Estimating heritability in plant breeding programs

Dissertation to obtain the doctoral degree of Agricultural Sciences (Dr. sc. agr.)

Faculty of Agricultural Sciences

University of Hohenheim

Biostatistics unit (340c), Institute of Crop Science (340)

submitted by

Paul Schmidt

from *Rostock, Germany*

2019

Die vorliegende Arbeit wurde am 14.08.2019 von der Fakultät Agrarwissenschaften der Universität Hohenheim als „Dissertation zur Erlangung des Grades eines Doktors der Agrarwissenschaften“ angenommen

Tag der mündlichen Prüfung: 22.11.2019

Leiter der Prüfung: Prof. Dr. Jörn Bennewitz
Berichterstatter 1. Prüfer: Prof. Dr. Hans-Peter Piepho
Mitberichterstatter 2. Prüfer: Prof. Dr. Jens Léon
Mitberichterstatter 3. Prüfer: Prof. Dr. Tobias Würschum

Table of contents

1. General introduction

Plant breeding has been practiced for thousands of years, since near the beginning of human civilization (Allard, 1976, Ch. 3). During this time, its aim has generally not changed: Improve the genetic composition of *genotypes* (*i.e.* cultivars/entries/lines/varieties) so that they fit human needs. What this means may vary depending on, *e.g.*, the crop, target region and time, but is generally associated with high (and stable) yields or other quantifiable traits of interest like quality or disease resistance.

What has obviously been improved and revolutionized countless times, however, are the methods used in plant breeding. In this context, one of the most important concepts of the last century was the *heritability*. Originally, it was defined as an intraclass correlation, *i.e.* a ratio of genotypic variance (σ_g^2) to phenotypic variance (σ_p^2) among individuals. A phenotype is the composite of an organism's observable traits and determined not only by its genotype, but also by environmental factors and the genotype-environment interactions. Thus, heritability expresses the extent to which a phenotype is genetically determined. Its notation as h^2 was introduced by Wright (1920) as the degree of determination by heredity. Today h^2 usually refers to *narrow-sense heritability*, whereas *broad-sense heritability* is denoted as H^2 (e.g. Xu, 2013). The difference between the two is that h^2 considers only the additive component of the genotypic variance, while H^2 also includes, *e.g.*, dominance and epistasis components.

Nyquist and Baker (1991) point out that when heritability was proposed, it was actually done so in the context of animal breeding and thus for individuals. Furthermore, it "was a function of variance components only and was invariant; it did not contain any properties or dimensions of the experiment." In plant breeding, however, it is usually not the individual plant that is considered. Instead, genotypes are evaluated in replicated experiments. As a result, *the phenotype* is an aggregated value (*i.e.* usually some sort of mean) over multiple observations and heritability is referred to as *heritability on an entry-mean basis*. This clearly distinguishes heritability in plant breeding from heritability in animal breeding or human genetics.

There are two main reasons why the concept of heritability became an important concept for plant breeders: On one hand, it is a descriptive measure used to assess the usefulness and precision of

results from cultivar evaluation trials. On the other hand, it can be plugged into *the breeder's equation*:

$$R = H^2 S\,, \tag{1}$$

where R is referred to as the *response to* selection or genetic *gain* and S is the mean phenotypic value of the selected genotypes, expressed as a deviation from the population mean. Thus, (1) allows for predicting R in a breeding program.

As a brief example, imagine a single trial, where a number of genotypes are tested in n_r replicates in a randomized complete block design. The observed data may be modelled as

$$y_{ik} = \mu + g_i + b_k + \varepsilon_{ik}\,, \tag{2}$$

where y_{ik} is the observation of the ith genotype in the kth replicate, μ is the intercept, g_i is the effect for the ith genotype, b_k is the effect for the kth block and ε_{ik} is the plot error effect corresponding to y_{ik}. Here, g_i is taken as random, so that a genotypic variance σ_g^2 and error variance σ_ε^2 can be estimated. To obtain an estimate for the phenotype of the ith genotype, on may obtain $\bar{y}_{i\cdot}$ as the arithmetic mean across its replicates. Furthermore, σ_g^2 and σ_ε^2 can be estimated from mean squares and their respective expected mean squares of a conventional analysis of variance (e.g. Yan, 2014, p. 17), so that the phenotypic variance can be defined as

$$\sigma_p^2 = \sigma_g^2 + \frac{\sigma_\varepsilon^2}{n_r}\,. \tag{3}$$

Finally, broad-sense heritability would be defined as

$$H^2 = \frac{\sigma_g^2}{\sigma_p^2} = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_\varepsilon^2}{n_r}}\,. \tag{4}$$

Nyquist and Baker (1991) point out what can be seen in (4): heritability on an entry-mean basis is "a function of the properties and dimensions of the experiment." Thus, for a plant breeder the concepts of heritability and experimental design are closely linked. Therefore, I use the next sections to give a short overview on experimental design during the time heritability was introduced, as well as the progress since then.

1.1 Plant breeding trials: the early days

When tracing back the history of plant breeding trials, one quickly finds that Ronald Aylmer Fisher is seen as "the undisputed creator of the modern field that statisticians call the design of experiments" (Savage, 1976). There have obviously been field trials before the 1920s, but it was Fisher*, e.g.,* with his 1935 book *The Design of Experiments* that led to specific layout and analysis techniques being widely accepted (Fienberg and Hinkley, 1980). Some of his contributions are the development of the analysis of variance (ANOVA) and concepts of the null hypothesis, blocking and randomization.

One assumption necessary for the traditional ANOVA approach of Fisher is a *balanced/orthogonal* dataset. When single observations are missing, they were traditionally imputed before ANOVA (see e.g. Cochran and Cox, 1992 Ch. 3.71). To obtain balanced data in the context of (2), we must have *complete replicates/blocks* and thus (i) a constant n_r across genotypes and (ii) blocks of size n_g (given genotype is the only factor). Regarding the design of single trials, this allows for either completely randomized designs, randomized complete block designs or Latin square designs. Fisher (1935, p. 109) points out that "in agricultural experimentation [a large number of treatments/genotypes to be tested in blocks] expresses itself very simply in the increased size of the blocks of land, each of which is to contain plots representative of all the different [treatment] combinations to be tested." When genotypes are tested in a series of trials at multiple locations and/or across several years, as is common practice in plant breeding, it becomes a *multi-environment trial* (MET), where a year × location combination is referred to as an *environment*. When a MET is conducted at n_e environments in a target population of environments (TPE), both n_e and n_r need to be constant across the n_g genotypes.

The fact that many of Fisher's ideas are still taught today as the basis of statistics in agriculture demonstrates what a fundamental impact they had and still have. Notice that the trial and analysis as described for (2), which I will from now on refer to as *Fisher's method*, is a representative example for the state of the art at that time. Furthermore, similar experiments are still conducted today which means that this framework has been applied for almost a century now.

"*Unfortunately, as so often happens,*

mathematical extensions, conceived as improvements,

appeared." (Yates, 1990, p. xxviii)

1.2 Plant breeding trials: recent developments

In this section, I will give a very brief overview of some relevant advancements in the methodology of experimental design for and statistical analysis of plant breeding trials that arose since the early 1900s. Yet, before I start, I would like to clarify that Fisher's method is not outdated in the sense that an alternative approach should generally be preferred. Instead, the advancements and extensions allowed for additional possibilities in the statistical analysis, which in turn led to more flexibility in the experimental design, such as, *e.g.,* non-constant n_e across genotypes. If such a more recent experimental design is used, however, *Fisher's method* is consequently no longer a valid choice for the respective analysis.

1.2.1 Experimental design across environments

MET in plant breeding are usually laid out not only across multiple trial locations, but also over multiple years,, *e.g.,* in a three-year-cycle. Yet, since the goal of plant breeding is to select superior genotypes efficiently, some genotypes are eliminated from the candidate set after their first year of testing. Furthermore, a new cycle with a new set of genotypes is introduced each year. Thus, even though data in this scenario is balanced within a single year, it is not so when considering the entire three-year-cycle due to incomplete genotype-year classifications. This is a very common cause for unbalanced data in plant breeding and it applies irrespective of the experimental design at the environment-level. Other, similar scenarios leading to unbalanced data are, *e.g.,* (i) incomplete genotype-location classifications due to varying genotype relevance between certain regions of the TPE or (ii) incomplete genotype-replicate classifications due to varying number of replicates between locations. Finally, the breeder may conduct MET with single replicates at each environment and thus treat the environment as a block (detailed insights in chapter 2). As a rule it can be said that the more environments there are in a MET, the more likely it is that data is unbalanced.

1.2.2 Linear mixed models

The analytical solution to this problem of unbalanced data came with linear mixed models using maximum likelihood (ML) for estimating variance components. Note that it was once again Fisher (1922, 1921) who proposed the likelihood function as well as the method of maximum likelihood. Later, Patterson and Thompson (1971) proposed using restricted maximum likelihood (REML) as an alternative to the ordinary ML, as the former reduces the bias of VC estimates for finite samples. Eisenhart (1947) was the first to use the word "mixed" in the context of linear models including both fixed and random effects. The mixed model equations were given by Henderson (1986) and further work was done by Searle et al. (2006). With mixed models and REML, analyzing unbalanced data was no longer a problem. A practically relevant example for this is that the software PLABSTAT (Plant Breeding Statistical Program), which uses Fisher's method and was and still is popular among plant breeders today, will impute missing values as long as they represent less than 13% of the dataset and subsequently conduct an ANOVA. Note that in an unbalanced dataset, all missing level classifications (such as genotype-year) are treated as missing values. When more than 13% of the data is missing, no ANOVA is conducted in PLABSTAT. Instead, the breeder is advised to switch to software that uses REML in order to obtain VC estimates (Utz, personal communication).

1.2.3 Experimental design on the environment level

The need to always form complete blocks/replicates was especially problematic for plant breeders, as they often have a relatively large number of candidate genotypes. Yates (1936) introduced the use of incomplete blocks and followed up with the potential ability to recover inter-block information (Yates, 1940b), which generally opened the door for new classes of experimental design such as *lattice squares* (Yates, 1940a). The proposition of*, e.g.,* augmented designs (Federer, 1956), α-lattice (Patterson and Williams, 1976), p-rep (Cullis et al., 2006) and augmented p-rep designs (Williams et al., 2011) followed. Moehring et al. (2014) showed that the latter two "clearly outperform replicated and classical augmented designs" in terms of efficiency. Hence, now the goal is to keep the number of replicates for a genotype as low as possible, while maintaining*, e.g.,* a small standard error of a difference (s.e.d.) for the genotype effects.

1.2.4 Geostatistics and variance structures

By discussing the idea of adjusting plot values by covariance on neighboring plot values instead of forming blocks, Papadakis (1937) and Bartlett (1938) laid the starting point for geostatistics/spatial error models, which experienced a comparable, parallel development since then. Some milestones are the nearest neighbor model (Wilkinson et al., 1983) and the extension to two-dimensional spatial trends (Cullis and Gleeson, 1991) even via smoothing methods such as P-splines (Rodríguez-Álvarez et al., 2018). Note that*, e.g.,* Moehring et al. (2014) and Damesa et al. (2018) state that using spatial error structures can be seen as a possibly advantageous add-on option to a randomization-based approach for designing field experiments. Yet Gilmour et al. (1997) treat it the other way around. They recognize three major sources of variation in field experiments: large-scale variation, extraneous variation and small-scale variation. They suggest accounting for the first via polynomials and spline smoothers and the second via spatial models and potentially including design effects.

Considering the analysis of MET, it had been pointed out from the beginning "that plot error variance is variable from one experiment to another […] and there is nothing that compels the variances of the GE interaction effects to be homogeneous" (Comstock and Moll, 1963). In this context, already Yates and Cochran (1938) discussed performing ANOVA with heterogeneous variances. Again, the mixed model framework turned out to be able to adequately model environment-specific variances (Cullis et al., 2006; Edwards and Jannink, 2006; So and Edwards, 2009; Smith and Cullis, 2018). In summary, multiple options for modeling the variance-covariance structure of the residuals at various levels of the MET are available to the plant breeder.

Furthermore, modeling genotype-by-environment interaction effects or more specifically their variance-covariance structure alone is a vast topic including*, e.g.,* Finlay-Wilkinson regression, additive main effects multiplicative interaction (AMMI) models and genotype and genotype-environment (GGE) interaction models (Finlay and Wilkinson, 1963; Gauch, 1992; Piepho, 1998; Smith et al., 2001). In addition, genotypic correlations between cultivation areas, that group together similar environments, can be estimated (Kleinknecht et al., 2013). Again, this can be summarized as the proposition of a number of variance-covariance structures, though this time for genotype-by-environment interaction effects and not the residual.

1.2.5 Genotypic variance structure

Examining (3) it can be seen that a single estimate for σ_g^2 is required, which corresponds to the single genotypic variance estimate we obtain via Fisher's method for (2). The implication is, however, that all genotypes have a homogeneous genotypic variance and are independent. Yet, especially in plant breeding programs, it seems unlikely that genotypes are independent. Instead, plant breeders usually have some population structure within their pool of candidate genotypes. There are approaches to model this, *e.g.,* via separating the genotypic effect into an overall population effect and genotype within population effect, or alternatively by using a factorial or diallel mating design in order to obtain estimates for the general combining ability (GCA) and specific combining ability (SCA) of a genotype (Bernardo, 1994). Another option is to assume a correlation structure for the genotypic effect by obtaining a genetic relationship matrix beforehand, which is usually based on either pedigree information or markers (Meuwissen et al., 2001; Pillen et al., 2003; VanRaden, 2008; Würschum, 2012; Crossa et al., 2017). Especially the latter is very popular nowadays, as many plant breeders perform some type of marker-assisted selection (*i.e.* genome-wide association study (GWAS) or genomic selection). Thus, once again a variance-covariance structure (that is different from ID) may be introduced to the mixed model of a plant breeder – this time for the genotype main effect.

1.2.6 Computers

A development that is indispensable for the popularity of mixed models using REML is the advent of computers, steady advancement regarding their computational power as well as the improvement of the software and applied algorithms. As Piepho et al. (2003) point out: "Prior to the advent of computers, the analysis of a mixed model was a daunting task. In fact, analysis was only feasible for simple, balanced designs. A full-fledged analysis of more complex data sets, *e.g.*, of an unbalanced series of experiments accommodating heterogeneity of variance at various levels (treatment by environment interaction and error) and spatial correlation at the field level, was beyond reach." It may be interesting to note here that "Fisher had little interest in computers, then in their infancy, and referred to their calculations as 'Mecano arithmetic'." (Yates, 1990, p. xxxi)

1.3 Objective of this study

It should have become clear at this point that since the proposition of heritability, multiple tremendous advancements took place in the design and analysis of plant breeding experiments. Today, a plant breeder is left with an extensive number of options to make the experimental design more efficient and data analysis more sophisticated than 100 years ago. Fortunately, the newer methods are indeed applied in practice and I will from now on refer to this combination of advanced experimental design with a mixed model analysis as the *modern plant breeding setting*. In fact, when asked to provide a recent dataset from a plant breeder that can correctly be analyzed via Fisher's method, one may find that this requires some searching.

It seems, however, that during this time of progress the notion of heritability did not receive the same amount of attention and extension. While, *e.g.*, Hanson and Robinson (1963) state that "the need for standardization of the concept of heritability is evident", the general approach towards heritability estimation in context of plant breeding and experimental designs remained unchanged. Considering the standard formula for heritability on an entry-mean basis (H^2_{Std}/h^2_{Std}) as shown in (4), there are mainly three challenges that arose with the extended mixed model framework: (i) unbalanced data, (ii) heterogeneous variances and (iii) covariances. In other words, it is not clear how to estimate heritability for, *e.g.*, a marker-assisted selection in an unbalanced MET, where trials are laid out as p-rep designs.

Yet, for several decades the relevant passages on the estimation of heritability on an entry-mean basis did not seem to go beyond "normally plots are replicated" (Wricke and Weber, 1986, p. 45) and even today H^2_{Std}/h^2_{Std} is taught at universities and sometimes applied in scientific publications, even though an analysis methodology more sophisticated than Fisher's method was used.

Therefore, the overall objective of this thesis is to propose new, generalized methods for estimating heritability appropriate for modern plant breeding programs and compare these to existing proposals. First, Chapter 2 exemplarily demonstrates some of the flexibility and benefit of the mixed model framework for typically unbalanced MET as described above. Here, a bivariate mixed model analyses is used to jointly analyze two MET for cultivar evaluation, which differ in multiple crucial aspects such as plot size, trial design and general purpose. Thus, the primary focus in this chapter lies on the application of linear mixed models in the *modern plant breeding setting*,

which serve as the basis to subsequently estimate heritability. Then in Chapter 3, six alternative methods for the estimation of broad-sense heritability on an entry-mean basis are applied and compared to the standard method. This is done for four different MET datasets for cultivar evaluation, which all display a typically unbalanced data structure, but differ in the genetic frameworks of their cultivars. Finally, a new approach to estimate heritability on an entry-difference basis is proposed in Chapter 4. Besides deriving this method, it is again exemplified and compared to other methods via analyzing four different datasets for cultivar evaluation. Here, however, the datasets differ in their complexity so that the first may correctly be analyzed via Fisher's method, while the last shows multiple aspects described above as the modern plant breeding setting.

2. More, larger, simpler: How comparable are on-farm and on-station trials for cultivar evaluation?

P. Schmidt[a], J. Möhring[a], R. J. Koch[b], and H.-P. Piepho[a]

[a] Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstrasse 23, 70599 Stuttgart, Germany

[b] Pioneer Hi-Bred Northern Europe Sales Division, Apensener St. 198, 21614 Buxtehude, Germany

Abstract

Traditionally, cultivar evaluation trials have been conducted as replicated small-plot, on-station trials at multiple locations and years. To this day, this is the method of choice for cultivar registration trials conducted by official federal institutes. Given a different purpose (e.g. marketing), cultivar evaluation may also be done as on-farm trials with single replicates and fewer plots laid out as large strips. Such trials are often conducted at a larger number of locations. It is not clear how comparable these two trial systems are. Our objective therefore was to compare the precision and accuracy of these two systems using yield data from both on-farm trials and from official on-station trials for winter oilseed rape (Brassica napus L.) across 8 yr. We set up multivariate mixed models to analyze the combined dataset of both trial systems and estimate heterogeneous variance components. Furthermore, based on 23 hybrid genotypes common to both datasets, we investigated the genetic correlation between systems and tested for genotype × system interaction effects. The results suggest that on-farm trials are comparable with on-station trials in terms of precision of a single plot, but that there are genotype × system interaction effects prohibiting the comparison of yield estimates for genotypes between systems. One potential explanation for this difference was identified as the system-specific group effect of semidwarf vs. long-strawed genotypes.

Status: published

Published online April 26, 2018

RESEARCH

More, Larger, Simpler: How Comparable Are On-Farm and On-Station Trials for Cultivar Evaluation?

P. Schmidt, J. Möhring, R. J. Koch, and H.-P. Piepho*

ABSTRACT

Traditionally, cultivar evaluation trials have been conducted as replicated small-plot, on-station trials at multiple locations and years. To this day, this is the method of choice for cultivar registration trials conducted by official federal institutes. Given a different purpose (e.g., marketing), cultivar evaluation may also be done as on-farm trials with single replicates and fewer plots laid out as large strips. Such trials are often conducted at a larger number of locations. It is not clear how comparable these two trial systems are. Our objective therefore was to compare the precision and accuracy of these two systems using yield data from both on-farm trials and from official on-station trials for winter oilseed rape (*Brassica napus* L.) across 8 yr. We set up multivariate mixed models to analyze the combined dataset of both trial systems and estimate heterogeneous variance components. Furthermore, based on 23 hybrid genotypes common to both datasets, we investigated the genetic correlation between systems and tested for genotype × system interaction effects. The results suggest that on-farm trials are comparable with on-station trials in terms of precision of a single plot, but that there are genotype × system interaction effects prohibiting the comparison of yield estimates for genotypes between systems. One potential explanation for this difference was identified as the system-specific group effect of semidwarf vs. long-strawed genotypes.

P. Schmidt, J. Möhring, and H.-P. Piepho, Biostatistics Unit, Institute of Crop Science, Univ. of Hohenheim, Fruwirthstrasse 23, 70599 Stuttgart, Germany; R.J. Koch, Pioneer Hi-Bred Northern Europe Sales Division, Apensener St. 198, 21614 Buxtehude, Germany. Received 14 Sept. 2017. Accepted 6 Mar. 2018. *Corresponding author (piepho@uni-hohenheim.de). Assigned to Associate Editor Lucía Gutiérrez.

Abbreviations: BLUE, best linear unbiased estimator; BLUP, best linear unbiased predictor; BSA, Bundessortenamt (Federal Plant Variety Office); MET, multi-environment trial; OF, on-farm; OS, on-station; RCBD, randomized complete block design; TPE, target population of environments; VC, variance component; VCU, value for cultivation and use.

Once a new crop cultivar has been bred, its value for cultivation and use (VCU) must be evaluated in cultivar evaluation trials. These trials should be designed to be as efficient as possible and yield valid results. In this context, "valid" means that differences between cultivars (genotypes) found in the trials should represent the actual differences in performance in the farmers' fields.

In many cases (e.g., for German official VCU trials), the method of choice can be described as replicated multi-environment, small-plot, on-station (OS) trials. This trial system usually comprises identical sets of genotypes that are tested at multiple locations and/or across several years, making it a multi-environment trial (MET), where a year × location combination is referred to as an environment. Furthermore, they are OS trials, as all field trials are planned and executed by trained personnel following established protocols and using field trial technology. A major reason for using specialized technology is the relatively small plot size of, for example, ~10 m^2 for winter oilseed rape (*Brassica napus* L.) trials (Bundessortenamt, 2000). German VCU trials are usually laid out as randomized complete block designs (RCBD) with three to four replicates (Fig. 1), as α-designs with three replicates, or as split-plot design with fertilizer treatments

Published in Crop Sci. 58:1–11 (2018).
doi: 10.2135/cropsci2017.09.0555

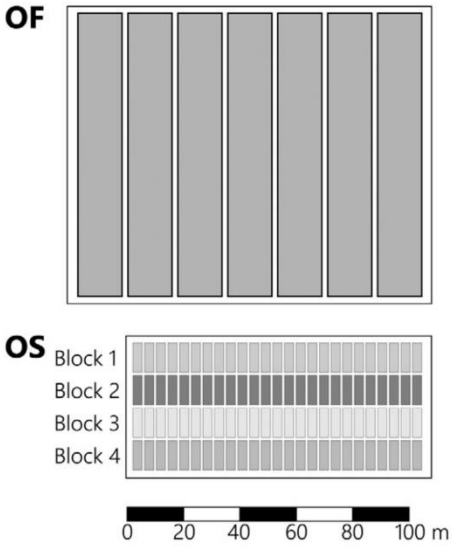

Fig. 1. Schematic layout of an on-farm (OF) trial with seven plots (top) and an on-station (OS) trial with four blocks, each with 25 plots (bottom), in a single environment.

randomized on main plots within complete blocks and genotypes randomized to subplots according to a RCBD (Laidig et al., 2008). Data within years are balanced with respect to genotypes and locations. Every year, the evaluation of a new set of genotypes starts. This set is tested in a 3-yr-cycle, during which some genotypes are discarded each year. These OS trials can be seen as the traditional approach, and they have been widely used and accepted for roughly a century now (Smith et al., 2005).

A more recent development is the use of on-farm (OF) trials (Bradley et al., 1988; Troyer, 1996; Troyer and Wellin, 2009; Yan et al., 2002). For a review of design and analysis of OF trials, see Piepho et al. (2011). Here, many aspects of the experimental design are substantially different from traditional approaches: plots are laid out in strips that can be hundreds of times larger than OS plots, and the number of plots per field site is often confined to <10 (Fig. 1). Therefore, the number of genotypes per environment becomes smaller.

Another difference from OS trials is the reduction to a single replication per environment. Thus, instead of replicating plots per environment, resources in OF trials are shifted towards increasing the plot size and the number of tested environments. In place of field trial technology, the farmer's machinery is used, setting its working width as plot widths for the respective environment. Typically, more genotypes are under investigation than can be grown at single OF field sites. This is handled by distributing genotypes across environments, with only subsets of genotypes tested in any one environment. Thus, the obtained datasets are not only unreplicated within environments, but they are also highly unbalanced with respect to the location × year and genotype × environment classifications.

Given the substantial differences in design between OF and OS trials, it is of interest to compare these two trial systems in terms of precision and accuracy. Yan et al. (2002) summarized it well in their article, which investigated analogous winter wheat (*Triticum aestivum* L.) trial systems in Canada: "Understanding the relationship between the two systems could have a significant impact on cultivar evaluation strategy. If they are highly correlated, either system would be sufficient; if they are complementary, both systems would be helpful; and if they are mutually exclusive, a decision must be made on which system is appropriate for cultivar evaluation."

For this purpose, we here consider German MET data on winter oilseed rape. The Pioneer Accurate Crop Testing System (PACTS) constitutes an OF trial system, which each year comprises ~20 genotypes in ~100 locations with single replicates in Germany. Data from these OF trials were compared with OS trials from official cultivar evaluations conducted by the German Federal Plant Variety Office (Bundessortenamt, Hannover; BSA). These OS trials comprise ~20 locations laid out as RCBD with three to four replicates in Germany each year and at the beginning of a cycle, including ~90 genotypes.

The aim of this article is to evaluate the precision, accuracy, and hence the overall usefulness of OF trials compared with OS trials regarding cultivar evaluation in winter oilseed rape.

MATERIALS AND METHODS

Data

Both METs (OF and OS) were conducted in Germany, and their datasets cover a period of 8 yr (2007–2014).

On-Farm Data

The OF data were obtained in trials overseen and conducted by Du Pont Pioneer in cooperation with local farmers. Their intention was on the one hand to assess the performance of a relatively small number of genotypes and on the other hand to present these genotypes for marketing purposes. All trials were sown with standard drilling technology either by Du Pont Pioneer staff or by the farmers themselves. The sown plot was at least 3 m wider than the harvested area of the plot. Crop husbandry measures such as fertilization and spraying were applied by the farmers according to their local practices. The farmers harvested each strip with their combine separately, whereas the respective yield was assessed and documented by Du Pont Pioneer staff with the company's technology as grain yield corrected for 9% moisture in 10^{-1} t ha^{-1}. This was done via three different weighing methods that all used electronic scales: (i) bagging scales with a weighing bag attached to an aluminum frame on top of a trailer, (ii) weighing plates that the trailer stands on, and (iii) trailers with an integrated weighing system. Within each environment, only a single weighing system and a single combine were used. The dataset comprises

6680 plot records, arising from a median of 100.5 environments per year and 801 environments in total (Table 1). There were 4 to 18 (median = 8) plots per environment, and genotypes were never replicated within an environment. Within each year up until 2012, one out of four possible genotype sets was tested at each environment. Although all four sets included a core set of seven genotypes with greatest commercial relevance in the respective year, three sets also comprised two to five additional and generally new genotypes. Starting in 2013, a fifth genotype set was added, exclusively containing genotypes with a herbicide tolerance. Accordingly, this set did not share genotypes with the other sets and was tested at environments treated with the corresponding herbicide. Finally, and for each environment individually, it was common practice to add one or two genotypes that were of interest to the respective farmer. These are subsequently referred to as external genotypes. Tall and semidwarf genotypes were present in every environment, and randomization plans kept them separate to improve comparisons within growth types. During the first 2 yr, no randomization within those groups was applied. For all the following years, one out of four randomization plans was used to lay out each trial. The median plot size was 710 m^2, with a minimum of 27 m^2 and a maximum of 3360 m^2. On a yearly basis, new genotypes were included, while others were excluded due to selection so that the number of years a genotype was tested ranged from 1 to 7 yr (median = 2). In total, the dataset comprises 110 genotypes (37 Pioneer, including 13 semidwarf; and 73 anonymized external genotypes). The number of environments in which a genotype was tested ranged from 1 to 618 (median = 76) for Pioneer and 1 to 402 (median = 3) for external genotypes, leading to 87.5% of all observations coming from Pioneer genotypes. On average, ~31 genotypes were tested each year (19 when excluding genotypes with only a single observation).

On-Station Data

The OS data were obtained in official trials run on behalf of and supervised by the BSA. Their results serve as the basis for variety registration decisions. Sowing and harvest were done using standard field trial technology. Trial plots were separated from neighboring plots by borders of the same genotype either via the plot-in-plot or the double-plot approach. A procedure where plants are physically separated (i.e., *Scheiteln*) was implemented ~2 wk before harvest. *Scheiteln* is necessary, because rapeseed plants of adjacent plots intertwine towards the end of a season. Concerning crop husbandry, it was attempted to recreate the farmers' local practices with the exception of fungicides, which generally were not applied. Semidwarf and tall genotypes were always grouped together. Whenever a semidwarf plot was adjacent to a tall genotype plot, an extra semidwarf plot was planted in between. The latter was not harvested but only served as a buffer. Grain yield was harvested and weighed according to BSA protocols (Bundessortenamt, 2014).

The OS data comprises 39,246 plot records coming from a median of 20.5 environments per year and 171 environments in total (Table 1). The BSA implements a procedure where genotypes are tested and selected in a 3-yr evaluation cycle. Since new genotypes are introduced on a yearly basis, stages of different evaluation cycles overlap and all stages (i.e., first year, second year, and third year) are present each year. Although genotype sets from different evaluation stages may be tested at the same environments, they are tested separately in adjacent trials. All trials were laid out as RCBD with either three or four replicates. The number of tested genotypes per environment ranges from 25 to 150 with a median of 70. In total, the OS dataset contains information on 727 genotypes with a median of 149.5 genotypes tested each year. The number of environments in which a genotype was tested ranged from 6 to 171 (median = 11).

Table 1. Summary of multi-environment trial data for on-farm and on-station datasets. Numbers in parentheses represent medians.

Parameter	On-farm data	On-station data
Years	2007–2014	2007–2014
Data points	6,680	39,246
Total no. of genotypes	110	727
Median plot size (m^2)	710	10
Environments per year	76–121 (100.5)	17–29 (20.5)
Replicates per environment	1	3–4 (3)
Genotypes per environment	4–18 (8)	25–150 (70)
Environments per genotype	1–618 (4)	6–171 (11)

Dataset Comparability

In summary, we have data of winter oilseed rape genotypes obtained within two different cultivar evaluation systems. Grain yield was used as the response, as it is the most important trait. It is important to note that the two datasets do not share a single location, yet both are the result of a cultivar evaluation in recent cultivation periods in the growing region of Germany. Therefore, the target population of environments (TPE) of both systems is comparable. Out of the total of 814 genotypes (801 tall, 13 semidwarf), we found that both datasets share a set of 23 (15 tall, 8 semidwarf) Pioneer genotypes (Fig. 2). External genotypes were anonymized in the OF data. Hence, any further shared genotypes among the external ones could not be identified, and the 23 genotypes will be referred to as the "identified shared" set. All Pioneer genotypes, but not all other genotypes, are hybrids.

Analysis

The OF and OS data were combined into a single dataset and then analyzed in a two-stage analysis (Möhring and Piepho, 2009; Piepho et al., 2012, 2016; Schulz-Streeck et al., 2013). In the first stage, OS data of each environment were analyzed separately with a linear model. Because the OF trials are unreplicated, their environments cannot be analyzed individually, and it is not possible to apply an analogous first-stage analysis. To resolve this problem, an alternative approach (described below) that allows for a joint analysis in the second stage was chosen. In the second stage, two different bivariate mixed models were fitted to analyze data across environments (Möhring and Piepho, 2009; Piepho et al., 2012, 2016; Schulz-Streeck et al., 2013). All analyses were done with ASREML-R (Gilmour et al., 2009) and SAS software 9.4 (SAS Institute, 2013).

First Stage and Weighting

In the first stage of our two-stage analysis, data per environment were analyzed by modeling effects that account for the respective design. As a result, adjusted means (i.e., best linear unbiased

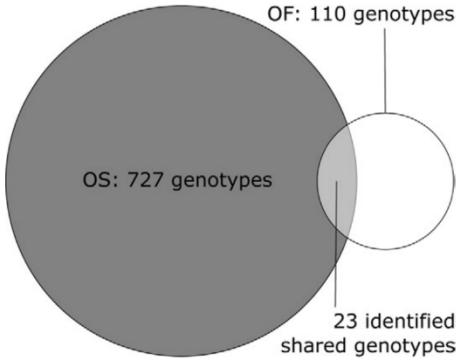

Fig. 2. Venn diagram of genotypes exclusive to and shared between the on-farm (OF) and on-station (OS) dataset.

estimators [BLUEs]) for the genotypes were obtained and could be used as the response in the second stage to perform an analysis across environments. The analysis for each year × location combination was performed for the OS data using the model

$$\mathbf{y} = G + T/R \qquad [1]$$

where $\mathbf{y}$ is the vector of observed plot yields, G represents the genotypes, T represents the trials, and R represents the replicates. The plot error associated with observation $\mathbf{y}$ is not listed in the model for simplicity but is part of the fitted model and was fitted as heterogeneous between trials. Here, we use the notation described in Piepho et al. (2003), where the dot operator (·) defines crossed effects ($A \cdot B$), the crossing operator (×) defines a full factorial model ($A \times B = A + B + A \cdot B$), and the nesting operator (/) indicates that a factor B is nested within another factor A ($A/B = A + A \cdot B$). The colon (:) is used to separate fixed (first) from random effects (last). Our first-stage model (Eq. [1]) takes all factors (except the plot error) as fixed. G was taken as fixed to obtain a vector of adjusted genotype means ($\overline{\mathbf{y}}_1$), which are submitted to the second stage. T and R were also taken as fixed. When submitting estimates from the first to the second stage, a weighting method was used to allow for variances of and covariances between adjusted genotype means, which can slightly improve precision (Möhring and Piepho, 2009). In this study, we used Smith's weights (Smith et al., 2001, 2005; Damesa et al., 2017), which can be obtained as the diagonal elements of the inverse of the variance–covariance matrix of the adjusted genotype means from the first stage.

For the OF data, each data point for a genotype in an environment is simply taken as the genotype's adjusted mean (i.e., $\mathbf{y} = \overline{\mathbf{y}}_1$). The effect for the main plot in the OF data caused by separating tall and semidwarf genotypes was fitted in the second stage.

Second Stage

One may set up a full model for analyzing MET data across environments in a single system and then extend this to cover both systems using a bivariate approach (Piepho and Möhring, 2011), allowing for heterogeneity of variance between systems for the same type of effect and for correlations between systems for any effect observed for both systems (Przystalski et al., 2008). Below, we will develop different models, starting from the case of a single system and then extending this to two systems. The general approach taken for this extension can be described in three steps:

1. Cross each effect of the single-system model (including the intercept) with a factor S for system, which in our case has two levels (OF and OS).
2. Assume heterogeneity of variance between systems for all effects.
3. For any effect observed for more than one system, allow for a covariance between systems.

In our case, a covariance is needed only when the effect does not involve locations because the systems do not share a single location. This is because locations are nested within systems and any effect crossed with location can occur in only one system. The factor S will be taken as fixed throughout because our interest is in comparing systems.

Analysis Ignoring Growth Types. The standard model for analyzing MET data with the addition of a random main-plot effect across environments in a single system is

$$\overline{\mathbf{y}}_1 = \mu + Y \times L \times G + Y \cdot L \cdot M \qquad [2]$$

where $\overline{\mathbf{y}}_1$ is the vector of adjusted genotype × environment means computed in the first stage, μ is the overall intercept, Y represents the years, L represents the locations, G represents the genotypes, and M represents the main plot (Table 2). Note that $Y \cdot L \cdot M$ is only fitted for the OF trials, where semidwarf and tall genotypes are allocated to two different main plots per trial. In this model, all effects (except μ) are taken as independent and identically distributed random variables with constant variance components (VCs).

As the OF trials involve no replication, we cannot perform a first-stage analysis, meaning the observed plot data is transferred as is from the first to the second stage, formally treating them as the adjusted means obtained from OS trials. We therefore denote these data as pseudoadjusted means. Moreover, we cannot dissect plot errors from the highest order interaction effect with OF data. To deal with this problem at the second stage of analysis, we fix the error variance of the OF data to a tiny value and estimate only the VC for the highest order interaction. This is accomplished by giving all the pseudoadjusted means $\overline{\mathbf{y}}_1$ from the first stage a very large weight (i.e., 100,000). The resulting variance estimate will confound the error variance and the actual interaction variance (see Discussion). Also, the analysis effectively assigns the same weight to each OF trial because a trial-specific error variance cannot be estimated.

