Heinrich Heine
Universität
Düsseldorf

Effective strategies incorporating genomic selection to improve short- and long-term genetic gain in plant breeding programs

Inaugural dissertation

for the attainment of the title of doctor
in the Faculty of Mathematics and Natural Sciences
at the Heinrich Heine University Düsseldorf

presented by

Po-Ya Wu
from New Taipei City, Taiwan

Düsseldorf, January 2024

From the institute for Quantitative Genetics
and Genomics of Plants
at the Heinrich Heine University Düsseldorf

Published by the permission of the
Faculty of Mathematics and Natural Sciences at
Heinrich Heine University Düsseldorf

Contributors:
1. Prof. Dr. Benjamin Stich
2. Prof. Dr. Friedrich Longin

Date of the oral examination: 13.05.2024

Declaration of the Doctoral Dissertation

I herewith declare under oath that this dissertation was the result of my own work without any unauthorized help in compliance with the "Principles for the Safeguarding of Good Scientific Practice at Heinrich Heine University Düsseldorf". This dissertation has never been submitted in this or similar format to any other institution. I have not previously failed a doctoral examination procedure.

Düsseldorf, 25.01.2024

_____

Po-Ya Wu

Contents

1 Summary **1**

2 General Introduction **3**

3 Optimal implementation of genomic selection in clone breeding programs – exemplified in potato: I. Effect of selection strategy, implementation stage, and selection intensity on short-term genetic gain[1] **18**

4 Optimal implementation of genomic selection in clone breeding programs - exemplified in potato: II. Effect of selection strategy and cross selection methods on long-term genetic gain[2] **47**

5 Improvement of prediction ability by integrating multi-omic datasets in barley[3] **91**

6 Structural variants in the barley gene pool: precision and sensitivity to detect them using short-read sequencing and their association with gene expression and phenotypic variation[4] **113**

7 List of publications **136**

8 Acknowledgements **137**

---

[1] **Wu, P.-Y.**, B. Stich, J. Renner, R. Muders, V. Prigge, D. van Inghelandt. 2023. The Plant Genome, e20327

[2] **Wu, P.-Y.**, B. Stich, J. Renner, R. Muders, V. Prigge, D. van Inghelandt. 2024. In preparation

[3] **Wu, P.-Y.**, B. Stich, M. Weisweiler, A. Shrestha, A. Erban, P. Westhoff, D. van Inghelandt. 2022. BMC geomics, 23(1), 200

[4] Weisweiler, M., C. Arlt*, **P.-Y. Wu***, D. van Inghelandt, T. Hartwig, B. Stich. 2022. Theoretical and Applied Genetics, 135(10), 3511-3529

*Contributed equally

1 Summary

Genomic selection (GS) based on single nucleotide polymorphisms (SNP) has emerged as a powerful tool to increase the genetic gain of complex traits in breeding programs of various animal and plant species. However, its optimal integration especially in clone breeding programs, and its combination with the cross-selection (CS) method in heterozygous and tetraploid crops to balance genetic gain and diversity in long-term breeding programs are still lacking. Another important aspect affecting the success of genetic gain is the degree of prediction accuracy/ability of a GS model. The use of additional or alternative layers of omics datasets closer to phenotypes as predictors may improve the prediction ability. The main objectives of this thesis were to (1) optimize potato breeding programs incorporating GS using computer simulation; and (2) improve the efficiency of GS using different omic datasets and structural variants as predictors compared to SNP array, taking barley as an experimental example. Both approaches have the final goal to further enhance the genetic gain in breeding programs. In the simulation results, implementing GS with optimal selection intensities had a higher short- and long-term genetic gain compared to the phenotypic selection solely. In addition, implementing GS in consecutive selection stages largely increased genetic gain compared to using GS in one stage. Furthermore, the results of my computer simulations suggest that the optimal selection intensities require to be adjusted under different scenarios considering cost, selection strategies, prediction accuracy of the GS model, etc. When studying the long-term selection response, the CS method considering additive and dominance effects to predict progeny mean based on simulated progenies (MEGV-O) reached the highest accuracy in predicting progeny mean and the highest long-term gain among the CS methods that only consider the progeny mean. However, it accompanied the loss of genetic variance quickly. The linear combination of usefulness criteria (UC) and genome-wide diversity, which was called EUCD, kept the same level of genetic gain compared to UC and MEGV-O. However, EUCD simultaneously kept a higher diversity as well as a certain degree of genetic variance compared to UC and MEGV-O. Therefore, these results of my thesis can provide breeders with a concrete method to improve their potato breeding programs and are presumably also helpful for other clone breeding programs. In the frame of the other aspect studied in this thesis, the prediction ability of the GS model using deleterious sequence variants, structural variants, transcriptome, and metabolome as a single predictor, was higher than

using SNP array on average across the assessed traits. Optimally combining the information of several layers of omic datasets in the GS model outperformed single predictors alone. Therefore, the results of my thesis will open the path to perform such analysis on a large scale segregating populations and even apply for potato breeding programs to boost genetic gain.

2 General Introduction

Potato (*Solanum tuberosum* L.) and barley (*Hordeum vulgare* L.) are two important crops, ranking fourth and seventh in their world-wide production with 375 and 155 million metric tons, respectively (FAO, 2022). In response to the demands of the growing global population, the shift of dietary to an increased meat consumption, the aim to increase biofuel production, and the reduction of arable land, makes the production of a sufficient amount of food from both crops a major challenge in current agriculture (Fróna et al., 2019). Model calculations predict that at least a doubling of the current production by 2050 is required (Ray et al., 2013). Furthermore, it is expected that climate change has a negative influence on crop production due to an increase in extreme temperatures and an alternation of rainfall patterns (Abberton et al., 2016). Therefore, developing new varieties with high and stable yields and stress tolerance for both crops is one of the important missions of plant breeding. Especially for potato, a low genetic gain in the past decades was observed (Stokstad, 2019; Ortiz et al., 2022) compared to most cereal crops (Figure 1). This is presumably because of its tetraploid and highly heterozygous genome (Lindhout et al., 2011; Jansky et al., 2016). In addition, its low multiplication coefficient (Grüneberg et al., 2009) leads to the availability of only few tubers per genotype for phenotypic assessment in the early breeding program (Gopal, 2006). The evaluation of traits related to productivity or quality (target traits) has to be postponed until the later breeding program with enough tubers, as these traits rely on multi-location field trials and/or destructive assessment.

According to the breeder's equation (Falconer and Mackay, 1996), the expected genetic gain is defined as $\Delta G = \frac{i \cdot h \cdot \sigma_G}{L}$, where i is the selection intensity, h the square root of heritability, σ_G the square root of genetic variance, and L the length of breeding cycle. Straightforwardly, genetic gain can be enhanced by increasing heritability, selection intensity, and genetic variance or shortening breeding cycles. Typically, heritability can be improved by increasing the number of locations, years, and replications in the field trials. Selection intensity can be enhanced by either increasing the testing population size or decreasing the number of selected candidates. Genetic variance can be increased by introducing novel germplasm (Xu et al., 2020). However, all these ways increase the required budget. To enhance genetic gain under a fixed budget, it is necessary to incorporate new approaches by adjusting the parameters in the breeder's equation. Genomic selection (GS), for example, could be one way.

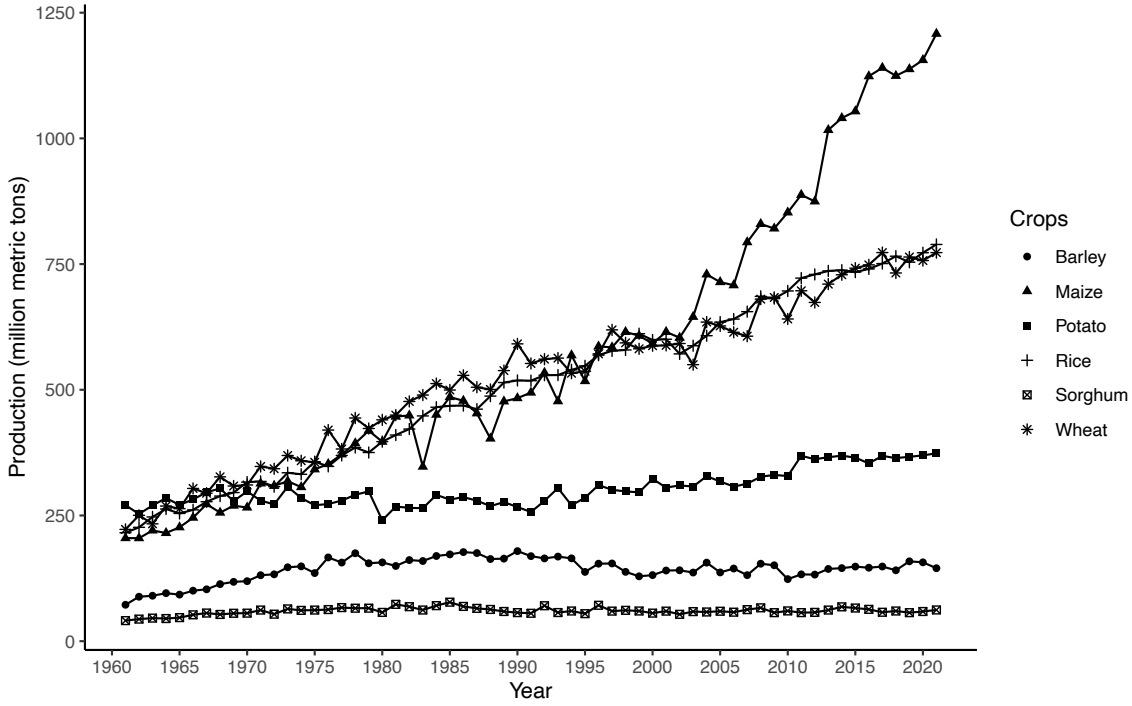

Figure 1. Trends in the production of maize, wheat, rice, potato, barley, and sorghum over the world from 1961 to 2021.

The concept of genomic selection

GS has become a powerful tool to increase genetic gain for complex traits in both livestock and plant breeding programs (Meuwissen et al., 2001; Desta and Ortiz, 2014). The concept of GS is to capture all effects of quantitative trait loci using dense single nucleotide polymorphism (SNP) markers across an entire genome. The GS model is first constructed by exploiting the known phenotypes and genotypes in a training set. Then, the resulting GS model can predict estimated genetic values (EGV) for individuals with only genotypes in an untested set. Thus, the superior individuals based on their EGV can be preselected before their phenotypes are measured in the field, which can potentially shorten the breeding cycles. In addition, GS can quickly and precisely identify the individuals carrying the most favorable alleles, especially if those traits (e.g. target trait) can not be easily measured or assessed at early stages. Meanwhile, the use of GS can reduce phenotyping costs as only genotypic information is required for the untested set. If genotyping is less costly than phenotyping, the total number of testing candidates can be increased if the number of selected individuals is kept constant, which can increase the selection intensity (Cobb et al., 2019). Therefore, using GS in breeding programs that rely solely on phenotypic selection (PS) has the potential to improve genetic gain.

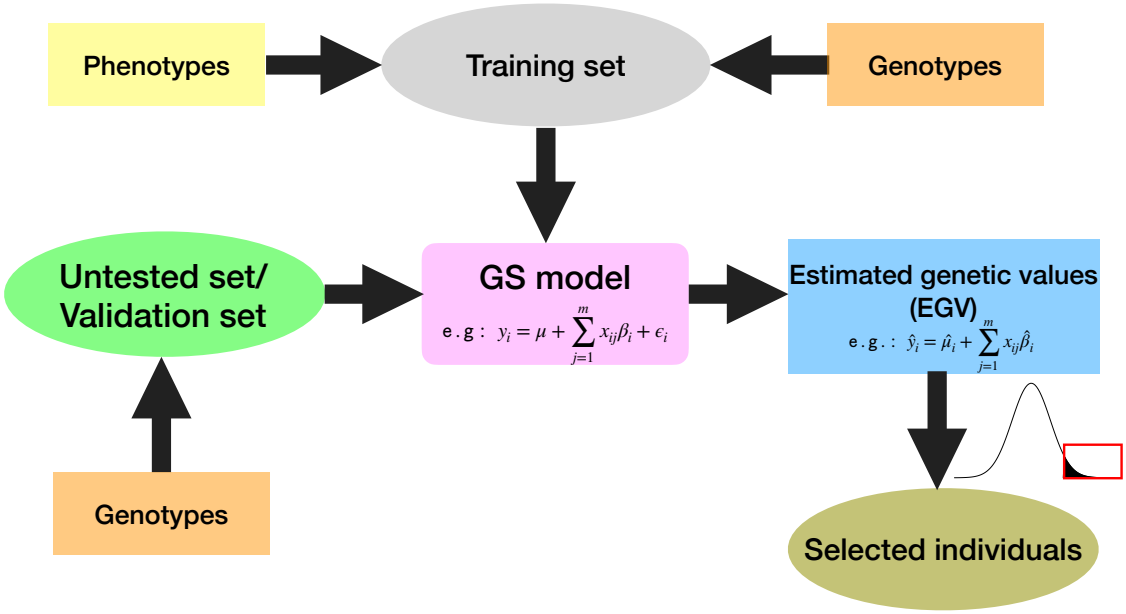

Figure 2. Genomic selection (GS) scheme.

Integrating genomic selection into breeding programs

Different strategies incorporating GS in a typical breeding program could affect the efficiency of the genetic gain. However, directly testing each proposed strategy using GS in practical plant breeding programs is a time-consuming and resource-intensive process to make the final decision. A computer simulation study is one of the ways to examine the feasibility of implementing GS in a breeding program. It can consider and vary several parameters: genetic architectures of traits, selection strategies integrating GS, etc, to maximize the breeding benefits to help draw conclusions before breeders conduct the experimental trials.

Several studies have investigated the potential of GS in many non-clonal crops via computer simulations, including barley, wheat, maize, rice, and sorghum (Iwata and Jannink, 2011; Marulanda et al., 2016; Gaynor et al., 2017; Muleta et al., 2019; Tessema et al., 2020; Bernardo, 2021; Fritsche-Neto et al., 2023). While some of them focused on the factors that affect the performance of the GS model, some of them comprehensively assessed the prospects of integrating GS into the breeding programs. Their results showed that GS can bring more genetic gain than PS.

