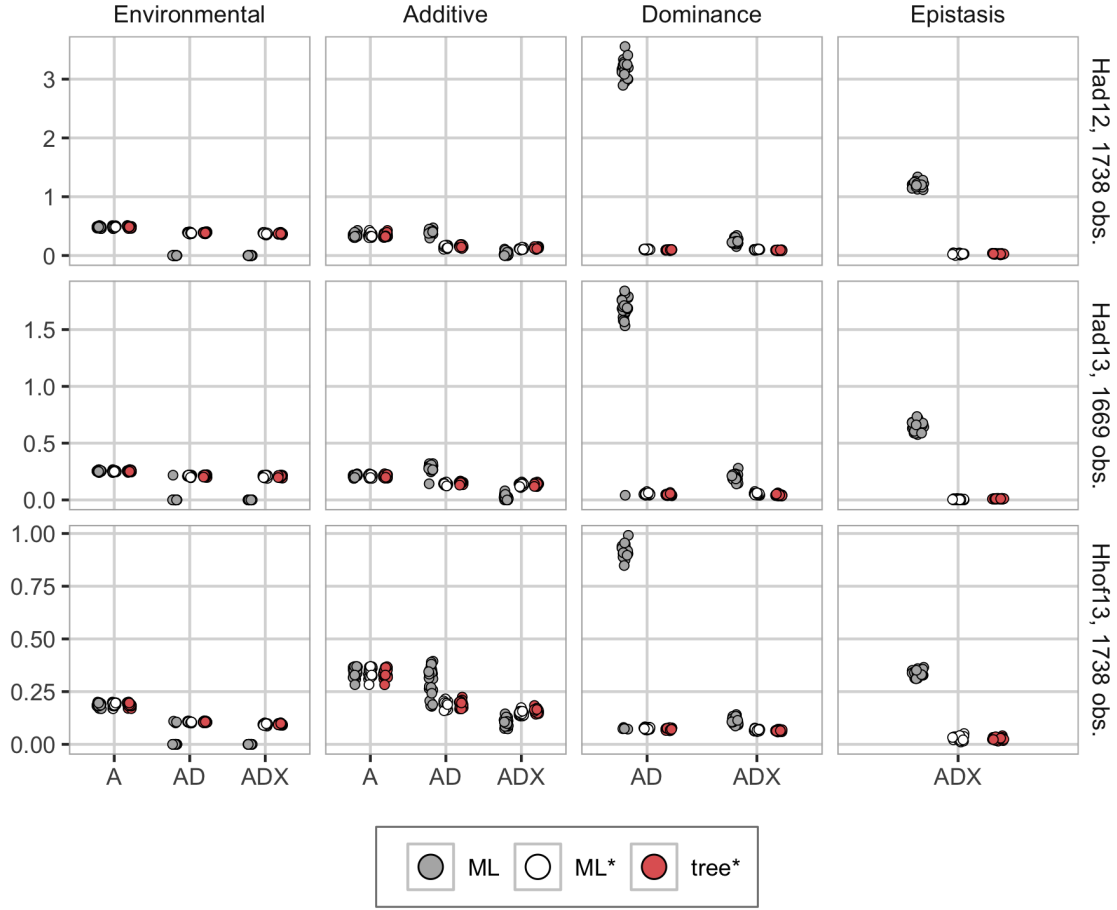
Figure S18: The ability of phenotype prediction in the real case study for all 11 trials, measured using continuous rank probability score (CRPS; low value is desired, boxplots show variation over cross-validation and folds). The number of observations in each dataset is indicated for each trial. The total number of parents and hybrids is 1,739.



(a) Trials where the approaches perform equally good.



(b) Trials where we have a lot of unobserved phenotypes, and ML is diverging. For Boh12 and Sel12, A-ML (additive model fitted with the maximum likelihood approach) is overestimating the additive variance to be so large (estimates over 900 and 400, respectively) that we have truncated the y-axes at 2.5 and 1.5, respectively, to highlight the other results.



(c) Trials where the maximum likelihood approach leads to overfitting.

Figure S19: Posterior median variances from (a) Ade13, Boh13, Hhof12, Hoh12, Hoh13 and Sel13, where both approaches perform equally good, (b) Boh12 and Sel12, where large amounts of the phenotypes are unobserved and the maximum likelihood approach is diverging, and (c) Had12, Had13 and Hhof13, where ML is overfitting. We have included the variances from the five 5-fold cross-validations, giving 25 estimates for each trial and model. We have removed results where the maximum likelihood optimizer did not converge. The y-axes of (b) are truncated to highlight the other results. The total number of parents and hybrids is 1,739.