Applying Steps 1 to 3 to Eq. [2], we obtain the corresponding model for multiple systems:

$$\overline{\mathbf{y}}_1 = S + S \cdot (Y \times L \times G + Y \cdot L \cdot M) \qquad [3]$$

where S represents the systems and all other terms are defined as for Eq. [2]. For all random effects, we assume heterogeneity of variance between systems. Additionally, a covariance between systems is allowed for $S \cdot Y$, $S \cdot G$, and $S \cdot Y \cdot G$ (Table 2).

Note that by estimating a covariance for, for example, $S \cdot G$ effects, we are assuming a correlation between $S \cdot G$ effects for the same genotype in the two different systems. The variance–covariance matrix of $S \cdot G$ for a single genotype is

Table 2. Full display of all models used in the second stage of the analysis in the notation described in Piepho et al. (2003).

Model	Full model†
Eq. [2]	$\bar{\mathbf{y}}_1 = \mu + Y \times L \times G + Y \cdot L \cdot M$ $= \mu : Y + L + Y \cdot L + G + Y \cdot G + L \cdot G + Y \cdot L \cdot G + Y \cdot L \cdot M$
Eq. [3]	$\bar{\mathbf{y}}_1 = S + S \cdot (Y \times L \times G + Y \cdot L \cdot M)$ $= S : \underline{S \cdot Y} + S \cdot L + S \cdot Y \cdot L + \underline{S \cdot G} + \underline{S \cdot Y \cdot G} + S \cdot L \cdot G + S \cdot Y \cdot L \cdot G + S \cdot Y \cdot L \cdot M$
Eq. [4]	$\bar{\mathbf{y}}_1 = \mu + Y \times L \times (D/G)$ $= \mu + D : Y + L + Y \cdot L + Y \cdot D + L \cdot D + Y \cdot L \cdot D + D \cdot G + Y \cdot D \cdot G + L \cdot D \cdot G + Y \cdot L \cdot D \cdot G$
Eq. [5]	$\bar{\mathbf{y}}_1 = S + S \cdot [Y \times L \times (D/G)]$ $= S \cdot (\mu + D + Y + L + Y \cdot D + L \cdot D + Y \cdot L \cdot D + D \cdot G + Y \cdot D \cdot G + L \cdot D \cdot G + Y \cdot L \cdot D \cdot G)$ $= S + S \cdot D : \underline{S \cdot Y} + S \cdot L + S \cdot Y \cdot L + \underline{S \cdot Y \cdot D} + S \cdot L \cdot D + S \cdot Y \cdot L \cdot D + \underline{S \cdot D \cdot G} + \underline{S \cdot Y \cdot D \cdot G} + S \cdot L \cdot D \cdot G + S \cdot Y \cdot L \cdot D \cdot G$

† In models, the dot operator (·) defines crossed effects ($A \cdot B$), the crossing operator (×) defines a full factorial model ($A \times B = A + B + A \cdot B$), and the nesting operator (/) indicates that a factor B is nested within another factor A ($A/B = A + A \cdot B$). The colon (:) is used to separate fixed (first) from random effects (last). $\bar{\mathbf{y}}_1$ is the vector of adjusted genotype × environment means computed in the first stage of the analysis, μ is the overall intercept, Y represents the years, L represents the locations, G represents the genotypes, M represents the main plot, S represents the systems and D represents the growth type. For all random effects crossed with the system factor, system-specific variances were fitted. Effects for which an additional covariance between systems is allowed are underscored.

$$\begin{pmatrix} \sigma^2_{S\cdot G-\mathrm{OF}} & \sigma_{S\cdot G-\mathrm{cov}} \\ \sigma_{S\cdot G-\mathrm{cov}} & \sigma^2_{S\cdot G-\mathrm{OS}} \end{pmatrix}$$

Accordingly, we can compute the correlation between the two systems as $\rho = \sigma_{\mathrm{cov}} \Big/ \sqrt{\sigma^2_{\mathrm{OS}} \cdot \sigma^2_{\mathrm{OF}}}$. In the extreme, this correlation could be unity if the effects were identical in the two systems, apart from scaling differences reflected by heterogeneity of variance. The same reasoning applies for the corresponding two correlated effects of the other two terms $S \cdot Y$ and $S \cdot Y \cdot G$. Since the correlation is a standardized measure that can be interpreted more intuitively, we always present the correlation coefficients instead of their corresponding covariance estimates. Note that by applying Eq. [3], we obtain estimates for the same variances as if we had analyzed the datasets separately with the standard Eq. [2]. In addition, we obtain three covariance–correlation estimates.

Analysis Accounting for Growth Types. After investigating the two systems' accuracies in the first analysis, we found a systematic difference between the estimates per system for the shared genotypes identified (see Results). This corroborated a previous conjecture of a discrepancy between the two systems regarding the relative performance of different growth types. We therefore conducted a second analysis where we considered a grouping of genotypes (G) into two categories, tall and semidwarf, represented by a fixed factor for growth type (D). For a single system, Eq. [2] extends to

$$\bar{\mathbf{y}}_1 = \mu + Y \times L \times (D/G) \qquad [4]$$

where all terms except D are defined as in Eq. [2] (Table 2). Note that since a main plot in the OF data groups genotypes of the same growth type, $Y \cdot L \cdot M$ is completely confounded with $Y \cdot L \cdot D$, which is why we only use the latter in this analysis. Applying Steps 1 to 3 to Eq. [4], we obtain the corresponding model for multiple systems:

$$\bar{\mathbf{y}}_1 = S + S \cdot \left[Y \times L \times (D/G) \right] \qquad [5]$$

where all terms are defined as in Eq. [2–4] (Table 2).

We applied Eq. [3] and [5] to estimate VCs and their SE, best linear unbiased predictors (BLUPs), and adjusted means (BLUEs), as well as to test fixed effects via Wald χ^2 tests. For visual comparisons of the two systems, we plotted BLUPs for $S \cdot G$ from Eq. [3] and for $S \cdot D \cdot G$ from Eq. [5]. Based on $S \cdot D$ and $S \cdot D \cdot G$, which are present in Eq. [5] only, we were able to add growth-type-specific lines: for a given growth type (semidwarf or tall), we fitted the slope for a line in the plot based on Eq. [5] as the first eigenvector obtained from a spectral decomposition of the variance–covariance matrix of $S \cdot D \cdot G$ (Jackson and Dunlevy, 1988). Finally, we computed adjusted means for $S \cdot D$. Thus, for a given growth type, we obtained two means, one for each system. These two means define a point through which the line was fitted in the plot.

Evaluation Criteria

Precision

To allow for a meaningful comparison of precision, VCs and their SE were estimated via Eq. [3] for each system. The main focus of our analysis was then on the size of the plot error variances. As stated above, the error variances in both second-stage models of this article were fixed to estimates from the first stage. Estimates for the overall plot error VC were obtained as follows: (i) for the OS data, the mean of the plot error variances across all environments was estimated by assuming a γ distribution and applying the generalized linear model

$$\log\left[E\left(\sigma^2_{\mathrm{plot}_i}\right)\right] = \lambda \qquad [6]$$

where $\sigma^2_{\mathrm{plot}_i}$ is the plot error variance for the ith environment, E denotes the marginal expectation, and λ is an intercept. The log function was chosen as the link function to link the expected value of the γ-distributed plot error variance estimates to the linear predictor (Cullis et al., 1996; Frensham et al., 1998). Via Eq. [6], we obtained an estimate for the mean plot error variance across environments ($\overline{\sigma}^2_{\mathrm{plot}}$), as well as its SE. (ii) As explained before, for the OF data, no first-stage analysis was conducted and no plot error variances per environment could be estimated. Hence, in the second-stage analysis, we obtained only a single VC that confounds the variances of error and the highest interaction term. We present this estimate as an upper bound to the error variance, denoting it as error variance for simplicity, and display no variance for the highest interaction effect. Additionally, a main plot effect is fitted for the OF data, whose VC represents a second plot error variance. For the overall comparison of error variances, we sum up the confounded OF plot error variance with the OF main plot error variance. This sum then serves as an upper limit of the true

OF error variance and is contrasted with the mean plot error variance obtained for the OS data.

Accuracy

As a measure to compare the accuracies of both systems, we first estimated the genetic correlation ρ_G (i.e., the pairwise correlation of the same genotype in the two systems). It was estimated in three different ways: (i) fitting Eq. [3] with $S \cdot G$ set as random to the full dataset to estimate ρ_{G1}; (ii) fitting Eq. [5] with $S \cdot D \cdot G$ set as random to the full dataset to estimate ρ_{G2}; and (iii) fitting Eq. [3], assuming $S \cdot G$ as random, to a reduced dataset where all data points coming from semidwarf genotypes were excluded to estimate ρ_{G3}. Note that the main plot effects cannot be estimated for this reduced dataset.

Second, to test for $S \cdot G$ and $S \cdot D$ interactions, we slightly modified Eq. [3] by taking S, G, and $S \cdot G$ as fixed and Eq. [5] by taking S, D, and $S \cdot D$ as fixed. Due to the computational burden, in this analysis, it was not possible to use the Kenward–Roger approximation (Kenward and Roger, 1997); instead, the number of denominator df was set to infinity, and thus a Wald χ^2 test was performed (Rao, 1973; Butler et al., 2009, p. 91).

RESULTS

Precision

Estimates for all VCs (and their correlations) are presented in Table 3 and Supplemental Fig. S1. In general, estimates are similar for both systems: the VC for location × year interactions dominates those of years and locations. Likewise, the sizes of the error VC estimates tend to be in between the relatively large environmental VC (Y, L, and $Y \cdot L$) and those of the rather small VC for genotype main and interaction effects (G, $G \cdot Y$, $G \cdot L$, and $G \cdot Y \cdot L$). This overall relationship between VC sizes corroborates findings by Laidig et al. (2008), who estimated VCs with a comparable model for winter oilseed rape BSA data in the period of 1991 to 2006.

Examining the differences, it can be seen that the OS/OF ratio of VCs for Y, L, and $Y \cdot L$ are around or below one, whereas those for G, $G \cdot Y$, and $G \cdot L$ are all >1.3 (Table 3).

Table 3. Variance component (VC) estimates with their SE for grain yield, VC ratios and correlations between VC of winter oilseed rape from German on-station and on-farm trial data in the period 2007–2014 obtained via model (3).

Factor†	VC estimate ± SE		On-station/ on-farm ratio	Correlation
	On-farm	On-station		
	10^{-2} t^2 ha^{-2}			
Y	11.1 ± 6.2	12.0 ± 7.3	1.09	0.51 ± 0.31
L	16.1 ± 2.7	8.2 ± 3.5	0.51	
$Y \cdot L$	32.4 ± 2.1	24.9 ± 3.6	0.77	
G	2.0 ± 0.5	3.1 ± 0.2	1.58	0.66 ± 0.19
$Y \cdot G$	0.7 ± 0.2	1.0 ± 0.1	1.37	0.60 ± 0.27
$L \cdot G$	0.3 ± 0.1	0.8 ± 0.2	2.59	
$Y \cdot L \cdot G$	–‡	3.3 ± 0.2		
$Y \cdot L \cdot M$	3.6 ± 0.2			
Error	5.4 ± 0.1‡	6.7 ± 0.2		

† Y = year, L = location, G = genotype, and M = main plot.

‡ Highest interaction term and error variance are confounded and presented as error variance.

Accuracy

Genetic Correlation

When applying Eq. [3] to the full dataset, ρ_{G1} was estimated to be a moderate 0.66 with a SE of 0.19 (Table 4). In contrast, the estimate for ρ_{G3} is very high (0.97), and the estimate for ρ_{G2} is even approaching unity.

Test of Interaction Effects

As can be seen in Table 5, for both $S \cdot G$ and $S \cdot D$, interaction effects were found to be significant.

DISCUSSION

Precision

Why We Chose the Plot Error Variance as an Evaluation Criterion for Precision

When deciding on a measure to compare the precision of OF and OS trials, several choices are available. One option is half the variance of a difference between two genotype means. For balanced data, this is given by (Talbot, 1984):

$$V_{gm} = \frac{\sigma^2_{YG}}{n_Y} + \frac{\sigma^2_{LG}}{n_L} + \frac{\sigma^2_{YLG}}{n_Y n_L} + \frac{\sigma^2_{plot}}{n_Y n_L n_R} \quad [7]$$

where V_{gm} is the variance of a genotype mean, n_Y, n_L, and n_R are the numbers of years, locations and replicates, respectively, σ^2_{GY}, σ^2_{GL}, and σ^2_{GYL} are VCs for $Y \cdot G$, $L \cdot G$, and $Y \cdot L \cdot G$, respectively, and σ^2_{plot} is the plot error variance. This formula holds true only for balanced data, but it is useful also when VCs were estimated from unbalanced data, because it shows the impact of design variables, such as the number of environments. Thus, comparing V_{gm} estimates from different METs would only be fair if the same amount of temporal and financial resources was available to conduct the trials for the same set of genotypes in the same cultivation area. In other words, irrespective of the VC, it must be clear that n_Y, n_L, and n_R, which are essentially limited by available resources, have a crucial impact on V_{gm}, which is why one would have to include resource information to put V_{gm} from different trials into perspective.

Moreover, it can be shown that using heritability as an indicator for the comparison is limited in a similar manner. This becomes clear when using the ad hoc measure of heritability (Holland et al., 2003; Piepho and Möhring, 2007), given as

$$\bar{H}^2_{Piepho} = \frac{\sigma^2_G}{\sigma^2_G + 0.5\bar{v}_{gd}} \quad [8]$$

where σ^2_G is the VC for the genotype main effect and $\bar{v}_{gd}$ is the mean variance of a difference of two adjusted genotype means. This latter quantity, and hence the heritability in Eq.

Table 4. Genetic correlation (ρ_G) estimates from Eq. [3] and [5] using the full and a reduced dataset.

Parameter	Genotypes in dataset	Model	Effect†	Genetic correlation	
				Estimate	SE
ρ_{G1}	All	Eq. [3]	$S \cdot G$	0.66	0.19
ρ_{G2}	All	Eq. [5]	$S \cdot D \cdot G$	1.00‡	–‡
ρ_{G3}	Semidwarf excluded	Eq. [3]	$S \cdot G$	0.97	0.06

† *S*, system; *G*, genotype; *D*, growth type.

‡ Fixed to 1 in the sense that the non-negativity constraints on variance components led to one of the heterogeneous variances of $S \cdot D \cdot G$ to be fixed to zero, which resulted in the correlation being one, given the nonzero covariance.

[8], can also be computed for unbalanced data. Note that $\overline{H}^2_{\text{Piepho}}$ coincides with the standard broad-sense heritability H^2 in the case of balanced data. Moreover, for balanced data, $0.5\overline{v}_{gd}$ coincides with V_{gm} in Eq. [5].

Furthermore, the VCs themselves may not allow for a meaningful comparison either. This is because purely environmental VCs are assumed to be similar in both analyses, since they were estimated in the same TPE (i.e., recent cultivation periods in Germany). Moreover, they are irrelevant for genotype mean comparisons, which only depend on VCs comprising the genotypic factor. These latter VCs, on the other hand, do depend on the set of genotypes tested, which cannot necessarily be assumed to be identical for the two datasets. The OF data used here mainly contain genotypes bred by a single breeding company, whereas the OS trials evaluate virtually all genotypes intended for registration in Germany and thus represent a much broader gene pool. This is also a reasonable explanation why the OS/OF ratios for the genotypic VCs are all larger than one. In other words, concerning our pursued comparison, we have purely environmental VCs on one hand, which are assumed to be similar and not directly relevant for cultivar evaluation. Genotypic VCs, on the other hand, are indeed relevant for yield performance comparisons, but contrasting them with the chosen dataset is not necessarily meaningful due to the different composition of tested genotype sets.

Conversely, the plot error variance is an appropriate indicator of a trial's precision, as it provides a measure for the trial's precision for individual plots. Restricted maximum likelihood estimators (Searle et al., 1992) of VCs typically have little bias, and in the present case, because of the large size of the datasets and the small number of fixed effects, the bias is not expected to strongly depend on the number of observations and replicates. Thus, this comparison among the two trial systems is feasible despite the differences in MET design.

Table 5. *F*-tests of fixed effects for Eq. [3] taking system (*S*), genotype (*G*) and $S \cdot G$ as fixed and for Eq. [5] taking *S*, growth type (*D*) and $S \cdot D$ as fixed. The number of denominator df was set to infinity, and thus a Wald χ^2 test was performed.

Model	Effect	df	*F*-value	*p*-value
Eq. [3]	*S*	2	2375.1	<0.001
	G	813	3883.4	<0.001
	$S \cdot G$	22	40.0	0.011
Eq. [5]	*S*	2	989.8	<0.001
	D	1	0.4	0.556
	$S \cdot D$	1	14.0	<0.001

Comparing Plot Error Variances

As stated above, for unreplicated trials it is impossible to estimate separate effects for the error and the highest interaction in the respective model. Instead, we obtain a single confounded effect and a corresponding confounded VC. Additionally, the VC for the main plot effect essentially represents a second error variance. This VC is also confounded because of the experimental design, more specifically due to the fact that an environment's main plot always groups all genotypes of the same growth type. Since the focus lies on the error variance, we can paraphrase this situation by saying that for Eq. [3], this means that (i) the OF error variance is potentially inflated by the VC of $S \cdot Y \cdot L \cdot G$, and (ii) the OF main plot VC also includes the VC of $S \cdot Y \cdot L \cdot D$. This implies that by taking the sum of the plot error variance ($5.4\ 10^{-2}$ t^2 ha^{-2}) and the main plot VC ($3.6\ 10^{-2}$ t^2 ha^{-2}) for OF data, we obtain an upper limit for the true error variance ($9.0\ 10^{-2}$ t^2 ha^{-2}). This upper limit is 34% larger than the estimated plot error variance for OS trials ($6.7\ 10^{-2}$ t^2 ha^{-2}). Note, however, that since the former is an upper limit, it would only apply if the respective confounded VCs (i.e., $S \cdot Y \cdot L \cdot G$ and $S \cdot Y \cdot L \cdot D$) were zero. This seems particularly unlikely for Eq. [3], where the corresponding $S \cdot Y \cdot L \cdot G$ VC estimate for OS data was $3.3 \pm 0.2\ 10^{-2}$ t^2 ha^{-2}.

Therefore, the results indicate that these two systems have comparable precision regarding a single plot. Keep in mind, however, that (i) as long as the OF trials only use a single replicate, the confounded VC will inevitably be included in the residual variance, which is expected to be 34% larger than that of OS trials, and (ii) OS trials have three to four replicates and are therefore more precise for a single trial.

Accuracy

Although comparing the precision of both trial systems is useful, this comparison misses out on the question whether there are systematic differences between their genotypic evaluations, which would adversely affect the accuracy. Therefore, an accuracy comparison serves to assess whether

the two systems evaluate genotype performances in the same manner or not. Note, however, that all results for the accuracy comparisons are based on—and thus are limited to—the 23 identified shared genotypes, which are all Pioneer hybrids (15 tall hybrids and 8 semidwarf hybrids).

As stated above, a genetic correlation estimate close to one would indicate that the evaluation of genotypes in both systems leads to equivalent outcomes. The estimate of the genetic correlation ρ_{G1} from Eq. [3], however, is only 0.66 (± 0.19). Combined with the significant test results for the $S \cdot G$ interaction effects (again using Eq. [3], but taking S, G, and $S \cdot G$ as fixed), this suggests that the two systems do not generally evaluate genotypes in the same way.

It is very striking by comparison, however, that the estimate for ρ_{G2} from Eq. [5] is very close to unity, and furthermore, that the $S \cdot D$ interaction effects in the modified Eq. [5] were found to be significant as well. To verify this considerably larger correlation estimate, we reparameterized the variance–covariance structure of $S \cdot D \cdot G$ by omitting its covariance and adding a random $D \cdot G$ main effect instead. As expected, the VCs for $S \cdot D \cdot G$ were estimated as zero for both systems, which confirmed the high genetic correlation. Furthermore, when fitting Eq. [3] to a dataset where all semidwarf genotypes were excluded, the estimate for ρ_{G3} was also close to one.

Note that by creating Eq. [5] according to Steps 1 to 3, we allowed for heterogeneous variance and a covariance between genotypes of different systems, but not between different growth types. We did consider including both variance structures by defining the variance–covariance structure for genotypes as the Kronecker product of unstructured correlation matrices for D and S, respectively. We found, however, that the Akaike information criterion of this model (−61491.99) is larger than that of our presented Eq. [5] (−61502.82), which led us to the decision against the former. Nevertheless, we would like to point out that this disregarded model provided estimates for the genetic correlation between systems for each growing type, respectively, both being >0.98. Furthermore, it could be argued that heterogeneous genetic variances between environments should be allowed (e.g., by applying factor-analytic variance structures). Due to the large number of environments in our dataset (>900), however, such an approach would lead to just as many additional VCs and to a relatively complex model that may be difficult to fit. Thus, we decided against such an approach to keep the focus on the heterogeneity between systems.

Altogether, this suggests that both systems do not evaluate the identified shared genotypes in a similar manner; there are notable interaction effects between the systems and growth types (Fig. 3). There can be multiple causes for this difference, as each system is a composition of a large number of measures and the two systems differ in several of them—not only regarding plot sizes, number of environments, and machinery, but also, for example, that fungicides are often applied in OF trials but never in OS trials. Thus, it is principally impossible to identify one or even multiple measures as the underlying biological cause for this interaction. This suggests that additional investigations and/or experiments on the issue may be worthwhile. Furthermore, it is not clear which system (if any) estimates the correct yield. On one hand, it is possible that OS trials underestimated semidwarf or overestimated tall hybrid genotypes compared with OF trials. On the other hand, OF trials could have overestimated the semidwarf hybrids or underestimated the tall hybrid genotypes compared with OS trials. Furthermore, note that the ambiguity of these statements is due to fact that we cannot know which of the two systems is closer to the truth, simply because we do not know the true genotypic performances. That being said, it might just as well be that the truth lies in between the findings of the two systems, but this issue is not the main concern here. Note also that only OS trials include nonhybrid tall genotypes, and that these nonhybrid tall genotypes were always separated from hybrid tall genotypes (until 2009 via border plots, as well as via the plot-in-plot or double-plot approach; starting in 2009, via plot-in-plot or double-plot).

The quintessence from these results is that there are strong indicators of a systematic difference between the two systems when collectively evaluating semidwarf and

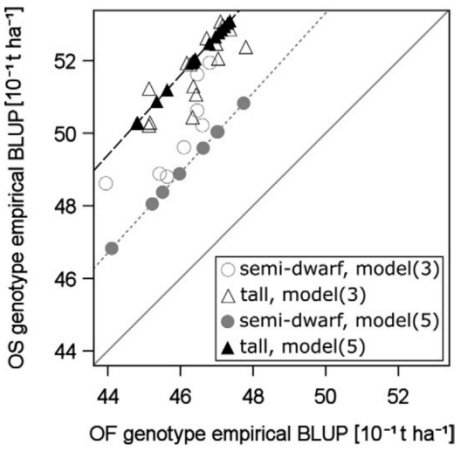

Fig. 3. System-specific best linear unbiased predictors (BLUPs) of winter oilseed rape grain yield for 23 identified shared genotypes estimated via Eq. [3] (unfilled symbols) and via Eq. [5] (filled symbols). Different symbols indicate whether a genotype is tall (triangles) or a semidwarf (circles). The solid reference line passes through the origin and has a slope of one. The grey dotted (semidwarf) and black dashed (tall) lines each pass through a point given by the adjusted system × growth type means and have a slope defined by the first eigenvector calculated from the estimated variance–covariance matrix of $S \cdot D \cdot G$ in Eq. [5]. OS, on-station; OF, on-farm.

tall hybrid genotypes, yet it must also be realized that when either a system-specific group effect accounting for these differences in the evaluation of semidwarf and tall genotypes was implemented (ρ_{G2}) or tall genotypes were analyzed separately (ρ_{G3}), the genotypic correlation estimates were almost one and thus hybrid genotype evaluations of both systems were very similar. This suggests that one cause for the difference between the systems is related to a systematic factor, the growth type.

Experimental Design Comparison

It can be argued that—metaphorically speaking—the overall intention of this article is to compare apples and oranges, because the two systems serve different purposes and there are several important differences between them. We understand and partly agree with this criticism, since there are indeed multiple differences between OF and OS trial systems that are not only considerable, but conceptual. As a result, any attempt to compare the two systems may shift the question of how these systems should be compared to whether they should be compared at all. At this point, however, we would like to argue that even though the comparison may not always be straightforward, it is nevertheless justified, valid, and worthwhile. To give deeper insight into some of the major differences between the systems' concepts, they are discussed individually below.

Unreplicated vs. Replicated Data

On-station trials are and have been the method of choice for the major part of agricultural field trials worldwide, and there are several reasons supporting this tradition. Probably the biggest influence on how to conduct single field trials came from the work of Fisher (1926), which is often quoted and taught as the very base of field trial design. For a detailed exposition of intent and implementation of the three main principles of experimental design (i.e., randomization, replication, and blocking), see Casler (2015). When considering METs rather than single-location trials, however, it is known that in terms of efficiency one should maximize the number of environments instead of the number of replicates per environment (Talbot, 1984; Casler, 2015); this follows directly from inspection of Eq. [7] and maximization for a fixed total number of plots. Nevertheless, Piepho et al. (2011) suggested that replication "should generally be adhered in OF experiments." The underlying reason for this advice is that, without replication, crucial consequences ensue. First of all, analyzing single environments becomes virtually impossible. Instead, environments are now treated as random blocks drawn from a larger TPE, and thus the data can indeed only be analyzed to provide genotypic means across environments. Second, losing a single environment's plot due to, for instance, weather damage means losing all information for the respective genotype in that environment. This, however, is not as devastating as in OS trials, since the number of tested environments is usually larger for OF trials. A third consequence is the confounding of the highest order interaction term with the plot error, which complicates the interpretation of VC estimates and is thus at the heart of this article. Furthermore, in case the VC that is confounded with the error variance is not zero, the analysis inevitably becomes less precise than it would be if they were not confounded. None of these consequences are genuinely desirable, yet it must be clear that they are not necessarily fatal.

In fact, by accepting the restrictions and further endorsing the shift towards more environments and higher practicability, head-to-head comparisons as an unreplicated field trial design approach became popular in plant breeding and genotype testing. With sometimes even the minimum of only two plots per environment comparing two genotypes, head-to-head comparisons can be seen as the extreme case in OF trials and have been applied by plant breeders for >30 yr now (Bradley et al., 1988).

Balanced vs. Unbalanced Data

In OS trials, usually all genotypes available in a given year are tested together in the same trials. By contrast, the OF trials considered in this paper often tested only a subset of all genotypes in each trial, because the capacity of each trial was limited. Using the same set of genotypes and locations each year would yield perfectly balanced data and allow for a conventional analysis via ANOVA techniques, which were indeed favored in the early times (Smith et al., 2005). With today's mixed model approaches and computational possibilities, however, the unbalance of data itself need not be avoided only for the sake of a conventional statistical analysis. As long as it can be assumed that data are missing at random or missing completely at random, mixed models can be used for data analyses (Piepho and Möhring, 2006; Little and Rubin, 2014). At the same time, it must be acknowledged that environments become incomplete blocks in these OF trials, and for a given set of test environments with fixed test capacity, it is useful to optimize the allocation of genotypes to environments regarding environments as incomplete blocks (Piepho et al., 2011). As a result, however, computation of means across environments entails substantial adjustments for environmental effects, which adversely affects the precision of genotype mean comparisons compared with a complete genotype × environment classification. Hence, from a design perspective, it is desirable to test as many genotypes as possible together in the same environments. It should be pointed out that, for OF trials, the number of genotypes tested at a single environment is more limited than for OS trials.

Strip Plot vs. Small Plot

The ability to use the farmers' machinery implies an increased practical relevance of OF trials (Piepho et al.,

2011); arguably, compared with OS trials, OF trials often show a broader validity than those drawn solely from "the narrow confines of a research institute setting" (University of Reading, 1998). Another crucial argument is the occurrence of border effects arising from competition (Büchse, 2002). This is especially true when plants in adjacent plots cannot be expected to be similarly competitive (e.g., due to different inherent growing sizes, root formation etc.). As stated above, however, it is common practice in the OS trials to use buffer plots between plots with semidwarf and tall genotypes to overcome this problem. Besides border effects due to competition between plants, there is also the possibility of front border effects. Front borders refer to plot borders adjacent to paths between plots. It can be argued that tall hybrids have a bigger photosynthetically active canopy by being able to lean into paths further than semidwarfs. Naturally, larger plots lead to borders taking up a smaller fraction of the plot area and accordingly diminish the influence of border effects.

Another critical aspect is the harvest of small plots with field trial technology and thus not conventional technology, especially when *Scheiteln* was applied. *Scheiteln* itself poses an unwanted but inevitable manipulation of the plants if not conducted with required care (UFOP, 2016). Moreover, the crop gets pushed down so that relatively low cutting needs to be done for harvest, which can lead to excessive strain on the straw walkers. In contrast, farmers always try to cut oilseed rape as high as possible to efficiently harvest oilseed grains.

Therefore, both differences in the purpose of cultivar evaluation trials and differences in plot sizes require differences between the two systems in the applied methods and technology. It seems probable that these changes play an important role in the observed difference in yield estimates.

CONCLUSION

In this article, we used bivariate mixed model analyses to compare unreplicated OF strip trials with traditional, replicated OS small-plot trials regarding their precision and accuracy for cultivar evaluation. Results indicate a comparable precision for a single plot of the two systems' yield estimates for winter oilseed rape genotypes in Germany. By contrast, we identified systematic growth type × system interaction effects for yield estimations. These interactions could not be attributed to a single cause, yet we found the growth type to be a factor that allowed us to model these interactions. After all, the two systems can currently only be expected to yield equivalent results within but not across growth types.

Conflict of Interest

The authors declare that there is no conflict of interest.

Supplemental Material Available

Supplemental material for this article is available online.

Acknowledgments

We thank Pioneer Hi-Bred Northern Europe for providing the dataset used in this study. This research was funded by the German Research Foundation (DFG Grant PI 377/18-1).

References

Bradley, J.P., K.H. Knittle, and A.F. Troyer. 1988. Statistical methods in seed corn product selection. J. Prod. Agric. 1:34–38. doi:10.2134/jpa1988.0034

Büchse, A. 2002. Optimierung der Versuchstechnik bei Winterraps. UFOP-Schriften 18. Union zur Förderung von Oel-un Proteinpflanzen, Bonn, Germany.

Bundessortenamt. 2000. Richtlinien für die Durchführung von landwirtschaftlichen Wertprüfungen und Sortenversuchen. Landbuch Verlag, Hannover, Germany. http://www.bundessortenamt.de/internet30/fileadmin/Files/PDF/Richtlinie_LW2000.pdf (accessed 9 June 2016).

Bundessortenamt. 2014. Richtlinien für die Durchführung von landwirtschaftlichen Wertprüfungen und Sortenversuchen. Bundessortenamt, Hannover, Germany.

Butler, D.G., B.R. Cullis, A.R. Gilmour, and B.J. Gogel. 2009. ASREML-R reference manual. Release 3.0. Tech. Rep. Queensland Dep. Prim. Ind., Brisbane, QLD.

Casler, M.D. 2015. Fundamentals of experimental design: Guidelines for designing successful experiments. Agron. J. 107:692–705. doi:10.2134/agronj2013.0114

Cullis, B.R., F.M. Thomson, J.A. Fisher, A.R. Gilmour, and R. Thompson. 1996. The analysis of the NSW wheat variety database. I. Modelling trial error variance. TAG. Theor. Appl. Genet. 92:21–27. doi:10.1007/BF00222947

Damesa, T., J. Möhring, M. Worku, and H.-P. Piepho. 2017. One step at a time: Stage-wise analysis of series of experiments. Agron. J. 109:845–857. doi:10.2134/agronj2016.07.0395

Fisher, R.A. 1926. The arrangement of field experiments. J. Min. Agric. Great Britain 33:503–513.

Frensham, A.B., A.R. Barr, B.R. Cullis, and S.D. Pelham. 1998. A mixed model analysis of 10 years of oat evaluation data: Use of agronomic information to explain genotype by environment interaction. Euphytica 99:43–56. doi:10.1023/A:1018395731621

Gilmour, A.R., B.J. Gogel, B.R. Cullis, and R. Thompson. 2009. ASReml user guide. Release 3.0. VSN Int., Hemel Hempstead, UK.

Holland, J.B., W.E. Nyquist, and C.T. Cervantes-Martínez. 2003. Estimating and interpreting heritability for plant breeding: An update. Plant Breed. Rev. 2003:9–112. doi:10.1002/9780470650202.ch2

Jackson, J.D., and J.A. Dunlevy. 1988. Orthogonal least squares and the interchangeability of alternative proxy variables in the social sciences. Statistician 37:7–14. doi:10.2307/2348374

Kenward, M.G., and J.H. Roger. 1997. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 53:983–997. doi:10.2307/2533558

Laidig, F., T. Drobek, and U. Meyer. 2008. Genotypic and environmental variability of yield for cultivars from 30 different crops in German official variety trials. Plant Breed. 127:541–547. doi:10.1111/j.1439-0523.2008.01564.x

Little, R.J.A., and D.B. Rubin. 2014. Statistical analysis with missing data. 2nd ed. Ser. Prob. Stat. Wiley, Hoboken, NJ.

Möhring, J., and H.-P. Piepho. 2009. Comparison of weighting in two-stage analysis of plant breeding trials. Crop Sci. 49:1977–1988. doi:10.2135/cropsci2009.02.0083

Piepho, H.-P., A. Büchse, and K. Emrich. 2003. A hitchhiker's guide to mixed models for randomized experiments. J. Agron. Crop Sci. 189:310–322. doi:10.1046/j.1439-037X.2003.00049.x

Piepho, H.-P., and J. Möhring. 2006. Selection in cultivar trials- is it ignorable? Crop Sci. 46:192–201. doi:10.2135/cropsci2005.04-0038

Piepho, H.-P., and J. Möhring. 2007. Computing heritability and selection response from unbalanced plant breeding trials. Genetics 177:1881–1888. doi:10.1534/genetics.107.074229

Piepho, H.-P., and J. Möhring. 2011. On estimation of genotypic correlations and their standard errors by multivariate REML using the MIXED procedure of the SAS system. Crop Sci. 51:2449–2454. doi:10.2135/cropsci2011.02.0088

Piepho, H.-P., J. Möhring, T. Schulz-Streeck, and J.O. Ogutu. 2012. A stage-wise approach for the analysis of multi-environment trials. Biom. J. 54:844–860. doi:10.1002/bimj.201100219

Piepho, H.-P., M.F. Nazir, M. Qamar, A. Rattu, Riaz-ud-Din, M. Hussain, et al. 2016. Stability analysis for a countrywide series of wheat trials in Pakistan. Crop Sci. 56:2465–2475. doi:10.2135/cropsci2015.12.0743

Piepho, H.-P., C. Richter, J. Spilke, K. Hartung, A. Kunick, and H. Thöle. 2011. Statistical aspects of on-farm experimentation. Crop Pasture Sci. 62:721–735. doi:10.1071/CP11175

Przystalski, M., A. Osman, E.M. Thiemt, B. Rolland, L. Ericson, H. Østergård, et al. 2008. Comparing the performance of cereal varieties in organic and non-organic cropping systems in different European countries. Euphytica 163:417–433. doi:10.1007/s10681-008-9715-4

Rao, C.R., editor. 1973. Linear statistical inference and its applications. John Wiley & Sons, Hoboken, NJ. doi:10.1002/9780470316436

SAS Institute. 2013. Base SAS 9.4 procedures guide: Statistical procedures. 2nd ed. SAS Inst., Cary, NC.