On the other hand, few studies have outlined the potential of GS to improve the genetic gain in clonal breeding programs (Slater et al., 2016; Stich and Inghelandt, 2018), but they were based on empirical data. Only Werner et al. (2023) comprehensively investigated via simulation study different strategies to implement GS in clone breeding programs exemplarily with genome parameters of strawberry.

Although the results of Werner et al. (2023) indicated that GS can bring more genetic gain compared to PS, they mainly focused on how to select the parents for the next crosses and only introduced GS in the first clonal stage. However, the *per-se* value of a clone is important to ensure the short-term selection response. A classical clonal breeding program however, consists of multiple stages that are named e.g. in potato as crossing (X), seedling (SL), single hills (SH), A clone (A), B clone (B), C clone (C), D clone (D) stages, until a new variety is released. This raises one question: at which stage should GS be implemented to reach the highest genetic gain? Furthermore, to avoid losing the individuals carrying the beneficial alleles, GS could be applied not only at one stage but at multiple stages whose phenotypes for the target traits are not still available. However, these aspects have not been studied so far. Therefore, one of the aspects I focused on in this thesis was to evaluate the effects of the implementation stage of GS on genetic gain in a clonal breeding program.

Optimal allocation of breeding resources

Optimal allocation of resources is of fundamental importance for the efficiency of breeding programs (Longin et al., 2006, 2007), and requires to be checked, if major changes are made to the breeding programs. This can lead to an optimization of the number of genotypes that are used in the evaluated stage and selected for the next breeding stage. In other words, selection intensity for each examined stage can be optimized to further reach the maximum genetic gain.

The budget is composed of the cost at each breeding stage, the number of assessed genotypes, the number of locations, and the number of replications. Under the fixed economic parameters (cost for genotyping and phenotyping, the number of locations, and the number of replications), numerous possible allocations, for example, via a grid search, require to be examined. Although the process of optimization requires, because of its high dimensionality, high computational resources, it can help to find an optimal compromise in the number of tested genotypes per stage to maximize the genetic gain.

Besides the considerations of the abovementioned parameters, different selection strategies, variance components of traits, and even changes in cost and budget lead to rearranging the optimal allocation of resources and simultaneously reaching different maximal genetic gain (Longin et al., 2015; Marulanda et al., 2016). The selection strategies include how to implement GS at different stages as well as PS only. Therefore, each combination of parameters is a unique scenario in the breeding program and requires to find its own optimal allocation of resources

maximizing the genetic gain.

These abovementioned studies have investigated the optimal allocation of resources either with or without incorporating GS in the breeding programs. However, they focused on the optimization of cereal breeding programs. To the best of my knowledge, no earlier study is available about the effect of the optimal allocation of resources on genetic gain in clone breeding programs either with or without incorporating GS, compared to the standard breeding program based on PS only. Thus, this is another aspect that I concentrated on in this study.

Balancing the genetic gain and diversity in the long-term breeding programs

GS enables to improve genetic gain via increasing selection intensity or shortening breeding cycles (Xu et al., 2020). However, it can accelerate the loss of genetic variability in a breeding program compared to PS (Jannink, 2010; Gaynor et al., 2017), leading to a reduction of the long-term genetic gain in the long-term breeding programs (Falconer and Mackay, 1996). This is because GS can precisely select individuals carrying favorable alleles, leading to the rapid fixation of alleles. Furthermore, these selected individuals have similar genetic backgrounds. Once they serve as candidate parents and directly intercross each other, it can be expected that inbreeding will occur, resulting in decreased genetic variation. Therefore, not only enhancing genetic gain but creating and preserving genetic diversity simultaneously must be considered in long-term breeding programs.

New genetic variability can come from (1) introducing new alleles using collections of genetic resources (Sanchez et al., 2023); and (2) generating new allelic combinations during meiotic recombinations, which occur after intercrossing two parents. Crossing two parents is the first important step in the breeding program. Thus, choosing appropriate new crosses for the next breeding cycle is an important issue, and this is one aspect that I focused on in this thesis.

Some researchers have proposed approaches to balance genetic gain and diversity while determining desirable new crosses. The usefulness criterion (UC) is one of the criteria used to predict the performance of a cross (Schnell and Utz, 1975), which considers the expected progeny mean (μ) and the expected response to selection in the progenies ($ih\sigma_G$): $\mu + ih\sigma_G$.

For inbred populations generated from inbred parents or for hybrids and outbreds in the absence of dominance effects, the progeny mean can be estimated as the mid-parental performance based on the EGV from the trained GS model. On the other side, for heterozygous parents in diploids, if dominance effects exist, so

the progeny mean can be estimated by a formula considering the dominance effects (Werner et al., 2023; Wolfe et al., 2021). Furthermore, the progeny variance can be estimated by the derived formulae either in inbred or outbred parents incorporating the estimated marker effects from the trained GS model (Bonk et al., 2016; Lehermeier et al., 2017; Osthushenrich et al., 2017; Wolfe et al., 2021). However, for tetraploid species with heterozygous parents, all these formulae cannot be applied, and progeny mean cannot be accurately predicted based on the mid-parental performance especially when dominance effects exist. Fortunately, some softwares are available, e.g. AlphaSimR (Gaynor et al., 2021), and enable to simulate virtual progenies of a cross, which can be used to estimate the progeny mean and variance by inferring the average and variance of the EGV. The use of simulated progenies to estimate progeny mean and variance could improve the prediction accuracy of progeny mean compared to mid-parental values and provide an alternative to predict progeny variance for tetraploid species with heterozygous parents, which need to be assessed.

In addition to the UC, optimal cross selection (OCS) can select a group of bi-parental crosses maximizing the expected progeny mean, while keeping a predefined level of genetic diversity (Kinghorn et al., 2009; Gorjanc et al., 2018). The quantification of genetic diversity is considered based on the selected individuals who serve as parents rather than each cross itself and is commonly measured with co-ancestry or expected heterozygosity (He) (Gorjanc et al., 2018; Allier et al., 2019). While the OCS has been confirmed to improve genetic gain and preserve a certain diversity (Gorjanc et al., 2018; Allier et al., 2019), it requires much more computational time to find the optimal crosses compared to the methods based on ranking the performance among all possible crosses. Especially for tetraploid species and for simulation with a high number of markers, the computational burden is extreme. To solve this issue, quantifying genome-wide diversity for each cross by He is a quick method that could be directly added to the UC of a cross, and may bring benefits to improve long-term genetic gain and preserve diversity simultaneously. To my knowledge, no earlier studies assessed the performance of a criterion including genome-wide diversity of a cross to determine new desirable crosses, and therefore, this was another point that I focused on in my research.

Another possibility to improve genetic gain: Further increase the prediction accuracy/ability of genomic selection

In addition to selection intensity, the degree of prediction accuracy of the GS model is an important parameter influencing the efficiency and success of GS in genetic gain. The prediction accuracy is defined as the correlation between true genetic values and EGV. However, in contrast to situation in simulation studies, the true genetic values for traits are unknown for breeding materials in a practical breeding program and, thus, the performance of the GS model in experimental datasets is measured as correlations between observed phenotypes and EGV, also called as prediction ability. Then, the prediction accuracy can be obtained by dividing the prediction ability by the square root of heritability (Legarra et al., 2008). A higher prediction ability/prediction accuracy indicates that the GS model can more precisely predict the true genetic values of individuals and capture more phenotypic variation for traits. Therefore, the improvement of prediction performance for the GS model to further increase genetic gain is an important issue, because prediction accuracy can replace the square root of heritability in the breeder's equation (Heffner et al., 2010).

Former studies have comprehensively investigated that prediction accuracy is influenced by several factors such as the relationship between training set and untested set, the size of the training set, the number of markers, statistical models, used breeding materials, heritability of the trait, etc (Xu et al., 2020; Krishnappa et al., 2021). The GS models used in these studies were based on SNP information generated from SNP array or genotyping by sequence (Crossa et al., 2017) as predictors. Besides SNP, large alternations ($>$ 49 bp) including deletions, insertions, duplications, inversions, and translocation occur in genomic sequence, and are commonly defined as structural variants (Mahmoud et al., 2019). These structural variants can affect the phenotypic variation in plants (Gabur et al., 2018). In addition, complex biological processes such as transcription, translation, and biochemical cascades resulting in various metabolites occur between DNA sequence and phenotypes (Guo et al., 2016). Therefore, the genetic variance of traits is expected to be captured not only solely by SNP information but also by other predictors that are closer to the phenotypes than SNP.

With the development of molecular technologies and bioinformatic tools, cheap and high-throughput gene expression and metabolite profiling can be used, and accurate structure variants can be detected. Especially, the characterization of a transcriptome based on mRNA sequencing has the advantage of extracting

different types of information such as SNP and small INDEL (= small insertions/deletions ranging from 2 to 49 bp in length, called sequence variants). Furthermore, it can quantify different transcript expression (TE). In contrast, the use of microarray can only measure gene expression (GE) and detect a significant proportion of genes that are not expressed in a subset of individuals within a species, which is called expression presence/absence variation (ePAV) and is known as dispensable transcriptome (Hirsch et al., 2014; Jin et al., 2016; Weisweiler et al., 2019).

Moreover, SNP can be categorized into two groups by using the SIFT algorithm (Ng and Henikoff, 2001): (1) tolerant SNP (tSNP), including SNP in non-gene coding regions as well as gene coding regions producing no change or a change in the amino acid but still keeping a protein's function unaltered; and (2) deleterious SNP (dSNP) – including SNP in gene coding region involving a change in the amino acid and affecting a protein's function. Thus, the performance of the GS model could be improved by using different layers of datasets closer to phenotypes as predictors than the SNP information, which could shorten the gap between genotypes and phenotypes and even capture higher-order epistatic interactions to precisely predict phenotypic variation of traits.

Some studies on the use of GE, ePAV, and metabolites to predict phenotypic traits in cereals reported lower or higher prediction abilities compared to SNP information, depending on the traits and species (Riedelsheimer et al., 2012; Guo et al., 2016; Schrag et al., 2018; Hu et al., 2019; Weisweiler et al., 2019; Gemmer et al., 2020; Longin et al., 2020). However, the use of dSNP, tSNP, TE, and structural variants as a predictor has not been examined before. Furthermore, the integration of at least two different layers of omic datasets (e.g. SNP + metabolites, or SNP + expression + metabolites, or sequence variants + ePAV + expression, etc) could enhance the prediction ability in comparison to the use of SNP information only (Guo et al., 2016; Schrag et al., 2018; Weisweiler et al., 2019; Longin et al., 2020). Thus, the integration of multiple layers of omic datasets such as sequence variants, ePAV, expression, metabolites, or even structural variants as predictors could outperform solely SNP information to predict the phenotypic variation of traits with GS models, and this should be evaluated.

Therefore, in this thesis, I evaluated the prediction of phenotypic variation using different omic datasets and structural variants based on 23 diverse inbred barley. This was considered as pilot research and could open the path towards performing such analyses on the large-scale population, e.g. on segregating related barley populations derived from the 23 inbreds studied here (Casale et al., 2022); and even apply such models for potato breeding programs in the future, helping the improvement of genetic gain.

Objectives of this thesis

The objectives of my thesis were to optimize breeding programs incorporating GS, especially in potato, as well as to improve the efficiency of GS using different omic datasets and structural variants as predictors compared to SNP array taking barley as an example. Both have the final goal to further enhance the genetic gain in breeding programs. In particular, the objectives were to:

1. investigate under a fixed budget and in comparison to PS how the weight of GS relative to PS, the stage of implementation of GS, the correlation between traits (auxiliary trait assessed in early generations and target trait), the variance components, and the prediction accuracy affect the short-term genetic gain of the target trait in potato breeding programs;

2. determine the optimal allocation of resources maximizing the short-term genetic gain of the target trait for each selection strategy and for varying cost scenarios in potato breeding programs;

3. assess how different cross-selection methods implementing GS affect both short- and long-term genetic gains in potato breeding programs compared to strategies using phenotypic values only;

4. make recommendations to breeders on how to implement GS incorporating an appropriate cross-selection method in potato breeding programs to improve genetic gain while preserving genetic diversity;

5. assess the prediction ability for the yield-related phenotypic traits using different omic datasets and structural variants as single predictors compared to SNP array in barley;

6. explore the predictive performance when using sequence variants, gene expression, and ePAV from simulated 3'end mRNA sequencing of different lengths as predictors in barley;

7. investigate the improvement in prediction ability when combining multiple omic datasets / structural variants information to predict phenotypic variation in barley breeding programs.

References

Abberton, M., Batley, J., Bentley, A., Bryant, J., Cai, H., Cockram, J., de Oliveira, A. C., Cseke, L. J., Dempewolf, H., Pace, C. D., Edwards, D., Gepts, P., Greenland, A., Hall, A. E., Henry, R., Hori, K., Howe, G. T., Hughes, S., Humphreys, M., Lightfoot, D., Marshall, A., Mayes, S., Nguyen, H. T., Ogbonnaya, F. C., Ortiz, R., Paterson, A. H., Tuberosa, R., Valliyodan, B., Varshney, R. K., and Yano, M. (2016). Global agricultural intensification during climate change: a role for genomics. *Plant Biotechnology Journal*, 14:1095–1098.

Allier, A., Lehermeier, C., Charcosset, A., Moreau, L., and Teyssèdre, S. (2019). Improving short-and long-term genetic gain by accounting for within-family variance in optimal cross-selection. *Frontiers in Genetics*, 10:1006.

Bernardo, R. (2021). Upgrading a maize breeding program via two-cycle genomewide selection: Same cost, same or less time, and larger gains. *Crop Science*, 61:2444–2455.

Bonk, S., Reichelt, M., Teuscher, F., Segelke, D., and Reinsch, N. (2016). Mendelian sampling covariability of marker effects and genetic values. *Genetics Selection Evolution*, 48:1–11.

Casale, F., Inghelandt, D. V., Weisweiler, M., Li, J., and Stich, B. (2022). Genomic prediction of the recombination rate variation in barley – a route to highly recombinogenic genotypes. *Plant Biotechnology Journal*, 20:676–690.

Cobb, J. N., Juma, R. U., Biswas, P. S., Arbelaez, J. D., Rutkoski, J., Atlin, G., Hagen, T., Quinn, M., and Ng, E. H. (2019). Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the breeder's equation. *Theoretical and Applied Genetics*, 132:627–645.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., and Varshney, R. K. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends in Plant Science*, 22:961–975.