Schulz-Streeck, T., J.O. Ogutu, and H.-P. Piepho. 2013. Comparisons of single-stage and two-stage approaches to genomic selection. TAG. Theor. Appl. Genet. 126:69–82. doi:10.1007/s00122-012-1960-1

Searle, S.R., G. Casella, and C.E. McCulloch. 1992. Variance components. Wiley, New York. doi:10.1002/9780470316856

Smith, A., B. Cullis, and A. Gilmour. 2001. The analysis of crop variety evaluation data in Australia. Aust. N. Z. J. Stat. 43:129–145. doi:10.1111/1467-842X.00163

Smith, A.B., B.R. Cullis, and R. Thompson. 2005. The analysis of crop cultivar breeding and evaluation trials: An overview of current mixed model approaches. J. Agric. Sci. 143:449–462. doi:10.1017/S0021859605005587

Talbot, M. 1984. Yield variability of crop varieties in the U.K. J. Agric. Sci. 102:315–321. doi:10.1017/S0021859600042635

Troyer, A.F. 1996. Breeding widely adapted, popular maize hybrids. Euphytica 92:163–174. doi:10.1007/BF00022842

Troyer, A.F., and E.J. Wellin. 2009. Heterosis decreasing in hybrids: Yield test inbreds. Crop Sci. 49:1969–1976. doi:10.2135/cropsci2009.04.0170

UFOP. 2016. Beiträge zum Sortenprüfwesen bei Öl- und Eiweißpflanzen für die deutsche Landwirtschaft. Union zur Förderung von Öl-und Proteinpflanzen, Bonn, Germany. http://www.ufop.de/agrar-info/erzeuger-info/raps/beitraege-zum-sortenpruefwesen-bei-oel-und-eiweisspflanzen-fuer-die-deutsche-landwirtschaft/ (accessed 8 June 2016).

University of Reading. 1998. On-farm trials: Some biometric guidelines. Stat. Serv. Ctr., Univ. of Reading, Reading, UK.

Yan, W., L.A. Hunt, P. Johnson, G. Stewart, and X. Lu. 2002. On-farm strip trials vs. replicated performance trials for cultivar evaluation. Crop Sci. 42:385–392. doi:10.2135/cropsci2002.0385

3. Estimating broad-sense heritability with unbalanced data from agricultural cultivar trials

P. Schmidt[a], J. Hartung[b], J. Rath[b], and H.-P. Piepho[a]

[a] Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstrasse 23, 70599 Stuttgart, Germany

[b] Deutsches Maiskomitee e.V., Brühler Straße 9, 53119 Bonn, Germany;

Abstract

Broad-sense heritability is defined as the proportion of phenotypic variance that is attributable to an overall variance for the genotype. It is often calculated as a measure (i) to quantify and eventually compare the precision of agricultural cultivar trials and/or (ii) to estimate the response to selection in plant breeding trials. In practice, most such trials are conducted at multiple environments *(i.e.* year-by-location-combinations) resulting in a multi-environment trial (MET) with unbalanced data, as*, e.g.,* not all cultivars are tested at each environment. Yet, the standard method for estimating heritability implicitly assumes balanced data, independent genotype effects and homogeneous variances. Therefore, we compared the estimates for broad-sense heritability computed via the standard method to those obtained via six alternative estimation methods (example codes: https://github.com/PaulSchmidtGit/Heritability). We did so by analyzing four cultivar MET, which all displayed a typically unbalanced data structure, but differed in the genetic frameworks of their cultivars. Results indicate that the standard method may over-estimate heritability for all datasets, while alternative methods show similar estimates per dataset and thus seem able to better handle this kind of unbalanced data. Finally, we show that in order to compare heritability estimates between different MET, genetic variance components estimates should be fixed to common values for both datasets.

Status: published

Published online February 14, 2019

RESEARCH

Estimating Broad-Sense Heritability with Unbalanced Data from Agricultural Cultivar Trials

P. Schmidt, J. Hartung, J. Rath, and H.-P. Piepho*

ABSTRACT

Broad-sense heritability is defined as the proportion of phenotypic variance that is attributable to an overall variance for the genotype. It is often calculated as a measure (i) to quantify and eventually compare the precision of agricultural cultivar trials, and/or (ii) to estimate the response to selection in plant breeding trials. In practice, most such trials are conducted at multiple environments (i.e., year–location combinations) resulting in a multienvironment trial (MET) with unbalanced data, as, for example, not all cultivars are tested at each environment. However, the standard method for estimating heritability implicitly assumes balanced data, independent genotype effects, and homogeneous variances. Therefore, we compared the estimates for broad-sense heritability computed via the standard method to those obtained via six alternative estimation methods (example codes: https://github.com/PaulSchmidtGit/Heritability). We did so by analyzing four cultivar METs, which all displayed a typically unbalanced data structure but differed in the genetic frameworks of their cultivars. Results indicate that the standard method may overestimate heritability for all datasets, whereas alternative methods show similar estimates per dataset and thus seem better able to handle this kind of unbalanced data. Finally, we show that to compare heritability estimates between different METs, genetic variance component estimates should be fixed to common values for both datasets.

P. Schmidt, J. Hartung, and H.-P. Piepho, Biostatistics Unit, Institute of Crop Science, Univ. of Hohenheim, Fruwirthstrasse 23, 70599 Stuttgart, Germany; J. Rath, Deutsches Maiskomitee, Brühler Straße 9, 53119 Bonn, Germany. Received 11 June 2018. Accepted 26 Nov. 2018. *Corresponding author (piepho@uni-hohenheim.de). Assigned to Associate Editor Marcio Resende Jr.

Abbreviations: BLUE, best linear unbiased estimator; BLUP, best linear unbiased predictor; BSA, Bundessortenamt; CRD, completely randomized design; MET, multienvironment trial; PC, Pro-Corn; RCBD, randomized complete block design; VC, variance component.

In plant breeding programs and cultivar evaluation trials, cultivars (genotypes) of interest are often grown and tested at multiple locations across several years. Such a series of trials is called a multienvironment trial (MET), where a year–location combination is referred to as an environment. To quantify and eventually compare the precision of METs, plant breeders often calculate narrow-sense heritability (h^2) or broad-sense heritability (H^2) on a genotype-mean basis. The latter is defined as the proportion of phenotypic variance that is attributable to an overall variance for the genotype, thus including additive, dominance, and epistatic variance (Holland et al., 2003; Falconer and Mackay, 2005). Moreover, there are usually additional interpretations associated with H^2: (i) it is equivalent to the coefficient of determination of a linear regression of the unobservable genotypic value on the observed phenotype, (ii) it is also the squared correlation between predicted phenotypic value and genotypic value, and (iii) it represents the proportion of the selection differential (S) that can be realized as the response to selection (R) (Falconer and Mackay, 2005). It is important to note that the necessity to estimate H^2 on a genotype-mean basis results from the fact that in plant breeding, genotypes are often tested across a wide range of environments in designed, replicated experiments.

Published in Crop Sci. 59:1–12 (2019).
doi: 10.2135/cropsci2018.06.0376

Thus, as opposed to common practice in animal breeding, the phenotypic value for a genotype in plant breeding is almost always some sort of mean value. Most formulas proposed for calculating H^2 implicitly assume balanced data and/or independent genotypic effects. Furthermore, the most common H^2 measure, given as $H^2_{\text{Standard}} = \sigma^2_g / \sigma^2_p$ and thus the ratio of genotypic and phenotypic variance, also assumes homogeneous variances. For example, for a balanced series of n_l locations in a single year laid out in randomized complete blocks with n_r replicates per location, the phenotypic values are simply the arithmetic means over all observations per genotype and the phenotypic variance is $\sigma^2_p = \sigma^2_g + \sigma^2_{gl} / n_l + \sigma^2_e / (n_l n_r)$, where σ^2_{gl} is the genotype × location interaction variance and σ^2_e is the residual plot error variance.

In practice, most METs are unbalanced. An incomplete genotype × environment classification (i.e., not all genotypes are tested every year and at the same locations) occurs, for example, due to selecting promising genotypes and dropping the rest from trials in the subsequent year, as well as due to adding new genotypes or testing just a portion of all genotypes in certain locations. Furthermore, when using p-rep designs (Cullis et al., 2006) or simply when plot data is lost, the number of replicates at each location varies between genotypes. Altogether, the questions of why and how data from a cultivar evaluation MET is unbalanced can have various answers, but the question of whether it is unbalanced can more often be answered with "yes." With necessity, unbalanced data result in heterogeneous variances and covariances of adjusted means, which often are the "phenotypes" (i.e., phenotypic values) on which decisions for (i) selection in breeding programs or (ii) registration in national cultivar lists are based.

Estimating H^2 via H^2_{Standard} can neither be done for unbalanced data nor heterogeneous complex variance–covariance structures. This problem is far from new, and it motivated the proposal of several alternative, generalized H^2 estimation methods (Laloë, 1993; Holland et al., 2003; Cullis et al., 2006; Helms and Hammond, 2006; Oakey et al., 2006; Piepho and Möhring, 2007). Nevertheless, H^2_{Standard} is still the most common method even for unbalanced trial data, whereas the alternative methods are rarely applied.

Therefore, this article reviews and illustrates six alternative estimation methods for H^2 on a genotype-mean basis, of which four are taken from publications, and two are, to our knowledge, unpublished approaches. We do so by analyzing four different datasets, all from METs for cultivar evaluation with maize (*Zea mays* L.) for biogas production in Germany. All datasets are unbalanced in a manner that is typical for cultivar evaluation trials. We focus on this type of unbalancedness when contrasting the six alternative H^2 methods to H^2_{Standard}. Although all four METs evaluate genotypes bred for similar use (i.e., maize for biogas production in Germany), they differ in both their objectives of cultivation (i.e., registration and post-registration recommendation) and their experimental design (e.g., numbers of locations per year and number of replicates per environment). Thus, we chose to compare H^2 estimates of these four datasets, as they represent slightly different, relevant types of a typically unbalanced MET for cultivar evaluation and therefore allow for a broader validity of the results.

MATERIALS AND METHODS

This section is subdivided into three subsections. First, relevant characteristics of the four datasets are outlined. Second, models and their estimates of two different two-stage mixed model analyses are described. Third, six alternative H^2 approaches are described.

Datasets

Four 2-yr MET datasets of maize genotypes bred for biogas production (BSA1, BSA2, PC1, and PC2) were analyzed. All four datasets were obtained from replicated small-plot cultivar evaluation trials in 2014 and 2015 conducted at multiple locations in Germany.

Two of the datasets (BSA1 and BSA2) were generated in trials of the Bundessortenamt (BSA, German Federal Plant Cultivar Office). Their goal is to evaluate the genotypes' value of cultivation and use prior to registration in the national list in Germany. In this system, a maize genotype is evaluated up to 2 yr before registration, and there is a new 2-yr-cycle starting each year. Thus, for each year, BSA1 and BSA2 include all observations across two cycles. The other two datasets (PC1 and PC2) were obtained in trials of Pro-Corn (PC, Bonn, Germany) in the framework of the EU-Prüfung from the Deutsches Maiskomitee (German Maize Committee). The aim of these trials is to carry out a preliminary evaluation of genotypes that are already registered in other EU countries, but not in Germany. After a 2-yr test cycle, the best candidates are eventually recommended for Anbaugebietsprüfung Biogas, which is a regional, post-registration cultivar evaluation trial in Germany.

The maize genotypes included in the datasets BSA1 and PC1 covered the range from mid-early to late maturity (S 230–S 270), as classified in the German silage maturity rating system, based on the whole dry matter content at harvest. Datasets BSA2 and PC2 show late maturity (S 270–S 320). Thus, genotypes are either not yet registered nationally (BSA-) or are registered outside Germany (PC-) and also show either mid-early maturity (-1) or late maturity (-2). Accordingly, the sets of genotypes are quite distinct between the four datasets: the numbers of genotypes per dataset ranged from 21 to 70 (Table 1) with zero to five common genotypes in any two datasets, respectively.

The number of environments per dataset ranged from 21 to 25 for all datasets except PC1, which comprises 67 environments. Note that we removed a maximum of five environments per dataset beforehand due to disproportionally high (i.e., >40%) mean dry matter contents. All PC trials were laid out as randomized complete block designs (RCBDs) with two replicates, except for two environments with three replicates. The BSA trials were

Table 1. Dimensions (number of genotypes, locations, and environments) and design (randomized complete block designs [RCBD], α-lattices) of the four multienvironment trial datasets (BSA1, BSA2, PC1, and PC2, where BSA indicates Bundessortenamt trial datasets, and PC indicates Pro-Corn trial datasets).

Dimension	BSA1	BSA2	PC1	PC2
No. of genotypes				
Total	70	29	32	21
2014	49	20	28	18
2015	46	22	23	15
Overlap	25	13	19	12
No. of locations				
Total	14	14	40	16
2014	11	13	36	14
2015	10	11	31	11
Overlap	7	10	28	9
No. of environments				
Total	21	24	67	25
Design				
RCBD†	1	13	67	25
α-lattice	20	11	–	–
No. of replicates per location				
2	–	–	66	24
3	21	25	1	1

† RCBD, randomized complete block design.

laid out as either RCBDs or α-lattice designs (with block size of six, seven, or eight), with three replicates. For the BSA1 data, 20 out of 21 environments were randomized according to an α-lattice, whereas for the BSA2 data, RCBDs and α-lattices were used in about half of the environments (Table 1), respectively. Plot sizes for all trials were very similar with a median plot size of 9 m^2 (minimum = 6.8 m^2, maximum = 10.5 m^2). In all four datasets, data within years are balanced with respect to genotypes and locations. A visualization of the size and degree of balance of the four datasets is given in Fig. 1.

Mixed Model analysis

We assume a standard linear mixed model for the observed data vector $\mathbf{y}$, which is of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \qquad [1]$$

where $\mathbf{y}$ is the vector of observations, $\boldsymbol{\beta}$ and $\mathbf{u}$ are vectors of fixed and random effects, respectively, $\mathbf{X}$ and $\mathbf{Z}$ are the associated design matrices, and $\mathbf{e}$ is a random residual vector. The random effects are assumed to be distributed as $\mathbf{u} \sim \text{MVN}(\mathbf{0}, \mathbf{G})$ and $\mathbf{e} \sim \text{MVN}(\mathbf{0}, \mathbf{R})$, such that $\mathbf{y} \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, where $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$ and MVN denotes the multivariate normal distribution with mean vector as the first argument and variance–covariance matrix as the second. Thus, $\mathbf{G}$, $\mathbf{R}$, and $\mathbf{V}$ are the variance–covariance matrices of random effects $\mathbf{u}$, error effects $\mathbf{e}$, and the vector of observations $\mathbf{y}$, respectively. The solutions to the mixed-model equations, yielding the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ ($\hat{\boldsymbol{\beta}}$) and the best linear unbiased predictor (BLUP) of $\mathbf{u}$ ($\hat{\mathbf{u}}$), are given by

$$\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{u}}\end{pmatrix} = \mathbf{C}^{-}\begin{pmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{y}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y}\end{pmatrix} \qquad [2]$$

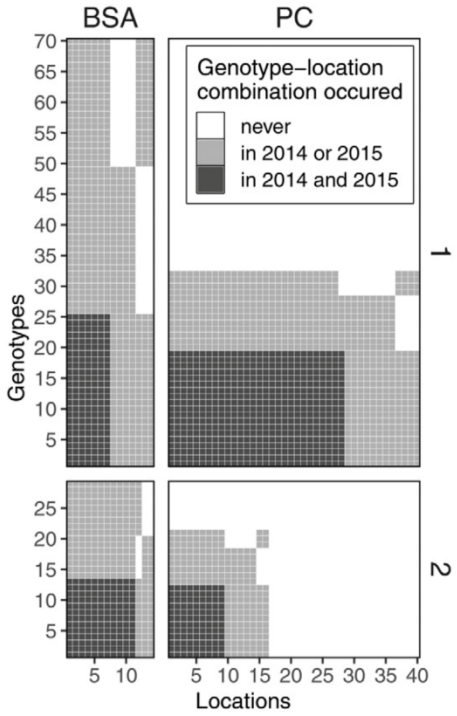

Fig. 1. Heat map visualization of the dataset sizes and degree of balancedness of the four datasets BSA1, BSA2, PC1, and PC2, where BSA indicates Bundessortenamt trial datasets, and PC indicates Pro-Corn trial datasets. Each square represents one genotype–location combination. A square's color indicates the number of years in which this genotype-by-location combination was tested (white = 0, light gray = 1, dark gray = 2). Example: of all four datasets, PC1 has the highest number of locations (40) and a total number of 32 genotypes. Of those, however, only 19 genotypes in 28 locations were tested in both years. Since data within a single year were balanced, these 19 × 28 genotype–location combinations represent the completely balanced subset of PC1.

where $\mathbf{C}^{-}$ is a g-inverse of $\mathbf{C}$ and

$$\mathbf{C} = \begin{pmatrix}\mathbf{C}_{11} & \mathbf{C}_{12}\\ \mathbf{C}_{21} & \mathbf{C}_{22}\end{pmatrix} = \begin{pmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1}\end{pmatrix} \qquad [3]$$

is the coefficient matrix of the mixed-model equations (McLean et al., 1991), and $\mathbf{C}_{22}$ has the same dimension as $\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}$. Data were analyzed using two mixed models fitted by restricted maximum likelihood method using ASREML-R 3.0 (Gilmour et al., 2009) and SAS 9.4 (SAS Institute, 2013). For practical reasons (i.e., computational burden), both mixed-model analyses are performed in two stages and thus in the framework of a multistage analysis, as will be further detailed below, and for each of the four datasets separately. Note that an appropriately conducted two-stage analysis is essentially equivalent to a single-stage analysis (Möhring and Piepho, 2009; Welham et al., 2010; Piepho et al., 2012; Damesa et al., 2017). The first stage is identical for both analyses of each of the four datasets. The only difference between the two analyses is that in the second stage, the genotype main effect was once taken

as random and once taken as fixed. This was necessary, since some H^2 estimation methods we used require both genotypic BLUPs, as well as genotypic BLUEs.

First Stage

In the first stage, each environment of the considered dataset is analyzed separately using the model

$$y_{ijk} = \mu_1 + g_{1i} + \text{rep}_j + \text{block}_{jk} + \text{plot}_{ijk} \quad [4]$$

where y_{ijk} is the dry matter yield (dt ha^{-1}) of the ith genotype in the kth block of the jth replicate, μ_1 is the first-stage intercept, g_{1i} is the effect for the ith genotype in the first stage, rep_j is the effect of the jth replicate, block_{jk} is the effect of the kth incomplete block of the jth replicate, and plot_{ijk} is the plot error effect corresponding to y_{ijk}. An incomplete block effect block_{jk} was only fitted when the trial was laid out as an α-design, in which case block_{jk} was taken as random to make use of inter-block information (Möhring et al., 2015), whereas the complete replicate effect rep_j was always taken as fixed, as no inter-replicate information was to be recovered. The genotype main effect g_{1i} was taken as fixed to obtain adjusted genotype means ($\overline{y}_i$) and their approximated variance–covariance matrix ($\hat{\mathbf{V}}$) in the first stage. Note that, although genotypes are modeled as random in the next stage, it is crucial to assume fixed genotypic effects at this stage to ensure equivalence with single-stage analysis (Piepho et al., 2012). Furthermore, an estimate of the plot error variance ($\hat{\sigma}^2_{\text{plot}}$) was obtained for each environment. Note that separate analyses and therefore separate effect estimates, as well as separate error variances, were obtained for each environment. Therefore, estimates forwarded to the second stage were indexed by year h and location m ($\overline{y}_{i_{hm}}$, $\hat{\mathbf{V}}_{hm}$, $\hat{\sigma}^2_{\text{plot}_{hm}}$).

When performing a two-stage analysis of plant breeding or cultivar evaluation trials, a proper weighting method can approximate the variance–covariance structure of adjusted means and therefore slightly improve the analysis (Möhring and Piepho, 2009). In this study, we used Smith's weights (Smith et al., 2001, 2005), which can be obtained as the diagonal elements of the inverse of $\hat{\mathbf{V}}_{hm}$. Note that using Smith's weights in a multistage analysis only approximates a fully efficient single-stage analysis, yet it approximates it quite well and is much easier to implement than the alternative of using a fully efficient weighting method (Damesa et al., 2017).

Second Stage

In the second stage, the model equation can be written as

$$\overline{y}_{ihm} = \mu + a_h + l_m + \text{al}_{hm} + g_i + \text{ga}_{ih} + \text{gl}_{im} + \text{gal}_{ihm} + \overline{e}_{ihm} \quad [5]$$

where $\overline{y}_{ihm}$ is the adjusted mean of the ith genotype in the hth year and the mth location obtained in the first stage, μ is the intercept, a_h is the main effect for the hth year, l_m is the main effect for the mth location, al_{hm} is the hmth year × location interaction effect, g_i is the main effect of the ith genotype, ga_{ih} is the ihth genotype × year interaction effect, gl_{im} is the imth genotype × location interaction effect, gal_{ihm} is the ihmth genotype × year × location interaction effect, and $\overline{e}_{ihm}$ is the error of the mean $\overline{y}_{ihm}$ obtained in the first stage.

Since some H^2 estimation methods we use require estimates from models where g_i is taken as random and as fixed, we applied Eq. [5] with two different assumptions about the genotype main effects g_i. The two models will be denoted here as Eq. [5r] and [5f], where r and f stand for random and fixed, respectively. In Eq. [5r], the effect g_i was taken as random, and thus all effects (except the intercept) were assumed as random. This model was used to obtain (i) an estimate for the intercept μ, which is denoted as $\hat{\mu}_r$, (ii) BLUPs for g_i (BLUP$_i$), (iii) the mean variance of all their pairwise differences ($\overline{v}_\Delta^{\text{BLUP}}$), (iv) the estimated variance–covariance matrix of random effects (**G**), (v) variance component (VC) estimates, and (vi) an estimate for $\mathbf{C}^-_{22}$ as the submatrix of the inverse of the mixed model equations coefficient matrix pertaining to the random effects **u** (see Eq. [2]). In Eq. [5f], the effect g_i was taken as fixed, and thus all effects except g_i (and the intercept) were taken as random. This model was used to obtain (i) adjusted means for g_i (BLUE$_i$), and (ii) the mean variance of all their pairwise differences among adjusted genotype means ($\overline{v}_\Delta^{\text{BLUE}}$).

To minimize the potential influence of an estimation error on VC estimation between the two models Eq. [5r] and Eq. [5f], we fixed the VC in the analysis of Eq. [5f] to estimates obtained by fitting Eq. [5r]; that is, all estimates except for σ^2_g, the VC for the genotype main effect g_i, since the latter is taken as a fixed effect in Eq. [5f] and thus only has a VC in Eq. [5r]. As a result, Eq. [5r] and Eq. [5f] are guaranteed to have the same estimates for σ^2_a, σ^2_l, σ^2_{al}, σ^2_{ga}, σ^2_{gl}, σ^2_{gal}, and $\sigma^2_{\bar{e}}$, which represent the VCs for the year and location main effects, for the year × location, genotype × year, genotype × location, genotype × year × location interaction effects, and the error term, respectively. All H^2 estimates presented in the sections below are entirely based on estimates created via Eq. [4], Eq. [5r], and Eq. [5f] (see Table 2).

Average Plot Error Variance

Variance components σ^2_a, σ^2_l, σ^2_{al}, σ^2_{ga}, σ^2_{gl}, and σ^2_{gal} were estimated via Eq. [5r], whereas heterogeneous estimates of $\sigma^2_{\text{plot}_{hm}}$—namely, one per year–location combination hm—were obtained via Eq. [4]. To calculate H^2, for example, with the standard formula, a single plot error variance estimate is required. Instead of simply taking the arithmetic mean of all $\sigma^2_{\text{plot}_{hm}}$, we allowed for heterogeneous plot error variances between environments and assumed a γ distribution by fitting a generalized linear model of the form

$$\log\left[E\left(\sigma^2_{\text{plot}_{hm}}\right)\right] = \lambda \quad [6]$$

where $\sigma^2_{\text{plot}_{hm}}$ is the plot error variance for the hth year and the mth location, and λ is an intercept. The log function was chosen to link the γ distributed plot error variances to the linear predictor (Cullis et al., 1996; Frensham et al., 1998). Via Eq. [6], we obtained an estimate for the average plot error variance as $\exp(\hat{\lambda})$, where $\hat{\lambda}$ is the estimate of λ. Note that in this article, whenever an overall plot error variance estimate is referred to, it is actually this estimate of $E\left(\sigma^2_{\text{plot}_{hm}}\right)$.

Broad-Sense Heritability Estimators

To compare the precisions of the four METs, we investigated H^2_{Standard} and six alternative H^2 estimation methods.

Table 2. Overview of all estimates obtained via the first-stage model (Eq. [4]), the second-stage model with random genotype main effect (Eq. [5r]), and the second-stage model with fixed genotype main effect (Eq. [5f] that are directly involved in the application the respective heritability (H^2) estimation method. A dot indicates that the respective estimate (row) is required for the respective H^2 estimation method (column).

Model	Estimate	Definition	H^2_{Standard}	H^2_{Holland}	H^2_{Piepho}	H^2_{Cullis}	H^2_{reg}	H^2_{sumdiv}	$H^2_{\hat{g}\hat{g}}$
First stage: For each year–location combination									
Eq. [4]	$\hat{\sigma}^2_{\text{plot}_{\text{ye}}}$	Plot error variance	●	●					
Second stage: Across year–location combinations									
Eq. [5r]	$\hat{\mu}_r$	Estimate of intercept					●	●	
	BLUP_i	Best linear unbiased predictor of genotype main effects					●	●	
	$\mathbf{G}$	Variance–covariance matrix of random effects							●
	$\mathbf{C}^-_{22}$	Submatrix of the inverse of the mixed model equations coefficient matrix pertaining to the random effects							●
	VC	Variance components	●	●	●	●			
	$\overline{\text{vd}}_{\text{BLUP}}$	Mean variance of a difference of two best linear unbiased predictors for the genotypes				●			
Eq. [5f]	BLUE_i	Adjusted means based on best linear unbiased estimators for genotype main effects					●	●	
	$\overline{\text{vd}}_{\text{BLUE}}$	Mean variance of a difference of two best linear unbiased estimators for the genotypes			●				

H^2_{Standard}

As stated before, heritability in the broad sense is originally defined as

$$H^2_{\text{Standard}} = \sigma^2_g \big/ \sigma^2_p \qquad [7]$$

where σ^2_g is the genotypic variance and σ^2_p is the phenotypic variance. Given the structure of our data and Eq. [5r], the latter can then be defined in the standard manner as

$$\sigma^2_p = \sigma^2_g + \frac{\sigma^2_{ga}}{n_a} + \frac{\sigma^2_{gl}}{n_l} + \frac{\sigma^2_{gal}}{n_a n_l} + \frac{\sigma^2_{plot}}{n_a n_l n_r} \qquad [8]$$

where σ^2_g, σ^2_{ga}, σ^2_{gl}, and σ^2_{gal} are the VCs of the genotype main effect, genotype × year interaction effect, genotype × location interaction effect, and genotype × year × location interaction effect, respectively, σ^2_{plot} is the (plot) error VC, and n_a, n_l, and n_r are the total numbers of years, locations, and replicates per environment, respectively. Note that Eq. [8] assumes not only a completely balanced dataset (e.g., that the same locations are used every year), but also independent and identically distributed effects and constancy of (the error) variance. A case like this will from now on be referred to as the simple, balanced setting. Such a simple, balanced setting at a single location can be attained either by a completely randomized design (CRD) or a randomized complete block design (RCBD) with independent genetic effects and no missing values. Whenever n_a, n_l, and/or n_r differed between genotypes due to an unbalanced design, the respective maximum value was taken.

Ad hoc H^2_{Holland}

To account for an unbalanced design where n_a, n_l, and/or n_r differ between genotypes, one may compute their harmonic means to get an approximation of the phenotypic variance as follows (Holland et al., 2003):

$$\sigma^2_{p_{\bar{n}}} = \sigma^2_g + \frac{\sigma^2_{ga}}{\overline{n}_a} + \frac{\sigma^2_{gl}}{\overline{n}_l} + \frac{\sigma^2_{gal}}{\overline{n}_{al}} + \frac{\sigma^2_{plot}}{\overline{n}_{alr}} \qquad [9]$$

with

$$\overline{n}_a = \frac{n_g}{\sum_{i=1}^{n} \frac{1}{n_{a_i}}},\ \overline{n}_l = \frac{n_g}{\sum_{i=1}^{n} \frac{1}{n_{l_i}}},\ \overline{n}_{al} = \frac{n_g}{\sum_{i=1}^{n} \frac{1}{n_{al_i}}},\ \overline{n}_{alr} = \frac{n_g}{\sum_{i=1}^{n} \frac{1}{n_{alr_i}}}$$

where n_g is the number of genotypes, n_{a_i}, n_{l_i}, n_{al_i}, n_{alr_i} are the total number of years, locations, environments, and plots per environment for the ith genotype, respectively. Thus, $\overline{n}_a$, $\overline{n}_l$, $\overline{n}_{al}$, and $\overline{n}_{alr}$ are the harmonic means for the numbers of years, locations, environments, and plots per environment across genotypes, respectively. Note that although the motivating, hypothetical MET in Holland et al. (2003, p. 65) is unbalanced, it is still assumed that $\overline{n}_e \overline{n}_r = \overline{n}_{er}$, where $\overline{n}_e$, $\overline{n}_r$, and $\overline{n}_{er}$ are the harmonic means of the number of environments, replicates, and replicates per environment, respectively. This is not the case for all METs, which is why we chose $\overline{n}_{al}$ as the divisor for σ^2_{gal}, and likewise $\overline{n}_{alr}$ for σ^2_{plot}. The corresponding H^2 estimating equation can be written as

$$H^2_{\text{Holland}} = \sigma^2_g \big/ \sigma^2_{p_{\bar{n}}} \qquad [10]$$

where σ^2_g is the genotypic variance and $\sigma^2_{p_{\bar{n}}}$ is estimated via Eq. [9].

H^2_{Piepho}

As another measure of H^2, one may compute (Holland et al., 2003; Piepho and Möhring, 2007)

$$H^2_{\text{Piepho}} = \frac{\sigma^2_g}{\sigma^2_g + \overline{v}^{\text{BLUE}}_{\Delta} \big/ 2} \qquad [11]$$

where σ^2_g is the genotypic variance and $\overline{v}^{\text{BLUE}}_{\Delta}$ is the mean variance of a difference of two genotypic BLUEs (i.e., adjusted genotype means; Table 2).

H^2_{Cullis}

Alternatively, one may compute a measure that makes use of genotypic BLUPs rather than genotypic BLUEs and thus solely rely on estimates from Eq. [5r] (Cullis et al., 2006):

$$H^2_{\text{Cullis}} = 1 - \frac{\overline{v}^{\text{BLUP}}_{\Delta}}{2\sigma^2_{\text{g}}} \qquad [12]$$

where σ^2_{g} is the genotypic variance and $\overline{v}^{\text{BLUP}}_{\Delta}$ is the mean variance of a difference of two genotypic BLUPs (Table 2).

H^2_{reg}

In the simple, balanced setting, the genotypic BLUPs and genotypic BLUEs have the following relationship (Walsh and Lynch, 2018):

$$\text{BLUP}_i = \alpha + \text{H}^2\text{BLUE}_i \qquad [13]$$

where H^2 is the heritability, α is an intercept and BLUP_i and BLUE_i are the BLUPs and BLUEs of the ith genotype main effects and genotype means, respectively (Table 2). Hence, this relationship is that of a simple linear regression with H^2 as the regression coefficient. Given the simple, balanced setting, α can be estimated by the negative of either (i) the mean of all BLUE_i multiplied with H^2, or (ii) the intercept of the model that was used to estimate BLUP_i (i.e., $\hat{\ }$) multiplied with H^2, or (iii) simply as the least squares estimate for a regression analysis based on Eq. [13]. In the simple, balanced setting, results will be identical by either method. With unbalanced data, however, the results will deviate. Nevertheless, one can make use of this relationship and perform a regression of genotypic BLUEs on genotypic BLUPs across genotypes to obtain an estimate for the regression coefficient, which then serves as a H^2 estimate. To allow for a regression through the origin, we centered the BLUE_i by subtracting $\hat{\mu}_{\text{r}}$, yielding this regression model:

$$\text{BLUP}_i = H^2_{\text{reg}}\left(\text{BLUE}_i - \hat{\mu}_{\text{r}}\right) + \varepsilon_i \qquad [14]$$

where BLUP_i, BLUE_i, and $\hat{\mu}_{\text{r}}$ are defined as above (Table 2), H^2_{reg} is the regression coefficient, and ε_i is the vector of errors of BLUP_i. One may also use this regression with unbalanced data, but it must be clear that the regression estimate of H^2_{reg} is then only an approximation of H^2. To our knowledge, this regression method has not been published in this form as an explicit method for estimating H^2, although similar ideas have been used in Yan (2014, Chapter 1.6) and DeLacy et al. (1996, Eq. [4.41]). However, this method is not unknown (Bruce Walsh [Professor, Ecology and Evolutionary Biology, University of Arizona], personal communication, 2017) and is sometimes applied in practice (Andres Gordillo [cereal breeder, KWS LOCHOW], personal communication, 2017).

H^2_{sumdiv}

As another rather simple way to compute a measure of H^2, which also exploits the shrinkage of BLUP relative to BLUE, we propose

$$H^2_{\text{sumdiv}} = \frac{\sum_{i=1}^{n_{\text{g}}}\left|\text{BLUP}_i\right|}{\sum_{i=1}^{n_{\text{g}}}\left|\text{BLUE}_i - \hat{\mu}_{\text{r}}\right|} \qquad [15]$$

where n_{g} is the number of genotypes, and BLUP_i, BLUE_i, and $\hat{\mu}_{\text{r}}$ are defined as above (Table 2). This approach is closely related to that of H^2_{reg} in Eq. [14], since all $X = \text{BLUE}_i - \hat{\mu}_{\text{r}}$ and $Y = \text{BLUP}_i$ are positive, the regression coefficient for a line through the origin is estimated as $\Sigma Y/\Sigma X$ (Snedecor and Cochran, 1980). Note that here not all X and Y necessarily have the same sign; therefore, H^2_{sumdiv} only approximates H^2_{reg}.

Simulated $H^2_{\mathbf{g}\hat{\mathbf{g}}}$

Lastly, we obtained a H^2 estimate via the simulation method presented in Piepho and Möhring (2007). This approach rests on the definition of H^2 as the squared correlation between the vectors of true ($\mathbf{g}$) and predicted ($\hat{\mathbf{g}}$) genotype effects, respectively, which we simulated from the mixed model fitted to a given dataset. Shortly put, we use $\mathbf{G}$ and $\mathbf{C}^{-}_{22}$ to set up $\boldsymbol{\Omega}$, which is the variance–covariance matrix of the joint distribution $w = (\mathbf{g}, \hat{\mathbf{g}})$ of the true simulated genotype effects $\mathbf{g}$ and their BLUP counterpart $\hat{\mathbf{g}}$. Conveniently, $\boldsymbol{\Omega}$ implicitly accounts for any experimental design effects of the MET, which makes it possible to simulate random draws of w given the same experimental design and genetic structure as was found for the given data. We ran 100,000 (= n_{sim}) simulations for each of the four datasets and subsequently calculated as many sample correlations of $\mathbf{g}$ and $\hat{\mathbf{g}}$. The simulated expected squared correlation of vectors of predicted and true genotypic effects, respectively, is then computed as

$$r^2 = n^{-1}_{\text{sim}} \sum_{q=1}^{n_{\text{sim}}} r^2_q \qquad [16]$$

where n_{sim} is the number of simulation runs and r^2_q is the squared sample correlation of $\mathbf{g}$ and $\hat{\mathbf{g}}$ of the qth simulation run. This method is not exactly identical to H^2_{Standard} in the simple, balanced setting, but it is asymptotically equivalent for an increasing number of genotypes. For example, in a balanced CRD or RCBD and with independent genotype effects, we find asymptotically that

$$H^2_{\mathbf{g}\hat{\mathbf{g}}} = E\left(r^2\right) \approx \left[\frac{\text{cov}\left(\mathbf{g}, \hat{\mathbf{g}}\right)}{\sqrt{\text{var}\left(\mathbf{g}\right)\text{var}\left(\hat{\mathbf{g}}\right)}}\right]^2 = \frac{\sigma^2_{\text{g}}}{\sigma^2_{\text{p}}} = H^2_{\text{Standard}} \qquad [17]$$

In this article, we consider $H^2_{\mathbf{g}\hat{\mathbf{g}}}$ as the gold standard, as it captures the variance–covariance structure in the data in a more comprehensive manner than the other methods. In conclusion, (i) we find $H^2_{\text{Standard}} = H^2_{\text{Holland}} = H^2_{\text{Piepho}} = H^2_{\text{Cullis}} = H^2_{\text{reg}} = H^2_{\text{sumdiv}}$ in the simple, balanced setting, and (ii) for an increasing number of genotypes, we asymptotically find $H^2_{\text{Standard}} \approx H^2_{\mathbf{g}\hat{\mathbf{g}}}$. An overview of which estimates are required for which method can be found in Table 2.

Improving H^2 Estimate Comparability between Datasets by Fixing Variance Component Sets

The H^2 estimate for any dataset mostly depends on two properties of the underlying trial: (i) the trial design and dimension, and (ii) the genetic variances. Thus, if the goal of estimating H^2 is to compare the precision of genotype comparisons between trials with different genotypes, one should fix the genetic VCs across analyses. If this is not done, a strong assumption of

equivalent genotypic variances in all trials is implicitly made. To achieve this, we took the estimates for σ^2_g, σ^2_{ga}, σ^2_{gl}, σ^2_{gal}, and σ^2_{plot} that were obtained via Eq. [5r] for each of the four datasets, respectively, and fixed them during additional analyses of the respective other three datasets. Thus, each of the four datasets was analyzed with four different VC sets, leading to a total of 16 H^2 estimates per method. Note that this is the second time we suggested fixing VC estimates between model analyses to allow for a better comparability: (i) for the analysis of any single dataset between Eq. [5r] and Eq. [5f], and (ii) between the analysis of different datasets via Eq. [5] in general.

Based on the differences of $H^2_{\mathbf{gg}}$ from H^2_{Standard}, H^2_{Holland}, H^2_{Piepho}, H^2_{Cullis}, H^2_{reg}, and H^2_{sumdiv} across all 16 dataset–VC–combinations, respectively, we calculated the mean deviation (MD), which is the arithmetic mean of the respective differences, and mean squared deviation (MSD), which is the arithmetic mean of the respective squared differences. Thus, small values for MD and MSD between, for example, $H^2_{\mathbf{gg}}$ and H^2_{Standard}, would indicate that they find similar estimates for heritability.

Degrees of Unbalancedness in Simulated Dummy Datasets

To further investigate the influence of unbalanced MET data on H^2 estimates in a way that is more systematic and not restricted by our four example datasets, we additionally simulated 27 MET datasets. These datasets follow the typical pattern of unbalancedness present in our real datasets yet display more systematic and extreme forms of imbalance: all simulated datasets represent METs across 2 yr, with 30 genotypes per year and 72 plots per genotype and year. One aspect that was varied was the manner in which the 72 plots per genotype were distributed to locations and replicates per location each year: it was either laid out as 18 locations with four replicates (18loc4rep), 24 locations with three replicates (24loc3rep), or as 36 locations with two replicates (36loc2rep). Second, the percentage of genotypes that were present in both years was set to either 100, 33, or 0%. Note that an overlap of 33% between the 2 yr means that 50% of the genotypes present in Year 1 are also present in Year 2. Furthermore, it should be clear that these three scenarios lead to different total numbers of genotypes across both years (i.e., 30, 45, and 60), whereas the number of genotypes per year is constant. Lastly, the percentage of locations that were present in both years was analogously varied between 100, 33, and 0%. We analyzed all 27 simulated datasets by applying Eq. [5r] and fixing all VCs to those obtained via applying Eq. [5r] to all four real datasets (BSA1, BSA2, PC1, and PC2). Finally, we computed H^2_{Standard} and H^2_{Cullis}. Note that we will only show results from the simulated dataset analyses using VC estimates of the PC1 dataset to avoid an unnecessary large result output. Results were similar for the analyses using VC estimates of BSA1, BSA2, and PC2. A similar argument can be made for showing results for only one alternative H^2 method, namely H^2_{Cullis}. It must be realized that the aim of this simulation is not to directly simulate true H^2 values to compare them with estimates of the alternative methods, since doing so would not allow for a meaningful outcome. The reason for this is once again the lack of an unequivocal definition for H^2 covering departures from a simple, balanced setting. Trying to simulate H^2 in the general case leads to the exact same problem as trying to define H^2 in the general case. Accordingly, this leads to a situation where any decision on how to directly simulate H^2 values will simultaneously and inevitably determine which of the alternative H^2 methods will produce the most accurate estimates. Instead, this simulation systematically varies the degree of data unbalancedness while using the VCs estimated from our real datasets.

RESULTS

VC and Plot Error

Variance component estimates are presented in Fig. 2. The purely environmental VCs (σ^2_a, σ^2_l, and σ^2_{al}) dominate, yet it should be noted that these three VCs are largely irrelevant for the comparison of genotypes and any of the evaluation criteria presented in this article. Thus, it is only the VCs involving **g** as well as the residual error variance that matter.

The genotype main and interaction effect VCs (σ^2_g, σ^2_{ga}, σ^2_{gl}, and σ^2_{gal}) show the smallest estimates. The VC estimates and their ratios for the genotype main effect, as well as the sum of all four VCs, tend to be slightly larger for the BSA datasets, especially for BSA1 (Table 3). The σ^2_{plot} estimates for all four datasets give a relatively homogenous picture and lie between those of the purely environmental VCs and those of the genotype main and interaction VCs.

Heritability Estimates

All H^2 estimates for BSA1, BSA2, PC1, and PC2 are presented in Fig. 3 and 4. Irrespective of whether the analysis is run with fixed VCs or not, H^2_{Standard} always gives the highest estimate, whereas all other estimates are more or less similar and ~10% lower than that of H^2_{Standard}. Considering the analyses without fixation of VC, H^2_{Standard} for BSA1 is the highest overall estimate, yet almost all other methods display the largest estimate for PC1 (Fig. 3).

When genetic VCs are fixed to a certain set, the trial design of PC1 always yields the highest H^2 estimate, irrespective of the method and origin of VCs (Fig. 4). When VCs are fixed to those originating from the BSA1 dataset, all H^2 estimates as a whole are ~5% higher than those for any of the other three VC sets (Fig. 4A).

We found that H^2_{Cullis} showed the smallest MSD from $H^2_{\mathbf{gg}}$ (i.e., 0.014×10^{-3}), followed by H^2_{Holland} and H^2_{Piepho} with 0.022×10^{-3} and 0.026×10^{-3}, respectively (Table 4). Regarding MD, H^2_{Holland} and H^2_{sumdiv} displayed the smallest values.

Simulated Dummy Datasets

In Fig. 5 and Supplemental Table S1, it can be seen that for all completely balanced datasets, H^2_{Standard} estimates are identical to those of H^2_{Cullis}. When the ratio of overlapping locations and/or genotypes between years is decreased and thus the unbalancedness of the data is increased, however, a discrepancy between the estimates of the two H^2 methods arises. This divergence is larger for unbalancedness in terms of genotype overlap than for location overlap.

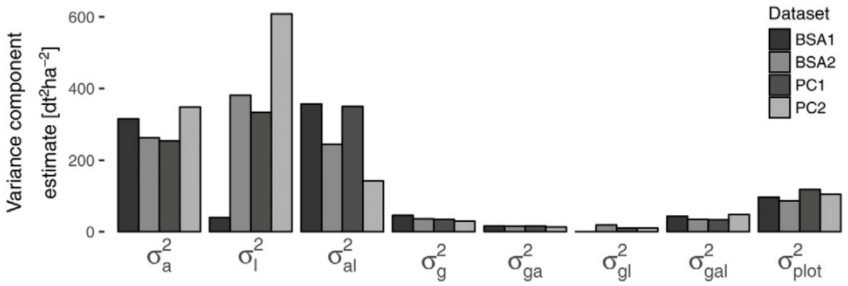

Fig. 2. Variance component (VC) estimates (in dt^2 ha^{-2}) for dry matter yield (in dt ha^{-1}) of four German cultivar evaluation trials (BSA1, BSA2, PC1, and PC2, where BSA indicates Bundessortenamt trial datasets, and PC indicates Pro-Corn trial datasets) for maize for biogas production across the years 2014 and 2015 obtained via Eq. [5r]. Variables σ^2_g, σ^2_a, σ^2_l, σ^2_{al}, σ^2_{ga}, σ^2_{gl}, σ^2_{gal}, and σ^2_{plot} represent the VCs for the genotype, year, and location main effects, for year × location, genotype × year, genotype × location, and genotype × year × location interaction effects, and error term, respectively.

Across all overlap scenarios (i.e., each of the nine location–genotype–overlap combinations), similar increases in H^2 are found when raising the number of locations per year: when shifting from 18loc4rep to 36loc2rep, $H^2_{Standard}$ and H^2_{Cullis} showed average increases of 0.0095 (minimum = 0.007, maximum = 0.013) and 0.013 (minimum = 0.011, maximum = 0.014), respectively.

DISCUSSION

The Influence of Genetic Variances on H^2 Estimates

The VC estimates relevant for the H^2 methods in this article could be grouped according to their size as $\sigma^2_{plot} > \sigma^2_g$ and $\sigma^2_{gal} > \sigma^2_{ga}$ and σ^2_{gl}, where σ^2_g and σ^2_{gal} are approximately 30 to 50% and σ^2_{ga} and σ^2_{gl} are 10 to 20% the size of σ^2_{plot}. This is in agreement with results from Laidig et al. (2008), who used a comparable analysis on data from BSA genotype trials from 1991 to 2006 on dry matter yield (dt ha^{-1}). They found similar relative sizes for named VCs not only for forage and grain maize, but also for other crops like winter oil seed rape (*Brassica napus* L.), summer and winter barley (*Hordeum vulgare* L.), sunflower (*Helianthus annuus* L.), and winter triticale (×*Triticosecale* Wittmack). Other crops like winter wheat (*Triticum aestivum* L.) or winter rye (*Secale cereal* L.) show similar or lower σ^2_{plot} compared with σ^2_g (Laidig et al., 2008). The relative sizes of VCs listed above are crucial for H^2 estimation. This becomes most visible when examining

Table 3. Selected variance component (VC) estimates relevant for estimating H^2 (in dt^2 ha^{-2}) for dry matter yield of four German cultivar evaluation trials (BSA1, BSA2, PC1, and PC2, where BSA indicates Bundessortenamt trial datasets, and PC indicates Pro-Corn trial datasets) for maize for biogas production across the years 2014 and 2015 obtained via Eq. [5r]. Variables σ^2_g, σ^2_{ga}, σ^2_{gl}, σ^2_{gal}, and σ^2_{plot} represent the VCs for the genotype main effect, the genotype × year, genotype × location, genotype × year × location interaction effects, and the (plot) error term, respectively. In parentheses are their respective relative contribution to the sum of all VCs per dataset. The remaining VC estimates for the year (σ^2_a) and location (σ^2_l) main effect, as well as for the year × location interaction effect (σ^2_{al}), are not presented.

Component	BSA1	BSA2	PC1	PC2
	dt^2 ha^{-2} (%)			
σ^2_g	46.1 (5.0)	36.4 (3.4)	34.4 (3.0)	29.4 (2.3)
σ^2_{ga}	16.6 (1.8)	15.9 (1.5)	16.4 (1.4)	13.3 (1.0)
σ^2_{gl}	0 (0.0)	18.7 (1.7)	10.2 (0.9)	10.2 (0.8)
σ^2_{gal}	42.9 (4.7)	34.5 (3.2)	33.1 (2.9)	47.9 (3.7)
Sum†	105.6 (11.5)	105.5 (9.8)	94.1 (8.2)	100.8 (7.8)
σ^2_{plot}	96.9 (10.6)	86.7 (8.0)	117.9 (10.3)	104.3 (8.0)

† Sum of σ^2_g, σ^2_{ga}, σ^2_{gl}, and σ^2_{gal}.

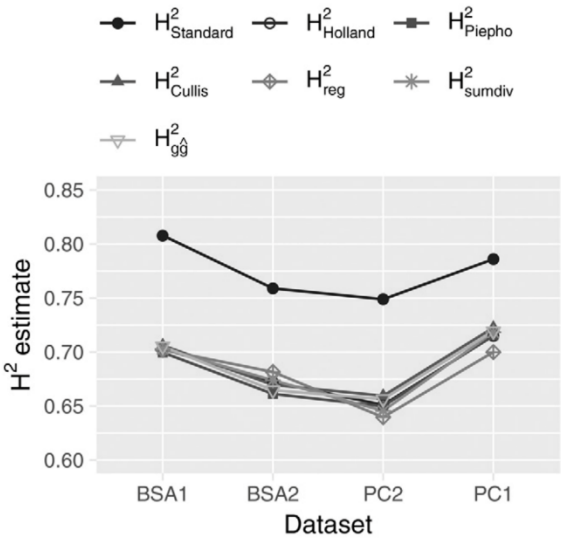

Fig. 3. Heritability (H^2) estimates for dry matter yield (dt ha^{-1}) of maize for biogas production from four German cultivar evaluation trials across the years 2014 and 2015 (BSA1, BSA2, PC1, and PC2, where BSA indicates Bundessortenamt trial datasets, and PC indicates Pro-Corn trial datasets). The seven H^2 methods are represented by different symbols and lines are drawn between estimates of the same method for easier visual identification. Datasets on the *x* axis are ordered in ascending order of the number of environments present.

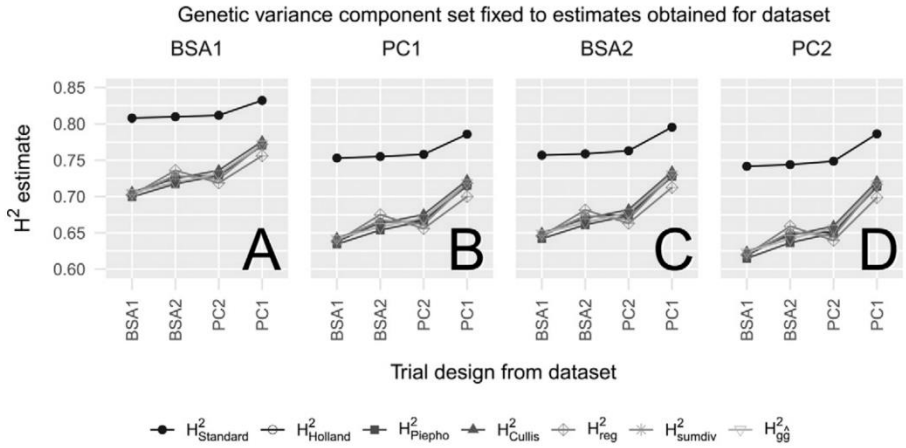

Fig. 4. Heritability (H^2) estimates for dry matter yield (dt ha^{-1}) of maize for biogas production from four German cultivar evaluation trials across the years 2014 and 2015 ([A] BSA1, [B] BSA2, [C] PC1, and [D] PC2). All estimates within Panels A, B, C, or D result from analyses (Eq. [5r] and Eq. [5f]) where the genetic variance components (σ^2_g, σ^2_{ga}, σ^2_{gl}, and σ^2_{gal}) were fixed to those originally found for the dataset written on top, respectively. The seven H^2 methods are represented by different symbols, and lines are drawn between estimates of the same method for easier visual identification. Datasets on the *x* axis are ordered in ascending order of the number of environments present. Note that those H^2 estimates, where data and fixed variance components come from the same dataset, are the same as in Fig. 3.

Eq. [8], yet it is true for all methods (e.g., for the calculation of $\overline{v}_{\Delta}^{\mathrm{BLUE}}$ needed for H^2_{Piepho}, where the computation of standard errors directly involves the VC estimates). Although it is true that the genetic VC in cultivar evaluation trials often follow a certain pattern, it must be clear that there are no two trials that are exactly identical in terms of their genetic variance estimates. Thus, any H^2 estimate will always include information about both the design of the trial and the genetic variances of the investigated genotype set. Therefore, for an appropriate comparison of different designs of trials, we suggest that VC sets should be fixed between datasets, as presented in this article. This approach seems especially advisable here, since there is almost no genotype overlap between the datasets, which among other things could have led to the different VC estimates per dataset (Table 3). Thus, by fixing VCs estimated for one dataset to the analysis of another, it is as if one genotype set was evaluated in another MET. It becomes clear how much impact this approach has on H^2 by comparing the results in Fig. 3 and 4. The results in Fig. 3 are based on analyses with their respective inherited VCs. In these analyses, we found that datasets of medium to early maturity (i.e., BSA1 and PC1) have similar H^2 estimates. Given these findings, one may conclude that the trial designs of BSA1 and PC1 are equally good and better than those of BSA2 and PC2. In contrast, in Fig. 4, it turns out that any trial design yields higher H^2 estimates when VCs are fixed to those estimated for BSA1. In other words, the BSA1 genotype set generally yields higher H^2 estimates, no matter which MET it is tested in. However, with any fixed VC set, the trial design of PC1 yields the highest H^2, whereas BSA1 is now actually lowest.

In conclusion, it can be said that within each Fig. 4A to 4D, respectively, a comparison of trial designs can be made, whereas between Fig. 4A to 4D, a rough comparison of genetic VCs is shown. Figure 3, however, only gives a convoluted picture, since the different genetic variances of each trial mask the comparison of their designs, yet an appropriate evaluation and comparison of the datasets is only possible when the genetic variances are fixed to any one set of VC estimates.

Table 4. Mean deviation (MD) and mean squared deviation (MSD) of H^2_{gg} from each of the H^2 estimation methods, respectively. Contrasted methods are sorted by their MSD.

Contrasted method	MD	MSD
H^2_{Cullis}	3.3×10^{-3}	0.014×10^{-3}
H^2_{Holland}	-0.7×10^{-3}	0.022×10^{-3}
H^2_{Piepho}	-4.6×10^{-3}	0.026×10^{-3}
H^2_{sumdiv}	-0.8×10^{-3}	0.045×10^{-3}
H^2_{reg}	-4.9×10^{-3}	0.211×10^{-3}
H^2_{Standard}	89.3×10^{-3}	8.256×10^{-3}

The Problems with H^2_{Standard}

In Fig. 3 and 4, two things stand out: (i) H^2_{Standard} is always clearly higher than H^2 estimates of all other methods, and (ii) all alternative methods yield more or less similar results. We suspect that H^2_{Standard} is overestimating H^2, while all of the alternative methods presented here seem to be able to handle this kind of unbalanced data. One reason for the potential overestimation of H^2_{Standard} is that values for n_a, n_l, and/or n_r are set to the respective maximum value present in the dataset. This means, for example, for the PC datasets,

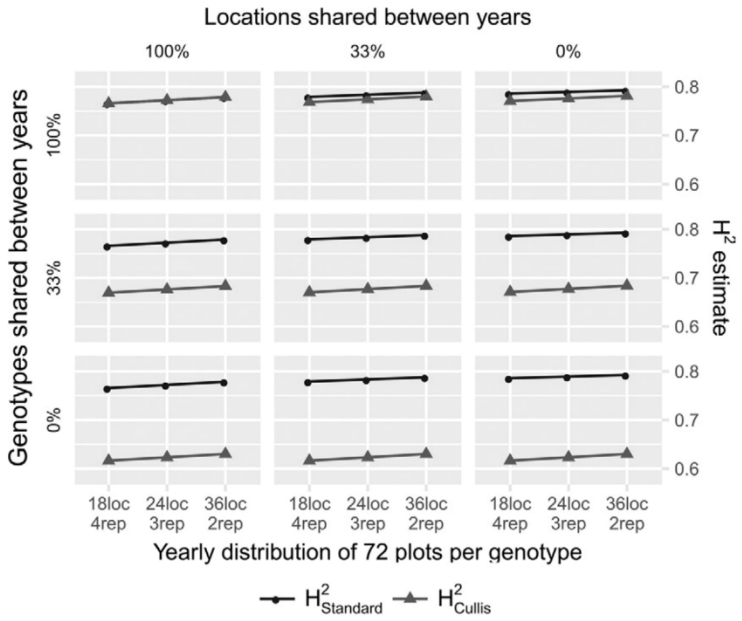

Fig. 5. Heritability (H^2) estimates of 27 simulated multienvironment trial datasets, each across 2 yr, with 30 genotypes per year and 72 plots per genotype and year, respectively. The columns and rows represent different percentages of overlap between the 2 yr regarding locations and genotypes, respectively. Note that an overlap of 33% between the 2 yr means that 50% of the locations or genotypes present in Year 1 are also present in Year 2. On the *x* axes, different distributions of the 72 plots per genotype and year to locations and replicates per locations are shown: 18 locations with four replicates (18loc4rep), 24 locations with three replicates (24loc3rep), or 36 locations with two replicates (36loc2rep). Analyses were done via Eq. [5r], fixing all variance components (VCs) to those obtained via applying the same model to the PC1 dataset. The two H^2 estimation methods are represented by different colors and symbols. Lines are drawn between estimates of the same method for easier visual identification.

where only a single environment has three instead of the otherwise two replicates, n_r is set to three and thus σ^2_{plot} is divided by a larger denominator in Eq. [8]. According to our experience, choosing n_a, n_l, and n_r in this way is indeed most common to estimate $H^2_{Standard}$. For example, it is estimated this way in PLABSTAT (Plant Breeding Statistical Program), a statistical software package written for the analysis of agronomy and plant breeding experiments (Utz, 2011).

Looking at the results of the simulation in Fig. 5, it becomes clear how $H^2_{Standard}$ and the alternative method H^2_{Cullis} are affected quite differently by systematically varying degrees of unbalancedness. The VC fixed for the analyses here were those obtained for the PC1 dataset, but the trends are similar for VCs from the other datasets (results not shown). When reducing the genotype overlap between years, $H^2_{Standard}$ is not affected at all: the estimates within each column of Fig. 5 are identical for the standard method. This is because Eq. [7] and Eq. [8] do not capture this type of unbalancedness of data or even just the number of genotypes. Since H^2_{Cullis} is based on the precision of pairwise comparisons of genotypic BLUPs, however, its estimates decrease on average by 0.151 when going from 100 to 0% genotype overlap between years. This is due to the large number of indirect comparisons required for comparing genotypes that were not grown in the same years. When the location overlap between years is reduced, the estimates for $H^2_{Standard}$ increase slightly more than those of H^2_{Cullis}. Since all VCs are fixed, the only thing changing in the estimation of the former is n_l in Eq. [8]. Taking the 18loc4rep scenario as an example, n_l is 18 in the 100% location overlap scenarios but is set to 27 in the 33% and to 36 in the 0% location overlap scenarios, respectively. As a result, the estimates for σ^2_{gl}, σ^2_{gal}, and σ^2_{plot} are divided by larger denominators in Eq. [8], decreasing the size of σ^2_p and therefore increasing the estimate for $H^2_{Standard}$ in Eq. [7].

Comparing Alternative H^2 Methods

Estimates obtained via alternative H^2 methods in this article are always noticeably smaller than those of $H^2_{Standard}$, but rather similar to each other. This suggests that given this type of unbalancedness in the data, $H^2_{Standard}$ tends to overestimate H^2, whereas the alternative methods give a better approximation. Although it is not even clear how to define the true H^2 in this context, we argue that H^2_{gg} can be considered as the gold standard, simply because the use of $\mathbf{\Omega}$ implicitly accounts for any experimental design and captures the variance–covariance structure in the data in a more comprehensive manner than the other methods. The results on the MSD from H^2_{gg} suggest that H^2_{reg} and H^2_{sumdiv} are inferior

to the already published H^2_{Cullis}, H^2_{Holland}, and H^2_{Piepho}. It should be noted, however, that all MSD (except for that for H^2_{Standard}) are relatively small, and thus it is questionable if the difference between them is actually very relevant. Thus, we recommend that any of the alternative H^2 methods presented here should be used instead of H^2_{Standard} when analyzing trials with a comparable data structure. On the one hand, it can be argued that H^2_{Piepho}, H^2_{reg}, and H^2_{sumdiv} are more cumbersome to use, as they require two models (i.e., Eq. [5f] and Eq. [5r]) with different assumptions about the genetic main effect. On the other hand, H^2_{sumdiv} and/or $H^2_{\mathbf{g\dot{g}}}$ require estimates (Table 2) that are not always easily obtained, depending on the statistical software available.

More Severely Unbalanced Datasets

As stated before, the METs analyzed here exhibit a data structure and thus type of unbalancedness that is common in cultivar evaluation trials. It should be noted, however, that these datasets are actually still relatively unproblematic in terms of the degree of unbalancedness. As can be seen in Fig. 1, it is primarily a matter of missing combinations of genotypes and environments. Within a single year, genotype–location combinations are in fact almost perfectly balanced, and regarding the PC datasets, all trials except one were laid out as RCBD with two replicates, respectively. We suspect that this is the reason why the different alternative methods give such a homogeneous picture—especially the surprisingly well-performing H^2_{Holland}. In other MET, datasets may display a much higher degree of unbalancedness due to a completely different trial design and/or genotype set for each environment, respectively. Moreover, if field trial designs are analyzed with spatial models (e.g., incomplete block designs, georeferenced data) and exploit relationship data via kinship matrices, neither the residuals nor the genotype effects are independent, even for balanced data. Furthermore, data may be analyzed via additive main effect and multiplicative interaction (AMMI) or genotype and genotype × environment interaction (GGE) models (Gauch, 1992; Hadasch et al., 2017), and/or there may be more than one genotypic main effect due to the use of so-called check cultivars. It is not clear how the alternative H^2 estimators perform in such scenarios.

Other Precision Measures

Considering the uncertainty of how to define, estimate, and then assess H^2 in certain situations, one may raise the question what other measures could be used instead. Regarding primarily the precision of a single plot and given that the magnitudes of VCs are comparable between datasets, the plot error variance σ^2_{plot} itself is an appropriate indicator (Schmidt et al., 2018). Concerning the precision of cultivar (or treatment) comparison in the MET, $\overline{v}^{\text{BLUE}}_{\Delta}$ or $\overline{v}^{\text{BLUP}}_{\Delta}$ can be evaluated as measures in their own right. Another quality measure that is often used in the framework of genomic selection is the prediction error obtained via cross validation.

It is obvious from the published literature that H^2 has played and still plays a dominant role, especially in breeding trials. It is therefore unlikely to be replaced. After all, this is what motivated this article: when in practice H^2 estimates are required even though data structures are often inappropriate for standard estimators, alternative H^2 estimators are preferred.

Comparing the Four Datasets

Although literature on METs is available (Smith et al., 2005), there are ongoing debates about several aspects, such as the experimental design (Yan et al., 2002; Schmidt et al., 2018), the number of replicates per environment (Talbot, 1984; Möhring et al., 2014), and the number of environments and their distribution to locations and years (Yan et al., 2015; Kleinknecht et al., 2016). As a consequence, layouts of METs vary between experimenters. One particular difference between the datasets investigated here were the number of genotypes, environments, and replicates per environment, respectively. In general, the aim of the PC trials was to have fewer replicates and more environments, which mostly came to fruition for PC1. Although the BSA datasets always had three replicates at <15 locations per year, they also needed to evaluate more genotypes, especially BSA1. Furthermore, it must be realized that since the genotypes in PC trials are already registered, whereas those tested in BSA trials are not, the variation in the genotypes is expected to be much higher for the latter. It was indeed the case that the estimates for σ^2_g as well as the sum of all genetic VCs were larger for the BSA datasets. On the whole, results suggest that PC1—the dataset with the most environments—had the highest H^2 estimates across all fixed VC sets. On another note, one must keep in mind that although the dry matter yield analyzed here may be the most important trait in European breeding for silage maize, it is certainly not the only relevant trait when it comes to a holistic cultivar evaluation. In the context of maize for biogas production, it is obvious that other traits of a cultivar, such as the specific biogas yield, play an important role (Rath et al., 2013). After all, it must always be remembered that a trial's design depends on the respective research question, and it is not the goal of this article to present a resource allocation strategy. Accordingly, there may be other points to consider in the decision-making process than just efficiency in terms of H^2, which could ultimately lead to using, for example, a third or even just a single replicate in a MET.

CONCLUSION

In this article, we compared the estimates for broad-sense heritability on an entry-mean basis computed via the standard method with those obtained via six alternative estimation methods using four unbalanced MET datasets. Although the standard method appears to overestimate heritability, all of the alternative methods presented here show similar estimates and thus seem to be able to handle this kind of unbalanced

data. Finally, we show that to compare heritability estimates between different MET, genetic VCs should be fixed to the same value for both METs.

Conflict of Interest

The authors declare that there is no conflict of interest.

Supplemental Material Available

Supplemental material is available online for this article.

Acknowledgments

This research was funded by the German Research Foundation (DFG Grant PI 377/18-1). We thank the Bundessortenamt (Hannover, Germany) for kindly providing the BSA1 and BSA2 datasets.

References

Cullis, B.R., A.B. Smith, and N.E. Coombes. 2006. On the design of early generation variety trials with correlated data. J. Agric. Biol. Environ. Stat. 11:381–393. doi:10.1198/108571106X154443

Cullis, B.R., F.M. Thomson, J.A. Fisher, A.R. Gilmour, and R. Thompson. 1996. The analysis of the NSW wheat variety database. I. Modelling trial error variance. Theor. Appl. Genet. 92:21–27. doi:10.1007/BF00222947

Damesa, T.M., J. Möhring, M. Worku, and H.-P. Piepho. 2017. One step at a time: Stage-wise analysis of a series of experiments. Agron. J. 109:845–857. doi:10.2134/agronj2016.07.0395

DeLacy, I.H., K.E. Basford, M. Cooper, J.K. Bull, and C.G. McLaren. 1996. Analysis of multi-environment trials: An historical perspective. In: M. Cooper and G.L. Hammer, editors, Plant adaptation and crop improvement. CABI, Wallingford, UK. p. 39–124.

Falconer, D.S., and T.F.C. Mackay. 2005. Introduction to quantitative genetics. 4th ed. Pearson Prentice Hall, Upper Saddle River, NJ.

Frensham, A.B., A.R. Barr, B.R. Cullis, and S.D. Pelham. 1998. A mixed model analysis of 10 years of oat evaluation data: Use of agronomic information to explain genotype by environment interaction. Euphytica 99:43–56. doi:10.1023/A:1018395731621

Gauch, H.G. 1992. Statistical analysis of regional yield trials: AMMI analysis of factorial designs. Elsevier, Amsterdam.

Gilmour, A.R., B.J. Gogel, B.R. Cullis, and R. Thompson. 2009. ASReml user guide Release 3.0. VSN Int., Hemel Hempstead, UK.

Hadasch, S., J. Forkman, and H.-P. Piepho. 2017. Cross-validation in AMMI and GGE models: A comparison of methods. Crop Sci. 57:264–274. doi:10.2135/cropsci2016.07.0613

Helms, T.C., and J.J. Hammond. 2006. Genetic gain equation with correlated genotype × environment effects. Crop Sci. 46:1137–1142. doi:10.2135/cropsci2005.07-0212

Holland, J.B., W.E. Nyquist, and C.T. Cervantes-Martínez. 2003. Estimating and interpreting heritability for plant breeding: An update. Plant Breed. Rev. 2003:9–112. doi:10.1002/9780470650202.ch2

Kleinknecht, K., J. Möhring, F. Laidig, U. Meyer, and H.P. Piepho. 2016. A simulation-based approach for evaluating the efficiency of multienvironment trial designs. Crop Sci. 56:2237–2250. doi:10.2135/cropsci2015.07.0405

Laidig, F., T. Drobek, and U. Meyer. 2008. Genotypic and environmental variability of yield for cultivars from 30 different crops in German official variety trials. Plant Breed. 127:541–547. doi:10.1111/j.1439-0523.2008.01564.x

Laloë, D. 1993. Precision and information in linear models of genetic evaluation. Genet. Sel. Evol. 25:557. doi:10.1186/1297-9686-25-6-557

McLean, R.A., W.L. Sanders, and W.W. Stroup. 1991. A unified approach to mixed linear models. Am. Stat. 45:54–64. doi:10.1080/00031305.1991.10475767

Möhring, J., and H.-P. Piepho. 2009. Comparison of weighting in two-stage analysis of plant breeding trials. Crop Sci. 49:1977–1988. doi:10.2135/cropsci2009.02.0083

Möhring, J., E. Williams, and H.-P. Piepho. 2015. Inter-block information: To recover or not to recover it? Theor. Appl. Genet. 128:1541–1554. doi:10.1007/s00122-015-2530-0

Möhring, J., E.R. Williams, and H.-P. Piepho. 2014. Efficiency of augmented p-rep designs in multi-environmental trials. Theor. Appl. Genet. 127:1049–1060. doi:10.1007/s00122-014-2278-y

Oakey, H., A. Verbyla, W. Pitchford, B. Cullis, and H. Kuchel. 2006. Joint modeling of additive and non-additive genetic line effects in single field trials. Theor. Appl. Genet. 113:809–819. doi:10.1007/s00122-006-0333-z

Piepho, H.-P., and J. Möhring. 2007. Computing heritability and selection response from unbalanced plant breeding trials. Genetics 177:1881–1888. doi:10.1534/genetics.107.074229

Piepho, H.-P., J. Möhring, T. Schulz-Streeck, and J.O. Ogutu. 2012. A stage-wise approach for the analysis of multi-environment trials. Biom. J. 54:844–860. doi:10.1002/bimj.201100219

Rath, J., H. Heuwinkel, and A. Herrmann. 2013. Specific biogas yield of maize can be predicted by the interaction of four biochemical constituents. BioEnergy Res. 6:939–952. doi:10.1007/s12155-013-9318-3

SAS Institute. 2013. Base SAS 9.4 procedures guide: Statistical procedures. 2nd ed. SAS Institute, Cary, NC.

Schmidt, P., J. Möhring, R.J. Koch, and H.-P. Piepho. 2018. More, larger, simpler: How comparable are on-farm and on-station trials for cultivar evaluation? Crop Sci. 58:1508–1518. doi:10.2135/cropsci2017.09.0555

Smith, A., B. Cullis, and A. Gilmour. 2001. The analysis of crop variety evaluation data in Australia. Aust. N. Z. J. Stat. 43:129–145. doi:10.1111/1467-842X.00163

Smith, A., B. Cullis, and R. Thompson. 2005. The analysis of crop cultivar breeding and evaluation trials: An overview of current mixed model approaches. J. Agric. Sci. 143:449–462. doi:10.1017/S0021859605005587

Snedecor, G.W., and W.G. Cochran. 1980. Statistical methods. 7th ed. Iowa State Univ. Press, Ames, IA.

Talbot, M. 1984. Yield variability of crop varieties in the U.K. J. Agric. Sci. 102:315–321. doi:10.1017/S0021859600042635

Utz, H.F. 2011. PLABSTAT: A computer program for statistical analysis of plant breeding experiments. Version 3A. Inst. Plant Breed., Seed Sci. Population Genet., Univ. Hohenheim, Stuttgart, Germany.

Walsh, B., and M. Lynch. 2018. Evolution and selection of quantitative traits. 1st ed. Oxford Univ. Press, Oxford, UK. doi:10.1093/oso/9780198830870.001.0001

Welham, S.J., B.J. Gogel, A.B. Smith, R. Thompson, and B.R. Cullis. 2010. A comparison of analysis methods for late-stage variety evaluation trials. Aust. N. Z. J. Stat. 52:125–149. doi:10.1111/j.1467-842X.2010.00570.x

Yan, W. 2014. Crop variety trials: Data management and analysis. John Wiley & Sons, Chichester, UK. doi:10.1002/9781118688571

Yan, W., J. Frégeau-Reid, R. Martin, D. Pageau, and J. Mitchell-Fetch. 2015. How many test locations and replications are needed in crop variety trials for a target region? Euphytica 202:361–372. doi:10.1007/s10681-014-1253-7

Yan, W., L.A. Hunt, P. Johnson, G. Stewart, and X. Lu. 2002. On-farm strip trials vs. replicated performance trials for cultivar evaluation. Crop Sci. 42:385–392. doi:10.2135/cropsci2002.3850

4. Heritability in plant breeding on a genotype-difference basis

P. Schmidt[1], J. Hartung[1], J. Bennewitz[2] and H.-P. Piepho[1]*

[1] Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstrasse 23, 70599 Stuttgart, Germany;

[2] Institute of Animal Science, University of Hohenheim, Garbenstrasse 17, 70599 Stuttgart, Germany;

Abstract

In plant breeding, heritability is often calculated (i) as a measure of precision of trials and/or (ii) to compute the response to selection. It is usually estimated on an entry-mean basis, since *the phenotype* is usually an aggregated value, as genotypes are replicated in trials, which stands in contrast with animal breeding and human genetics. When this was first proposed, assumptions such as balanced data and independent genotypic effects were made that are often violated in modern plant breeding trials/analyses. Due to this, multiple alternative methods have been proposed, aiming to generalize heritability on an entry-mean basis. In this study, we propose an extension of the concept for heritability on an entry-mean to an entry-difference basis, which allows for a more detailed insight and is more meaningful in the context of selection in plant breeding, because the correlation among entry means can be accounted for. We show that under certain circumstances our method reduces to other popular generalized methods for heritability estimation on an entry-mean basis. The approach is exemplified via four examples that show different levels of complexity, where we compare six methods for heritability estimation on an entry-mean basis to our approach (example codes: https://github.com/PaulSchmidtGit/Heritability). Results suggest that heritability on an entry-difference basis is a well-suited alternative for obtaining an overall heritability estimate and in addition provides one heritability per genotype as well as one per difference between genotypes.