Desta, Z. A. and Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends in Plant Science*, 19:592–601.

Falconer, D. S. and Mackay, T. F. C. (1996). *Introduction to quantitative genetics*. Longman group, 4 edition.

FAO (2022). FAOSTAT.

Fritsche-Neto, R., Ali, J., Asis, E. J. D., Allahgholipour, M., and Labroo, M. R. (2023). Improving hybrid rice breeding programs via stochastic simulations: number of parents, number of hybrids, tester update, and genomic prediction of hybrid performance. *Theoretical and Applied Genetics*, 137:1–12.

Fróna, D., Szenderák, J., and Harangi-Rákos, M. (2019). The challenge of feeding the world. *Sustainability (Switzerland)*, 11:5816.

Gabur, I., Chawla, H. S., Snowdon, R. J., and Parkin, I. A. (2018). Connecting genome structural variation with complex traits in crop plants. *Theoretical and Applied Genetics*, 132:733–750.

Gaynor, R. C., Gorjanc, G., Bentley, A. R., Ober, E. S., Howell, P., Jackson, R., Mackay, I. J., and Hickey, J. M. (2017). A two-part strategy for using genomic selection to develop inbred lines. *Crop Science*, 57:2372–2386.

Gaynor, R. C., Gorjanc, G., and Hickey, J. M. (2021). Alphasimr: an r package for breeding program simulations. *G3: Genes—Genomes—Genetics*, 11.

Gemmer, M. R., Richter, C., Jiang, Y., Schmutzer, T., Raorane, M. L., Junker, B., Pillen, K., and Maurer, A. (2020). Can metabolic prediction be an alternative to genomic prediction in barley? *PLOS ONE*, 15:e0234052.

Gopal, J. (2006). *Considerations for successful breeding*, pages 77–108. CRC Press, 1st edition.

Gorjanc, G., Gaynor, R. C., and Hickey, J. M. (2018). Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. *Theoretical and Applied Genetics*, 131:1953–1966.

Grüneberg, W., Mwanga, R., Andrade, M., and Espinoza, J. (2009). *Selection methods. Part 5: breeding clonally propagated crops.*, pages 275–322. Food and Agriculture Organization of the United Nations (FAO).

Guo, Z., Magwire, M. M., Basten, C. J., Xu, Z., and Wang, D. (2016). Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. *Theoretical and Applied Genetics*, 129:2413–2427.

Heffner, E. L., Lorenz, A. J., Jannink, J. L., and Sorrells, M. E. (2010). Plant breeding with genomic selection: gain per unit time and cost. *Crop Science*, 50:1681–1690.

Hirsch, C. N., Foerster, J. M., Johnson, J. M., Sekhon, R. S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M. A., Barry, K., de Leon, N., Kaeppler, S. M., and Buell, C. R. (2014). Insights into the maize pan-genome and pan-transcriptome. *The Plant Cell*, 26:121–135.

Hu, X., Xie, W., Wu, C., and Xu, S. (2019). A directed learning strategy integrating multiple omic data improves genomic prediction. *Plant Biotechnology Journal*, 17:2011–2020.

Iwata, H. and Jannink, J. L. (2011). Accuracy of genomic selection prediction in barley breeding programs: a simulation study based on the real single nucleotide polymorphism data of barley breeding lines. *Crop Science*, 51:1915–1927.

Jannink, J. L. (2010). Dynamics of long-term genomic selection. *Genetics Selection Evolution*, 42:1–11.

Jansky, S. H., Charkowski, A. O., Douches, D. S., Gusmini, G., Richael, C., Bethke, P. C., Spooner, D. M., Novy, R. G., Jong, H. D., Jong, W. S. D., Bamberg, J. B., Thompson, A. L., Bizimungu, B., Holm, D. G., Brown, C. R., Haynes, K. G., Sathuvalli, V. R., Veilleux, R. E., Miller, J. C., Bradeen, J. M., and Jiang, J. (2016). Reinventing potato as a diploid inbred line–based crop. *Crop Science*, 56:1412–1422.

Jin, M., Liu, H., He, C., Fu, J., Xiao, Y., Wang, Y., Xie, W., Wang, G., and Yan, J. (2016). Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. *Scientific Reports*, 6:1–12.

Kinghorn, B. P., Banks, R., Gondro, C., Kremer, V. D., Meszaros, S. A., Newman, S., Shepherd, R. K., Vagg, R. D., and van der Werf, J. H. J. (2009). Strategies to exploit genetic variation while maintaining diversity. *Adaptation and Fitness in Animal Populations*, pages 191–200.

Krishnappa, G., Savadi, S., Tyagi, B. S., Singh, S. K., Mamrutha, H. M., Kumar, S., Mishra, C. N., Khan, H., Gangadhara, K., Uday, G., Singh, G., and Singh, G. P. (2021). Integrated genomic selection for rapid improvement of crops. *Genomics*, 113:1070–1086.

Legarra, A., Robert-Granié, C., Manfredi, E., and Elsen, J. M. (2008). Performance of genomic selection in mice. *Genetics*, 180:611–618.

Lehermeier, C., Teyssèdre, S., and Schön, C. C. (2017). Genetic gain increases by applying the usefulness criterion with improved variance prediction in selection of crosses. *Genetics*, 207:1651–1661.

Lindhout, P., Meijer, D., Schotte, T., Hutten, R. C., Visser, R. G., and van Eck, H. J. (2011). Towards F1 hybrid seed potato breeding. *Potato Research*, 54:301–312.

Longin, C. F. H., Mi, X., and Würschum, T. (2015). Genomic selection in wheat: optimum allocation of test resources and comparison of breeding strategies for line and hybrid breeding. *Theoretical and Applied Genetics*, 128:1297–1306.

Longin, C. F. H., Utz, H. F., Melchinger, A. E., and Reif, J. C. (2007). Hybrid maize breeding with doubled haploids: II. Optimum type and number of testers in two-stage selection for general combining ability. *Theoretical and Applied Genetics*, 114:393–402.

Longin, C. F. H., Utz, H. F., Reif, J. C., Schipprack, W., and Melchinger, A. E. (2006). Hybrid maize breeding with doubled haploids: I. One-stage versus two-stage selection for testcross performance. *Theoretical and Applied Genetics*, 112:903–912.

Longin, F., Beck, H., Gütler, H., Heilig, W., Kleinert, M., Rapp, M., Philipp, N., Erban, A., Brilhaus, D., Mettler-Altmann, T., and Stich, B. (2020). Aroma and quality of breads baked from old and modern wheat varieties and their prediction from genomic and flour-based metabolite profiles. *Food Research International*, 129.

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., and Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. *Genome Biology*, 20:1–14.

Marulanda, J. J., Mi, X., Melchinger, A. E., Xu, J. L., Würschum, T., and Longin, C. F. H. (2016). Optimum breeding strategies using genomic selection for hybrid breeding in wheat, maize, rye, barley, rice and triticale. *Theoretical and Applied Genetics*, 129:1901–1913.

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

Muleta, K. T., Pressoir, G., and Morris, G. P. (2019). Optimizing genomic selection for a sorghum breeding program in haiti: a simulation study. *G3: Genes—Genomes—Genetics*, 9:391–401.

Ng, P. C. and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Research*, 11:863–874.

Ortiz, R., Reslow, F., Cuevas, J., and Crossa, J. (2022). Genetic gains in potato breeding as measured by field testing of cultivars released during the last 200 years in the nordic region of europe. *The Journal of Agricultural Science*, pages 1–7.

Osthushenrich, T., Frisch, M., and Herzog, E. (2017). Genomic selection of crossing partners on basis of the expected mean and variance of their derived lines. *PLOS ONE*, 12:e0188839.

Ray, D. K., Mueller, N. D., West, P. C., and Foley, J. A. (2013). Yield trends are insufficient to double global crop production by 2050. *PLOS ONE*, 8:e66428.

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., Altmann, T., Stitt, M., Willmitzer, L., and Melchinger, A. E. (2012). Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nature Genetics*, 44:217–220.

Sanchez, D., Sadoun, S. B., Mary-Huard, T., Allier, A., Moreau, L., and Charcosset, A. (2023). Improving the use of plant genetic resources to sustain breeding programs' efficiency. *Proceedings of the National Academy of Sciences of the United States of America*, 120:e2205780119.

Schnell, F. and Utz, H. (1975). F1 leistung und elternwahl in der zuchtung von selbstbefruchtern. pages 243–248.

Schrag, T. A., Westhues, M., Schipprack, W., Seifert, F., Thiemann, A., Scholten, S., and Melchinger, A. E. (2018). Beyond genomic prediction: combining different types of omics data can improve prediction of hybrid performance in maize. *Genetics*, 208:1373–1385.

Slater, A. T., Cogan, N. O., Forster, J. W., Hayes, B. J., and Daetwyler, H. D. (2016). Improving genetic gain with genomic selection in autotetraploid potato. *The Plant Genome*, 9:0.

Stich, B. and Inghelandt, D. V. (2018). Prospects and potential uses of genomic prediction of key performance traits in tetraploid potato. *Frontiers in Plant Science*, 9:159.

Stokstad, E. (2019). The new potato. *Science*, 363:574–577.

Tessema, B. B., Liu, H., Sørensen, A. C., Andersen, J. R., and Jensen, J. (2020). Strategies using genomic selection to increase genetic gain in breeding programs for wheat. *Frontiers in Genetics*, 11:1538.

Weisweiler, M., de Montaigu, A., Ries, D., Pfeifer, M., and Stich, B. (2019). Transcriptomic and presence/absence variation in the barley genome assessed from multi-tissue mrna sequencing and their power to predict phenotypic traits. *BMC Genomics*, 20:787.

Werner, C. R., Gaynor, R. C., Sargent, D. J., Lillo, A., Gorjanc, G., and Hickey, J. M. (2023). Genomic selection strategies for clonally propagated crops. *Theoretical and Applied Genetics*, 136:1–17.

Wolfe, M. D., Chan, A. W., Kulakow, P., Rabbi, I., and Jannink, J. L. (2021). Genomic mating in outbred species: predicting cross usefulness with additive and total genetic covariance matrices. *Genetics*, 219.

Xu, Y., Liu, X., Fu, J., Wang, H., Wang, J., Huang, C., Prasanna, B. M., Olsen, M. S., Wang, G., and Zhang, A. (2020). Enhancing genetic gain through genomic selection: from livestock to plants. *Plant Communications*, 1:100005.

3 Optimal implementation of genomic selection in clone breeding programs – exemplified in potato: I. Effect of selection strategy, implementation stage, and selection intensity on short-term genetic gain

This manuscript was published in The plant genome in May, 2023.

Authors:
Po-Ya Wu, Benjamin Stich, Juliane Renner, Katja Muders, Vanessa Prigge, and Delphine van Inghelandt.

Own contribution: First author. I performed the data analyses and wrote the manuscript.

Received: 13 December 2022 | Accepted: 3 March 2023

DOI: 10.1002/tpg2.20327

The Plant Genome

ORIGINAL ARTICLE

Optimal implementation of genomic selection in clone breeding programs—Exemplified in potato: I. Effect of selection strategy, implementation stage, and selection intensity on short-term genetic gain

Po-Ya Wu[1] | **Benjamin Stich**[1,2,3] | **Juliane Renner**[4] | **Katja Muders**[5] | **Vanessa Prigge**[6] | **Delphine van Inghelandt**[1]

[1]Institute of Quantitative Genetics and Genomics of Plants, Heinrich Heine University, Düsseldorf, Germany

[2]Cluster of Excellence on Plant Sciences (CEPLAS), Heinrich Heine University, Düsseldorf, Germany

[3]Max Planck Institute for Plant Breeding Research, Köln, Germany

[4]Böhm-Nordkartoffel Agrarproduktion GmbH & Co. OHG, Hohenmocker, Germany

[5]NORIKA GmbH, Sanitz, Germany

[6]SaKa Pflanzenzucht GmbH & Co. KG, Windeby, Germany

Correspondence
Delphine van Inghelandt, Institute of Quantitative Genetics and Genomics of Plants, Heinrich Heine University, 40225 Düsseldorf, Germany.
Email: inghelan@hhu.de

Funding information
Fachagentur Nachwachsende Rohstoffe, Grant/Award Number: 22011818

[Correction added on 25 May 2023, after first online publication: Italic "r" changed to roman "r" throughout text.]

Abstract

Genomic selection (GS) is used in many animal and plant breeding programs to enhance genetic gain for complex traits. However, its optimal integration in clone breeding programs, for example potato, that up to now relied on phenotypic selection (PS) requires further research. In this study, we performed computer simulations based on an empirical genomic dataset of tetraploid potato to (i) investigate under a fixed budget how the weight of GS relative to PS, the stage of implementing GS, the correlation between an auxiliary trait and the target trait, the variance components, and the prediction accuracy affect the genetic gain of the target trait, (ii) determine the optimal allocation of resources maximizing the genetic gain of the target trait, and (iii) make recommendations to breeders how to implement GS in clone and especially potato breeding programs. In our simulation results, any selection strategy involving GS had a higher short-term genetic gain for the target trait than Standard-PS. In addition, we showed that implementing GS in consecutive selection stages can largely

Abbreviations: α_k, weights of genomic selection relative to phenotypic selection; A, A clone stage; B, B clone stage; C, C clone stage; D, D clone stage; EGV, estimated genetic values; ΔG, genetic gain; GS, genomic selection; i, intensity of selection; L, location; LSD, least significant difference; N, number of tested clones; N_{GS}, number of genotyped clones; Optimal-GS, GS strategies with optimum allocation of resources; Optimal-PS, standard potato breeding program relying exclusively on PS with optimum allocation of resources; p, selected proportion; PA, prediction accuracy; PS, phenotypic selection; $P_{T_{a\,(t)}}$, phenotypic value of the auxiliary (target) trait; QTL, quantitative trait loci; r, genetic correlations between the two traits; SH, single hills stage; SL, seedling stage; SNP, single nucleotide polymorphism; Standard-PS, standard potato breeding program relying exclusively on PS; Standard-GS, standard potato breeding program incorporating with GS; T_a, auxiliary traits; T_t, target trait; TGV, true genetic value; VC, variance components.

enhance short-term genetic gain and recommend the breeders to implement GS at single hills and A clone stages. Furthermore, we observed for selection strategies involving GS that the optimal allocation of resources maximizing the genetic gain of the target trait differed considerably from those typically used in potato breeding programs and, thus, require the adjustment of the selection and phenotyping intensities. The trends are described in our study. Therefore, our study provides new insight for breeders regarding how to optimally implement GS in a commercial potato breeding program to improve the short-term genetic gain for their target trait.