Status: published

GENETICS | INVESTIGATION

Heritability in Plant Breeding on a Genotype-Difference Basis

Paul Schmidt,* Jens Hartung,* Jörn Bennewitz,† and Hans-Peter Piepho*,1
*Biostatistics Unit, Institute of Crop Science and †Institute of Animal Science, University of Hohenheim, Stuttgart, 70599 ,Germany
ORCID ID: 0000-0003-1528-2082 (P.S.)

ABSTRACT In plant breeding, heritability is often calculated (i) as a measure of precision of trials and/or (ii) to compute the response to selection. It is usually estimated on an entry-mean basis, since *the phenotype* is usually an aggregated value, as genotypes are replicated in trials, which stands in contrast with animal breeding and human genetics. When this was first proposed, assumptions such as balanced data and independent genotypic effects were made that are often violated in modern plant breeding trials/analyses. Due to this, multiple alternative methods have been proposed, aiming to generalize heritability on an entry-mean basis. In this study, we propose an extension of the concept for heritability on an entry-mean to an entry-difference basis, which allows for more detailed insight and is more meaningful in the context of selection in plant breeding, because the correlation among entry means can be accounted for. We show that under certain circumstances our method reduces to other popular generalized methods for heritability estimation on an entry-mean basis. The approach is exemplified via four examples that show different levels of complexity, where we compare six methods for heritability estimation on an entry-mean basis to our approach (example codes: https://github.com/PaulSchmidtGit/Heritability). Results suggest that heritability on an entry-difference basis is a well-suited alternative for obtaining an overall heritability estimate, and in addition provides one heritability per genotype as well as one per difference between genotypes.

KEYWORDS heritability; plant breeding; mixed models

THE idea behind measures of heritability as used in plant breeding is relatively simple: they express the proportion of the total phenotypic variance that is attributable to the average effects of genes, which in turn determines the degree of resemblance between relatives (Falconer and Mackay 2005, chapter 10). Lourenço *et al.* (2017) phrase it as "the extent to which a phenotype is genetically determined." A phenotype is the composite of an organism's observable traits. It results from (i) the expression of the organism's genotype, (ii) the influence of environmental factors, and (iii) the interactions between both. Thus, heritability investigates the relationship between observed/phenotypic values with phenotypic variance σ_p^2 and their respective underlying true genotypic values $(\boldsymbol{g})$ with genotypic variance σ_g^2. We can furthermore dissect $\boldsymbol{g}$ and σ_g^2 into additive, dominance, and epistasis components to extract average effects of alleles and breeding values $(\boldsymbol{a})$, with variance σ_a^2. Depending on whether genotypic values or breeding values are considered, we refer to broad-sense heritability (H^2) or narrow-sense heritability (h^2), respectively (*e.g.*, Xu 2013). Accordingly, there is a clear distinction between H^2 and h^2. However, note that the approach presented in this article is relevant for both measures, which is why we will refer to *heritability* in general throughout this article, unless it is necessary to refer to one of the two specifically. The true genotypic/breeding values (and their variances) are of course unknown, but can be estimated/predicted from phenotypic data.

Piepho and Möhring (2007) pointed out that heritability was originally proposed in the context of animal breeding where the observations are collected from individuals, and are used in mixed models to estimate genetic parameters or breeding values by utilizing the additive genetic relationship of all individuals. It is important to note that in animal breeding, each individual usually has its own and unique genotype. Hence, an *animal* and a *genotype* refer to the same thing: a single individual. Naturally, this means that

doi: https://doi.org/10.1534/genetics.119.302134
Manuscript received March 18, 2019; accepted for publication June 17, 2019; published Early Online June 27, 2019.
Supplemental material available at FigShare: https://doi.org/10.25386/genetics.7370168.
[1]Corresponding author: Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstrasse 23, Stuttgart, 70599, Germany. E-mail: piepho@uni-hohenheim.de

genotypes/animals cannot be truly replicated in or across environments. Instead, the same genotype/animal may only be measured repeatedly over time (Mrode and Thompson 2014, chapter 1.3.2) and/or via its progeny/relatives (Mrode and Thompson 2014, chapter 3.4). Repeated observations of individuals are either modeled by adding a random uncorrelated permanent environment effect to the mixed model or by treating repeated observations as different traits.

However, in plant breeding, we usually have crop cultivars/lines/varieties that are represented by a large group of individual plants with exactly the same genotype. This is mainly because most crop cultivars are clones, inbred lines, or hybrids. Thus, in contrast to animal breeding, a *genotype* in plant breeding does not refer to just a single individual, but to all individual plants belonging to the same cultivar and thus sharing the same genotype. In other words, a genotype in animal breeding is a single, genetically unique individual, whereas a genotype in plant breeding is usually a group of genetically identical individuals (though there are exceptions; see the *Discussion*). This allows for true replication of a genotype in and across multiple environments, even at a single time point and thus without repeated measurements in time. Notice that plant breeders may also decide to have repeated measures over time and/or account for observations of related genotypes, but the salient feature of plant genotypes as opposed to animal genotypes is the availability of true replication.

As a consequence, the multiple observations of the same cultivar are usually aggregated to obtain a single phenotypic value. Levels of aggregation range from individual plants to means of many plants, with the same genotype tested across several locations and years in designed experiments. The need for this aggregation in plant breeding results is an additional step that is not necessary in animal breeding. Since we do not consider analyses of individual plants in this article, the phenotypic value for a genotype will always be assumed to be some sort of mean value. In this context, heritability is referred to as heritability on an entry-mean (*i.e.*, genotype-mean) basis, which stands in contrast with heritability in animal breeding and human genetics. This difference can also be deduced from the level of the heritability, which is usually much lower for complex traits in humans and animals compared to the entry-mean heritability of complex traits in plant breeding. Nowadays, the aggregation of the multiple observations into a mean phenotypic value per genotype is usually either done by best linear unbiased estimation (BLUE) or best linear unbiased prediction (BLUP). Note that the use of BLUPs here is not the same as in the framework of breeding value estimation typical for animal breeders (see *Materials and Methods*). The choice between BLUE and BLUP in plant breeding depends on the goal of the analysis and both are commonly used in practice [see Piepho *et al.* (2008), for more information and a review on the decision-making].

In early work on measures of heritability for plant breeding trials (*e.g.*, Hanson and Robinson 1963), it was assumed that (i) the trial design is completely balanced/orthogonal, (ii) genotypic effects are independent, and (iii) variances and covariances are constant, so that the arithmetic mean across all observations for a genotype is the natural choice to aggregate to a single phenotypic value. We will from now on refer to such a scenario as the *simple*, *balanced setting*. Only in a simple, balanced setting does heritability have multiple direct interpretations. Specifically, H^2 is (I) the ratio of genetic variance to phenotypic variance, (II) the slope coefficient of a linear regression of the genotypic values on the phenotypic values [γ_{ap}], (III) the squared correlation between genotypic values and phenotypic values [r^2_{ap}], and (IV) the ratio of response to selection [ΔG] to selection differential [S] [see Falconer and Mackay (2005), pages 160 and 186]. For the simple, balanced setting, these interpretations can be represented by

$$H^2 = \overbrace{\frac{\sigma^2_g}{\sigma^2_p}}^{I} = \overbrace{\gamma_{gp}}^{II} = \overbrace{r^2_{gp}}^{III} = \overbrace{\frac{\Delta G}{S}}^{IV}, \tag{1}$$

where ΔG is often also referred to as the genetic gain and S is the mean phenotypic value of the selected genotypes, expressed as a deviation from the population mean. Note that when heritability is estimated via (1)-IV, it is referred to as the realized heritability, as it requires ΔG to be known, which means that selection has already occurred and that the offspring have been observed. By rewriting (1)-IV, we obtain the more directly interpretable *breeder's equation*:

$$\Delta G = H^2 S, \tag{2}$$

which allows the prediction of the expected response to selection and is the key reason why heritability plays such an important role in plant (and animal) breeding. Hanson and Robinson (1963, page 128) get to the heart of it: "Heritability has value primarily as a method of quantifying the concept of whether progress from selection for a plant character is relatively easy or difficult to make in a breeding program. A plant breeder, through experience, can perhaps rate a series of characters on their response to selection. Heritability gives a numerical description of this concept."

The standard method of estimating H^2 for phenotypic mean values is by plugging estimates for σ^2_g and σ^2_p into (1)-I. Given the simple, balanced setting, these variances can easily be estimated from mean squares and their respective expected mean squares of a conventional ANOVA (Yan 2014, chapter 1). For a single environment, where n_g genotypes are tested in n_r replicates in a completely randomized design (CRD), the observed data may be modeled as

$$y_{ik} = \mu + g_i + \varepsilon_{ik}, \tag{3}$$

where y_{ik} is the kth observation of the ith genotype, μ is the intercept, g_i is the effect for the ith genotype, and ε_{ik} is the plot error effect corresponding to y_{ik}. The phenotypic value of

genotype i can be obtained as the arithmetic mean across replicates, $\bar{y}_{i\cdot}$, which is also the BLUE in this simple, balanced setting. Assuming that g_i and ε_{ik} are random, with independent distribution, zero mean, and variances σ^2_g and σ^2_ε, respectively, we can calculate the phenotypic variance of $\bar{y}_{i\cdot}$ as:

$$\sigma^2_p = \sigma^2_g + \frac{\sigma^2_\varepsilon}{n_r}. \tag{4}$$

When genotypes are tested in a multienvironment trial (MET), where an environment denotes a year-by-location combination and CRDs are used at all environments, the observed data may be modeled as

$$y_{ikt} = \mu + g_i + e_t + (ge)_{it} + \varepsilon_{ikt}, \tag{5}$$

where y_{ikt} is the kth observation of the ith genotype at the tth environment, μ is the intercept, g_i is the main effect for the ith genotype, e_t is the main effect for the tth environment, $(ge)_{it}$ is the itth genotype-by-environment interaction effect, and ε_{ikt} is the plot error effect corresponding to y_{ikt}. Again, since we have a simple, balanced setting, we can obtain phenotypic values as $\bar{y}_{i\cdot\cdot} = BLUE(\mu + g_i)$. Assuming that g_i, e_t, $(ge)_{it}$, and ε_{ikt} are random, with independent distributions, zero mean, and variances σ^2_g, σ^2_e, σ^2_{ge} and σ^2_ε, respectively, one may then calculate the phenotypic variance as

$$\sigma^2_p = \sigma^2_g + \frac{\sigma^2_{ge}}{n_e} + \frac{\sigma^2_\varepsilon}{n_e n_r}, \tag{6}$$

where n_e is the number of environments (Hallauer *et al.* 2010, page 59). This approach assumes a single variance for genotype-by-environment interactions $(g \times e)$, even when multiple locations were tested across multiple years. In the latter case, one may instead model the environmental effects via separate year, and location main and interaction effects as

$$y_{ikmq} = \mu + g_i + f_m + l_q + (fl)_{mq} + (gf)_{im} + (gl)_{iq} + (gfl)_{imq} + \varepsilon_{ikmq}, \tag{7}$$

where y_{ikmq} is the kth observation of the ith genotype in the mth year and the qth location, μ is the intercept, g_i is the main effect of the ith genotype, f_m is the main effect for the mth year, l_q is the main effect for the qth location, $(fl)_{mq}$ is the mqth year-by-location interaction effect, $(gf)_{im}$ is the imth genotype-by-year interaction effect, $(gl)_{iq}$ is the iqth genotype-by-location interaction effect, $(gfl)_{imq}$ is the imqth genotype-by-year-by-location interaction effect, and ε_{ikmq} is the plot error effect corresponding to y_{ikmq}. Once more, we have the phenotypic values for each genotype as $\bar{y}_{i\cdots} = BLUE(\mu + g_i)$. Assuming $g_i, f_m, l_q, (fl)_{mq}, (gf)_{im}, (gl)_{iq}$, $(gfl)_{imq}$, and ε_{ikmq} are random, with independent distribution, zero mean, and variances σ^2_g, σ^2_f, σ^2_l, σ^2_{fl}, σ^2_{gf}, σ^2_{gl}, σ^2_{gfl}, and σ^2_ε, respectively, we can calculate the phenotypic variance as:

$$\sigma^2_p = \sigma^2_g + \frac{\sigma^2_{gl}}{n_l} + \frac{\sigma^2_{gf}}{n_f} + \frac{\sigma^2_{gfl}}{n_l n_f} + \frac{\sigma^2_\varepsilon}{n_l n_f n_r}, \tag{8}$$

where n_f and n_l are the number of years and locations, respectively (Becker 2011; Yan 2014). Note that when a MET is conducted either at multiple locations within a single year or at a single location but across multiple years, (8) simplifies to (6). We will refer to heritability estimates involving (4), (6), or (8) and (1)-I as the *standard heritability* (H^2_{Std}).

By examining (4), (6), and (8), it can be seen that the calculation of σ^2_p always involves the genotypic variance and all $g \times e$ variances. However, it does not incorporate purely environmental variance components, such as σ^2_e in the context of (6), and σ^2_f, σ^2_l, and σ^2_{fl} in the context of (8) [but see Yan (2014, page 3)]. Hence, referring to σ^2_p as the phenotypic variance may be considered misleading, as it neglects the purely environmental effects and therefore is definitely not the variance of phenotypic mean values. However, a key feature of the perspective on heritability put forward in this paper is that σ^2_p, as defined in (3)–(8), actually coincides with half the variance of a difference between two phenotypic mean values in the simple, balanced setting. Taking (5) as an example, the variance of a genotype mean is

$$var(\bar{y}_{i\cdot\cdot}) = \sigma^2_g + \frac{\sigma^2_e}{n_e} + \frac{\sigma^2_{ge}}{n_e} + \frac{\sigma^2_\varepsilon}{n_e n_r} \neq \sigma^2_p, \tag{9}$$

whereas the variance of a difference between means of genotypes i and j is $var(\bar{y}_{i\cdot\cdot}) + var(\bar{y}_{j\cdot\cdot}) - 2cov(\bar{y}_{i\cdot\cdot}, \bar{y}_{j\cdot\cdot})$, which in this simple, balanced setting reduces to

$$var(\bar{y}_{i\cdot\cdot} - \bar{y}_{j\cdot\cdot}) = 2\left(\sigma^2_g + \frac{\sigma^2_{ge}}{n_e} + \frac{\sigma^2_\varepsilon}{n_e n_r}\right) = 2\sigma^2_p. \tag{10}$$

It can be argued that this focus on genotype "differences" makes more sense than a focus on genotype means or effects themselves, since the goal is to select the best-performing genotype(s) to maximize ΔG and the ranking of genotypes is uniquely determined by all pairwise differences, whereas the individual genotypic effects and the absolute (mean) performance level as such do not inform about the ranking (Searle *et al.* 1992; Piepho *et al.* 2008). Accordingly, the correct ranking of genotypes and thus the precision of estimating genotype differences is more relevant than the precision of the genotype effect estimates. Furthermore, we found that especially in some older publications, the definitions of heritability refer to differences: Knight (1948) defines heritability as "the portion of the observed variance for which differences in heredity are responsible," while Hanson and Robinson (1963, page 125) state that the "concept of heritability originated as an attempt to describe whether differences actually observed between individuals arose from the differences in genetic makeup between the individuals or

resulted from different environmental forces." Finally, it is no coincidence that two published alternative heritability estimation methods also involve the variance of a difference between genotypes: Cullis *et al.* (2006) proposed to calculate heritability as

$$H^2_{Cullis} = 1 - \frac{\bar{v}^{BLUP}_{\Delta\cdot\cdot}}{2\sigma^2_g}, \tag{11}$$

where $\bar{v}^{BLUP}_{\Delta\cdot\cdot}$ is the mean variance of a difference of two BLUPs for the genotypic effect and σ^2_g is the genotypic variance. The BLUE counterpart was suggested by Piepho and Möhring (2007) as

$$H^2_{Piepho} = \frac{\sigma^2_g}{\sigma^2_g + \bar{v}^{BLUE}_{\Delta\cdot\cdot}/2}, \tag{12}$$

where $\bar{v}^{BLUE}_{\Delta\cdot\cdot}$ is the mean variance of a difference of two genotypic BLUEs and σ^2_g is the genotypic variance. Both measures reduce to H^2_{Std} in the simple, balanced setting.

These considerations suggest that heritability is best defined in terms of pairwise comparisons among genotypes, and this is in fact the key perspective taken in this paper. We aim to provide two things. First, an elaboration of how a heritability on an entry-difference basis (H^2_Δ/h^2_Δ) can be defined. Second, a comparison of how the application of H^2_Δ/h^2_Δ compares to other published methods for estimating heritability on an entry-mean basis. To do so, we give a derivation of H^2_Δ and subsequently apply it alongside six other heritability measures in four example analyses with real data.

Materials and Methods

Mixed model theory

We assume a standard linear mixed model for the observed data vector $\boldsymbol{y}$, which is of the form

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon}, \tag{13}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are vectors of fixed and random effects, respectively, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the associated design matrices, and $\boldsymbol{\varepsilon}$ is a vector of random residual errors. The random effects $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ are assumed to be independently distributed as $\boldsymbol{u} \sim MVN(0, \boldsymbol{G})$ and $\boldsymbol{\varepsilon} \sim MVN(0, \boldsymbol{R})$, such that $\boldsymbol{y} \sim MVN(\boldsymbol{X\beta}, \boldsymbol{V})$, where $MVN(.,.)$ denotes the multivariate normal distribution with a mean vector given as the first argument and a variance–covariance matrix as the second. To obtain estimates for $\boldsymbol{\beta}$ as well as predictions for $\boldsymbol{u}$, the mixed model equations (Henderson 1986; Searle *et al.* 1992) can be solved as

$$\begin{bmatrix} BLUE(\boldsymbol{\beta}) \\ BLUP(\boldsymbol{u}) \end{bmatrix} = \boldsymbol{C} \begin{bmatrix} \boldsymbol{X'R^{-1}y} \\ \boldsymbol{Z'R^{-1}y} \end{bmatrix}, \tag{14}$$

where

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'R^{-1}X} & \boldsymbol{X'R^{-1}Z} \\ \boldsymbol{Z'R^{-1}X} & \boldsymbol{Z'R^{-1}Z} + \boldsymbol{G}^{-1} \end{bmatrix}^{-}. \tag{15}$$

Note that we will explicitly use the words "estimate/estimator" and "predict/predictor" for BLUEs and BLUPs, respectively. In the context of variety trials, genotype effects $\boldsymbol{g}$ may be considered to be fixed or random. The choice between these two options depends on the goal of the analysis and both are commonly used in practice [see Piepho *et al.* (2008) for a review on decision-making as well as Appendix A]. Accordingly, we consider both as complementing cases in this article.

If $\boldsymbol{g}$ is assumed to be a fixed effect, we obtain genotypic BLUEs ($\hat{\boldsymbol{g}}^{BLUE}$) as a subset of $BLUE(\boldsymbol{\beta})$ with $var(\hat{\boldsymbol{g}}^{BLUE}|\boldsymbol{g}) = \boldsymbol{C}_{11(g)}$ as the submatrix of $\boldsymbol{C}_{11}$ associated with $\boldsymbol{g}$. If $\boldsymbol{g}$ is assumed to be a random effect, we obtain genotypic BLUPs ($\hat{\boldsymbol{g}}^{BLUP}$) as a subset of $BLUP(\boldsymbol{u})$ with $var(\hat{\boldsymbol{g}}^{BLUP}|\boldsymbol{g}) = \boldsymbol{C}_{22(g)}$ as the submatrix of $\boldsymbol{C}_{22}$ associated with $\boldsymbol{g}$. Note that in the case where $\boldsymbol{g}$ is the only fixed/random effect in the model, its estimators/predictors and their variance matrices are not subsets of, but equal to, $BLUE(\boldsymbol{\beta})/BLUP(\boldsymbol{u})$ and $\boldsymbol{C}_{11}/\boldsymbol{C}_{22}$, respectively. Notice further that BLUPs can always be computed from entry means (BLUEs) using a stage-wise approach (see Appendix A). Only in the simple, balanced setting is the correlation of genotypic BLUEs and BLUPs equal to unity, and thus the ranking of genotypes does not change. On a side note, it should be mentioned that one may want to use linear combinations of $BLUE(\boldsymbol{\beta})$, such as, *e.g.*, *least squares means* (SAS Institute Inc. 2013) or *estimated marginal means* (Searle *et al.* 2012) as phenotypic mean values instead. While this certainly results in different estimated values and (co)variances, it does not affect the heritability measures proposed in this article. This is because for all methods that rely on $\hat{\boldsymbol{g}}^{BLUE}$, the focus lies on differences, and any additional variance brought in via the linear combination with nongenetic effects cancels out when computing differences.

Example data sets

To exemplify the approach proposed in this article, we use four examples that show different levels of complexity. In examples 1 and 2, a single-environment data set (Wright 2017) is analyzed, while in example 3 a subset (*i.e.*, a single environment) of the MET data set in example 4 (Hadasch *et al.* 2016) is analyzed.

Example 1: The R-package *agridat* (Wright 2017) provides yield data from an oat trial at a single environment. The trial had 24 genotypes and was laid out as an α-design, with three complete replicates and six incomplete blocks of size four within each replicate. Yet, in this first example, we will analyze the data as if the trial were laid out as CRD, and thus ignore the information about the complete replicates and incomplete blocks. Therefore, the model is simply

$$y_{ik} = \mu + g_i + \varepsilon_{ik}, \tag{16}$$

where y_{ik} is the kth observation of the ith genotype, μ is the intercept, g_i is the effect for the ith genotype, and ε_{ik} is the plot error effect corresponding to y_{ik} with $var(\boldsymbol{\varepsilon}) = \boldsymbol{I}_{72}\sigma^2_\varepsilon$. Accordingly, this example is a simple, balanced setting.

Example 2: Here, we analyzed the same data set as in example 1, but included effects accounting for the trial's design in the model

$$y_{iko} = \mu + g_i + r_k + b_{ko} + \varepsilon_{iko}, \tag{17}$$

where y_{iko} is the observation of the ith genotype in the oth block of the kth replicate, μ is the intercept, g_i is the effect for the ith genotype, r_k is the effect for the kth replicate, b_{ko} is the effect for the oth block in the kth replicate, and e_{iko} is the plot error effect corresponding to y_{iko} with $var(\boldsymbol{\varepsilon}) = \boldsymbol{I}_{72}\sigma^2_{\varepsilon}$. To recover interblock information, b_{ko} was modeled as random with $var(\boldsymbol{b}) = \boldsymbol{I}_{18}\sigma^2_b$, whereas r_k was taken as fixed.

Example 3: This data set was taken from Hadasch *et al.* (2016). It comprises plot data of 89 lettuce varieties (*i.e.*, "geno1"–"geno89") tested at three environments (*i.e.*, "env1"–"env3"), each laid out as a randomized complete block design (RCBD). The measured trait was resistance to downy mildew scored on a scale ranging from 0 to 5. We note that despite the acknowledged discreteness of this response variable, residual plots did not indicate an appreciable deviation of residuals from the normal distribution. All 89 varieties were genotyped with a total of 300 markers [*i.e.*, 95 single nucleotide polymorphisms and 205 amplified fragment length polymorphism markers, see Hayes *et al.* (2014) for details] so that a marker matrix $\boldsymbol{M}_{89\times 300}$ was provided. The biallelic marker M_{iw} for the ith genotype, and the wth marker with alleles A_1 (*i.e.*, the reference allele) and A_2, was coded as 1 for A_1A_1, -1 for A_2A_2, and 0 for A_1A_2 and A_2A_1. The kinship matrix $\boldsymbol{K}$ was obtained as

$$\boldsymbol{K} = \boldsymbol{M}\boldsymbol{M}'. \tag{18}$$

We note that for our purposes, the scaling of $\boldsymbol{M}$ is largely immaterial, but see Appendix D for different scaling options of $\boldsymbol{M}$ and the associated interpretations of genetic variance. In this example, we analyzed only the second environment ("env2") and thus have 89 genotypes at a single environment laid out as RCBD with three replicates, which we modeled as

$$y_{ik} = \mu + a_i + r_k + \varepsilon_{ik}, \tag{19}$$

where y_{ik} is the observation of the ith genotype in the kth replicate, μ is the intercept, a_i is the additive effect for the ith genotype, r_k is the effect for the kth replicate, and ε_{ik} is the plot error effect corresponding to y_{ik} with $var(\boldsymbol{\varepsilon}) = \boldsymbol{I}_{267}\sigma^2_{\varepsilon}$. For simplicity, we assume here that there are no genetic effects for dominance or epistasis, but an extension of the model to cover this case is straightforward (Wellmann and Bennewitz 2011; Dias *et al.* 2018; Viana *et al.* 2018).

Example 4: In contrast to example 3, we here analyzed all three environments jointly. The data set is not completely balanced, as (i) environment env1 has two complete replicates, while environments env2 and env3 have three, (ii) there are no observations for genotypes geno38 and geno49 at environment env1, and (iii) there are no observations for genotype geno81 at either env1 or env3. The linear mixed model used for this analysis was

$$y_{ikt} = \mu + a_i + e_t + (ge)_{it} + r_{kt} + \varepsilon_{ikt}, \tag{20}$$

where y_{ikt} is the observation of the ith genotype in the kth replicate at the tth environment, μ is the intercept, a_i is the additive effect for the ith genotype, e_t is the effect for the tth environment, $(ge)_{it}$ is the itth genotype-by-environment interaction, r_{kt} is the effect for the kth replicate at the tth environment, and ε_{ikt} is the plot error effect corresponding to y_{ikt}. We allowed for heterogeneous error variances between environments, assuming $var(\boldsymbol{\varepsilon}) = \oplus_{t=1}^{n_e}\boldsymbol{I}_{\boldsymbol{n_{obs(t)}}}\sigma^2_{\varepsilon_t}$, where $\oplus$ is the direct sum operator, and $n_{obs(t)}$ and $\sigma^2_{\varepsilon_t}$ are the number of observations and plot error variance at the tth environment, respectively. Finally, e_t and r_k were taken as fixed effects, while $(ge)_{it}$ was taken as random with $var(\boldsymbol{ge}) = \boldsymbol{I}_{263}\sigma^2_{ge}$.

Modeling of genotypic main effect in all examples: Note that some of the heritability measures used in this article require estimates from both the model with a fixed genotype main effect as well as the model with a random genotype main effect. For example, H^2_{Piepho} in (12) needs an estimate for $\bar{v}^{BLUE}_{\Delta\cdot\cdot}$ as well as for σ^2_g. Thus, in this article we always fit both potential models in parallel, and accordingly obtain both $\hat{\boldsymbol{g}}^{BLUE}$ and $\hat{\boldsymbol{g}}^{BLUP}$ for the same example. For models where $\boldsymbol{g}$ is assumed to be a random effect, we assume $var(\boldsymbol{g}) = \boldsymbol{I}_{\boldsymbol{n_g}}\sigma^2_g$. Additionally, we make use of the marker information in examples 3 and 4, and thus estimate additive genotypic effects $\boldsymbol{a}$ by assuming $var(\boldsymbol{a}) = \boldsymbol{K}\sigma^2_a$.

Further, notice that for each example the variance components in the model with the fixed genotype main effect are fixed to the corresponding estimates in the model where $\boldsymbol{g}$ is assumed to be a random effect. This is done to rule out the potential influence of an estimation error on variance component estimation between the two models (Schmidt *et al.* 2019).

Pairwise heritability

As shown in (1), heritability can be interpreted and thus estimated in multiple ways. To derive a *pairwise heritability* or *heritability on an entry-difference basis* ($H^2_{\Delta}/h^2_{\Delta}$), we can exploit and slightly alter (1)-III so that it applies to phenotypic and genotypic differences instead of their respective mean values. We here first derive H^2_{Δ} under the assumption of independent genotypic effects with constant variance and afterward generalize H^2_{Δ} to correlated genotypic effects with nonconstant (co)variances. For both scenarios, we differentiate between $\hat{\boldsymbol{g}}^{BLUE}$ and $\hat{\boldsymbol{g}}^{BLUP}$.

Independent genotypic effects with constant variance: Given independent genotypic effects with constant variance σ^2_g (as in the simple, balanced setting), and irrespective of

whether $\hat{\boldsymbol{g}}^{BLUE}$ or $\hat{\boldsymbol{g}}^{BLUP}$ is used, the variance of the true difference between genotypes i and j is

$$var(g_i - g_j) = 2\sigma_g^2. \tag{21}$$

As stated before, the conditional variance, or prediction error variance, of the predicted BLUPs given the observed data is $var(\hat{\boldsymbol{g}}^{BLUP}|\boldsymbol{g}) = \boldsymbol{C}_{22(\boldsymbol{g})}$. However, notice that the marginal variance of $\hat{\boldsymbol{g}}^{BLUP}$is $var(\hat{\boldsymbol{g}}^{BLUP}) = \boldsymbol{G}_{(\boldsymbol{g})} - \boldsymbol{C}_{22(\boldsymbol{g})}$, where $\boldsymbol{G}_{(\boldsymbol{g})}$ is the submatrix of $\boldsymbol{G}$ associated with $\boldsymbol{g}$ (see Searle *et al.* 1992, chapter 7.4.d.). As a result, we find that the variance of a difference between two BLUPs of genotypes i and j is

$$var\left(\hat{g}_i^{BLUP} - \hat{g}_j^{BLUP}\right) = 2\sigma_g^2 - v_{\Delta ij}^{BLUP}, \tag{22}$$

where $v_{\Delta ij}^{BLUP}$ is the prediction error variance of a difference between BLUPs of genotypes i and j, and can be obtained via $\boldsymbol{C}_{22(\boldsymbol{g})}$. Furthermore, we find the covariance to be

$$cov\left(g_i - g_j, \hat{g}_i^{BLUP} - \hat{g}_j^{BLUP}\right) = 2\sigma_g^2 - v_{\Delta ij}^{BLUP}. \tag{23}$$

Hence, making use of (21), (22), and (23), the squared correlation between the true difference and its predictor, which is the heritability of the predictor of the difference $(g_i - g_j)$, is found to be

$$H_{\Delta ij_{BLUP}}^2 = \left(\frac{2\sigma_g^2 - v_{\Delta ij}^{BLUP}}{\sqrt{2\sigma_g^2\left(2\sigma_g^2 - v_{\Delta ij}^{BLUP}\right)}}\right)^2 = \frac{2\sigma_g^2 - v_{\Delta ij}^{BLUP}}{2\sigma_g^2}. \tag{24}$$

This pairwise heritability is a quantity that explicitly accounts for unbalanced data, is analogous to the coefficient of determination (CD) (Piepho 2019), and can be reported in its own right. Yet, as the number of genotypes increases, the number of genotype differences n_Δ can quickly get very large $(n_\Delta = n_g[n_g - 1]/2)$. Thus, it is not a convenient statistic to report. However, the statistic can be averaged per genotype, so that a genotype-specific average heritability is obtained, *i.e.*

$$\bar{H}_{\Delta i\cdot BLUP}^2 = \frac{1}{n_g - 1}\sum_{j \neq i} H_{\Delta ij_{BLUP}}^2. \tag{25}$$

Note that (25) is similar in spirit to the repeatability for an individual genotype's BLUP as is commonly reported in animal breeding (Laloë 1993; Mrode and Thompson 2014). Finally, if we go further and instead average (24) across all pairs of genotypes, we obtain a single average pairwise heritability as

$$\bar{H}_{\Delta\cdot\cdot BLUP}^2 = \frac{2}{n_g(n_g - 1)}\sum_i\sum_{j<i} H_{\Delta ij_{BLUP}}^2. \tag{26}$$

Hence, $\bar{H}_{\Delta\cdot\cdot BLUP}^2$ has the interpretation of an arithmetic mean of all pairwise heritabilities $H_{\Delta ij_{BLUP}}^2$ and is (given independent genotypic effects with constant variance) the algebraically equivalent simplification of H_{Cullis}^2, since the pairwise heritability in (24) has a constant denominator.

If $\hat{\boldsymbol{g}}^{BLUE}$ is used, we take the marginal variance of a difference between two BLUEs of genotypes i and j as

$$var\left(\hat{g}_i^{BLUE} - \hat{g}_j^{BLUE}\right) = 2\sigma_g^2 + v_{\Delta ij}^{BLUE}, \tag{27}$$

where $v_{\Delta ij}^{BLUE}$ is the variance of a difference between BLUEs of genotypes i and j, and can be obtained via $\boldsymbol{C}_{11(\boldsymbol{g})}$. Moreover,

$$cov\left(g_i - g_j, \hat{g}_i^{BLUE} - \hat{g}_j^{BLUE}\right) = 2\sigma_g^2. \tag{28}$$

Analogous to (24), only this time applying (21), (27), and (28), the squared correlation between the true difference and its estimator, which is the heritability of the estimator of the difference $(g_i - g_j)$, is found to be

$$H_{\Delta ij_{BLUE}}^2 = \left(\frac{2\sigma_g^2}{\sqrt{2\sigma_g^2\left(2\sigma_g^2 + v_{\Delta ij}^{BLUE}\right)}}\right)^2 = \frac{2\sigma_g^2}{2\sigma_g^2 + v_{\Delta ij}^{BLUE}}. \tag{29}$$

In accordance with (25) and (26), we could now take the arithmetic mean of $H_{\Delta ij_{BLUE}}^2$ per genotype and across all pairs. However, the inconvenient difference compared to $\bar{H}_{\Delta\cdot\cdot BLUP}^2$ in (24) is that the pairwise heritabilities in (29) do not have a constant denominator. As a result, taking the arithmetic mean across all $H_{\Delta ij_{BLUE}}^2$ does not lead to the algebraically equivalent simplification of H_{Piepho}^2. However, if we take the harmonic mean instead of the arithmetic mean for all pairwise heritabilities as

$$\bar{H}_{\Delta\cdot\cdot BLUE}^2 = \frac{n_g(n_g - 1)}{2\sum_i\sum_{i<j}\frac{1}{H_{\Delta ij_{BLUE}}^2}}, \tag{30}$$

the resulting average does indeed coincide with H_{Piepho}^2. Correspondingly, we take the harmonic mean per genotype as:

$$\bar{H}_{\Delta i\cdot BLUE}^2 = \frac{n_g - 1}{\sum_{i<j}\frac{1}{H_{\Delta ij_{BLUE}}^2}}. \tag{31}$$

Notice that since we here assume independent genotypic effects with constant variance, both $\bar{H}_{\Delta\cdot\cdot BLUP}^2$ and $\bar{H}_{\Delta\cdot\cdot BLUE}^2$ can, in fact, also be obtained by averaging the numerator and denominator in (24) and (29) separately across pairs (see Appendix B).

Generalization to correlated genotypic effects with nonconstant variances: So far, we have assumed independent genotypic effects with constant variance σ_g^2. Yet, the approach outlined above naturally extends to the case where effects have nonconstant variance, (*i.e.*, $var(g_i) = \sigma_{g(i,i)}^2$) and/or are correlated (*i.e.*, $cov(g_i, g_j) = \sigma_{g(i,j)}$), *e.g.*, due to pedigree- or marker-based kinship. The variance of the true difference between genotypes i and j then becomes

$$var(g_i - g_j) = \sigma^2_{g(i,i)} + \sigma^2_{g(j,j)} - 2\sigma_{g(i,j)}. \quad (32)$$

When BLUP is used, the pairwise heritability in (24) accordingly generalizes to

$$H^2_{\Delta ij_{BLUP}} = \frac{var(g_i - g_j) - v^{BLUP}_{\Delta ij}}{var(g_i - g_j)} = \frac{\sigma^2_{g(i,i)} + \sigma^2_{g(j,j)} - 2\sigma_{g(i,j)} - v^{BLUP}_{\Delta ij}}{\sigma^2_{g(i,i)} + \sigma^2_{g(j,j)} - 2\sigma_{g(i,j)}}. \quad (33)$$

When BLUE is used, the pairwise heritability in (29) generalizes to

$$H^2_{\Delta ij_{BLUE}} = \frac{var(g_i - g_j)}{var(g_i - g_j) + v^{BLUE}_{\Delta ij}} = \frac{\sigma^2_{g(i,i)} + \sigma^2_{g(j,j)} - 2\sigma_{g(i,j)}}{\sigma^2_{g(i,i)} + \sigma^2_{g(j,j)} - 2\sigma_{g(i,j)} + v^{BLUE}_{\Delta ij}}. \quad (34)$$

These pairwise heritabilities can be averaged as before, *i.e.*, as per Equations (25) and (26) for BLUP, and Equations (31) and (30) for BLUE, to obtain $H^2_{\Delta i\cdot}$ and $H^2_{\Delta\cdot\cdot}$. Unfortunately, the average over all pairwise heritabilities cannot generally be simplified algebraically for either BLUE or BLUP, since in both cases neither the denominator nor the numerator of the pairwise heritability is constant. But a simplification is forthcoming if, as an approximation, we average the numerator and denominator of the pairwise heritability separately (see Appendix C).

Other proposals to compute heritability

Besides H^2_{Std}, H^2_{Cullis} (11), and H^2_{Piepho} (12), we calculated three additional generalized heritability measures that are in common usage.

H^2 Oakey: An alternative measure of heritability, which is sometimes referred to as *the generalized heritability*, can be traced back to Laloë (1993) and was proposed by Oakey *et al.* (2006) in the context of plant breeding. It has gained popularity due to its ability to account for heterozygosity/covariances in $\boldsymbol{G}_{(g)}$ (*e.g.*, Mathews *et al.* 2008; Rodríguez-Álvarez *et al.* 2018). In this approach, contrasts of the true and predicted genotypic effects are defined as $\boldsymbol{c}'\boldsymbol{g}$ and $\boldsymbol{c}'\hat{\boldsymbol{g}}^{BLUP}$, respectively. Note that, unlike for H^2_{Δ} where all differences/contrasts between genotype pairs are considered, the contrast vector $\boldsymbol{c}$ can be any linear combination of genotypic effects where the elements of $\boldsymbol{c}$ sum to 0. This allows different genotypic effects to have different heritabilities, while at the same time reducing to a single scalar quantity. Similar to our derivation, they make use of (1)-III, with the difference that it applies to $\boldsymbol{c}'\boldsymbol{g}$ and $\boldsymbol{c}'\hat{\boldsymbol{g}}^{BLUP}$, and thus represents the squared correlation between the true and predicted genotypic contrasts defined by $\boldsymbol{c}$. The vector $\boldsymbol{c}$ is then chosen so that this squared correlation (*i.e.*, the heritability) is maximized. *Components of the full heritability* are defined as

$$\lambda = \left(\frac{cov(\boldsymbol{c}'\boldsymbol{g},\ \boldsymbol{c}'\hat{\boldsymbol{g}}^{BLUP})}{\sqrt{var(\boldsymbol{c}'\boldsymbol{g})var(\boldsymbol{c}'\hat{\boldsymbol{g}}^{BLUP})}} \right)^2 = \frac{\boldsymbol{c}'\boldsymbol{G}_{(g)}\left(\boldsymbol{I}_{n_g} - \boldsymbol{G}^{-1}_{(g)}\boldsymbol{C}_{22(g)}\right)\boldsymbol{c}}{\boldsymbol{c}'\boldsymbol{G}_{(g)}\boldsymbol{c}} = \frac{\boldsymbol{c}'\boldsymbol{G}_{(g)}\boldsymbol{D}\boldsymbol{c}}{\boldsymbol{c}'\boldsymbol{G}_{(g)}\boldsymbol{c}}, \quad (35)$$

where $\boldsymbol{c}$ is an eigenvector of the matrix $\boldsymbol{D} = \boldsymbol{I}_{n_g} - \boldsymbol{G}^{-1}_{(g)}\boldsymbol{C}_{22(g)}$ with constraint $\boldsymbol{c}'\boldsymbol{G}_{(g)}\boldsymbol{c} = 1$ and λ is the associated eigenvalue. The largest among the eigenvalues, λ_ζ ($\zeta = \{1, \ldots, n_\lambda\}$), equals the maximized squared correlation of $\boldsymbol{c}'\boldsymbol{g}$ and $\boldsymbol{c}'\hat{\boldsymbol{g}}^{BLUP}$. Note that $n_\lambda = n_g$ and due to constraints on $\hat{\boldsymbol{g}}^{BLUP}$, $n_z < n_g$ eigenvalues will be zero. The generalized heritability is then defined as the mean of all nonzero eigenvalues:

$$H^2_{Oakey} = \frac{\sum_{\zeta=1}^{n_g}\lambda_\zeta}{n_g - n_z} = \frac{\sum_{\zeta=n_z+1}^{n_g}\lambda_\zeta}{n_g - n_z}. \quad (36)$$

In more recent work, Rodríguez-Álvarez *et al.* (2018) show how random spatial variation in plant breeding experiments can be analyzed using tensor product P-splines, which is a two-dimensional smoothing technique. In this context, they denote the trace of $\boldsymbol{D}$ as the *effective dimension* of the genotypic effects and reexpress H^2_{Oakey} as:

$$H^2_{Oakey} = \frac{trace(\boldsymbol{D})}{n_g - n_z}. \quad (37)$$

They point out that the notion of effective dimension is well known in the smoothing context, where it can be interpreted as a complexity measure for a given model and its component effects.

Simulated: This method was proposed by Piepho and Möhring (2007), and is also based on the squared correlation between $\boldsymbol{g}$ and $\hat{\boldsymbol{g}}^{BLUP}$ in (1)-III. It first defines the variance–covariance matrix of all random effects $\boldsymbol{u}$ and the target genotypic effects $\boldsymbol{g}$ as

$$var\begin{pmatrix}\boldsymbol{g}\\ \boldsymbol{u}\end{pmatrix} = \begin{pmatrix}\boldsymbol{G}_{(g)} & \boldsymbol{U}\\ \boldsymbol{U}' & \boldsymbol{G}\end{pmatrix}, \quad (38)$$

so that $\boldsymbol{U} = cov(\boldsymbol{g}, \boldsymbol{u})$. They then define $\boldsymbol{\Omega}$ as the variance–covariance matrix of the joint distribution of the true and predicted genotypic effects:

$$\boldsymbol{\Omega} = var\begin{pmatrix}\boldsymbol{g}\\ \hat{\boldsymbol{g}}^{BLUP}\end{pmatrix} = \begin{pmatrix}\boldsymbol{G}_{(g)} & \boldsymbol{F}\boldsymbol{G}^{-1}(\boldsymbol{G} - \boldsymbol{C}_{22})\boldsymbol{G}^{-1}\boldsymbol{F}'\\ \boldsymbol{F}\boldsymbol{G}^{-1}(\boldsymbol{G} - \boldsymbol{C}_{22})\boldsymbol{G}^{-1}\boldsymbol{F}' & \boldsymbol{F}\boldsymbol{G}^{-1}(\boldsymbol{G} - \boldsymbol{C}_{22})\boldsymbol{G}^{-1}\boldsymbol{F}'\end{pmatrix}, \quad (39)$$

Table 1 Summary of heritability measures

Heritability measure			Accounts for heteroscedasticity/covariances		
Overall	Per genotype	Per pair	In $\boldsymbol{G}_{(g)}$	In $\boldsymbol{C}_{11(g)}$	In $\boldsymbol{C}_{22(g)}$
H^2_{Std}	—	—	No/no	No/no	—
$\bar{H}^2_{\Delta\cdot\cdot BLUE}$	$\bar{H}^2_{\Delta i\cdot BLUE}$	$H^2_{\Delta ij BLUE}$	Yes/yes	Yes/yes	—
H^2_{Piepho}	—	—	No/no	Yes/yes	—
$\bar{H}^2_{\Delta\cdot\cdot BLUP}$	$\bar{H}^2_{\Delta i\cdot BLUP}$	$H^2_{\Delta ij BLUP}$	Yes/yes	—	Yes/yes
H^2_{Cullis}	—	—	No/no	—	Yes/yes
H^2_{Oakey}	—	—	Yes/yes	—	Yes/yes
$\bar{r}^2_{\cdot}$	r^2_i	—	Yes/no	—	Yes/no
H^2_{sim}	—	—	Yes/yes	—	Yes/yes

where $\boldsymbol{F} = \begin{pmatrix} \boldsymbol{G}_{(g)} & \boldsymbol{U} \end{pmatrix}$. Subsequently, $\boldsymbol{\Omega}$ can be decomposed as $\boldsymbol{\Omega} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}'$ by singular value decomposition or Cholesky decomposition. Finally, values for $\boldsymbol{g}$ and $\hat{\boldsymbol{g}}^{BLUP}$ can be simulated (*i.e.*, $\boldsymbol{g}_{\boldsymbol{sim}}$ and $\hat{\boldsymbol{g}}^{\boldsymbol{BLUP}}_{\boldsymbol{sim}}$) for an experiment, with the same design and genotypic relationships as those underlying the actual data $\boldsymbol{y}$, as

$$\boldsymbol{d}_{\boldsymbol{sim}} = \begin{pmatrix} \boldsymbol{g}_{\boldsymbol{sim}} \\ \hat{\boldsymbol{g}}^{\boldsymbol{BLUP}}_{\boldsymbol{sim}} \end{pmatrix} = \boldsymbol{\Gamma}\boldsymbol{z}_{\boldsymbol{sim}}, \tag{40}$$

where $\boldsymbol{z}_{\boldsymbol{sim}}$ is a vector with $2n_g$ simulated random independent standard normal deviates. Thus, for each of the $n_{sim} = \{1, \ldots, s\}$ simulation runs, a new vector $\boldsymbol{z}_{\boldsymbol{sim}}$ of independent standard normal deviates is randomly generated, so that we can obtain $\boldsymbol{d}_{\boldsymbol{sim}}$, and compute the squared sample correlation (r^2_s) of $\boldsymbol{g}_{\boldsymbol{sim}}$ and $\hat{\boldsymbol{g}}^{\boldsymbol{BLUP}}_{\boldsymbol{sim}}$. By running a large number of simulations, we can obtain the simulated expected squared correlation of predicted and true genotypic effects as

$$H^2_{sim} = \frac{1}{n_{sim}} \sum_{s=1}^{n_{sim}} r^2_s. \tag{41}$$

Notice that even in the *simple*, *balanced setting*, this method is not exactly identical to H^2_{Std}, but it is asymptotically equivalent for an increasing number of genotypes:

$$H^2_{sim} = E\left(r^2_{g,\hat{g}}\right) \approx \left(\frac{cov(\boldsymbol{g}, \hat{\boldsymbol{g}}^{BLUP})}{\sqrt{var(\boldsymbol{g})var(\hat{\boldsymbol{g}}^{BLUP})}} \right)^2 = H^2_{Std}. \tag{42}$$

It can be argued that H^2_{sim} is preferable over, *e.g.*, H^2_{Piepho} or H^2_{Cullis}, as it captures the entire variance–covariance structure in the data in a more comprehensive manner (Piepho and Möhring 2007; Schmidt *et al.* 2019). Further notice that this approach also allows direct simulation of the response to selection (see Appendix E).

Reliability: In animal breeding and thus mostly in the context of a single observation per genotype, the reliability (r^2), also known as the CD, is a popular statistic expressing the squared correlation between predicted and true breeding value, and is closely related to heritability (Laloë *et al.* 1996; Laloë and Phocas 2003; Kuehn *et al.* 2007; Piepho 2019). We can estimate the reliability for the (genotypic/breeding value of the) ith entry as

$$r^2_i = 1 - \frac{var\left(\hat{g}^{BLUP}_i\right)}{var(g_i)}, \tag{43}$$

where $var(\hat{g}^{BLUP}_i)$ is the ith diagonal element of the $\boldsymbol{C}_{22(\boldsymbol{g})}$ matrix and $var(g_i)$ is ith diagonal element of the $\boldsymbol{G}_{(\boldsymbol{g})}$ matrix (Mrode and Thompson 2014, chapter 9.3.4.). Accordingly, r^2_i does not account for the off-diagonal elements of either $\boldsymbol{C}_{22(\boldsymbol{g})}$ or $\boldsymbol{G}_{(\boldsymbol{g})}$ [but see Mrode and Thompson (2014), Appendix D]. We can further obtain the mean reliability as

$$\bar{r}^2_{\cdot} = \frac{1}{n_g} \sum_{i=1}^{n_g} r^2_i. \tag{44}$$

Computation of heritability measures: As summarized in Table 1, we computed eight overall heritability measures (H^2_{Std},$\bar{H}^2_{\Delta\cdot\cdot BLUE}$, H^2_{Piepho}, $\bar{H}^2_{\Delta\cdot\cdot BLUP}$, H^2_{Cullis}, H^2_{Oakey}, $\bar{r}^2_{\cdot}$, and H^2_{Sim}), three genotype-wise heritability measures ($\bar{H}^2_{\Delta i\cdot BLUE}$, $\bar{H}^2_{\Delta i\cdot BLUP}$, and r^2_i), and two pair-wise heritability measures ($H^2_{\Delta ij BLUE}$ and $H^2_{\Delta ij BLUP}$) for each of the four examples. When assuming $var(\boldsymbol{a}) = \boldsymbol{K}\sigma^2_a$ for examples 3 and 4, we are estimating narrow-sense heritabilities, otherwise broad-sense heritabilities were computed. The former is not possible for H^2_{Std}, H^2_{Cullis}, and H^2_{Piepho}, however, as these measures implicitly assume a constant genotypic variance and no covariances. We therefore based estimates of these heritability measures on the alternative model that assumes $var(\boldsymbol{g}) = \boldsymbol{I}_{\boldsymbol{n_g}}\sigma^2_g$ and accordingly completely ignores marker information. Finally, we used an *ad hoc* solution to compute H^2_{Std} via (6) in example 4, where we had slightly unbalanced data and environment-specific error variance estimates. Although only 86 out of 89 genotypes were present at all three environments, and one environment only had two replicates instead of three, we considered $n_e = n_r = 3$. Thus, we took the respective maximum value, which is often done in practice. Furthermore, we computed the mean error variance as $\sigma^2_{\bar{\varepsilon}} = \sum \sigma^2_{\varepsilon_i}/n_e$.

Software: All statistical analyses were done by REML with the mixed model package ASReml-R version 3.0 (Gilmour *et al.*

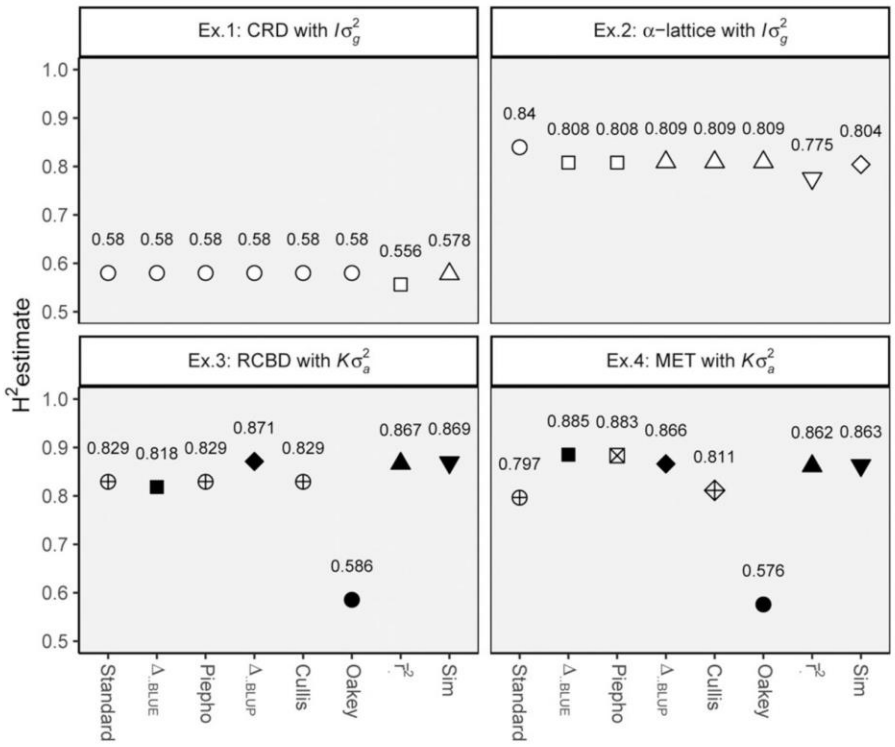

Figure 1 Overall heritability estimates for each method and example. For clearer comparison, (i) estimated values are shown above each symbol and (ii) within each experiment identical values show the same symbol. Black symbols indicate that $var(\boldsymbol{a}) = \boldsymbol{K}\sigma_a^2$ was assumed for the estimation, whereas for white symbols $var(\boldsymbol{g}) = \boldsymbol{I}\sigma_g^2$ was assumed. BLUE, best linear unbiased estimation; BLUP, best linear unbiased prediction; CRD, completely randomized design; Ex, example; RCBD, randomized complete block design.

2009) in the R Statistical Computing Environment (R Core Team 2015) and SAS 9.4 (SAS Institute Inc. 2013).

Data availability

The data used in examples 1 and 2 are the "john.alpha" data set provided in the R package agridat (Wright 2017), which is available at https://cran.r-project.org/web/packages/agridat/index.html. Supplemental material and the data used in examples 3 and 4 are available at https://figshare.com/articles/Lettuce_trial_phenotypic_and_marker_data_/8299493. R code is available at https://github.com/PaulSchmidtGit/Heritability.

Results

In this section, we present overall heritability estimates first, followed by genotype-wise and finally pairwise heritability estimates. The results are accompanied by minimal interpretations, whereas an elaboration on general reasons for differences between estimates of the different methods is provided in the *Discussion*.

Overall heritability

Figure 1 shows estimates for overall heritability obtained via all eight methods for all examples. As expected for the simple, balanced setting of example 1, all estimates are identical, except for the slightly smaller estimates obtained for $\bar{r}_\cdot^2$ and H_{sim}^2. A similar picture is found for example 2, with the exception of H_{Std}^2 displaying a notably larger estimate than the rest, which is due to the method's inability to account for the variance of the random incomplete block effect. Examples 3 and 4 show a more heterogeneous picture, with h_{Oakey}^2 displaying estimates that are strikingly lower than the rest.

Genotype-wise heritability

In example 1, estimates for the pairwise heritability measures $H_{\Delta ij_{BLUE}}^2$ and $H_{\Delta ij_{BLUP}}^2$ were constant across all pairs (≈0.580), and hence equal to $\bar{H}_{\Delta i\cdot_{BLUE}}^2$, $\bar{H}_{\Delta i\cdot_{BLUP}}^2$, $\bar{H}_{\Delta \cdot\cdot_{BLUE}}^2$, and $\bar{H}_{\Delta \cdot\cdot_{BLUP}}^2$, which in turn were equal to H_{Std}^2, H_{Cullis}^2, and H_{Oakey}^2. Analogously, the slightly lower r_i^2 estimates were also constant (≈0.556) across genotypes and, therefore, equal to $\bar{r}_\cdot^2$.

In example 2, estimates for genotype-wise heritability measures were no longer constant, neither between nor within each method. Instead, each method found two unique, but relatively similar, estimates and each one for half of the genotypes, respectively, *i.e.*, 0.80792 and 0.80801 for $\bar{H}_{\Delta i\cdot_{BLUE}}^2$, 0.80911 and 0.80916 for $\bar{H}_{\Delta i\cdot_{BLUP}}^2$, and 0.77537 and 0.77547 for r_i^2.

Figure 2 details the heritability estimates per genotype for examples 3 and 4. It can be seen that the estimates are much more heterogeneous compared to examples 1 and 2. In fact, all of the 89 genotype-wise estimates per method were different from each other in both example 3 and example 4. In both of the latter, and irrespective of the heritability method, genotypes geno49, geno82, and geno88 showed noticeably lower estimates than the rest. This is related to the fact that the additive variances of these three genotypes stand out with a size of only ≈30% of the average (results not shown).

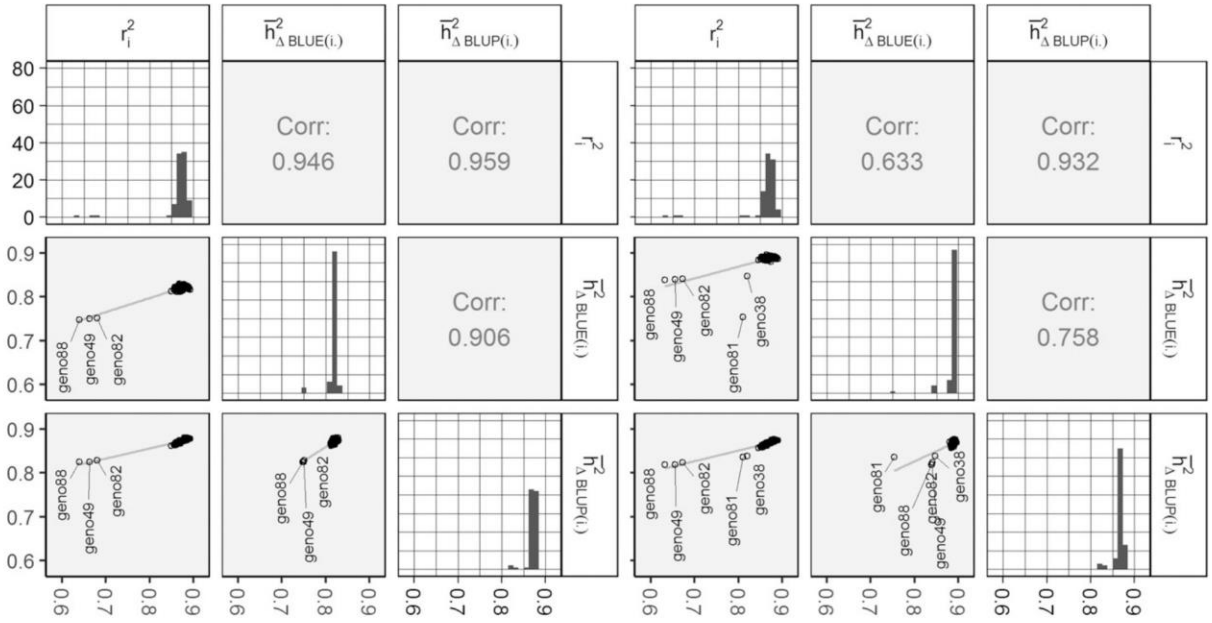

Figure 2 Summary plot for genotype-wise heritability estimates for example 3 (left) and example 4 (right). Plots on the bottom left show scatter plots with simple linear regression lines and genotype-labels for outlying estimates. The plots on the diagonal show histograms for the respective column. The upper right corner displays the Pearson correlation estimate. BLUE, best linear unbiased estimation; BLUP, best linear unbiased prediction; Corr, correlation.

We additionally found relatively low estimates for geno49 and geno81 (*i.e.*, two of the genotypes with missing observations) in example 4 (Figure 2).

In example 3, estimates between the three methods all showed correlations larger than 0.9, with $\bar{h}^2_{\Delta i\cdot_{BLUP}}$ and r_i^2 displaying the highest correlation of ∼0.959. In example 4, the correlation between these two was still high (≈0.932), while the other two correlation estimates dropped below 0.76 (Figure 2).

Pairwise heritability

Figure 3 details the estimated pairwise heritability measures for all examples. As mentioned above, estimates for $H^2_{\Delta ij_{BLUE}}$ and $H^2_{\Delta ij_{BLUP}}$ in example 1 were constant across all $n_\Delta = 276$ genotype pairs (≈0.580). In example 2, estimates for each method split into two clusters, respectively, yet even the overall estimate ranges were relatively small (0.802–0.817 for $H^2_{\Delta ij_{BLUE}}$ and 0.803–0.818 for $H^2_{\Delta ij_{BLUP}}$). The heterogeneous estimates in example 2 are due to the recovery of interblock information via the random incomplete block effect in (17). It results from the varying number of times two genotypes (or even their neighboring genotypes) appear together in the same incomplete block and thus can be compared directly (John and Williams 1995).

Estimates in example 3 mostly form a single cluster for both methods, respectively, ranging from ∼0.70–0.88 for $H^2_{\Delta ij_{BLUE}}$ and 0.80–0.92 for $H^2_{\Delta ij_{BLUP}}$, so that the cluster is located left of/above the diagonal in Figure 3. There are three exceptionally low estimates for both methods, which are those of the differences geno49–geno82, geno49–geno88, and geno82–geno88, *i.e.*, the three genotypes with relatively small additive variances as mentioned before. Generally speaking, genotype pairs with a lower additive genotypic covariance $cov(g_i, g_j)$ tend to show higher estimates for both $H^2_{\Delta ij_{BLUE}}$ and $H^2_{\Delta ij_{BLUP}}$ (Figure 3).

In most aspects, results from example 4 are similar to those in example 3. The most striking difference is the second cluster of lower pairwise heritability estimates. As highlighted in Figure 3, this cluster consists only of the 88 differences from/to genotype geno81, *i.e.*, the genotype present only at a single of the three environments. Notice that while not highlighted, the differences from/to the two genotypes present at only two of the environments (*i.e.*, geno38 and geno49) also form similar clusters, but they are merely located at the lower left border of the main cluster and thus do not stand out as drastically. Furthermore, the main cluster here lies more or less on the diagonal. Finally, the range of estimates for $cov(g_i, g_j)$ is smaller than in example 3. This is mostly due to the $g \times e$ variance that for example 4 in (20) could be estimated separately via the genotype-by-environment interaction effect $(ge)_{it}$ as $var(\boldsymbol{ge}) = \boldsymbol{I}_{263}\sigma^2_{ge}$, whereas fo example 3 the $g \times e$ variance is confounded within the variance of the genotype main effect, since there is no genotype-by-environment interaction effect

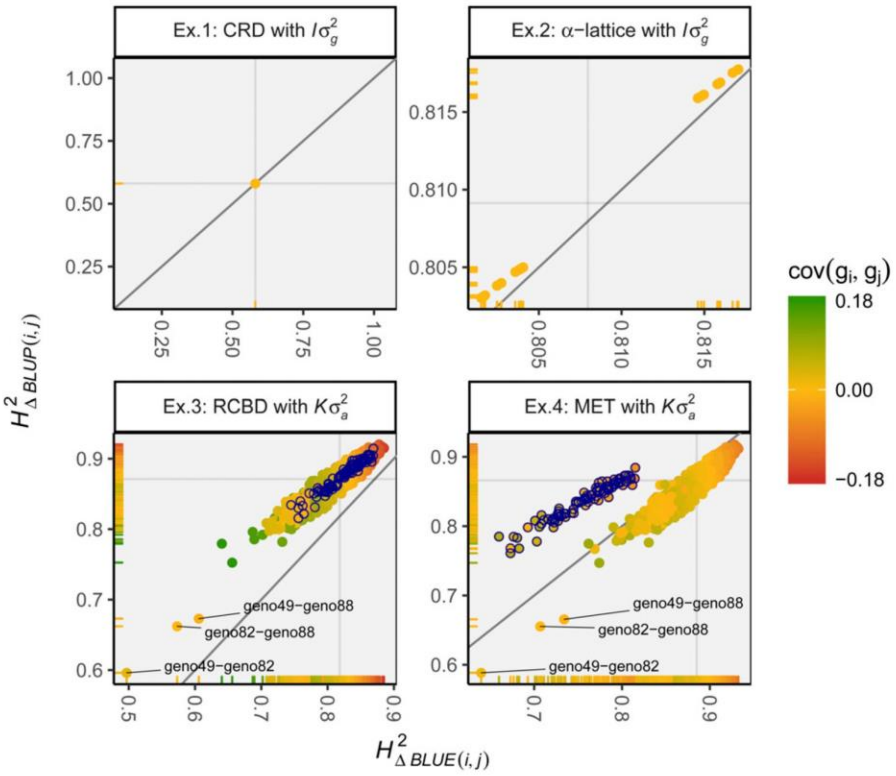

Figure 3 All pairwise heritability estimates. Color indicates covariance in $\boldsymbol{G}_{(\boldsymbol{g})}$ for the respective pair. Circles with blue border are pairs with genotype "geno81," which is only present at a single environment. Vertical and horizontal gray lines represent $\bar{H}^2_{\Delta\cdot\cdot BLUE}$ and $\bar{H}^2_{\Delta\cdot\cdot BLUP}$ for examples 1 and 2, and $\bar{h}^2_{\Delta\cdot\cdot BLUE}$ and $\bar{h}^2_{\Delta\cdot\cdot BLUP}$ for examples 3 and 4. BLUE, best linear unbiased estimation; BLUP, best linear unbiased prediction; CRD, completely randomized design; Ex, example; MET, multienvironment trial; RCBD, randomized complete block design.

in (19), as data from only a single environment are present.

Discussion

Variations of the mixed model for replicated plant data

A consequence of having multiple individuals represent the same genotype is that the definition of the genetic variance σ_g^2 requires careful consideration; rather unproblematic are genetically completely homogeneous cultivars such as inbred lines, doubled-haploid (DH) lines, clones, or hybrids. Yet, it should be noted that when it comes to genetically heterogeneous population cultivars in rye, for example, the matter is no longer straightforward (Bernal-Vasquez *et al.* 2017). Note that estimates for $\bar{h}^2_{\Delta\cdot\cdot BLUE}$, $\bar{h}^2_{\Delta\cdot\cdot BLUP}$ increased with decreasing covariance $cov(g_i, g_j)$ as seen in the *Results* section for pairwise heritability. Note further that when we assumed $var(\boldsymbol{a}) = \boldsymbol{K}\sigma_a^2$ for the (additive) genotype main effect in the analysis of example 4, for simplicity we did not correspondingly make use of $\boldsymbol{K}$ for modeling the genotype–environment interaction effect, but instead assumed $var(\boldsymbol{ge}) = \boldsymbol{I}_{263}\sigma_{ge}^2$. Yet, to account for marker-by-environment interaction, one may instead assume *e.g.*, $var(\boldsymbol{ge}) = \boldsymbol{K}\otimes\boldsymbol{I}_3\sigma_{ge}^2$, where ⊗ denotes the Kronecker product. Although Schulz-Streeck *et al.* (2012) found that the improvement of accounting for marker-by-environment interaction can be negligible, the reader may easily decide to include $\boldsymbol{K}$ for random genotype-interaction effects and it may actually be advisable if the variances for those effects are expected to be relatively large (Bernal-Vasquez *et al.* 2017).

Relatedness of heritability methods

When both $\boldsymbol{G}_{(\boldsymbol{g})}$ and $\boldsymbol{C}_{22(\boldsymbol{g})}$ are proportional to identity matrices, all heritability measures give identical estimates. However, this only holds for a linear random model, since even a single fixed effect [*e.g.*, μ in (16)] results in nonzero off-diagonal elements in $\boldsymbol{C}_{22(\boldsymbol{g})}$. Yet, it should be noted that such a purely random model is not practically relevant in plant (and animal) breeding.

When $\boldsymbol{G}_{(\boldsymbol{g})} = \boldsymbol{I}_{n_g}\sigma_g^2$ and $\boldsymbol{C}_{22(\boldsymbol{g})}$ have compound symmetry structure (*i.e.*, simple, balanced setting), all methods except H^2_{Sim} and $\bar{r}^2_{\cdot}$ give identical estimates, which we confirmed in the simple, balanced setting of example 1. Furthermore, both H^2_{Δ} methods display that same value across all genotypes and pairs. Since we always had to assume $\boldsymbol{G}_{(\boldsymbol{g})} = \boldsymbol{I}_{n_g}\sigma_g^2$ for H^2_{Std}, H^2_{Piepho}, and H^2_{Cullis}, example 3 technically reduces to a simple, balanced setting for these measures as well, since it involves a single environment laid out as RCBD with balanced data. Accordingly, these three methods yielded identical estimates in that scenario as well (Figure 1). The estimates for $\bar{r}^2_{\cdot}$ differ, because off-diagonal elements are ignored in its estimation. However, notice that with increasing n_g, the off-diagonal

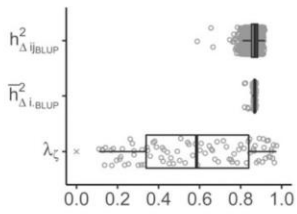

Figure 4 Individual estimates and boxplots for $h^2_{\Delta ij_{BLUP}}$, $\bar{h}^2_{\Delta i\cdot_{BLUP}}$, and λ_ζ (*i.e.*, eigenvalues needed for h^2_{Oakey}) from example 4. The zero eigenvalue is denoted by the symbol ×. BLUP, best linear unbiased prediction.

elements of $\boldsymbol{C}_{22(\boldsymbol{g})}$ approach 0, and thus $\bar{r}^2_{\cdot\cdot}$ approaches the other measures and, as stated before, so does H^2_{Sim}.

In the case where $\boldsymbol{G}_{(\boldsymbol{g})} = \boldsymbol{I}_{\boldsymbol{n_g}}\sigma^2_g$ and $\boldsymbol{C}_{22(\boldsymbol{g})}$ is an unstructured matrix, $\bar{H}^2_{\Delta\cdot\cdot_{BLUP}}$ and H^2_{Cullis} give identical estimates. This is analogously true for $\bar{H}^2_{\Delta\cdot\cdot_{BLUE}}$ and H^2_{Piepho}, and for $\boldsymbol{C}_{11}$ (rather than $\boldsymbol{C}_{22(\boldsymbol{g})}$) being an unstructured matrix. Rodríguez-Álvarez *et al.* (2018) show a connection between $\bar{r}^2_{\cdot\cdot}$ and H^2_{Oakey}. They point out that given $\boldsymbol{G}_{(\boldsymbol{g})_{\boldsymbol{n_g}}} = \boldsymbol{I}_{\boldsymbol{n_g}}\sigma^2_g$, the mean reliability can be expressed as $\bar{r}^2_{\cdot\cdot} = \frac{1}{n_g}\sum_{i=1} r^2_i = trace(\boldsymbol{I}_{\boldsymbol{n_g}} - \boldsymbol{G}^{-1}_{(\boldsymbol{g})}\boldsymbol{C}_{22(\boldsymbol{g})})/n_g = trace(\boldsymbol{D})/n_g$ and thus as a special case of H^2_{Oakey}, ignoring the number of zero eigenvalues. This is in agreement with Laloë (1993), and our results confirm this relationship for examples 1 and 2 (results not shown). However, we did not find it to be true for examples 3 and 4, where $\boldsymbol{G}_{(\boldsymbol{g})} \neq \boldsymbol{I}_{\boldsymbol{n_g}}\sigma^2_g$. Note that in their article Rodríguez-Álvarez *et al.* (2018) refer to the mean reliability ($\bar{r}^2_{\cdot\cdot}$) as H^2_C in Cullis *et al.* (2006), even though the latter method is based on differences [see (11)], whereas the former is based on individual genotypes [see (43)].

Interestingly, though not practically very relevant, r^2_i and $\bar{H}^2_{\Delta i\cdot_{BLUP}}$ give identical estimates per genotype for diagonal $\boldsymbol{G}_{(\boldsymbol{g})}$ and $\boldsymbol{C}_{22(\boldsymbol{g})}$.

Finally, when $\boldsymbol{G}_{(\boldsymbol{g})}$ displays covariances (*e.g.*, as $\boldsymbol{G}_{(\boldsymbol{g})} = \boldsymbol{K}\sigma^2_g$) and $\boldsymbol{C}_{22(\boldsymbol{g})}$ is an unstructured matrix, none of the five measures that are appropriate for this case (*i.e.*, $\bar{h}^2_{\Delta\cdot\cdot_{BLUE}}$, $\bar{h}^2_{\Delta\cdot\cdot_{BLUP}}$, h^2_{Oakey}, $\bar{r}^2_{\cdot\cdot}$, and h^2_{Sim}) give identical results. Nevertheless, with the exception of h^2_{Oakey}, heritability estimates within examples 3 and 4 were relatively similar (Figure 1). This is true even for the methods that did not include kinship information (*i.e.*, H^2_{Std}, H^2_{Piepho}, and H^2_{Cullis}). Yet, when Ould Estaghvirou *et al.* (2013) simulated phenotypic and marker data for a single location, they found that these three methods gave lower estimates than the rest.