1 | INTRODUCTION

Potato (*Solanum tuberosum* L.) is with respect to the production volume one of the most important food crops in the world after sugarcane, maize, wheat, and rice (http://www.fao.org/faostat/en/). However, in contrast to other crops, only a low genetic gain was observed for yield in the past decades (Stokstad, 2019; Ortiz et al., 2022). The selection gain is, compared to the one in homozygous diploid species, limited by the high heterozygosity and tetraploidy of potato (Lindhout et al., 2011; Jansky et al., 2016). In addition, potato has a low multiplication coefficient (Grüneberg et al., 2009), which leads to the availability of only one or few tubers per genotype for phenotypic evaluation at early stages in the breeding program (Gopal, 2006). This delays the evaluation of traits related to productivity (such as tuber yield) or quality, as they rely on multi-location field trials and/or destructive assessment, and these can only be performed after one to several multiplication steps. As a consequence, only traits which can be assessed based on a low number of plants can be considered in the early stages of potato breeding programs. In contrast, target traits whose evaluation requires many plants and/or environments can only be selected for in later stages of the breeding program. Instead, early indirect selection on the auxiliary trait can be performed. However, the correlation between the latter and the target trait shows a high range of variability depending on the considered traits, and can even be negative. This can limit the benefit of the early indirect selection on the auxiliary trait. Furthermore, the evaluation of target traits in potato is more expensive compared to their evaluation in non-clonal crops as a considerably lower level of mechanization is currently possible. Therefore, clone and especially potato breeding programs would highly benefit from the possibility to select for target traits at early stages of the breeding program, for example, with the implementation of genomic selection (GS).

GS proved to enhance genetic gain for complex traits in both animal and plant breeding programs (Meuwissen et al., 2001; Desta & Ortiz, 2014). This is because GS allows to predict the performance of target traits without phenotypic evaluation in early stages. The selection on target traits at early stages using estimated genetic values (EGV) avoids discarding those individuals with desirable alleles for the trait, which will increase the genetic gain per year. In addition, the performance prediction of target traits without phenotypic evaluation in early stages has the potential to reduce the length of the breeding cycle. One parameter that influences the potential of GS is the prediction accuracy.

Several empirical studies have explored the potential of implementing GS in potato breeding for different traits by determining the prediction accuracy (Slater et al., 2016; Sverrisdóttir et al., 2017; Enciso-Rodriguez et al., 2018; Endelman et al., 2018; Stich & Van Inghelandt, 2018; Sverrisdóttir et al., 2018; Caruana et al., 2019; Byrne et al., 2020; Gemenet et al., 2020; Sood et al., 2020; Wilson et al., 2021). Different degrees of prediction accuracies from low to high depending on the studied traits have been reported, which could be caused by the different genetic architectures, prediction models, but also the considered genetic material. However, only few studies evaluated the effect of GS on the genetic gain for the studied traits. One of them was Slater et al. (2016), who estimated that the genetic gain after implementing GS for complex traits was higher than that of phenotypic selection (PS). The results of Stich and Van Inghelandt (2018) suggested that for some traits GS leads to a higher gain of selection than PS even without reducing the cycle length. However, no earlier study considered directly the aspect that PS and GS need to be compared at a fixed budget. Furthermore, when implementing GS in a clone breeding program, the selected proportion of PS on the early trait will be partially shifted to GS on the target trait. This shift can be realized to different degrees and the resulting selected proportion for PS or GS might influence the efficiency of the selection strategy. Therefore, for the implementation of GS in clone breeding programs not only the prediction accuracy of the GS model but also its relative weight to PS has to be examined. Furthermore, these aspects are influenced by the correlation between the early and the target trait and also the variance components of the considered trait have an influence on the genetic gain. However, the influences of these parameters and their interaction on the genetic gain in clone breeding programs have not been investigated until now.

Werner et al. (2023) investigated different strategies to implement GS in clone breeding programs exemplarily with genome parameters of strawberry. They evaluated the performance of a breeding program that introduced GS in the first clonal stage and mainly focused on how to select parents for the next crosses and drive population improvement to enhance long-term genetic gain. However, in a classical clone breeding program, there are several stages where GS could be implemented and their effect on the gain of selection have not been studied so far.

Another aspect that needs to be decided during the implementation of GS in clone or potato breeding programs is the number of stages in which GS is applied. Once the clones are genotyped for the first GS application, the possibility of reusing the same EGV to perform GS in two or more stages is given. A similar idea was proposed by Spindel et al. (2015) for a rice breeding program but has neither been assessed by theoretical considerations nor by computer simulations nor any empirical experiments. To the best of our knowledge, no earlier study has investigated at which stage and in how many selection stages GS should be implemented in clonal crops to maximize the short-term genetic gain under a given budget.

Optimum allocation of resources under a given budget is essential to improve the efficiency of breeding programs (Longin et al., 2006). However, most studies on the implementation of GS in breeding programs neglected this effect. Longin et al. (2015) and Marulanda et al. (2016) assessed this point for cereal breeding programs. However, to the best of our knowledge, no earlier study is available about the effect of the implementation of GS on the optimum allocation of resources in clone breeding programs.

The objectives of this study were to (i) investigate under a fixed budget how the weight of GS relative to PS, the stage of implementation of GS, the correlation between traits (auxiliary trait assessed in early generations and target trait), the variance components, and the prediction accuracy affect the short-term genetic gain of the target trait in potato breeding programs compared to PS, (ii) determine the optimal allocation of resources maximizing the short-term genetic gain of the target trait in each selection strategy and for varying cost scenarios, and (iii) make recommendations to breeders how to implement GS in clone breeding programs.

Core Ideas

- Genomic selection strategies can improve the genetic gain of clone breeding programs versus phenotypic selection.
- Implementing genomic selection in consecutive selection stages can largely enhance short-term genetic gain.
- Optimal implementation of genomic selection requires changes in the allocations of resources.

2 | MATERIALS AND METHODS

2.1 | Empirical basis of the computer simulations

Our simulations were based on an empirical genomic dataset of tetraploid potato. This empirical genomic dataset comprised 19,649,193 sequence variants revealed in a diversity panel of 100 tetraploid potato clones (Baig et al. in preparation). The unphased sequence variants included single nucleotide polymorphism (SNP) and insertion/deletion (InDel) polymorphisms. Sequence variants with a minor allele frequency <0.05 and missing rate >0.1 were removed. The 100 clones were used as parents of the simulated progenies and will be called parental clones hereafter.

The progenies were simulated using AlphaSimR (Gaynor et al., 2021). For this, the genetic map information of all genomic variants was estimated using a Marey map (for details see Method S1 and Figure S1). Subsequently, the genomic information for each variant served as input for the simulations.

2.2 | Simulation of initial population

To stick to the size of commercial breeding programs (Breeders personal communication, Table 1) an initial population of 300,000 clones was simulated like described here under. From all possible crosses in the half-diallel among the 100 parental clones, 300 were randomly selected. For each of these 300 crosses, 1000 F1 progenies were simulated using AlphaSimR. The two steps of this procedure (the random selection of 300 crosses and the simulation of their progenies) were repeated 1000 times independently.

2.3 | Simulation of true genetic and phenotypic values

2.3.1 | Target trait (T_t)

In our study, a genetically complex target trait representing the weighted sum of all market relevant quantitative traits was considered and will be named T_t hereafter. A random set of 2000 sequence variants were considered as quantitative trait loci (QTL) for T_t. The true additive effects of the 2000

[Correction added on 25 May 2023, after first online publication: Sentence changed from "...was repeated 1000 times independently." to "...were repeated 1000 times independently."]

TABLE 1 Dimensioning of a standard potato breeding program that exclusively relies on phenotypic selection.

Stage	Number of clones	Number of locations	Phenotyping cost per clone and plot (€)	Cost per stage (€)
Seedling	300,000	1	1.4	420,000
Single hills	100,000	1	1.4	140,000
A clone	10,000	1	1.4	14,000
B clone	1500	2	25	75,000
C clone	300	3	25	22,500
D clone	60	4	25	6000
Sum				677,500

QTL were drawn from a gamma distribution (cf. Hayes & Goddard, 2001) with $k = 2$ and $\theta = 0.2$, where k and θ are shape and scale parameters, respectively. To control the degree of dominance δ between 0 and 1 for each QTL, the ratios of dominance to additive effect were produced from a beta distribution with the two shape parameters $a = 2$ and $b = 2$. The true dominance effect at each QTL was then calculated by multiplying the true additive effect by the QTL specific δ (Figure S2). For each QTL, all possible genotype classes were AAAA, AAAB, AABB, and BBBB, which were, respectively, coded from 0, 1, 2, 3, and 4 for additive effect; and 0, 1, 1, 1, and 0 for dominance effect. Finally, the true genetic value for T_t (TGV_{T_t}) was calculated for each clone by summing up the true additive and dominance effects at the 2000 QTL.

In order to simulate phenotypic values, two ratios of variance components (VC) were assumed for T_t: $\sigma^2_G : \sigma^2_{G\times L} : \sigma^2_\epsilon = 1 : 1 : 0.5$ (VC1) and $1 : 0.5 : 0.5$ (VC2), where σ^2_G denoted the genotypic variance, $\sigma^2_{G\times L}$ the variance of interaction between genotype and location, and σ^2_ϵ the error variance. The genotypic variance was estimated by the sample variance of TGV_{T_t} in the initial population. The phenotypic value for the target trait was then calculated as $P_{T_t} = TGV_{T_t} + \epsilon_{T_t}$, where ϵ_{T_t} was the non-genetic value following a normal distribution $N(0, \sigma^2_{\epsilon_{T_t}})$, with

$$\sigma^2_{\epsilon_{T_t}} = \frac{\sigma^2_{G\times L}}{L_j} + \frac{\sigma^2_\epsilon}{L_j R_j} \quad (1)$$

representing the non-genetic variance, in which L_j was the number of locations at stage j, and R_j the number of repetitions at stage j. We set the number of replications to one ($R_j = 1$) in each location (cf. Melchinger et al., 2005).

2.3.2 | Phenotypic trait assessed in early generations of the breeding program (T_a)

The weighted sum of the auxiliary traits measured in the first three generations of the breeding program will be referred to as T_a hereafter. To control the genetic correlations between T_a and T_t (r), the true genetic values for T_a were generated by $TGV_{T_a} = TGV_{T_t} + \epsilon_r$, where ϵ_r was the residual value following a normal distribution $N(0, \sigma^2_{\epsilon_r})$, with

$$\sigma^2_{\epsilon_r} = \frac{1}{n-2}\frac{1-r^2}{r^2}\sum_{i=1}^{n}(TGV_{T_t(i)} - \overline{TGV}_{T_t})^2 \quad (2)$$

determined by the degree of r, where n was the number of clones for the initial population, $TGV_{T_t(i)}$ the TGV for T_t of the i^{th} clone, and $\overline{TGV}_{T_t}$ the average of TGV_{T_t} in the initial population. Then, the phenotypic value for T_a was calculated as $P_{T_a} = TGV_{T_a} + \epsilon_{T_a}$, where ϵ_{T_a} was a non-genetic value following a normal distribution $N(0, \frac{1-H^2_{T_a}}{H^2_{T_a}}\sigma^2_{G_{T_a}})$, in which $H^2_{T_a}$ was the broad-sense heritability for T_a, and $\sigma^2_{G_{T_a}}$ the genetic variance of T_a and estimated by the sample variance of TGV_{T_a} in the initial population. In this study, $H^2_{T_a}$ was set as 0.6.

2.4 | Simulation of estimated genetic values

In this study, we assumed that a GS model was trained for T_t on earlier cycles of the breeding program, and that this model has the prediction accuracy PA. The estimated genetic values (EGV) of T_t obtained from the GS model were estimated by $EGV_{T_t} = TGV_{T_t} + \epsilon_{PA}$, where ϵ_{PA} was the residual value following a normal distribution $N(0, \sigma^2_{\epsilon_{PA}})$, with

$$\sigma^2_{\epsilon_{PA}} = \frac{1}{n-2}\frac{1-PA^2}{PA^2}\sum_{i=1}^{n'}(TGV_{T_t(i)} - \overline{TGV}_{T_t})^2 \quad (3)$$

determined by the level of PA, where n' was the number of genotyped clones ($= N_{GS}$), $TGV_{T_t(i)}$ the TGV of the target trait at the i^{th} genotyped clone, and $\overline{TGV}_{T_t}$ the average of TGV_{T_t} on all N_{GS} genotyped clones.