The most striking results are those for h^2_{Oakey} in examples 3 and 4. Notice that Lourenço *et al.* (2017) also found outstandingly low estimates for h^2_{Oakey} in models with $\boldsymbol{G}_{(\boldsymbol{g})} = \boldsymbol{K}\sigma^2_g$ for two real data sets and one simulated data set. This raises the question whether H^2_{Oakey} is suited for cases where $\boldsymbol{G}_{(\boldsymbol{g})} = \boldsymbol{K}\sigma^2_g$. While it is true that H^2_{Oakey} estimates are equal to H^2_{Std} in the simple, balanced setting (of example 1), this is only a necessary, but not a sufficient, condition for H^2_{Oakey} to be generally applicable also with correlated genotypes. It can be argued that H^2_{Oakey} and $\bar{H}^2_{\Delta\cdot\cdot}$ are similar in the sense that both are based on contrasts between genotypes. However, the important difference is that for $\bar{H}^2_{\Delta\cdot\cdot}$ the contrasts of interest (*i.e.*, all genotype pairs) are chosen according to the breeders' goals, while the contrasts in H^2_{Oakey} are determined by the data set. Laloë (1993) shows that the smallest and largest nonzero eigenvalues, λ_ζ, are the lower and upper limit for the heritability of all possible contrasts of genotypes. Accordingly, they are also the lower and upper limit for all $H^2_{\Delta ij}/h^2_{\Delta ij}$, which is confirmed by our results (Figure 4). It may be noted that H^2_{Oakey} is a simple average of "canonical" heritabilities (eigenvalues) corresponding to canonical contrasts and derived from the canonical decomposition of covariance matrices. As such, low heritabilities indicate a low design efficiency, as the smallest eigenvalue appears to be a good indicator of the robustness of the design and a measure of the part of the genetic trend that can be predicted (Laloë and Phocas 2003). In contrast, a pairwise difference is a combination of low and high heritability contrasts that corresponds to the weighted mean of canonical heritabilities. This may go some way toward explaining the observed discrepancies in estimates of h^2_{Oakey} and the alternatives considered here.

Genotype-specific information gained from H^2_Δ

Figure 2 and Figure 3 are good examples of how H^2_Δ/h^2_Δ can give a more in-depth insight into the outcome of a plant breeding trial. More precisely, the results show that investigating genotype-specific, or even genotype-contrast-specific, heritability measures can (i) summarize how variable/ambiguous an overall heritability estimate is and (ii) point out individual genotypes that stand out.

The first point may seem obvious by now, but it must be realized that whenever only a single (overall) heritability estimate is obtained, there is no information whatsoever about the dispersion of that estimate, only about its central tendency. We argue that investigating $\bar{H}^2_{\Delta i\cdot}/\bar{h}^2_{\Delta i\cdot}$ and/or $H^2_{\Delta ij}/h^2_{\Delta ij}$ gives more comprehensive insight, since they explicitly show the variability across genotypes and thus, bringing us to the second point, point out exceptional genotypes such as geno38, geno49, geno81, geno82, and geno88 in example 4. Notice that we identified two separate reasons for why these five genotypes were exceptional (low additive variance and missing observations; see *Results* section). Moreover, looking at the cluster of estimates for geno81 in example 4 in Figure 3, it becomes clear how a single genotype can display notably lower pairwise heritability estimates than the rest. These estimates are obviously decreasing the overall heritability estimate for this example, which would be especially unfavorable in a scenario where the breeder does not have great interest in this particular genotype. In the end, none of this would become apparent by only looking at overall heritability estimates (Figure 1).

What do we really want heritability for and how does H^2_Δ help?

There are two main reasons why heritability on an entry-mean basis is of interest in plant breeding. On one hand, it is plugged

into the breeder's Equation (2) to predict the response to selection. On the other hand, it is a descriptive measure used to assess the usefulness and precision of results from cultivar evaluation trials. In the simple, balanced setting, heritability is suited for both purposes. However, whenever we depart from the simple, balanced setting, H^2_{Std} is no longer suited for either of the two purposes. Furthermore, any alternative measure can ultimately only aim to generalize heritability as a descriptive measure. Regarding the purpose of heritability as a mean to estimate the response to selection, we reiterate the view of Piepho and Möhring (2007): instead of trying to approximate heritability using some *ad hoc* measure, one should simulate the response to selection directly. As stated before, this can be done via the same simulation-based approach used to obtain H^2_{Sim} and is exemplified in their work (also see Appendix E).

In terms of the descriptive function of heritability, we believe that H^2_{Δ} is a valuable extension of heritability on an entry-mean basis, because its genotype-wise and pair-wise estimates give more detailed insight, and ultimately allow for better decision-making by the breeder. Furthermore, we would like to argue that its derivation, calculation, and interpretation is rather intuitive, which makes us optimistic about its acceptance in practice. Our view is further supported by Laloë (1993), who also suggested extending reliability to pairwise contrasts in the animal breeding framework.

Conclusions

Irrespective of whether a breeder ultimately decides to apply H^2_{Δ}, we think that understanding the idea behind it alone raises awareness for the mentioned problems with heritability outside the scope of simple, balanced settings and thus can potentially improve the use of heritability in breeding programs. In the end, all heritability measures aim to inform about the same underlying subject matter, and, albeit via different methodologies, they all do. What should therefore be held above all, especially given the mentioned ambiguity problems in a nonsimple, unbalanced setting, is to report how heritability was estimated. Hanson and Robinson (1963, page 612) put it in this crucial context of reproducible science by pointing out that "One must extrapolate in the real world, and one must use estimates of heritability derived by someone else, especially with new crops in new environments. Therefore, it is important to specify exactly how published estimates were obtained in order that others may extrapolate."

Acknowledgments

We thank Ivan Simko and Ryan J. Hayes (US Department of Agriculture-Agricultural Research Service, Crop Improvement and Protection Research Unit, Salinas, CA) for providing the lettuce data. We also wish to thank two anonymous reviewers for their very helpful comments. This research was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft grant PI 377/18-1).

Literature Cited

Becker, H., 2011 *Pflanzenzüchtung*. Ed. 2 Ulmer, Stuttgart, Germany.

Bernal-Vasquez, A.-M., A. Gordillo, M. Schmidt, and H.-P. Piepho, 2017 Genomic prediction in early selection stages using multi-year data in a hybrid rye breeding program. BMC Genet. 18: 51. https://doi.org/10.1186/s12863-017-0512-8

Cullis, B. R., A. B. Smith, and N. E. Coombes, 2006 On the design of early generation variety trials with correlated data. J. Agric. Biol. Environ. Stat. 11: 381–393. https://doi.org/10.1198/108571106X154443

Dias, K. O. D. G., S. A. Gezan, C. T. Guimarães, A. Nazarian, L. da Costa *et al.*, 2018 Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. Heredity 121: 24–37. https://doi.org/10.1038/s41437-018-0053-6

Falconer, D. S., and T. F. C. Mackay, 2005 *Introduction to Quantitative Genetics*, Ed. 4. Pearson Prentice Hall, Harlow.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson, 2009 *ASReml user guide release 3.0*. VSN International Ltd., Hemel Hempstead.

Hadasch, S., I. Simko, R. J. Hayes, J. O. Ogutu, and H.-P. Piepho, 2016 Comparing the predictive abilities of phenotypic and marker-assisted selection methods in a biparental lettuce population. Plant Genome 9. Available at: https://dl.sciencesocieties.org/publications/tpg/abstracts/9/1/plantgenome2015.03.0014. https://doi.org/10.3835/plantgenome2015.03.0014

Hallauer, A. R., M. J. Carena, and J. B. Miranda Filho, 2010 *Quantitative Genetics in Maize Breeding*, Vol. 3. Springer, New York.

Hanson, W., and H. Robinson, 1963 *Statistical Genetics and Plant Breeding*. National Academy of Sciences - National Research Council, Washington, DC.

Hayes, R. J., C. H. Galeano, Y. Luo, R. Antonise, and I. Simko, 2014 Inheritance of decay of fresh-cut lettuce in a recombinant inbred line population from Salinas 88 × La Brillante. J. Am. Soc. Hortic. Sci. 139: 388–398. https://doi.org/10.21273/JASHS.139.4.388

Henderson, C. R., 1986 Statistical methods in animal improvement. Historical overview, pp. 2–14 in *Advances in Statistical Methods for Genetic Improvement of Livestock*, Vol. 18, edited by D. Gianola, and K. Hammond. Springer, Berlin. https://doi.org/10.1007/978-3-642-74487-7_1

John, N. A., and E. R. Williams, 1995 *Cyclic and Computer Generated Designs*, Ed. 2. Chapman and Hall, London. https://doi.org/10.1007/978-1-4899-7220-0

Knight, R. L., 1948 Dictionary of Genetics, Chronica Botánica Co., Waltham, Massachusetts.

Kuehn, L. A., R. M. Lewis, and D. R. Notter, 2007 Managing the risk of comparing estimated breeding values across flocks or herds through connectedness: a review and application. Genet. Sel. Evol. 39: 225–247. https://doi.org/10.1051/gse:2007001

Laloë, D., 1993 Precision and information in linear models of genetic evaluation. Genet. Sel. Evol. 25: 557. https://doi.org/10.1186/1297-9686-25-6-557

Laloë, D., and F. Phocas, 2003 A proposal of criteria of robustness analysis in genetic evaluation. Livest. Prod. Sci. 80: 241–256. https://doi.org/10.1016/S0301-6226(02)00092-1

Laloë, D., F. Phocas, and F. Menissier, 1996 Considerations on measures of precision and connectedness in mixed linear

models of genetic evaluation. Genet. Sel. Evol. 28: 359. https://doi.org/10.1186/1297-9686-28-4-359

Lourenço, V. M., P. C. Rodrigues, A. M. Pires, and H.-P. Piepho, 2017 A robust DF-REML framework for variance components estimation in genetic studies. Bioinformatics 33: 3584–3594. https://doi.org/10.1093/bioinformatics/btx457

Mathews, K. L., M. Malosetti, S. Chapman, L. McIntyre, M. Reynolds *et al.*, 2008 Multi-environment QTL mixed models for drought stress adaptation in wheat. Theor. Appl. Genet. 117: 1077–1091. https://doi.org/10.1007/s00122-008-0846-8

Mrode, R. A., and R. Thompson, 2014 *Linear Models for the Prediction of Animal Breeding Values*, Ed. 3. Centre for Agriculture and Bioscience International, Wallingford. https://doi.org/10.1079/9781780643915.0000

Oakey, H., A. Verbyla, W. Pitchford, B. Cullis, and H. Kuchel, 2006 Joint modeling of additive and non-additive genetic line effects in single field trials. Theor. Appl. Genet. 113: 809–819. https://doi.org/10.1007/s00122-006-0333-z

Ould Estaghvirou, S. B., J. O. Ogutu, T. Schulz-Streeck, C. Knaak, M. Ouzunova *et al.*, 2013 Evaluation of approaches for estimating the accuracy of genomic prediction in plant breeding. BMC Genomics 14: 860. https://doi.org/10.1186/1471-2164-14-860

Piepho, H.-P., 2019 A coefficient of determination (R^2) for generalized linear mixed models. Biom. J. 61: 860–872. https://doi.org/10.1002/bimj.201800270

Piepho, H.-P., and J. Möhring, 2007 Computing heritability and selection response from unbalanced plant breeding trials. Genetics 177: 1881–1888. https://doi.org/10.1534/genetics.107.074229

Piepho, H.-P., J. Möhring, A. E. Melchinger, and A. Büchse, 2008 BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161: 209–228. https://doi.org/10.1007/s10681-007-9449-8

Piepho, H.-P., J. Möhring, T. Schulz-Streeck, and J. O. Ogutu, 2012a A stage-wise approach for the analysis of multi-environment trials. Biom. J. 54: 844–860. https://doi.org/10.1002/bimj.201100219

Piepho, H.-P., J. O. Ogutu, T. Schulz-Streeck, B. Estaghvirou, A. Gordillo *et al.*, 2012b Efficient computation of ridge-regression best linear unbiased prediction in genomic selection in plant breeding. Crop Sci. 52: 1093–1104. https://doi.org/10.2135/cropsci2011.11.0592

R Core Team, 2015 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Rodríguez-Álvarez, M. X., M. P. Boer, F. A. van Eeuwijk, and P. H. C. Eilers, 2018 Correcting for spatial heterogeneity in plant breeding experiments with P-splines. Spat. Stat. 23: 52–71. https://doi.org/10.1016/j.spasta.2017.10.003

SAS Institute Inc, 2013 *Base SAS 9.4 Procedures Guide: Statistical Procedures*, *2*. SAS Institute Inc., Cary, NC.

SAS Institute Inc., 2017 *SAS/STAT® 14.3 User's Guide*. SAS Institute Inc., Cary, NC.

Schmidt, P., J. Hartung, J. Rath, and H.-P. Piepho, 2019 Estimating broad-sense heritability with unbalanced data from agricultural cultivar trials. Crop Sci. 59: 525–536. https://doi.org/10.2135/cropsci2018.06.0376

Schulz-Streeck, T., J. O. Ogutu, Z. Karaman, C. Knaak, and H.-P. Piepho, 2012 Genomic selection using multiple populations. Crop Sci. 52: 2453–2461. https://doi.org/10.2135/cropsci2012.03.0160

Searle, S. R., G. Casella, and C. E. McCulloch, 1992 *Variance Components*. Wiley, New York. https://doi.org/10.1002/9780470316856

Searle, S. R., F. M. Speed, and G. A. Milliken, 2012 Population marginal means in the linear model. An alternative to least squares means. Am. Stat. 34: 216–221. https://doi.org/10.1080/00031305.1980.10483031

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

Viana, J. M. S., H. D. Pereira, G. B. Mundim, H.-P. Piepho, and F. F. E. Silva, 2018 Efficiency of genomic prediction of non-assessed single crosses. Heredity 120: 283–295. https://doi.org/10.1038/s41437-017-0027-0

Wellmann, R., and J. Bennewitz, 2011 The contribution of dominance to the understanding of quantitative genetic variation. Genet. Res. 93: 139–154. https://doi.org/10.1017/S0016672310000649

Wright, K., 2017 agridat: Agricultural Datasets. R package version 1.13. dataset: john.alpha. https://CRAN.R-project.org/package=agridat.

Xu, S., 2013 *Principles of Statistical Genomics*. Springer, New York. https://doi.org/10.1007/978-0-387-70807-2

Yan, W., 2014 *Crop Variety Trials: Data Management and Analysis*. John Wiley & Sons Inc., New York. https://doi.org/10.1002/9781118688571

Communicating editor: A. Charcosset

Appendix

Appendix A: Computing $\hat{g}^{BLUP}$ From $\hat{g}^{BLUE}$ in a Stage-Wise Approach

Consider the analysis of a series of trials where the aim is to obtain $\hat{g}^{BLUP}$. One may use a single-stage analysis with a random genotype main effect. Alternatively, it is also possible to use a stage-wise analysis, where the first stage obtains $\hat{g}^{BLUE}$ from which $\hat{g}^{BLUP}$ are predicted in the second stage. Both approaches are equivalent, if variance components are known. In practice, small differences are encountered, as variances need to be estimated (Piepho *et al.* 2012a).

Single-stage analysis

Assume that we can write the general single-stage model in (13) as

$$\boldsymbol{y} = \boldsymbol{X}_A\boldsymbol{\beta}_A + \boldsymbol{Z}_A\boldsymbol{u}_A + \boldsymbol{Z}_B\boldsymbol{u}_B + \boldsymbol{Z}_W\boldsymbol{u}_W + \boldsymbol{\varepsilon}, \tag{45}$$

where $\boldsymbol{\beta}_A$ is a vector of fixed across-environment effects (subscript A), $\boldsymbol{u}_A$ is a vector of random across-environment effects with $\boldsymbol{u}_A \sim (0, \boldsymbol{G}_A)$, $\boldsymbol{u}_B$ is a vector of random between-environment effects (subscript B) with $\boldsymbol{u}_B \sim (0, \boldsymbol{G}_B)$, $\boldsymbol{u}_W$ is a vector of random within-environment effects (subscript W) with $\boldsymbol{u}_W \sim (0, \boldsymbol{G}_W)$, $\boldsymbol{\varepsilon}$ is a vector of plot errors with $\boldsymbol{\varepsilon} \sim (0, \boldsymbol{R})$, and $\boldsymbol{y}$ is the observed data vector with $\boldsymbol{y} \sim (0, \boldsymbol{V})$, where $\boldsymbol{V} = \boldsymbol{Z}_A\boldsymbol{G}_A\boldsymbol{Z}_A^T + \boldsymbol{Z}_B\boldsymbol{G}_B\boldsymbol{Z}_B^T + \boldsymbol{Z}_W\boldsymbol{G}_W\boldsymbol{Z}_W^T + \boldsymbol{R}$ [see Piepho *et al.* (2012a)]. Thus, $\boldsymbol{G}$ here is a block diagonal matrix with blocks $\boldsymbol{G}_A$, $\boldsymbol{G}_B$, and $\boldsymbol{G}_W$ on its diagonal. Given a MET similar to example 4, $\boldsymbol{X}_A\boldsymbol{\beta}_A$ would be a general intercept, $\boldsymbol{u}_A$ the random genotype main effect (*i.e.* $\boldsymbol{g}$), $\boldsymbol{u}_B$ the random environment main effect and random genotype-by-environment interaction effect, $\boldsymbol{u}_W$ the random block effect at each environment, and $\boldsymbol{\varepsilon}$ the plot error. Accordingly, $\boldsymbol{G}_A$ would correspond to the kinship matrix $\boldsymbol{K}$, whereas $\boldsymbol{G}_B$ and $\boldsymbol{G}_W$ would be proportional to identity matrices.

Two-stage analysis

Model (45) has a two-stage representation, with the first stage model given by

$$\boldsymbol{y} = \boldsymbol{X}_{A1}\boldsymbol{\beta}_{A1} + \boldsymbol{Z}_B\boldsymbol{u}_B + \boldsymbol{Z}_W\boldsymbol{u}_W + \boldsymbol{\varepsilon}_1, \tag{46}$$

where effects are defined as in (45), while the subscript 1 denotes that the corresponding parameters are estimated in this first stage and thus $\boldsymbol{V}_1 = \boldsymbol{Z}_B\boldsymbol{G}_B\boldsymbol{Z}_B^T + \boldsymbol{Z}_W\boldsymbol{G}_W\boldsymbol{Z}_W^T + \boldsymbol{R}_1$. Notice that in contrast to (45), $\boldsymbol{Z}_A\boldsymbol{u}_A$ is missing here, since the genotype main effect needs to be taken as fixed at this stage (Piepho *et al.* 2012a) and is therefore comprised in $\boldsymbol{X}_{A1}\boldsymbol{\beta}_{A1}$. For simplicity, we can assume that here the genotypic main effects are the only fixed effects so that $\boldsymbol{\beta}_{A1}$ are the genotype means. Accordingly, we can obtain genotypic BLUEs as

$$BLUE(\boldsymbol{\beta}_{A1}) = \hat{\boldsymbol{g}}^{BLUE} = \left(\boldsymbol{X}_{A1}^{'}\ \boldsymbol{V}_1^{-1}\boldsymbol{X}_{A1}\right)^{-1}\boldsymbol{X}_{A1}^{'}\boldsymbol{V}_1^{-1}\boldsymbol{y}, \tag{47}$$

with variance

$$var\left(\hat{\boldsymbol{g}}^{BLUE}\right) = \left(\boldsymbol{X}_{A1}^{'}\boldsymbol{V}_1^{-1}\boldsymbol{X}_{A1}\right)^{-1} = \boldsymbol{R}_2. \tag{48}$$

In the second stage we then have

$$\hat{\boldsymbol{g}}^{BLUE} = \boldsymbol{X}_{A2}\boldsymbol{\beta}_{A2} + \boldsymbol{Z}_{A2}\boldsymbol{u}_A + \boldsymbol{\varepsilon}_2, \tag{49}$$

where $\boldsymbol{\varepsilon}_2 \sim (0, \boldsymbol{R}_2)$ and just like in (45), the random genotype main effect $\boldsymbol{g} = \boldsymbol{u}_A \sim (0, \boldsymbol{G}_A)$ where $\boldsymbol{G}_A$ corresponds to the kinship matrix $\boldsymbol{K}$. Notice that $\boldsymbol{X}_{A2}\boldsymbol{\beta}_{A2} = 1_{n_g}\mu$ and $\boldsymbol{Z}_{A2} = \boldsymbol{I}_{n_g}$. Accordingly, we have $\boldsymbol{V}_2 = \boldsymbol{Z}_{A2}\boldsymbol{G}_A\boldsymbol{Z}_{A2}^T + \boldsymbol{R}_2$ and for the mixed model equations we find

$$\begin{bmatrix} \boldsymbol{X}_{A2}^{'}\boldsymbol{R}_2^{-1}\boldsymbol{X}_{A2}^{'} & \boldsymbol{X}_{A2}^{'}\boldsymbol{R}_2^{-1}\boldsymbol{Z}_{A2} \\ \boldsymbol{Z}_{A2}^{'}\boldsymbol{R}_2^{-1}\boldsymbol{X}_{A2} & \boldsymbol{Z}_{A2}^{'}\boldsymbol{R}_2^{-1}\boldsymbol{Z}_{A2} + \boldsymbol{G}_A^{-1} \end{bmatrix}^{-} \begin{bmatrix} \hat{\boldsymbol{\beta}}_{A2} \\ \hat{\boldsymbol{g}}^{BLUP} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_{A2}^{'}\boldsymbol{R}_2^{-1}\hat{\boldsymbol{g}}^{BLUE} \\ \boldsymbol{Z}_{A2}^{'}\boldsymbol{R}_2^{-1}\hat{\boldsymbol{g}}^{BLUE} \end{bmatrix}. \tag{50}$$

It is important to note that we can write

$$\boldsymbol{X}_A = \boldsymbol{X}_{A1}\boldsymbol{X}_{A2} \tag{51}$$

and

$$\boldsymbol{Z}_A = \boldsymbol{X}_{A1}\boldsymbol{Z}_{A2}. \tag{52}$$

Finally, plugging in (48) and (49) into (50), while making use of (51) and (52), it can be seen that the mixed model equation solutions for the second stage are equivalent to those for the single-stage analysis.

Appendix B: An Alternative Estimation for $\bar{H}^2_{\Delta\cdot\cdot}$ in the Case of Independent Genotypic Effects with Constant Variance by Averaging Separately Across Numerator and Denominator

Given independent genotypic effects with constant variance, averaging numerator and denominator of $H^2_{\Delta ij_{BLUP}}$ in (24) separately first leads to the algebraically equivalent simplification

$$\bar{H}^2_{\Delta\cdot\cdot_{BLUP}} = \frac{\sum_i\sum_{j<i}\mathrm{var}(g_i - g_j) - v^{BLUP}_{\Delta ij}}{\sum_i\sum_{j<i}\mathrm{var}(g_i - g_j)} = \frac{2\sigma^2_g - \frac{2}{n_g(n_g-1)}\sum_i\sum_{j<i}v^{BLUP}_{\Delta ij}}{2\sigma^2_g} = \frac{2\sigma^2_g - \bar{v}^{BLUP}_{\Delta\cdot\cdot}}{2\sigma^2_g} = 1 - \frac{\bar{v}^{BLUP}_{\Delta\cdot\cdot}}{2\sigma^2_g} = H^2_{Cullis}. \tag{53}$$

Analogously, we find for $H^2_{\Delta ij_{BLUE}}$ in (29) that

$$\bar{H}^2_{\Delta\cdot\cdot_{BLUE}} = \frac{\sum_i\sum_{j<i}\mathrm{var}(g_i - g_j)}{\sum_i\sum_{j<i}\mathrm{var}(g_i - g_j) + v^{BLUE}_{\Delta ij}} = \frac{2\sigma^2_g}{2\sigma^2_g + \frac{2}{n_g(n_g-1)}\sum_i\sum_{j<i}v^{BLUE}_{\Delta ij}} = \frac{2\sigma^2_g}{2\sigma^2_g + \bar{v}^{BLUE}_{\Delta\cdot\cdot}} = \frac{\sigma^2_g}{\sigma^2_g + \bar{v}^{BLUE}_{\Delta\cdot\cdot}/2} = H^2_{Piepho}. \tag{54}$$

Appendix C: An Approximation for $\bar{H}^2_{\Delta\cdot\cdot}$ in the General Case

It can be shown (Ould Estaghvirou *et al.* 2013, Additional Files 1) that

$$\sum_i\sum_{j<i}v^{BLUE}_{\Delta ij} = trace\left(n_g\boldsymbol{P}_{\boldsymbol{\mu}}\boldsymbol{C}_{11(\boldsymbol{g})}\right), \tag{55}$$

where $\boldsymbol{P}_{\boldsymbol{\mu}} = \boldsymbol{I}_{n\boldsymbol{g}} - n_g^{-1}1_{n_g}1_{n_g}^T$ with 1_{n_g} being a vector of ones. Thus, $\boldsymbol{P}_{\boldsymbol{\mu}}$ is a matrix that centers for the overall mean, such that $\bar{v}^{BLUE}_{\Delta\cdot\cdot} = \frac{2}{n_g(n_g-1)}trace(n_g\boldsymbol{P}_{\boldsymbol{\mu}}\boldsymbol{C}_{11(\boldsymbol{g})})$. Analogously, we have

$$\sum_i\sum_{j<i}v^{BLUP}_{\Delta ij} = trace\left(n_g\boldsymbol{P}_{\boldsymbol{\mu}}\boldsymbol{C}_{22(\boldsymbol{g})}\right), \tag{56}$$

such that $\bar{v}^{BLUP}_{\Delta\cdot\cdot} = \frac{2}{n_g(n_g-1)}trace(n_g\boldsymbol{P}_{\boldsymbol{\mu}}\boldsymbol{C}_{22(\boldsymbol{g})})$ and finally

$$\sum_i\sum_{j<i}var(g_i - g_j) = trace\left(n_g\boldsymbol{P}_{\boldsymbol{\mu}}\boldsymbol{G}_{(\boldsymbol{g})}\right), \tag{57}$$

such that $\bar{v}^{g}_{\Delta\cdot\cdot} = \frac{2}{n_g(n_g-1)}trace(n_g\boldsymbol{P}_{\boldsymbol{\mu}}\boldsymbol{G}_{(\boldsymbol{g})})$. By inserting (55), (56), and (57) into (34), we get

$$\bar{\bar{H}}^2_{\Delta\cdot\cdot_{BLUE}} = \frac{trace\left(\boldsymbol{P}_{\boldsymbol{\mu}}\boldsymbol{G}_{(\boldsymbol{g})}\right)}{trace\left(\boldsymbol{P}_{\boldsymbol{\mu}}\boldsymbol{G}_{(\boldsymbol{g})}\right) + trace\left(\boldsymbol{P}_{\boldsymbol{\mu}}\boldsymbol{C}_{11(\boldsymbol{g})}\right)}, \tag{58}$$

where, loosely speaking, $trace(\boldsymbol{P}_{\boldsymbol{\mu}}\boldsymbol{G}_{(\boldsymbol{g})})$ captures the genotypic variance, $trace(\boldsymbol{P}_{\boldsymbol{\mu}}\boldsymbol{C}_{11(\boldsymbol{g})})$ captures the environmental variance, and hence $trace(\boldsymbol{P}_{\boldsymbol{\mu}}\boldsymbol{G}_{(\boldsymbol{g})}) + trace(\boldsymbol{P}_{\boldsymbol{\mu}}\boldsymbol{C}_{11(\boldsymbol{g})})$ captures the phenotypic variance. Notice that (58) coincides with Method 4 of Ould Estaghvirou *et al.* (2013). Based on (33), we find correspondingly

$$\bar{\bar{H}}^2_{\Delta\cdot\cdot_{BLUP}} = \frac{trace\left(\boldsymbol{P}_{\boldsymbol{\mu}}\boldsymbol{G}_{(\boldsymbol{g})}\right) - trace\left(\boldsymbol{P}_{\boldsymbol{\mu}}\boldsymbol{C}_{22(\boldsymbol{g})}\right)}{trace\left(\boldsymbol{P}_{\boldsymbol{\mu}}\boldsymbol{G}_{(\boldsymbol{g})}\right)}. \tag{59}$$

Appendix D: A Standardization of *M* To Directly Capture the Total Genotypic Variance

As shown in Appendix C, $\bar{v}^{g}_{\Delta..}$ captures the average genotypic variance of a difference between two genotypes. In accordance with the notion of this article, we could obtain half the average variance of a difference as an estimate for *the total genotypic variance*. Notice that this approach is in line with the average semi variance $(\sigma^2_{g(asv)})$ used by Piepho (2019) and can be defined as

$$\sigma^2_{g(asv)} = 0.5\bar{v}^{g}_{\Delta..} = \frac{trace\left(\boldsymbol{P_\mu G_{(g)}}\right)}{n_g - 1}. \tag{60}$$

In the case of $\boldsymbol{G_{(g)}} = \boldsymbol{MM'}\sigma^2_a = \boldsymbol{K}\sigma^2_a$, we can do some useful rearrangements:

$$\begin{aligned}\sigma^2_{g(asv)} &= \frac{trace\left(\boldsymbol{P_\mu MM'}\right)\sigma^2_g}{n_g - 1} = \frac{trace\left(\boldsymbol{P_\mu P_\mu MM'}\right)\sigma^2_g}{n_g - 1} \\ &= \frac{trace\left(\boldsymbol{P_\mu MM'P_\mu}\right)\sigma^2_g}{n_g - 1} = \frac{trace\left(\tilde{\boldsymbol{M}}\tilde{\boldsymbol{M}}'\right)\sigma^2_g}{n_g - 1},\end{aligned} \tag{61}$$

where $\tilde{\boldsymbol{M}} = \boldsymbol{P_\mu M}$ is column mean-centered. Notice that $\tilde{\boldsymbol{M}}$ is equal to what is referred to as $\boldsymbol{Z}$ in VanRaden (2008) (page 4416). Notice further that the standardization presented here works for an arbitrary coding of markers and arbitrary scale. As a result, the trace of $\tilde{\boldsymbol{G}}_{(g)} = \tilde{\boldsymbol{M}}\tilde{\boldsymbol{M}}'\widetilde{\sigma^2_a} = \tilde{\boldsymbol{K}}\widetilde{\sigma^2_a}$ directly returns $\hat{\sigma}^2_{g(asv)}$. We can standardize further:

$$\overline{\tilde{M}} = \frac{\tilde{\boldsymbol{M}}}{\sqrt{\frac{trace\left(\tilde{\boldsymbol{M}}\tilde{\boldsymbol{M}}'\right)}{n_g - 1}}} \tag{62}$$

and define $\overline{\tilde{\boldsymbol{G}}}_{(g)} = \overline{\tilde{\boldsymbol{M}}\tilde{\boldsymbol{M}}}'\overline{\widetilde{\sigma^2_a}} = \overline{\tilde{\boldsymbol{K}}}\overline{\widetilde{\sigma^2_a}}$. Now the total genotypic variance simplifies to

$$\sigma^2_{g(asv)} = trace(\boldsymbol{P_\mu}\frac{\overline{\tilde{\boldsymbol{M}}\tilde{\boldsymbol{M}}}'\overline{\widetilde{\sigma^2_a}})}{n_g - 1} = \overline{\widetilde{\sigma^2_a}}, \tag{63}$$

which simplifies (58) and (59) to

$$\bar{h}^2_{\Delta BLUE} = \frac{\overline{\widetilde{\sigma^2_a}}}{\overline{\widetilde{\sigma^2_a}} + trace\left(n_g\boldsymbol{P_\mu C}_{11(\boldsymbol{g})}\right)} \tag{64}$$

and

$$\bar{h}^2_{\Delta BLUP} = \frac{\overline{\widetilde{\sigma^2_a}} - trace\left(n_g\boldsymbol{P_\mu C}_{22(\boldsymbol{g})}\right)}{\overline{\widetilde{\sigma^2_a}}}. \tag{65}$$

Thus, if $\overline{\tilde{\boldsymbol{M}}}$ is standardized as shown above, the genotypic variance we are trying to capture is conveniently estimated directly as the variance for $\boldsymbol{a}$. However, it should be noted that both $\tilde{\boldsymbol{K}}$ and $\overline{\tilde{\boldsymbol{K}}}$ are singular, and thus may lead to problems depending on the statistical software being used: ASReml-R 3.0 (Gilmour *et al.* 2009) will show an error message, while PROC MIXED in SAS 9.4 (SAS Institute Inc. 2013) will effectively compute a modified version of the mixed model equations (Piepho *et al.* 2012b; SAS Institute Inc. 2017).

We would like to point out that neither $\tilde{\boldsymbol{K}}$ nor $\overline{\tilde{\boldsymbol{K}}}$ are equal to the genomic relationship matrices proposed in VanRaden (2008). Their first method obtains the general relationship matrix as:

$$\boldsymbol{K_{VanRaden1}} = \frac{\tilde{\boldsymbol{M}}\tilde{\boldsymbol{M}}'}{\sum_{w=1}^{n_m} 2\boldsymbol{p_w}(1 - \boldsymbol{p_w})}, \tag{66}$$

where $\boldsymbol{p_w}$ are the mean frequencies of the second allele across all genotypes at marker $\boldsymbol{w} = \{1, \ldots, n_m\}$ and can also be expressed as $\boldsymbol{p_w} = \frac{1'_{n_g}\boldsymbol{M}}{2n_g}$. Their second method obtains the matrix as $\boldsymbol{K_{VanRaden2}} = \tilde{\boldsymbol{M}}\boldsymbol{L}\tilde{\boldsymbol{M}}'$ where $\boldsymbol{L}$ is a diagonal with elements

$$L_{w,w} = \frac{1}{n_m[2\boldsymbol{p_w}(1-\boldsymbol{p_w})]}. \tag{67}$$

Since we have $\tilde{\boldsymbol{K}} = \tilde{\boldsymbol{M}}\tilde{\boldsymbol{M}}'$, we find that opposed to $\tilde{\boldsymbol{K}}$ and therefore $\overline{\overline{\boldsymbol{K}}}$, (66) and (67) additionally require a division involving allele frequencies. It may also be pointed out that in a population with Hardy–Weinberg equilibrium, $trace(\boldsymbol{MM}')$ corresponds to the heterozygosity of the markers. However, it should be kept in mind that completely genetically homogeneous cultivars—such as inbred lines, DH-lines, clones, or hybrids—are definitely not in Hardy–Weinberg equilibrium. Finally, note that if we approximate $\bar{\bar{h}}^2_{\Delta\cdot\cdot_{BLUE}}$ via (59) and implement $\hat{\sigma}^2_{g(asv)}$ into h^2_{Piepho}, we find that $\bar{\bar{h}}^2_{\Delta\cdot\cdot_{BLUE}} = h^2_{Piepho}$ even in the general case. Analogously, we have $\bar{\bar{h}}^2_{\Delta\cdot\cdot_{BLUP}} = h^2_{Cullis}$.

Appendix E: Directly Simulate Response to Selection

As proposed by Piepho and Möhring (2007), one may directly simulate the response to selection via the same approach used to obtain H^2_{Sim} in this article. Since in (40) we simulate n_g pairs of true ($\boldsymbol{g_{sim}}$) and predicted ($\hat{\boldsymbol{g}}^{\boldsymbol{BLUP}}_{\boldsymbol{sim}}$) genotypic values, we can select a subset with the best values for $\hat{\boldsymbol{g}}^{\boldsymbol{BLUP}}_{\boldsymbol{sim}}$ and compute the response to selection for the sth simulation run as

$$R_s = \frac{\sum_{i\in\varphi_s}\boldsymbol{g_{sim}}}{n_{sel}}, \tag{68}$$

where φ_s is the selected subset of genotypes in the sth simulation run and n_{sel} is the number of selected genotypes (*i.e.* $n_{sel} \leq n_g$). Analogous to (41), we can then obtain the simulated expected response to selection as

$$R_{sim} = n^{-1}_{sim}\sum_{s=1}^{n_{sim}} R_s. \tag{69}$$

Note that, in principle, the simulation-based approach could also be adapted to simulate the joint distribution of true ($\boldsymbol{g_{sim}}$) and estimated ($\hat{\boldsymbol{g}}^{\boldsymbol{BLUE}}_{\boldsymbol{sim}}$) genotypic values. Furthermore, one could then compare the simulated expected response to selection values from both simulation approaches to determine whether to ultimately model the genotypic main effect as fixed or random. Yet, not least because of the ability to account for kinship information, we generally expect an analysis with random genotypic effects to be more useful than one with fixed genotypic effects [see Piepho *et al.* (2008)].

5. General discussion

In the last chapters, the fact that the standard method for computing heritability on an entry-mean basis is lacking the ability to account for, *e.g.,* heterogeneous variances or covariances in both the phenotypic and genotypic component has been pointed out multiple times. Yet at the same time, the sheer number of proposed alternative heritability methods that try to rectify this shortage shows that a generalization of heritability estimation in a modern plant breeding setting is on one hand in demand, but on the other hand not straight-forward. In fact, this thesis is not resolving the issue either. I argue that this is because the issue may simply be irresolvable - at least in the sense that there can never be an ultimate extension for estimating a generalized heritability on an entry-mean basis, which fulfills all of the following criteria:

(1) it has all the same interpretations as H^2_{Std}/h^2_{Std} in the simple, balanced setting,

(2) it can be used as both, an estimate to be plugged into the breeder's equation in order to obtain R and a descriptive measure as is possible with H^2_{Std}/h^2_{Std} in the simple, balanced case and

(3) it can generally be accepted by the academic and/or plant breeding community as the method of choice.