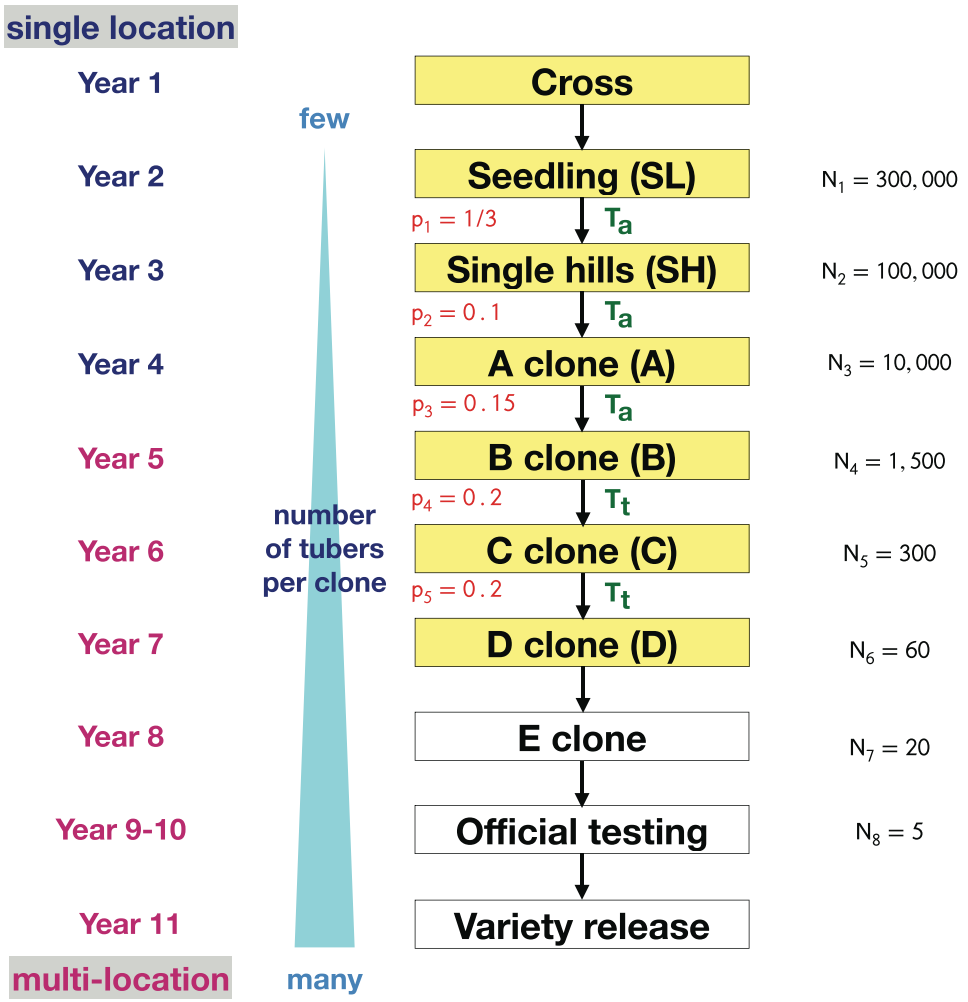

FIGURE 1 The standard clone breeding program examined in this study that relies exclusively on phenotypic selection. $_1$–p_5 are the selected proportions from SL to SH, SH to A, A to B, B to C, and C to D, respectively, where SL, SH, A, B, C, and D represent the stages of seedling, single hills, A, B, C, and D clones. T_a represented the integral of early measured traits and T_t the integral of the target traits. The yellow marked stages are those that were examined in our study.

2.5 | Selection strategies

2.5.1 | Standard breeding program

A standard potato breeding program relying exclusively on PS (Standard-PS) was considered as benchmark (Figure 1). To simplify the comparison between PS and GS strategies, we considered in this study six testing stages in the potato breeding program. The six testing stages were seedling, single hills, and A, B, C, and D clone stages, abbreviated in the following as SL, SH, A, B, C, and D, respectively. The number of tested clones (N) and locations (L) for each testing stage are shown in Table 1. The selected proportions from SL to SH (p_1), SH to A (p_2), A to B (p_3), B to C (p_4), and C to D (p_5) were set to $\frac{1}{3}$, 0.1, 0.15, 0.2, and 0.2, respectively, as estimates from typical commercial potato breeding programs (Breeders personal communication). The selection in the early stages (SL, SH, and A) was based on the phenotypic value of the auxiliary trait P_{T_a}, and for the late stages (B, C, and D) on the phenotypic value of the target trait P_{T_t} (Figure 1).

2.5.2 | Breeding programs involving genomic selection

Three GS strategies were evaluated in which GS was implemented at the (1) seedling, (2) single hills, and (3) A clone stage, abbreviated as GS-SL, GS-SH, and GS-A, respectively. All selection steps of the GS strategies were similar to those of the standard breeding program except the following modifications (Figure 2). Here, the strategy GS-SL will be taken

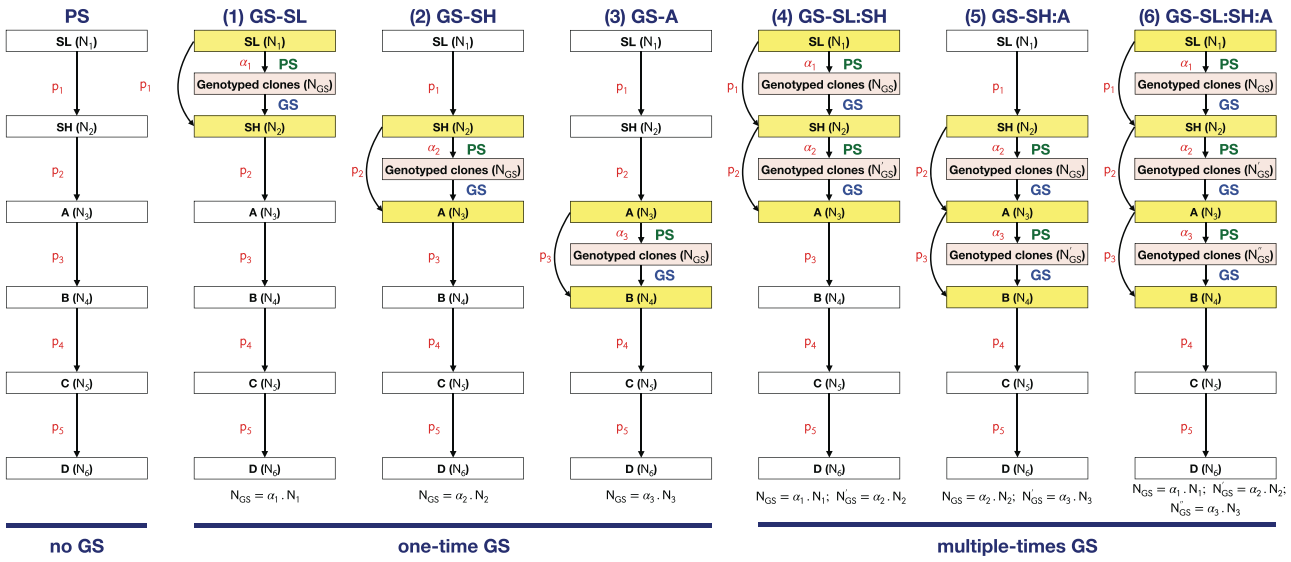

FIGURE 2 Graphical illustration of the standard as well as the six selection strategies that include genomic selection that were examined in our study. $_1$–p_5 are the selected proportions from SL to SH, SH to A, A to B, B to C, and C to D, respectively, where SL, SH, A, B, C, and D represent the stages of seedling, single hills, A, B, C, and D clones. α_k is the proportion of clones selected by PS to be genotyped in stage k and N_k is the number of clones of the respective stage.

as an example for the description. In the seedling stage, N_1 clones were evaluated for P_{T_a}. From these N_1 clones, the N_{GS} ones with a higher P_{T_a} were genotyped. α_1 was defined as ratio of N_{GS} to N_1, that is, the proportion of clones selected by PS to be genotyped. Then, N_2 clones were selected based on the EGV_{T_t} in the N_{GS} genotyped clones for the single hills stage. Afterward, the selection process in the following stages was the same as in Standard-PS. For the other two GS strategies, GS-SH and GS-A, the selection was performed accordingly. For each stage k in which GS was applied, the corresponding α_k was larger than p_k, where p_k $(= \frac{N_k}{N_{k+1}})$ was the selected proportion between the two stages to which GS was applied. k was set to 1, 2, and 3 for the strategies (1) GS-SL, (2) GS-SH, and (3) GS-A, respectively (Figure 2).

To evaluate whether adopting the same GS model for selection on T_t in several stages improves the short-term genetic gain compared to using GS only once, we evaluated three additional strategies (Figure 2):

(4) GS-SL:SH—GS was applied not only at seedling stage but also at single hills stage;

(5) GS-SH:A—GS was applied not only at single hills stage but also at A clone stage; and

(6) GS-SL:SH:A—GS was applied at seedling, single hills, and A clone stages.

For these three GS strategies, genotyping of N_{GS} clones only took place when GS was used for the first time. When GS was used a second or third time, the same EGV_{T_t} for the tested clones from the initial GS model were used for the selection.

2.6 | Economic settings and additional quantitative genetic parameters

In this study, the costs for phenotypic evaluation of T_a and T_t in one environment were assumed to be 1.4 and 25 €, respectively. The costs for genotypic evaluation per clone were assumed as 25 € (Table 1). To compare the short-term genetic gain of T_t (ΔG) between Standard-PS and several GS strategies, the budget across different selection strategies was fixed to 677,500 €. Therefore, the number of tested clones in seedling stage (N_1) must be adjusted/reduced when introducing GS into a breeding program to compensate for the additional genotyping cost. In the first part of the simulations, the selected proportions were fixed to those of Standard-PS. This was realized in our study by randomly sampling the reduced N_1 from the initial population with an equal sample size for each cross population.

We were interested in how different values of r, PA, VC, and L influence ΔG. Therefore, three different levels of r (-0.15, 0.15, and 0.3), PA (0.3, 0.5, and 0.7), and two different ratios of VC for T_t (see above) were examined in our simulations. The selection of clones based on T_t that was assessed in field experiments in more than one location happened at B and C clone stages. Thus, we varied the number of locations from 2 to 4 and 3 to 6 in increments of 1 for B and C clone stages, respectively, and designated them as L_4 and L_5. Furthermore, to investigate how different levels of α_k affect ΔG, we varied α_k from 0.4 to 0.9 in increments of 0.1 for the strategies GS-SL, GS-SL:SH, and GS-SL:SH:A, and from 0.2

to 0.9 in increments of 0.1 for the other strategies. ΔG was calculated as the difference in mean genetic values of T_t between the D clone and the seedling stage (cf. Longin et al., 2015; Marulanda et al., 2016).

2.7 | Optimum allocation of resources

In the below described simulations, we relaxed the restrictions of the above described simulations that the selected proportions were fixed to those of Standard-PS. To determine the optimum allocation of resources maximizing ΔG under a given budget, a general linear cost function to aggregate all costs across all stages in the breeding program was created:

$$\begin{aligned}\text{Budget} &= \sum_{j=1}^{6} N_j \times \text{cost}_{\text{pheno}(j)} \times L_j + N_{GS} \times \text{cost}_{\text{geno}} \\ &= \sum_{j=1}^{5} \frac{N_6}{\Pi_{k=j}^{5} p_k} \text{cost}_{\text{pheno}(j)} L_j + N_6 \text{cost}_{\text{pheno}(6)} L_6 \\ &+ \frac{N_6 \text{cost}_{\text{geno}} \alpha_m}{\Pi_{k=m}^{5} p_k},\end{aligned} \quad (4)$$

where N_j was the number of clones at stage j, $\text{cost}_{\text{pheno}(j)}$ the cost for phenotypic evaluation at stage j, N_{GS} the number of genotyped clones, and $\text{cost}_{\text{geno}}$ the genotyping cost (for details see Method S2). In addition, p_k was the selected proportion from stage $j(m)$ to stage $j(m)+1$, where m was the stage in which GS was applied first. For more details, $m = 1$ referred to GS-SL, GS-SL:SH, and GS-SL:SH:A; $m = 2$ for GS-SH and GS-SH:A; and $m = 3$ for GS-A. The GS strategies with optimum allocation of resources will be named Optimal-GS hereafter.

The optimum allocation was determined by a grid search across the permissible space of p_2 to p_5 and α_k for a set of given input parameters. The latter included the number of tested clones at D clone stage (N_6), the GS strategy, the phenotyping and genotyping costs, L, r, VC of T_t, $H^2_{T_a}$, and the total budget. We set N_6 to 60. In the grid search, any p_k varied between 0.1 and 0.5 in increments of 0.05 to avoid too strong/weak selections. α_k was chosen as described above. Consequently, in each permissible allocation, p_1 was completely determined by Equation (4) under the constrained budget and the given input parameters. Subsequently, the mean genetic gain across 1000 simulation runs was calculated for each permissible allocation of the grid search. To obtain reliable estimates of the optimal allocation of resources, we performed a least significant difference (LSD) test on ΔG across all permissible allocations of the grid search within a specific scenario. We selected the significant group showing the maximum ΔG among all permissible sets and then considered the average of the allocations as optimal result.

The above described simulations required for some grid search sets (those with low p_1 to p_3 but high p_4 and p_5) with more than 300,000 clones in the seedling stage. Thus, the size of the initial population was increased to 900,000 clones.

To investigate whether an increase of phenotyping cost of T_a and the genotyping cost have an influence on the optimal allocation of resources, we considered three different phenotyping costs for T_a (0.7, 1.05, and 1.4 €), and three different genotyping costs (15, 25, and 40 €).

3 | RESULTS

The mean genetic gain (ΔG) and genetic variance (σ^2_G) of the target trait at D clone were assessed considering different values of r, PA, α_k, as well as different selection strategies. To easily compare among the examined strategies, the budget, the selection proportion between stages p_1–p_5 and the number of test locations were fixed according to those of the Standard-PS strategy.

Increasing r and PA either individually or simultaneously led to a higher ΔG (Figure 3 and Figure S3). Regardless of PA and r, any selection strategy incorporating GS was superior to the Standard-PS strategy with respect to ΔG (Figure 3). Low or negative values for r and high PA increased this tendency even more. The least improvement of ΔG relative to Standard-PS was observed across all scenarios for the strategy GS-SL. The strategies GS-A and GS-SH resulted in considerably higher values for ΔG relative to PS and under the scenarios with low r but high PA, the latter strategy was significantly superior to the former.

Implementing GS in successive stages (GS-SL:SH, GS-SH:A, and GS-SL:SH:A) had an advantage over the strategies using GS one time, except for the scenario with the lowest PA (=0.3) but the highest r (=0.3). The ranking of performance among these strategies was GS-SL:SH:A > GS-SH:A > GS-SL:SH. The difference among these strategies was lower, if r increased or PA decreased.