What this thesis *does* provide, however, is a substantial investigation into

(i) why the assumptions for H^2_{Std}/h^2_{Std} are seldom met in the modern plant breeding setting, so that a generalized heritability is needed,

(ii) why generalizing heritability is problematic,

(iii) how to extend the notion of heritability to an entry-difference basis and

(iv) how existing proposals as well as newly proposed methods by our group generalize heritability and how they compare to each other[1].

[1] This includes the example codes for calculating the different methods via multiple mixed model software packages that are made available at https://github.com/PaulSchmidtGit/Heritability

In the end, a recommendation is given on how heritability and R should be used in the general case. While the answers to (i)-(iv), as well as the recommendation are given and discussed in Chapters 1 – 4, some final remarks on aspects of these points are given in this Chapter 5.

5.1 The true heritability – "Why don't you just simulate it?"

On multiple occasions, may it be by a reviewer or someone at a conference, I was asked why I did not simply simulate a dataset with a given true heritability and then check which of the alternative methods find estimates closest to this true heritability. At a first glance, this is a very appealing approach and we considered the idea at the beginning of this work. Unfortunately, this method is doomed to fail. In a simple, balanced setting, the *true heritability* is essentially H^2_{Std}/h^2_{Std}. It is of course possible to simulate a dataset that represents a simple, balanced setting with a given heritability. As shown in Chapter 4, however, most of the proposed generalized heritability methods will then reduce to H^2_{Std}/h^2_{Std} and thus give the exact same estimate. Clearly, this approach does not acquire any new information, since we simulate a simple, balanced setting and hence there is no need for a generalized method in the first place.

If the simulated dataset deviates from the simple, balanced setting due to*, e.g.,* unbalance or genotypic population structures, we can no longer define a true heritability (in the same way). The question on how to simulate a true heritability in such a scenario is essentially the same as how to calculate a generalized heritability. Any decision on how to simulate the data given some sort of underlying true heritability (*e.g.* as the correlation between simulated phenotypic and simulated genotypic values or as a function of variance components) is not only ambiguous, but also directly and inevitably determines which of the alternative methods of estimation will yield the best results. In other words, if I knew how to simulate a *true heritability* in a non-simple, unbalanced case, then I would also know the appropriate generalized heritability estimation method.

5.2 Another complex generalized heritability method – "Aren't we overdoing it?"

In Chapter 4 we introduce H^2_Δ/h^2_Δ – a heritability based on an entry-difference basis. Furthermore, we propose two versions of it, one based on genotypic BLUEs and one based on genotypic BLUPs. The measure explicitly accounts for unbalanced data and any variances and covariances in both the true genotypic component and estimated/predicted genotypic component (*i.e.* phenotypic component). Accordingly, this approach is arguably the final product of this work and in a sense

the successor/an extension of H^2_{Piepho}, which was proposed by our group more than ten years ago (Piepho and Möhring, 2007). After presenting our work on H^2_Δ/h^2_Δ in a talk at the EUCARPIA Biometrics 2018 conference, Fred van Eeuwijk (Professor at the Department of Plant Sciences at Wageningen University, NL) asked a question that articulated thoughts I had been having as well. He basically questioned whether the efforts taken (by academia) to overcome the issues with heritability estimation in the modern plant breeding setting are worthwhile, since ultimately each new proposition seems to be getting more complicated while at the same time it is still an approximation. As an uncompromising alternative, he suggested that we could simply discourage the use and estimation of heritability in cases that strongly deviate from the simple balanced setting. To put his comment into the right light, it should be noted that Fred van Eeuwijk's group had published a paper with a secondary focus on generalized heritability methods earlier the same year (Rodríguez-Álvarez et al., 2018).

The criticism is indeed a crucial one and at the very heart of this thesis. After all, one major reason for the popularity of heritability in plant breeding was its simplicity, both in calculating, as well as in interpreting it. Both of these features are lost in the modern plant breeding setting. It is important to note that in practice the use of more sophisticated methods does not merely lead to higher efforts, but also makes analysis and interpretation more prone to error. Thus, given such a setting, we need to take a step back and answer the fundamental question of what can actually be gained from calculating any generalized heritability estimate in addition to other, simpler, but already very informative measures like*, e.g.,* variance component (vc) estimates or (an average of) the standard errors of a difference (s.e.d.) between genotypic BLUEs/BLUPs. Hanson and Robinson (1963) nicely phrase the matter at hand: "In any research program involving the estimation of heritability it is essential at the outset to know exactly why such an estimate is required. This is a truism, but in many instances it would appear that the estimation of heritability is the prime object of the experiment, rather than a link in the chain of research. Unfortunately, this is a criticism which can be leveled against many statistical techniques; they come to be regarded as an end in themselves, rather than a means to an end. Instead, they should, of course, be regarded merely as tools - albeit useful, and often powerful, tools - in the hands of the research worker."

As stated in the previous chapters, there are two main uses for heritability and they are addressed consecutively in the following sections.

5.3. Heritability as a mean to predict response to selection

As stated before, a key motivation for estimating heritability is that it can be used to predict the response to selection (R) via the breeder's equation. Yet, heritability can no longer simply be "that fraction of the selection differential expected to be gained when selection is practiced on a defined reference unit" (Nyquist and Baker, 1991), if genetic effects no longer have a "common heritability" due to a non-simple, unbalanced setting. Applying the breeder's equation by using an approximate mean heritability across genotypes will naturally result in a prediction for R that is only as good as the approximated heritability estimate.

Therefore, I here repeat the conclusion of Chapter 4, which has its roots in Piepho and Möhring (2007): In order to predict R in a modern plant breeding setting, one should not use the breeder's equation with some *ad hoc* measure of heritability. Instead, one should simulate R directly. This can be done*, e.g.,* via the simulation approach proposed by Piepho and Möhring (2007). Accordingly, we recommend avoiding the breeder's equation as a whole and thus there is no benefit of estimating (a generalized) heritability in this respect.

5.4. Heritability as a descriptive measure of precision and usefulness

When it comes to the descriptive function of heritability, there is arguably still benefit in estimating it, yet again it can only be as good as the approximation that is used in the framework of the respective method.

One may group all heritability methods regarding two criteria. The first one asks whether the method is accounting for heteroscedasticity and/or covariances in the genotypic and/or phenotypic component. To our knowledge, H^2_{Oakey}, H^2_{sim} and H^2_{Δ} are the only ones attempting to account for all of these features (see Table 1 in Chapter 4).

As a second criterion, one may ask whether more than a single scalar quantity (*i.e.* a local or central tendency) is returned as *the* heritability or whether information that is more detailed (*i.e.* dispersion) can be obtained in addition. Strictly speaking, all three heritability measures mentioned above provide multiple estimates for heritability and thus allow for evaluating its

distribution. We argue, however, that those of H^2_{Δ} (*i.e.* $H^2_{\Delta \mathrm{i}}$. and $H^2_{\Delta \mathrm{ij}}$) are the most relevant: When calculating H^2_{sim}, there are as many heritability estimates as there are simulation runs. Yet each one is already an average correlation across all genotypes. H^2_{Oakey} gives *components of heritability* as the non-zero eigenvalues, but their interpretation is abstract (see discussion Chapter 4).

Consequently, we argue that if heritability is only used to describe precision, then H^2_{Δ} gives the most detailed insight. Furthermore, we feel that its motivation, derivation and interpretation is rather intuitive. Moreover, note that H^2_{Δ} can also be used for generalized linear (mixed) models. Notice that obtaining $H^2_{\Delta \mathrm{ij}}$ involves both, all s.e.d. between genotypic means as well as genotypic variance components. Thus, the gain from using $H^2_{\Delta \mathrm{ij}}$ here is that it combines these two important model outputs. This, in a sense, is very much in the spirit of an original motivation for heritability, as otherwise one could have simply shown the error variance as a measure for precision in simple, balanced settings. Accordingly, there is a benefit to calculating $H^2_{\Delta \mathrm{ij}}$ in addition to s.e.d. if the genotypic variance is not proportional to identity matrices.

5.5 Related topics and outlook

An issue that is directly related to this work is the "missing heritability", which in turn involves the question of how to define the genomic variance. The latter generally represents the genomic counterpart to the genetic variance and refers to the variance of a trait which can be explained by a linear regression on a set of markers (Los Campos et al., 2015). It was coined *missing* heritability when it was found that the genomic variance often only captures a fraction of the genetic variance (Maher, 2008). While there are possible explanations for this issue, for example the linkage disequilibrium between markers and loci, there has also been an recent conceptual inquiry about how genomic variance should be estimated (Lehermeier et al., 2017; Schreck and Schlather, 2018). Basically, it was pointed out that the assumptions of classical quantitative genetics theory are not in line with assumptions often made by applied linear mixed models including random genotypic effects that make use of a kinship matrix such as the one in Chapter 4. More specifically, for $\boldsymbol{Zu}$ - the random effects part of these models - it is assumed that $\boldsymbol{Z}$ is fixed, whereas $\boldsymbol{u}$ is random. In the classical quantitative genetics theory, however, it is the other way around: $\boldsymbol{Z}$ is stochastic and $\boldsymbol{u}$ are fixed values.

In the same vein lies the problem of choosing the "base population" when a pedigree matrix is used instead of a marker-based kinship matrix. Powell et al. (2010) point out that the genotypes in the base population, *i.e.* the founders, are usually assumed to be genetically independent and only their descendants are correlated due to their common pedigree. Yet, given that pedigree information is present, the decision on how many generations back the base population should go, is completely ambiguous, which impedes understanding and interpretation of the resulting genetic variance estimates. Moreover, the interest of a plant breeder does not usually lie in the genetic variance of a base population, but of the population under current selection. This thesis has entirely focused on the present population and on models where $\boldsymbol{Z}$ is fixed. Currently, there is ongoing work on whether the expected sample variance of the random genetic effects is suited as an estimate for the genetic variance, even with correlated genotypes (Schreck, 2018).

In the end, findings in this area may have a direct impact on the estimation an interpretation of heritability whenever genomic variance estimation is involved.

6. References

Allard, R. W. (1976): Principles of plant breeding. New York: Wiley.

Bartlett, M. S. (1938): The approximate recovery of information from replicated field experiments with large blocks. In *J. Agric. Sci.* 28 (03), p. 418. DOI: 10.1017/S0021859600050875.

Bernardo, R. (1994): Prediction of maize single-cross performance using RFLPs and information from related hybrids. In *Crop Science* 34 (1), p. 20. DOI: 10.2135/cropsci1994.0011183X003400010003x.

Cochran, W. G.; Cox, G. M. (1992): Experimental designs. 2. ed., Wiley classics library ed. New York: Wiley (Wiley classics library).

Comstock, R. E.; Moll, R. H. (1963): Statistical genetics and plant breeding. Genotype-environment interactions. p. 164-196.

Crossa, J.; Pérez-Rodríguez, P.; Cuevas, J.; Montesinos-López, O.; Jarquín, D.; Los Campos, G. de et al. (2017): Genomic selection in plant breeding. methods, models, and perspectives. In *Trends in plant science* 22 (11), pp. 961–975. DOI: 10.1016/j.tplants.2017.08.011.

Cullis, B. R.; Gleeson, A. C. (1991): Spatial analysis of field experiments - an extension to two dimensions. In *Biometrics* 47 (4), p. 1449. DOI: 10.2307/2532398.

Cullis, B. R.; Smith, A. B.; Coombes, N. E. (2006): On the design of early generation variety trials with correlated data. In *Journal of Agricultural, Biological, and Environmental Statistics* 11 (4), pp. 381–393. DOI: 10.1198/108571106X.

Damesa, T. M.; Möhring, J.; Forkman, J.; Piepho, H.-P. (2018): Modeling spatially correlated and heteroscedastic errors in ethiopian maize trials. In *Crop Science* 58 (4), p. 1575. DOI: 10.2135/cropsci2017.11.0693.

Edwards, J. W.; Jannink, J.-L. (2006): Bayesian modeling of heterogeneous error and genotype × environment interaction variances. In *Crop Science* 46 (2), p. 820. DOI: 10.2135/cropsci2005.0164.

Eisenhart, C. (1947): The assumptions underlying the analysis of variance. In *Biometrics* 3 (1), p. 1. DOI: 10.2307/3001534.

Federer, W. T. (1956): Augmented (or hoonuiaku) designs. In *Biometrics Unit* (Cornell Univ. Mimeo. BU-74-M).

Fienberg, S. E.; Hinkley, D. V. (Eds.) (1980): R.A. Fisher. An appreciation. New York, NY: Springer (Lecture Notes in Statistics, 1).

Finlay, K. W.; Wilkinson, G. N. (1963): The analysis of adaptation in a plant-breeding programme. In *Aust. J. Agric. Res.* 14 (6), p. 742. DOI: 10.1071/AR9630742.

Fisher, R. A. (1921): On the "probable error" of a coefficient of correlation deduced from a small sample. In *Metron* 1, pp. 3–32.

Fisher, R. A. (1922): On the mathematical foundations of theoretical statistics. In *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 222 (594-604), pp. 309–368. DOI: 10.1098/rsta.1922.0009.

Fisher, R. A. (1935): The design of experiments. Oxford, England: Oliver & Boyd.

Gauch, H. G. (1992): Statistical analysis of regional yield trials. AMMI analysis of factorial designs. Amsterdam u.a.: Elsevier.

Gilmour, A. R.; Cullis, B. R.; Verbyla, A. P.; Verbyla, A. P. (1997): Accounting for natural and extraneous variation in the analysis of field experiments. In *Journal of Agricultural, Biological, and Environmental Statistics* 2 (3), p. 269. DOI: 10.2307/1400446.

Hanson, W.; Robinson, H. (1963): Statistical genetics and plant breeding. Washington, D.C.: National academy of sciences - national research council.

Henderson, C. R. (1986): Statistical methods in animal improvement. Historical overview. In Daniel Gianola, Keith Hammond (Eds.): Advances in statistical methods for genetic improvement of livestock, vol. 18. Berlin: Springer (Advanced Series in Agricultural Sciences, 0172-4207, 18), pp. 2–14.

Kleinknecht, K.; Möhring, J.; Singh, K. P.; Zaidi, P. H.; Atlin, G. N.; Piepho, H. P. (2013): Comparison of the performance of best linear unbiased estimation and best linear unbiased prediction of genotype effects from zoned indian maize data. In *Crop Science* 53 (4), p. 1384. DOI: 10.2135/cropsci2013.02.0073.

Lehermeier, C.; Los Campos, G. de; Wimmer, V.; Schön, C.-C. (2017): Genomic variance estimates: With or without disequilibrium covariances? In *Journal of animal breeding and genetics* 134 (3), pp. 232–241. DOI: 10.1111/jbg.12268.

Los Campos, G. de; Sorensen, D.; Gianola, D. (2015): Genomic heritability: what is it? In *PLoS genetics* 11 (5), e1005048. DOI: 10.1371/journal.pgen.1005048.

Maher, B. (2008): Personal genomes: The case of the missing heritability. In *Nature* 456 (7218), pp. 18–21. DOI: 10.1038/456018a.

Meuwissen, T. H.; Hayes, B. J.; Goddard, M. E. (2001): Prediction of total genetic value using genome-wide dense marker maps. In *Genetics* 157 (4), pp. 1819–1829.

Moehring, J.; Williams, E. R.; Piepho, H.-P. (2014): Efficiency of augmented p-rep designs in multi-environmental trials. In *TAG. Theoretical and applied genetics* 127 (5), pp. 1049–1060. DOI: 10.1007/s00122-014-2278-y.

Nyquist, W. E.; Baker, R. J. (1991): Estimation of heritability and prediction of selection response in plant populations. In *Critical Reviews in Plant Sciences* 10 (3), pp. 235–322. DOI: 10.1080/07352689109382313.

Papadakis, J. S. (1937): Méthode statistique pour des expériences sur champ. In *Bull. Inst. Amél. Plantes á Salonique* 23.

Patterson, H. D.; Thompson, R. (1971): Recovery of inter-block information when block sizes are unequal. In *Biometrika* 58 (3), pp. 545–554. DOI: 10.1093/biomet/58.3.545.

Patterson, H. D.; Williams, E. R. (1976): A new class of resolvable incomplete block designs. In *Biometrika* 63 (1), pp. 83–92. DOI: 10.1093/biomet/63.1.83.

Piepho, H. P.; Buchse, A.; Emrich, K. (2003): A Hitchhiker's Guide to Mixed Models for Randomized Experiments. In *J Agron Crop Sci* 189 (5), pp. 310–322. DOI: 10.1046/j.1439-037X.2003.00049.x.

Piepho, H.-P. (1998): Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. In *Theor Appl Genet* 97 (1-2), pp. 195–201. DOI: 10.1007/s001220050885.

Piepho, H.-P.; Möhring, J. (2007): Computing heritability and selection response from unbalanced plant breeding trials. In *Genetics* 177 (3), pp. 1881–1888. DOI: 10.1534/genetics.107.074229.

Pillen, K.; Zacharias, A.; Léon, J. (2003): Advanced backcross QTL analysis in barley (Hordeum vulgare L.). In *TAG. Theoretical and applied genetics* 107 (2), pp. 340–352. DOI: 10.1007/s00122-003-1253-9.

Powell, J. E.; Visscher, P. M.; Goddard, M. E. (2010): Reconciling the analysis of IBD and IBS in complex trait studies. In *Nature Reviews Genetics* 11 (11), pp. 800–805. DOI: 10.1038/nrg2865.

Rodríguez-Álvarez, M. X.; Boer, M. P.; van Eeuwijk, F. A.; Eilers, P. H.C. (2018): Correcting for spatial heterogeneity in plant breeding experiments with P-splines. In *Spatial Statistics* 23, pp. 52–71. DOI: 10.1016/j.spasta.2017.10.003.

Savage, L. J. (1976): On rereading R. A. Fisher. In *The Annals of Statistics* 4 (3), pp. 441–500.

Schreck, N.; Schlather, M. (2018): From estimation to prediction of genomic variances: allowing for linkage disequilibrium and unbiasedness. In *bioRxiv: 282343*. DOI: 10.1101/282343.

Schreck, N. M. (2018): From estimation to prediction of genomic variances: allowing for linkage disequilibrium and unbiasedness. Dissertation. University, Mannheim. Available online at https://ub-madoc.bib.uni-mannheim.de/view/types/dissertation.html.

Searle, S. R.; Casella, G.; McCulloch, C. E. (2006): Variance components. New York: Wiley.

Smith, A.; Cullis, B.; Thompson, R. (2001): Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. In *Biometrics* 57 (4), pp. 1138–1147. DOI: 10.1111/j.0006-341X.2001.01138.x.

Smith, A. B.; Cullis, B. R. (2018): Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. In *Euphytica* 214 (8), p. 992. DOI: 10.1007/s10681-018-2220-5.

So, Y.-S.; Edwards, J. (2009): A comparison of mixed-model analyses of the Iowa crop performance test for corn. In *Crop Science* 49 (5), p. 1593. DOI: 10.2135/cropsci2008.09.0574.

Utz, H. F.: Personal communication. 2016. Professor, Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim.

VanRaden, P. M. (2008): Efficient methods to compute genomic predictions. In *Journal of dairy science* 91 (11), pp. 4414–4423. DOI: 10.3168/jds.2007-0980.

Wilkinson, G. N.; Eckert, S. R.; Hancok, T. W.; Mayo, O. (1983): Nearest neighbour (NN) analysis of field experiments. In *Journal of the Royal Statistical Society* Series B (Methodological), pp. 151–211.

Williams, E.; Piepho, H.-P.; Whitaker, D. (2011): Augmented p-rep designs. In *Biometrical journal* 53 (1), pp. 19–27. DOI: 10.1002/bimj.201000102.

Wricke, G.; Weber, E. (1986): Quantitative genetics and selection in plant breeding. Berlin, New York: Walter de Gruyter.

Wright, S. (1920): The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. In *Proceedings of the National Academy of Sciences* 6 (6), pp. 320–332. DOI: 10.1073/pnas.6.6.320.

Würschum, T. (2012): Mapping QTL for agronomic traits in breeding populations. In *TAG. Theoretical and applied genetics* 125 (2), pp. 201–210. DOI: 10.1007/s00122-012-1887-6.

Xu, S. (2013): Principles of statistical genomics. New York, NY: Springer. Available online at http://dx.doi.org/10.1007/978-0-387-70807-2.

Yan, W. (2014): Crop variety trials. Data management and analysis. Chichester, West Sussex, UK: John Wiley & Sons Inc. Available online at http://lib.myilibrary.com/detail.asp?id=584469.

Yates, F. (1936): Incomplete randomized blocks. In *Annals of Eugenics* 7 (2), pp. 121–140. DOI: 10.1111/j.1469-1809.1936.tb02134.x.

Yates, F. (1940a): Lattice squares. In *J. Agric. Sci.* 30 (04), p. 672. DOI: 10.1017/S0021859600048292.

Yates, F. (1940b): The recovery of inter-block information in balanced incomplete block designs. In *Annals of Eugenics* 10 (1), pp. 317–325. DOI: 10.1111/j.1469-1809.1940.tb02257.x.

Yates, F. (Ed.) (1990): Foreword in. Fisher, Ronald Aylmer; Bennett, J. H.; Yates, F. Statistical methods, experimental design, and scientific inference. Oxford: Oxford Univ. Press.

Yates, F.; Cochran, W. G. (1938): The analysis of groups of experiments. In *J. Agric. Sci.* 28 (04), p. 556. DOI: 10.1017/S0021859600050978.

7. Summary

Heritability is an important notion in, *e.g.,* human genetics, animal breeding and plant breeding, since the focus of these fields lies on the relationship between phenotypes and genotypes. A phenotype is the composite of an organism's observable traits, which is determined by its underlying genotype, by environmental factors and by genotype-environment interactions. For a set of genotypes, the notion of heritability expresses the proportion of the phenotypic variance that is attributable to the genotypic variance. Furthermore, as it is an intraclass correlation, heritability can also be interpreted as, *e.g.,* the squared correlation between phenotypic and genotypic values.

It is important to note that heritability was originally proposed in the context of animal breeding where it is the individual animal that represents the basic unit of observation. This stands in contrast to plant breeding, where multiple observations for the same genotype are obtained in replicated trials. Furthermore, trials are usually conducted as multi-environment trials (MET), where an environment denotes a year × location combination and represents a random sample from a target population of environments. Hence, the observations for each genotype first need to be aggregated in order to obtain a single phenotypic value, which is usually done by obtaining some sort of mean value across trials and replicates. As a consequence, heritability in the context of plant breeding is referred to as heritability on an entry-mean basis and its standard estimation method is a linear combination of variances and trial dimensions.

Ultimately, I find that there are two main uses for heritability in plant breeding: The first is to predict the response to selection and the second is as a descriptive measure for the usefulness and precision of cultivar trials. Heritability on an entry-mean basis is suited for both purposes as long as three main assumptions hold: (i) the trial design is completely balanced/orthogonal, (ii) genotypic effects are independent and (iii) variances and covariances are constant.

In the last decades, however, many advancements in the methodology of experimental design for and statistical analysis of plant breeding trials took place. As a consequence it is seldom the case that all three of above mentioned assumptions are met. Instead, the application of linear mixed models enables the breeder to straightforwardly analyze unbalanced data with complex

variance structures. Chapter 2 exemplarily demonstrates some of the flexibility and benefit of the mixed model framework for typically unbalanced MET by using a bivariate mixed model analyses to jointly analyze two MET for cultivar evaluation, which differ in multiple crucial aspects such as plot size, trial design and general purpose. Such an approach can lead to higher accuracy and precision of the analysis and thus more efficient and successful breeding programs.

It is not clear, however, how to define and estimate a generalized heritability on an entry-mean basis for such settings. Therefore, multiple alternative methods for the estimation of heritability on an entry-mean basis have been proposed. In Chapter 3, six alternative methods are applied to four typically unbalanced MET for cultivar evaluation and compared to the standard method. The outcome suggests that the standard method over-estimates heritability, while all of the alternative methods show similar, lower estimates and thus seem able to handle this kind of unbalanced data.

Finally, it is argued in Chapter 4 that heritability in plant breeding is not actually based on or aiming at entry-means, but on the differences between them. Moreover, an estimation method for this new proposal of heritability on an entry-difference basis ($H_{\Delta}^2/h_{\Delta}^2$) is derived and discussed, as well as exemplified and compared to other methods via analyzing four different datasets for cultivar evaluation which differ in their complexity. I argue that regarding the use of heritability as a descriptive measure, $H_{\Delta}^2/h_{\Delta}^2$, can on the one hand give a more detailed and meaningful insight than all other heritability methods and on the other hand reduces to other methods under certain circumstances. When it comes to the use of heritability as a means to predict the response to selection, the outcome of this work discourages this as a whole. Instead, response to selection should be simulated directly and thus without using any *ad hoc* heritability measure.

8. Zusammenfassung

In der Humangenetik, Tier- und Pflanzenzüchtung sowie anderen Forschungsdisziplinen, bei denen die Beziehung zwischen Genotypen und Phänotypen im Fokus steht, ist die Heritabilität eine wichtige Maßzahl. Der Phänotyp setzt sich aus einem oder mehreren beobachteten Merkmalen eines Organismus zusammen und wird durch den zugrunde liegenden Genotypen, durch Umwelteinflüsse, sowie durch Genotyp-Umwelt-Wechselwirkungseffekte bestimmt. Die Heritabilität gibt an, welcher Anteil der phänotypischen Varianz genetisch bedingt ist. Sie kann als quadrierte Korrelation zwischen phänotypischen und genotypischen Werte interpretiert werden.

Ursprünglich wurde die Heritabilität in der Tierzüchtung vorgeschlagen, in welcher das einzelne Tier die kleinste Beobachtungseinheit darstellt. Dies steht im Gegensatz zur Pflanzenzüchtung, in der meist wiederholte Versuche durchgeführt werden, so dass derselbe Genotyp reproduziert und mehrfach beobachtet werden kann. Hinzu kommt, dass die Versuche meist in Versuchsserien an mehreren Standorten und über mehrere Jahre hinweg durchgeführt werden. Um also einen phänotypischen Wert je Genotyp zu erhalten, müssen dessen Beobachtungen aggregiert werden, was meist durch eine Form von Mittelwertbildung geschieht. Aus diesem Grund wird Heritabilität in der Pflanzenzüchtung standardmäßig als Heritabilität auf Sortenmittelwertbasis.

Ich sehe zwei Hauptnutzen von Heritabilität in der Pflanzenzüchtung: Zum einen kann mit ihr der Selektionserfolg vorhergesagt werden und zum anderen dient sie als beschreibende Maßzahl für die Präzision und Brauchbarkeit eines Versuchs. Die Heritabilität auf Sortenmittelwertbasis ist für beide Zwecke geeignet solange folgende Bedingungen erfüllt sind: (i) Das Versuchsdesign ist vollkommen balanciert/orthogonal, (ii) die Genotyp-Effekte sind unabhängig und (iii) alle Varianzen, sowie Kovarianzen sind konstant.

In den letzten Jahrzehnten gab es mehrere Weiterentwicklungen in der Methodik des Versuchsdesigns sowie der statistischen Analyse von Pflanzenzüchtungsversuchen. Gemischte Modelle ermöglichen komplexe Varianzstrukturen und unbalancierte Daten auszuwerten In Kapitel 2 wird beispielhaft gezeigt, welche Möglichkeiten und Vorteile in der Anwendung von gemischten Modellen liegen, indem typisch unbalancierte Datensätze von zwei verschiedenen Sortenversuchsserien mithilfe eines bivariaten gemischten Modells gemeinsam ausgewertet

werden. Ansätze wie dieser können eine höhere Analyseexaktheit und -präzision erzielen und demnach die Effizienz und den Erfolg von Pflanzenzüchtungsprogrammen steigern. Gleichzeitig führt dies dazu, dass die oben genannten Bedingungen nur selten erfüllt sind.

In solchen Fällen ist dann nicht klar, wie eine Heritabilität auf Sortenmittelwertbasis definiert und geschätzt werden kann. Mehrere alternative Methoden wurden vorgeschlagen. In Kapitel 3 werden sechs dieser alternativen Methoden für vier typische Datensätze aus Sortenversuchsserien berechnet und miteinander, sowie mit der Standardmethode verglichen. Die Ergebnisse deuten darauf hin, dass letztere die Heritabilität überschätzt, während alle alternativen Methoden ähnliche, niedrigere Schätzungen zeigen. Dies lässt vermuten, dass diese Methoden besser für die vorliegenden, unbalancierten Daten geeignet sind.

Abschließend wird in Kapitel 4 gezeigt, dass Heritabilität in der Pflanzenzüchtung im Grunde nicht auf Genotypmittelwerten sondern auf deren Differenzen basiert. Hieraus wird eine Methode zur Berechnung einer generalisierten Heritabilität auf Sortendifferenzbasis (H_Δ^2/h_Δ^2) hergeleitet und diskutiert. Vier unterschiedlich komplexe Datensätze von Sortenversuchen werden verwendet und mit alternativen Heritabilitätsschätzern verglichen. Bezüglich der Verwendung der Heritabilität als beschreibende Maßzahl bietet H_Δ^2/h_Δ^2 einen ausführlicheren und bedeutsameren Einblick als die alternativen Heritabilitätsschätzer oder die Standardmethode. Hinzu kommt, dass H_Δ^2/h_Δ^2 die bisher bekannten Methoden nicht nur verallgemeinert, sondern in Spezialfällen exakt abbildet. Basierend auf den Resultaten der gesamten Arbeit rate ich von der Verwendung von Heritabilität als Mittel zur Vorhersage des Selektionserfolges ab. Der Selektionserfolg sollte stattdessen direkt simuliert werden, sodass die Nutzung einer *ad hoc* Schätzungsmethode der Heritabilität unnötig ist.

9. Acknowledgements

Foremost, I want to thank my supervisor Prof. Hans-Peter Piepho who guided me since my M.Sc. thesis and taught me a lot - not only in biostatistics. His ability to combine patience, kindness and diplomacy with large workloads, deadlines and uncompromising quality, while at the same time keeping an open door for spontaneous discussions, baffles and inspires me to this day. On top of all this, he not only tolerated, but supported ideas that are not by the book, like new (software) solutions to old problems or my visit at Purdue University. I truly believe for me personally, this was one of the best Ph.D. experiences I could have asked for.

Then, I need to give a big *Thank you* to Jens Hartung. Also starting during my M.Sc. thesis and without any obligation to do so, he gave me countless advice and hence in a way became a second supervisor to me. I am very thankful for this and can say without a doubt that the progress of my work would not have been the same without him in the office next door.

Obviously, I also need to thank my other colleagues and friends Steffen, Nhã and everyone else who made the atmosphere productive at work, harmonic in the breaks and fun in our free time.

Finally, I would like to thank Hoa and my parents, as well as my friends and family. It is their consistent love, support and reliability that allows me to wholeheartedly throw myself at challenges like this.

9. Eidesstattliche Versicherung

Annex 3

Declaration in lieu of an oath on independent work

according to Sec. 18(3) sentence 5 of the University of Hohenheim's Doctoral Regulations for the Faculties of Agricultural Sciences, Natural Sciences, and Business, Economics and Social Sciences

1. The dissertation submitted on the topic

Estimating heritability in plant breeding programs

.....

.....

is work done independently by me.

2. I only used the sources and aids listed and did not make use of any impermissible assistance from third parties. In particular, I marked all content taken word-for-word or paraphrased from other works.

3. I did not use the assistance of a commercial doctoral placement or advising agency.

4. I am aware of the importance of the declaration in lieu of oath and the criminal consequences of false or incomplete declarations in lieu of oath.

I confirm that the declaration above is correct. I declare in lieu of oath that I have declared only the truth to the best of my knowledge and have not omitted anything.

_____ _____

Place, Date Signature

10. Curriculum vitae (Tabellarischer Lebenslauf)

Name:	Paul Schmidt
Date of birth:	November 27, 1989
Place of birth:	Rostock
Mail:	schmidtpaul@hotmail.de
Phone:	+49 (0)172 3091577
Address:	Kaltenkirchener Str. 8, 22769 Hamburg

Professional experience

Since 11/2018	Data Scientist at BioMath GmbH, 18119 Rostock
Since 09/2015	PhD student at the Institute of Biostatistics, University of Hohenheim
01/2015 – 08/2015	Research associate at BioMath GmbH, 18119 Rostock

Education

10/2012 – 12/2014	Master program *Crop Science* [English], with emphasis on plant breeding, University of Hohenheim. Degree: Master of Science
10/2009 – 09/2012	Bachelor program *Agrobiology* [German], with emphasis on crop sciences, University of Hohenheim. Degree: Bachelor of Science
2000 – 2009	High School (Innerstädtisches Gymnasium Rostock)
2006 – 2007	Student exchange year in Taylorsville, North Carolina, USA including High-School-Diploma (Alexander Central High School)

.....

Ort, Datum, Unterschrift

11. List of publications

List of scientific publications included in this doctoral thesis

Schmidt, P., Hartung, J., Bennewitz, J., and Piepho, H.-P., 2019. Heritability in plant breeding on a genotype-difference basis. (submitted)

Schmidt, P., Hartung, J., Rath, J., Piepho, H.-P. (2019): Estimating broad-sense heritability with unbalanced data from agricultural cultivar trials. In Crop Science 59 (2), p. 0525–536. DOI: 10.2135/cropsci2018.06.0376.

Schmidt, P., Möhring, J., Koch, R. J. and Piepho, H.-P., 2018b. More, larger, simpler: how comparable are on-farm and on-station trials for cultivar evaluation? Crop Science 58(4):1508. doi: 10.2135/cropsci2017.09.0555.

Complete list of scientific publications

Schmidt, K., Schmidtke, J., Schmidt, P., Kohl, C., Wilhelm, R., Schiemann, J., van der Voet, H. and Steinberg, P., 2017. Variability of control data and relevance of observed group differences in five oral toxicity studies with genetically modified maize MON810 in rats. Archives of toxicology 91(4):1977–2006. doi: 10.1007/s00204-016-1857-x.

Schmidt, P., Hartung, J., Bennewitz, J., and Piepho, H.-P., 2019. Heritability in plant breeding on a genotype-difference basis. (submitted)

Schmidt, P., Hartung, J., Rath, J., Piepho, H.-P. (2019): Estimating broad-sense heritability with unbalanced data from agricultural cultivar trials. In Crop Science 59 (2), p. 0525–536. DOI: 10.2135/cropsci2018.06.0376.

Schmidt, P., Möhring, J., Koch, R. J. and Piepho, H.-P., 2018b. More, larger, simpler: how comparable are on-farm and on-station trials for cultivar evaluation? Crop Science 58(4):1508. doi: 10.2135/cropsci2017.09.0555.

Tulinská, J., Adel-Patient, k., Bernard, H., Líšková, A., Kuricová, M., Ilavská, S., Horváthová, M., Kebis, A., Rollerová, E., Babincová, J., Aláčová, R., Wal, J.-M., Schmidt, K., Schmidtke, J., Schmidt, P., Kohl, C., Wilhelm, R., Schiemann, J. and Steinberg, P., 2018. Humoral and cellular immune response in Wistar Han RCC rats fed two genetically modified maize MON810 varieties for 90 days (EU 7th

Framework Programme project GRACE). Archives of Toxicology 92(7):2385–2399. doi: 10.1007/s00204-018-2230-z.

Zeljenková, D., Aláčová, R., Ondrejková, J., Ambrušová, K., Bartušová, M., Kebis, A., Kovrižnych, J., Rollerová, E., Szabová, E., Wimmerová, S., Černák, M., Krivošíková, Z., Kuricová, M., Líšková, A., Spustová, V., Tulinská, J., Levkut, M., Révajová, V., Ševčíková, Z., Schmidt, K., Schmidtke, J., Schmidt, P., La Paz, J. L., Corujo, M., Pla, M., Kleter, G. A., Kok, E. J., Sharbati, J., Bohmer, M., Bohmer, N., Einspanier, R., Adel-Patient, K., Spök, A., Pöting, A., Kohl, C., Wilhelm, R., Schiemann, J. and Steinberg, P., 2016. One-year oral toxicity study on a genetically modified maize MON810 variety in Wistar Han RCC rats (EU 7th Framework Programme project GRACE). Archives of Toxicology 90(10):2531–2562. doi: 10.1007/s00204-016-1798-4.