For all GS strategies, higher α_k values led to reductions in the number of clones available in the seedling stage (Figure S4), but increased ΔG (Figure 3). For all except eight scenarios, the highest ΔG was observed if α_k was at its maximum (0.9). The remaining scenarios in which the maximum ΔG were observed for α_k=0.7 or 0.8 instead of 0.9, however, showed ΔG values that were not significantly different from the ΔG values observed for α_k=0.9 (data not shown). Only for GS-SL:SH:A an exception was observed from this trend, namely that the maximal ΔG was observed for α_k=0.5 for the scenario with r=0.3 and PA=0.3. In accordance with the above described observations regarding the differences among selection strategies, also the differences among ΔG for

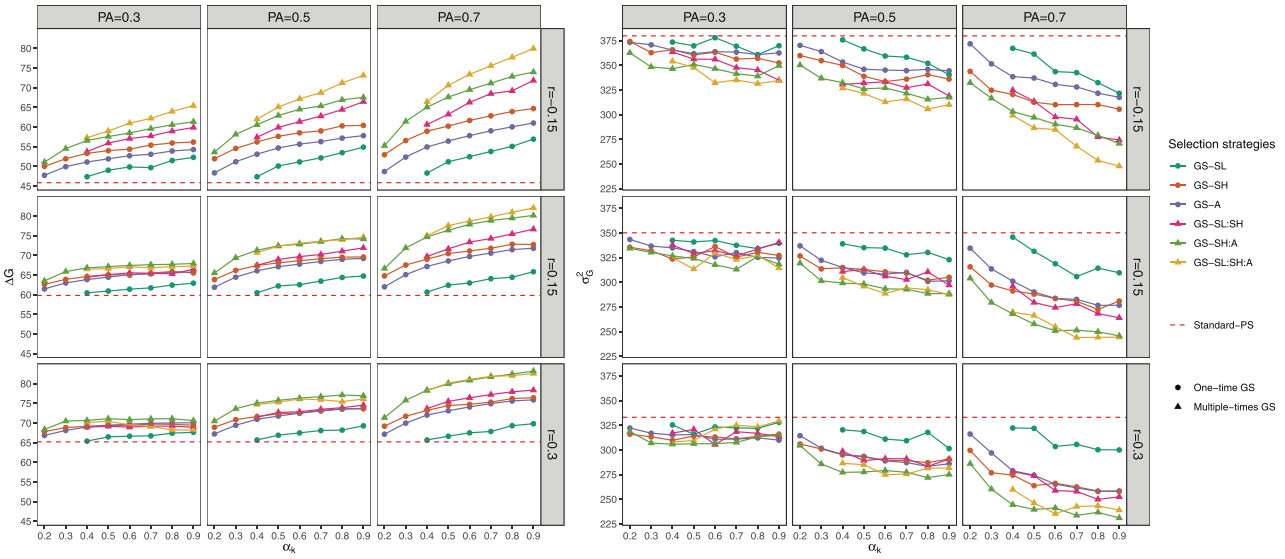

FIGURE 3 Genetic gain (ΔG, left) and genetic variance (σ^2_G, right) for the target trait on average across 1000 simulation runs at D clone stage for different weights of genomic selection (GS) relative to phenotypic selection (α_k), different selection strategies, different correlations between the traits (r = -0.15, 0.15, and 0.3), prediction accuracies (PA = 0.3, 0.5, and 0.7), and for the ratio of variance components VC1 ($\sigma^2_G : \sigma^2_{G\times L} : \sigma^2_\epsilon = 1 : 1 : 0.5$) The details regarding the selection strategies are shown in Figure 2.

the different levels of α_k were low for the scenarios with high r and/or low PA.

In all the above described simulations of the selection strategies that exploit GS in several stages, α_k was the same for each stage in which GS was applied. However, for these strategies, we also evaluated whether varying α_k had an influence on ΔG. For the strategies GS-SL:SH and GS-SH:A, a higher ΔG was observed with an increase of both α_k values (that is, α_1 and α_2 or α_2 and α_3) (Figure S5). The combination of two α_k values that resulted in the highest ΔG was 0.84 and 0.79 or 0.86 and 0.86 for the respective strategies. A similar trend was observed for GS-SL:SH:A (Figure S6). However, for the scenarios with high r (=0.3), intermediate values of α_1 were sufficient to result with high values of α_2 and α_3 in the maximal values of ΔG of 0.4–0.5 (Table S1).

The effect of variation of selection strategies, α_k, r, and PA on the genetic variance were opposite to their effect on genetic gain (Figure 3). The scenarios with a higher genetic gain showed a lower genetic variance.

We also investigated the effects of different ratios of variance components (VC1 and VC2) and number of locations for phenotypic evaluation (L_4 and L_5) on ΔG. The ranking of the selection strategies with respect to ΔG was not affected by the studied ratios of VC (Figure 3 and Figure S7). When $\sigma^2_{G\times L}$ was halved (i.e., VC2 vs. VC1), ΔG increased from 3% to 8% depending on the selection strategies, PA, r, and α_k (Figure S8). Although increasing L caused a decrease in the number of clones that are available at the seedling stage to compensate for additional phenotyping costs, ΔG significantly increased with increasing number of locations that were used for the evaluation of B and C clones (Figure 4). This trend was independent of selection strategies, PA, r, and α_k. In all scenarios, the highest ΔG was observed with the highest number of locations in the B and C clone stages, that is, $L_4 = 4$ and $L_5 = 6$. In these cases, ΔG was increased by 8% compared to Standard-PS with $(L_4, L_5) = (2, 3)$.

The optimal allocation of resources was assessed via a grid search across p_1–p_5 and $\alpha_k, k \in [1, 3]$ in a scenario with VC1, budget, L, and N_6 as in the Standard-PS scenario. The optimum allocation of resources led also for the PS to an increase of ΔG (Optimal-PS) compared with the Standard-PS (Figure 5). On average across all evaluated scenarios, the strategy GS-SL had the worst performance out of the strategies incorporating GS. In a scenario with r < 0 and PA > 0.5, any selection strategy with GS revealed a higher ΔG than the Optimal-PS. The strategy GS-SL:SH:A only outperformed the other selection strategies if r = -0.15. In contrast, the strategy GS-SH:A or GS-A resulted in the highest ΔG if r was >-0.15. On average across all the examined scenarios, the strategy GS-SH:A resulted in the highest and most stable ΔG values.

With the exception of one specific scenario, a high α_k was required for each selection strategy to reach the maximal ΔG value (Table 2, Tables S2 and S3). This exception was the strategy GS-SL in case of a positive r for which α_k ranging from 0.21 to 0.61 resulted in the maximal ΔG values. Furthermore, to achieve maximum ΔG values, the selected proportions for the last two stages (p_4 and p_5) were low (0.17) on average across all scenarios. The level of the optimal p_k

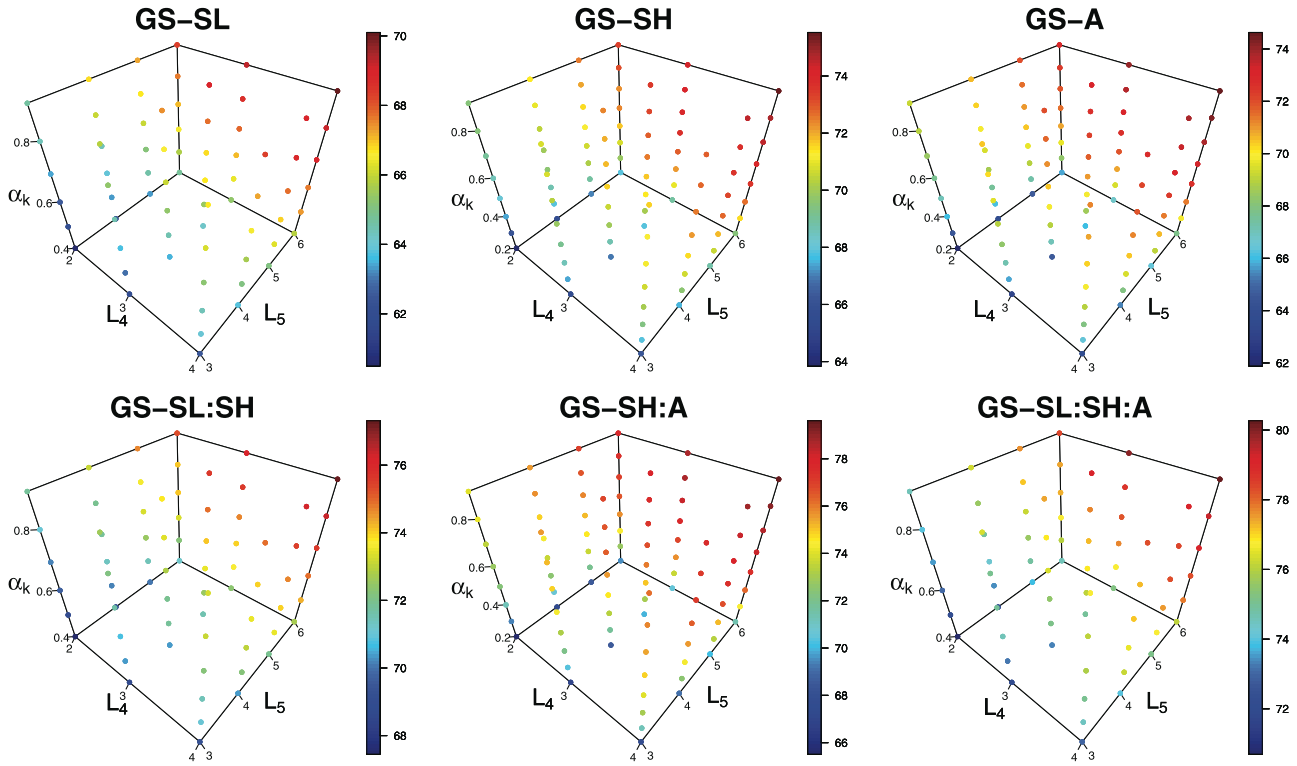

FIGURE 4 Genetic gain for the target trait (ΔG) on average across 1000 simulation runs at the D clone stage for six different selection strategies with genomic selection (GS) for varying numbers of locations in the B and C clone stages (L_4 and L_5) and different weights of genomic selection (GS) relative to phenotypic selection (α_k) when the correlation between the two traits was set to 0.15 and prediction accuracy was set to 0.5.

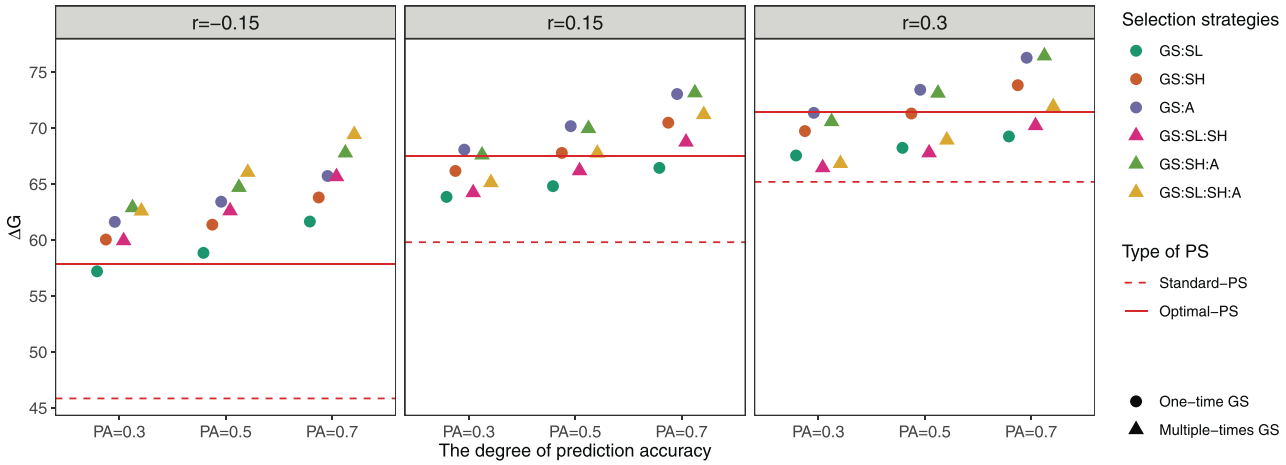

FIGURE 5 Genetic gain of the target trait (ΔG) after optimally allocated resources for different correlations between the traits (r = -0.15, 0.15, and 0.3) and different prediction accuracies (PA = 0.3, 0.5, and 0.7). The presented ΔG values are the average of the genetic gains from the grid search sets that revealed no significant ($p < 0.05$) difference compared to the set with maximum genetic gain.

was influenced by the level of r as well as by the stage in which GS was implemented. In general, high optimal p_1 values were observed with a negative correlation in comparison with the scenarios with a positive correlation. Furthermore, we observed for all strategies with implementation of GS that the selection proportion for that stage in which GS was applied was lower than the one observed at the same stage in the other strategies. This trend was more pronounced for scenarios with high PA. For instance, p_2 (p_3) for the strategy GS-SH (GS-A) was on average across all scenarios about 0.25 (0.21) lower than the one for the strategies excluding GS-SH (GS-A) with 0.42 (0.45).

The effects of different phenotyping and genotyping costs on the maximum ΔG were assessed exemplarily for

TABLE 2 Optimum allocation of resources to maximize genetic gain of the target trait (ΔG) for the different selection strategies and correlations between the two traits (r = -0.15, 0.15, and 0.3). The prediction accuracy was 0.5 and the phenotyping cost of early measured trait 1.4 € and genotyping cost 25 €. p_1 to p_5, α_k, and N_1 are the selected proportion per stage, the weight of genomic selection relative to phenotypic selection, and the number of clones at the seedling stage, respectively. For description of selection strategies see text.

Correlations	Selection strategies	ΔG [1]	$SD_{\Delta G}$ [2]	p_1	p_2	p_3	p_4	p_5	α_k	N_1
−0.15	PS	57.87 (g)	5.04	0.39	0.36	0.31	0.10	0.10	-	152,995.09
	GS-SL	58.86 (f)	5.18	0.30	0.50	0.50	0.16	0.23	0.87	23,709.50
	GS-SH	61.38 (e)	5.56	0.44	0.29	0.50	0.13	0.19	0.88	43,099.80
	GS-A	63.43 (c)	5.88	0.46	0.48	0.21	0.10	0.20	0.90	67,708.00
	GS-SL:SH	62.61 (d)	5.71	0.38	0.45	0.50	0.17	0.21	0.90	22,501.43
	GS-SH:A	64.70 (b)	6.03	0.48	0.38	0.37	0.14	0.19	0.90	40,018.33
	GS-SL:SH:A	66.05 (a)	6.22	0.43	0.47	0.47	0.16	0.20	0.90	21,914.47
0.15	PS	67.54 (b)	6.45	0.28	0.38	0.38	0.10	0.10	-	170,906.06
	GS-SL	64.82 (d)	6.06	0.16	0.50	0.50	0.16	0.21	0.40	50,256.67
	GS-SH	67.79 (b)	6.44	0.24	0.23	0.50	0.16	0.19	0.74	93,815.07
	GS-A	70.18 (a)	6.75	0.32	0.45	0.19	0.13	0.18	0.82	108,386.59
	GS-SL:SH	66.19 (c)	6.21	0.39	0.44	0.50	0.16	0.20	0.86	23,237.68
	GS-SH:A	69.96 (a)	6.76	0.19	0.38	0.39	0.16	0.18	0.89	95,290.54
	GS-SL:SH:A	67.76 (b)	6.50	0.41	0.46	0.46	0.16	0.21	0.86	23,206.52
0.3	PS	71.42 (b)	7.05	0.23	0.39	0.42	0.10	0.10	-	178,386.46
	GS-SL	68.24 (d)	6.54	0.13	0.49	0.49	0.17	0.20	0.28	65,661.65
	GS-SH	71.31 (b)	6.94	0.17	0.18	0.49	0.18	0.21	0.66	135,331.14
	GS-A	73.43 (a)	7.18	0.22	0.41	0.16	0.17	0.19	0.77	159,402.35
	GS-SL:SH	67.79 (d)	6.46	0.33	0.39	0.49	0.18	0.21	0.68	30,172.78
	GS-SH:A	73.12 (a)	7.15	0.13	0.37	0.37	0.17	0.19	0.86	123,779.24
	GS-SL:SH:A	68.93 (c)	6.62	0.40	0.44	0.44	0.16	0.20	0.75	26,376.25

[1]The letters in parentheses after ΔG represent the significance groups ($p < 0.05$) across these selection strategies within a specific correlation.
[2] $SD_{\Delta G}$ is the standard deviation of ΔG across 1000 simulation runs.

strategy GS-SH:A and for intermediate levels of PA (=0.5) and r (=0.15) (Table 3). ΔG increased by 1%, if the costs of phenotyping T_a reduced from 1.4 to 0.7 €. An increase of ΔG of 4 % was observed if the genotyping costs were reduced from 40 to 15 €.

4 | DISCUSSION

GS has been implemented in many commercial crop breeding programs nowadays (Krishnappa et al., 2021). However, implementation of GS in clonally propagated species is lagging behind, despite the expected advantages. This might be on one side because genomic resources are less developed in clonally propagated species compared to species bred as hybrids or inbred lines. Furthermore, a lower number of breeding methodological studies is dedicated to clonally propagated crops compared to inbred or hybrid species. Therefore, we evaluated the prospects to integrate GS into commercial potato breeding programs and assessed which parameters are crucial for its implementation.

4.1 | Comparison of selection strategies

We have studied the implementation of GS in a standard clone breeding program with minimal changes of the breeding program. This procedure was chosen as we expect that this will be the way how commercial clone breeding programs will deal with this possibility or challenge. However, we are aware that GS might result in even higher gains of selection if applied in a less conservative setting where the possibilities of reducing the length of breeding cycles are exploited. In addition, we assumed in this study that a GS model has been trained for the target trait on earlier cycles of the breeding program, and thus, the prediction accuracy was given. However, to keep this accuracy at a high level, the GS model should be re-trained and updated at each new breeding cycle. One possibility is that the clones selected as parents are used to update the GS model. These aspects will be considered in a companion study.

In this study, all evaluated selection strategies that make use of GS resulted in higher ΔG compared to the Standard-PS strategy if other parameters such as budget, variance components and selected proportions were held constant (Figure 3).

TABLE 3 Optimum allocation of resources to maximize genetic gain of the target trait (ΔG) across different cost scenarios when genomic selection was applied in single hills and A clone stages (GS-SH:A). The correlation between the two traits was 0.15 and the prediction accuracy 0.5. p_1 to p_5, α_k, and N_1 are the selected proportion per stage, the weight of genomic selection relative to phenotypic selection, and the number of clones at the seedling stage, respectively.

$Cost_{T_a}$ [1]	$Cost_{geno}$ [1]	ΔG [2]	$SD_{\Delta G}$ [3]	p_1	p_2	p_3	p_4	p_5	α_k	N_1
0.70	15	72.33 (a)	7.08	0.17	0.34	0.34	0.14	0.16	0.87	171,397.94
0.70	25	70.76 (c)	6.87	0.15	0.37	0.37	0.15	0.18	0.87	133,082.10
0.70	40	69.11 (e)	6.66	0.12	0.41	0.40	0.16	0.20	0.89	106,152.00
1.05	15	71.85 (ab)	7.00	0.20	0.35	0.36	0.14	0.17	0.88	135,898.99
1.05	25	70.39 (cd)	6.83	0.16	0.38	0.38	0.16	0.17	0.88	113,093.84
1.05	40	68.61 (ef)	6.57	0.15	0.40	0.41	0.18	0.19	0.87	87,752.05
1.40	15	71.39 (b)	6.95	0.23	0.35	0.37	0.14	0.17	0.88	110,223.57
1.40	25	69.96 (d)	6.76	0.19	0.38	0.39	0.16	0.18	0.89	95,290.54
1.40	40	68.41 (f)	6.52	0.17	0.41	0.41	0.16	0.21	0.90	76,796.40

[1] $Cost_{T_a}$ is the phenotyping cost of early measured trait, and $Cost_{geno}$ the genotyping cost per clone.
[2]The letters in parentheses after ΔG represent the significance groups ($p < 0.05$) across these cost scenarios.
[3] $SD_{\Delta G}$ is the standard deviation of ΔG across 1000 simulation runs.

This is in accordance with the theory about indirect selection response. This theory suggests that GS strategies should be superior to the Standard-PS if $PA > r \cdot H_{T_a}$, keeping the intensity of selection for GS ($i_{EGV_{T_t}}$) and PS (i_{T_a}) equal. Furthermore, the theory suggests that this trend should be even more pronounced, if $i_{T_a} < i_{EGV_{T_t}}$. This is what we have observed in our simulations, namely that the difference between ΔG of GS and PS was increased, if α_k increases.

Among the examined strategies using GS in only one stage, the ranking with respect to maximum ΔG was GS-SH > GS-A > GS-SL, independently of PA, r, and α_k (Figure 3). The observation that GS-SH resulted in a higher ΔG than GS-A can be explained by superiority of early selection on T_t because thereby one can avoid discarding clones with top performance for T_t in the early stages. Our observation of an increased advantage of GS-SH over GS-A if r decreased confirmed this explanation.

Following this argumentation, one could have expected GS-SL to be the strategy with the highest ΔG, especially if r is negative. This is because a direct selection of seedlings for EGV_{T_t} should be more efficient than selecting them based on P_{T_a} that negatively correlated with TGV_{T_t}. Therefore, the observation of GS-SL as the most disadvantageous GS method (Figure 3) was surprising at a first glance. However, in this strategy after one step of GS all further selection steps are exclusively made based on P_{T_a} and this hampers the selection of those individuals with beneficial alleles for T_t. Thus, the individuals with the highest TGV_{T_t} that were selected by GS in the seedling stage are probably discarded in the following selection steps from single hills to B clone stages. Another explanation for the observation of GS-SL as the most disadvantageous GS method is that the selection of the seedling stage based on GS leads to a dramatic reduction of population size in the seedling stage to keep the budget constant despite the burden of high genotyping costs (Figure S4). Our observations suggest that alternative prediction and selection methods to GS need to be developed for the first stage of clone breeding programs that result in a much lower cost per clone in order to exploit the potential of predictive breeding.

Among all examined selection strategies, those that applied GS several times are for all combinations of α_k, VC, and L superior to the ones using GS in only one stage of the breeding program (Figure 3), even without recalibrating the GS model. This superiority is most probably due to the possibility to select several times on EGV_{T_t} without having extra genotyping costs.

Among the strategies that used GS multiple times, the highest ΔG was observed for the strategies GS-SL:SH:A and GS-SH:A (Figure 3). The ranking of these two strategies was influenced by the genetic situation. GS-SL:SH:A outperformed GS-SH:A under low r and high PA. Therefore, we advice using GS-SL:SH:A in a very favorable GS environment (high PA and low r), and GS-SH:A in a favorable PS environment (low PA and high r).

In the scenario discussed in the previous paragraph, the selection intensities of the individual stages were kept equal to those of the Standard-PS strategy. However, theoretical considerations suggest that the implementation of GS requires an adaptation of the selection intensities as well as the phenotyping intensities. These are discussed in the next paragraph.

4.2 | Optimal allocation of resources

We observed a significantly higher ΔG for the Optimal-PS compared to the Standard-PS strategy (Figure 5). Smaller values for p_4 and p_5 (i.e., higher selection intensities) in

Optimal-PS (0.10) were observed compared to those in Standard-PS (0.20) (Table 2, Tables S2 and S3). This can be explained by the fact that at the B and C clone stages, the selection is exclusively based on P_{T_t} in a direct selection. Therefore, when increasing the selection intensities in these stages, ΔG is increasing as well.

The correlation between T_a and T_t also influences the optimal selection intensity. We observed a higher p_1, that is, a lower selection intensity, when r= -0.15 compared to the scenario with positive values for r (Table 2, Tables S2 and S3). This can be interpreted such that in cases of a negative r, i_{T_a} needs to be reduced to avoid discarding too many clones based on P_{T_a} that have a high TGV_{T_t}.

Furthermore, p_k values were lower for those stages of the breeding program at which GS was applied compared to the same stage in a selection strategy without GS (Table 2, Tables S2 and S3). The explanation for this observation can be that a low number of clones are enough to identify those with the best TGV_{T_t} if the more precise GS is applied. This finding illustrates that either an increased prediction accuracy or $i_{EGV_{T_t}}$ or both simultaneously can enhance ΔG.

We observed for most considered simulation scenarios no significant difference of ΔG between the Optimal-GS strategies and Standard-GS strategies (Figures 3 and 5). However, this comparison was not the purpose of our study. The simulations with varying selection intensities required to fix the final number of clones (N_6). We have decided to fix N_6 to that of the Standard-PS in order to allow a fair comparison of ΔG. In contrast, the purpose of the simulations of the standard strategies (PS but also GS) was based on keeping the selection intensities fixed between PS and GS strategies. The latter, however, results in considerably lower numbers of clones at the D clone stage (N_6) which increases ΔG (cf. Longin et al., 2006).

The ranking of the optimized selection strategies with respect to ΔG was with the exception of GS-SH and GS-A identical to the one observed for the Standard-GS strategies (Figure 5). One explanation for the rank change of GS-SH and GS-A might be the stronger selection applied at A clone stage in GS-A compared to GS-SH (Table 2, Tables S2 and S3). This indicates that a higher selection intensity in a later stage can improve ΔG more than an earlier selection on EGV_{T_t}.

4.3 | Impact of novel technical developments in the field of genomics or phenomics on the selection strategy

Another possibility to increase the selection intensity for improvement of short-term genetic gain is to generate more selection candidates while keeping the number of selected individuals constant (Cobb et al., 2019). Under a fixed budget, a reduction of either genotyping or phenotyping costs could increase the population size. With the development of high-throughput phenotyping and genotyping techniques, both their costs could gradually decrease (Araus & Cairns, 2014; Ragoussis, 2009). Consequently, we considered three different levels of phenotyping and genotyping costs and investigated how they affect the genetic gain in the context of optimal allocation of resources with the strategy GS-SH:A. The reduction of cost increased the population size at the seedling stage as well as enhanced the selection intensities p_2 and p_3 (when implementing GS), and p_4 and p_5 (direct selection on T_t). The increasing ΔG value observed in our study with a decrease in either genotyping or phenotyping cost (Table 3) confirmed this hypothesis. Furthermore, our findings are in line with a former study in wheat (Marulanda et al., 2016), which showed an increased ΔG and a higher number of test candidates as the cost for hybrid seed production or double haploids decreased. In summary, changes in correlation between the two selected traits, prediction accuracy, stage of implementation, and costs for genotyping and phenotyping have a crucial influence on the optimal allocation of resources to maximize the short-term genetic gain, accentuating the necessity for clone and especially potato breeders to regularly and carefully re-adjust their selection strategy.

4.4 | Impact of GS on genetic variance

Not only the genetic gain is important for the evaluation of the GS strategy, but also the genetic variance reduction of T_t. As expected, all the selection strategies showed a decrease in the genetic variance after selection (Figure S9). This tendency increased when GS was implemented. This is in accordance with former studies (Gaynor et al., 2017; Muleta et al., 2019) which showed a greater loss of genetic variance over time using GS compared to PS. In our study, the genetic variance decreased particularly at the stage of implementation (k), but not to the same extent for all strategies (Figure 3 and Figure S9). This trend can be explained by the Bulmer effect (Bulmer, 1971), which reduces the proportion of genetic variance due to linkage disequilibrium between trait coding polymorphisms (Van Grevenhof et al., 2012). This is in accordance with results of Jannink (2010), who showed that GS can accelerate the fixation of favorable alleles for T_t compared to PS resulting in a loss of genetic variance for the trait. The reduction of genetic variance, however, limits the ΔG for long-term improvement. Therefore, maintaining diversity of the population in the breeding materials is one possibility to slow down this drawback to improve long-term genetic gain in breeding programs (Gorjanc et al., 2018). However, for commercial breeding programs a balance between short- and long-term gain of selection is required, which needs further research.

4.5 | Conclusions

The present study demonstrated that implementing GS in a typical clone breeding program improves the gain of selection even without exploiting the possibilities to reduce the length of the breeding cycles. Furthermore, we showed that the integration of GS in consecutive selection stages can largely enhance the gain from selection compared to the use in only one stage. In detail, the strategy GS-SL:SH:A is highly recommended if the correlation between T_a and T_t is negative. Otherwise, GS-SH:A can be the most efficient strategy. However, with the consideration of optimal resource allocation, the superiority of multiple GS over single GS is not obvious anymore and their ranking depends on PA and r. Furthermore, we observed that the implementation of GS in potato breeding programs requires the adjustment of the selection intensities as well as the phenotyping intensities compared to those typically used in breeding programs exploiting exclusively PS. Finally, we outlined how to adjust the selection intensities in potato breeding programs after implementing GS.

ACKNOWLEDGMENTS

Computational infrastructure and support were provided by the Centre for Information and Media Technology (ZIM) at Heinrich Heine University Düsseldorf. This study was funded by the Federal Ministry of Food and Agriculture/Fachagentur Nachwachsende Rohstoffe (grantID 22011818, PotatoTools). The funders had no influence on study design, the collection, analysis and interpretation of data, the writing of the manuscript, and the decision to submit the manuscript for publication.

DATA AVAILABILITY STATEMENT

The datasets generated or analyzed during this study and R scripts are available in the Github repository: https://github.com/poyawu/optimize_potato_breeding_program_I.

AUTHOR CONTRIBUTIONS

Po-Ya Wu: Conceptualization, Data curation, Formal analysis, Writing – original draft. **Benjamin Stich**: Conceptualization, Funding acquisition, Project administration, Review & editing. **Juliane Renner**: Funding acquisition, Resources, Review & editing. **Katja Muders**: Funding acquisition, Resources, Review & editing. **Vanessa Prigge**: Funding acquisition, Resources, Review & editing. **Delphine Van Inghelandt**: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – original draft, Review & editing.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

ORCID

Po-Ya Wu https://orcid.org/0000-0002-7342-2867
Benjamin Stich https://orcid.org/0000-0001-6791-8068
Delphine van Inghelandt https://orcid.org/0000-0002-2819-843X

REFERENCES

Araus, J. L., & Cairns, J. E. (2014). Field high-throughput phenotyping: The new crop breeding frontier. *Trends in Plant Science*, *19*(1), 52–61.

Bulmer, M. G. (1971). The effect of selection on genetic variability. *The American Naturalist*, *105*(943), 201–211.

Byrne, S., Meade, F., Mesiti, F., Griffin, D., Kennedy, C., & Milbourne, D. (2020). Genome-wide association and genomic prediction for fry color in potato. *Agronomy*, *10*(1), 90.

Caruana, B. M., Pembleton, L. W., Constable, F., Rodoni, B., Slater, A. T., & Cogan, N. O. (2019). Validation of genotyping by sequencing using transcriptomics for diversity and application of genomic selection in tetraploid potato. *Frontiers in Plant Science*, *10*, 670.

Cobb, J. N., Juma, R. U., Biswas, P. S., Arbelaez, J. D., Rutkoski, J., Atlin, G., Hagen, T., Quinn, M., & Ng, E. H. (2019). Enhancing the rate of genetic gain in public-sector plant breeding programs: Lessons from the breeder's equation. *Theoretical and Applied Genetics*, *132*(3), 627–645.

Desta, Z. A., & Ortiz, R. (2014). Genomic selection: Genome-wide prediction in plant improvement. *Trends in Plant Science*, *19*(9), 592–601.

Enciso-Rodriguez, F., Douches, D., Lopez-Cruz, M., Coombs, J., & de los Campos, G. (2018). Genomic selection for late blight and common scab resistance in tetraploid potato (*Solanum tuberosum*). *G3: Genes, Genomes, Genetics*, *8*(7), 2471.

Endelman, J. B., Carley, C. A., Bethke, P. C., Coombs, J. J., Clough, M. E., da Silva, W. L., De Jong, W. S., Douches, D. S., Frederick, C. M., & Yencho, G. C. (2018). Genetic variance partitioning and genome-wide prediction with allele dosage information in autotetraploid potato. *Genetics*, *209*(1), 77–87.

Gaynor, R. C., Gorjanc, G., Bentley, A. R., Ober, E. S., Howell, P., Jackson, R., Mackay, I. J., & Hickey, J. M. (2017). A two-part strategy for using genomic selection to develop inbred lines. *Crop Science*, *57*(5), 2372–2386.

Gaynor, R. C., Gorjanc, G., & Hickey, J. M. (2021). AlphaSimR: An R package for breeding program simulations. *G3: Genes, Genomes, Genetics*, *11* (2), jkaa017.

Gemenet, D. C., Lindqvist-Kreuze, H., De Boeck, B., da Silva Pereira, G., Mollinari, M., Zeng, Z. B., Craig Yencho, G., & Campos, H. (2020). Sequencing depth and genotype quality: Accuracy and breeding operation considerations for genomic selection applications in autopolyploid crops. *Theoretical and Applied Genetics*, *133*(12), 3345–3363.

Gopal, J. (2006). Considerations for successful breeding. In J. Gopal & S. M. P. Khurana (Eds.), *Handbook of potato production, improvement, and postharvest management* (1st ed., pp. 77–108). CRC Press.

Gorjanc, G., Gaynor, R. C., & Hickey, J. M. (2018). Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. *Theoretical and Applied Genetics*, *131*(9), 1953–1966.

Grüneberg, W., Mwanga, R., Andrade, M., & Espinoza, J. (2009). Selection methods. Part 5: Breeding clonally propagated crops. In S. Ceccarelli, E. P. Guimarães, & E. Weltzien (Eds.), *Plant breed-*

ing and farmer participation (pp. 275–322). Food and Agriculture Organization of the United Nations (FAO).

Hayes, B., & Goddard, M. E. (2001). The distribution of the effects of genes affecting quantitative traits in livestock. *Genetics Selection Evolution*, *33*, 209–229.

Jannink, J. L. (2010). Dynamics of long-term genomic selection. *Genetics Selection Evolution*, *42*(1), 1–11.

Jansky, S. H., Charkowski, A. O., Douches, D. S., Gusmini, G., Richael, C., Bethke, P. C., Spooner, D. M., Novy, R. G., De Jong, H., & Jiang, J. (2016). Reinventing potato as a diploid inbred line–based crop. *Crop Science*, *56*(4), 1412–1422.

Krishnappa, G., Savadi, S., Tyagi, B. S., Singh, S. K., Mamrutha, H. M., Kumar, S., Mishra, C. N., Khan, H., Gangadhara, K., Uday, G., Singh, G., & Singh, G. P. (2021). Integrated genomic selection for rapid improvement of crops. *Genomics*, *113*(3), 1070–1086.

Lindhout, P., Meijer, D., Schotte, T., Hutten, R. C., Visser, R. G., & van Eck, H. J. (2011). Towards F 1 hybrid seed potato breeding. *Potato Research*, *54*(4), 301–312.

Longin, C. F. H., Mi, X., & Würschum, T. (2015). Genomic selection in wheat: Optimum allocation of test resources and comparison of breeding strategies for line and hybrid breeding. *Theoretical and Applied Genetics*, *128*(7), 1297–1306.

Longin, C. F. H., Utz, H. F., Reif, J. C., Schipprack, W., & Melchinger, A. E. (2006). Hybrid maize breeding with doubled haploids: I. One-stage versus two-stage selection for testcross performance. *Theoretical and Applied Genetics*, *112*(5), 903–912.

Marulanda, J. J., Mi, X., Melchinger, A. E., Xu, J. L., Würschum, T., & Longin, C. F. H. (2016). Optimum breeding strategies using genomic selection for hybrid breeding in wheat, maize, rye, barley, rice and triticale. *Theoretical and Applied Genetics*, *129*(10), 1901–1913.

Melchinger, A. E., Longin, C. F., Utz, H. F., & Reif, J. C. (2005). Hybrid maize breeding with doubled haploid lines: Quantitative genetic and selection theory for optimum allocation of resources. In *Proceedings of the 41st Annual Illinois Corn Breeders School* (pp. 8–21).

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

Muleta, K. T., Pressoir, G., & Morris, G. P. (2019). Optimizing genomic selection for a sorghum breeding program in Haiti: A simulation study. *G3: Genes, Genomes, Genetics*, *9*(2), 391–401.

Ortiz, R., Reslow, F., Cuevas, J., & Crossa, J. (2022). Genetic gains in potato breeding as measured by field testing of cultivars released during the last 200 years in the Nordic Region of Europe. *The Journal of Agricultural Science*, 1–7.

Ragoussis, J. (2009). Genotyping technologies for genetic research. *Annual Review of Genomics and Human Genetics*, *10*, 117–133.

Slater, A. T., Cogan, N. O., Forster, J. W., Hayes, B. J., & Daetwyler, H. D. (2016). Improving genetic gain with genomic selection in autotetraploid potato. *The Plant Genome*, *9*(3), 1–15.

Sood, S., Lin, Z., Caruana, B., Slater, A. T., & Daetwyler, H. D. (2020). Making the most of all data: Combining non-genotyped and genotyped potato individuals with HBLUP. *Plant Genome*, *13*(3), 1–12.

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., Atlin, G., Jannink, J. L., & McCouch, S. R. (2015). Genomic selection and association mapping in rice (Oryza sativa): Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLOS Genetics*, *11*(2), e1004982.

Stich, B., & Van Inghelandt, D. (2018). Prospects and potential uses of genomic prediction of key performance traits in tetraploid potato. *Frontiers in Plant Science*, *9*, 159.

Stokstad, E. (2019). The new potato. *Science*, *363*(6427), 574–577.

Sverrisdóttir, E., Byrne, S., Sundmark, E. H. R., Johnsen, H. Ø., Kirk, H. G., Asp, T., Janss, L., & Nielsen, K. L. (2017). Genomic prediction of starch content and chipping quality in tetraploid potato using genotyping-by-sequencing. *Theoretical and Applied Genetics*, *130*(10), 2091–2108.

Sverrisdóttir, E., Sundmark, E. H. R., Johnsen, H. Ø., Kirk, H. G., Asp, T., Janss, L., Bryan, G., & Nielsen, K. L. (2018). The value of expanding the training population to improve genomic selection models in tetraploid potato. *Frontiers in Plant Science*, *9*, 1118.

Van Grevenhof, E. M., Van Arendonk, J. A., & Bijma, P. (2012). Response to genomic selection: The Bulmer effect and the potential of genomic selection when the number of phenotypic records is limiting. *Genetics Selection Evolution*, *44*(1), 1–10.

Werner, C. R., Gaynor, R. C., Sargent, D. J., Lillo, A., Gorjanc, G., & Hickey, J. M. (2023). Genomic selection strategies for clonally propagated crops. *Theoretical and Applied Genetics*, *136*(4), 1–17.

Wilson, S., Zheng, C., Maliepaard, C., Mulder, H. A., Visser, R. G., van der Burgt, A., & van Eeuwijk, F. (2021). Understanding the effectiveness of genomic prediction in tetraploid potato. *Frontiers in Plant Science*, *12*, 1634.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wu, P.-Y., Stich, B., Renner, J., Muders, R., Prigge, V., & van Inghelandt, D. (2023). Optimal implementation of genomic selection in clone breeding programs—Exemplified in potato: I. Effect of selection strategy, implementation stage, and selection intensity on short-term genetic gain. *The Plant Genome*, *16,* e20327. https://doi.org/10.1002/tpg2.20327

[Correction added on 25 May 2023, after first online publication: Author name corrected from Wu, P. to Wu, P.-Y.]

SUPPORTING INFORMATION

Method S1: The establishment of the Marey map

The Marey map (Chakravarti, 1991) was established from two public datasets: i. Bourke et al. (2015) – genetic map of 3,273 markers; ii. Vos et al. (2015) – physical position of 3,273 markers. First, the markers with unknown physical positions and linkage group discrepancies between both datasets were removed. As a result, 3,206 markers were retained for further analyses. Second, the position of the genetic map with inverse order against the physical map was adjusted within each homolog of a chromosome and within each parent. Third, in order to increase the power of the cubic smoothing spline, we aggregated the SNP information of the four homologs and two parents by shifting the position of the genetic map with the known centromere information. The centromere information was taken from Table S2 of Bourke et al. (2015). The resulting Marey map, physical position (Mb) against genetic position (cM), was shown in Figure S1. Then, a cubic smoothing spline was used to fit the coordinates of the Marey map for each chromosome. To avoid the computational burden, we randomly selected from all possible variants one every 2.5 kilobases resulting in 287,858 sequence variants. Their genetic map positions were predicted based on the fitted cubic smoothing spline.

Method S2: The derivation of cost function

The initial cost function can be expressed by

$$\text{Budget} = \sum_{j=1}^{6} \mathrm{N}_j \times \mathrm{cost}_{\mathrm{pheno}(j)} \times \mathrm{L}_j + \mathrm{N}_{\mathrm{GS}} \times \mathrm{cost}_{\mathrm{geno}},$$

where we let N_1-N_5 replace by related to N_6 and the selected proportions, then $\mathrm{N}_1 = \frac{\mathrm{N}_6}{\mathrm{p}_1\mathrm{p}_2\mathrm{p}_3\mathrm{p}_4\mathrm{p}_5}$, $\mathrm{N}_2 = \frac{\mathrm{N}_6}{\mathrm{p}_2\mathrm{p}_3\mathrm{p}_4\mathrm{p}_5}$, $\mathrm{N}_3 = \frac{\mathrm{N}_6}{\mathrm{p}_3\mathrm{p}_4\mathrm{p}_5}$, $\mathrm{N}_4 = \frac{\mathrm{N}_6}{\mathrm{p}_4\mathrm{p}_5}$, and $\mathrm{N}_5 = \frac{\mathrm{N}_6}{\mathrm{p}_5}$. Furthermore,

N_{GS} can be also replaced by related to N_6, the selected proportions, and the proportion of selected clones to genotype (α_m), where m was the stage that GS was applied first. For more details, $m = 1$ referred to GS-SL, GS-SL:SH and GS-SL:SH:A; $m = 2$ for GS-SH and GS-SH:A; and $m = 3$ for GS-A. That is, $N_{GS} = \frac{N_6 \alpha_m}{\Pi_{k=m}^{5} p_k}$. Therefore, the initial cost function can be modified by

$$\text{Budget} = \sum_{j=1}^{5} \frac{N_6}{\Pi_{k=j}^{5} p_k} \text{cost}_{\text{pheno}(j)} L_j + N_6 \text{cost}_{\text{pheno}(6)} L_6 + \frac{N_6 \text{cost}_{\text{geno}} \alpha_m}{\Pi_{k=m}^{5} p_k}.$$

REFERENCES

Bourke, P. M., Voorrips, R. E., Visser, R. G. F., and Maliepaard, C. (2015). The double-reduction landscape in tetraploid potato as revealed by a high-density linkage map. *Genetics*, 201(3):853–863.

Chakravarti, A. (1991). A graphical representation of genetic and physical maps: the Marey map. *Genomics*, 11(1):219–222.

Vos, P. G., Uitdewilligen, J. G., Voorrips, R. E., Visser, R. G., and van Eck, H. J. (2015). Development and analysis of a 20K SNP array for potato (Solanum tuberosum): an insight into the breeding history. *Theoretical and Applied Genetics*, 128(12):2387–2401.

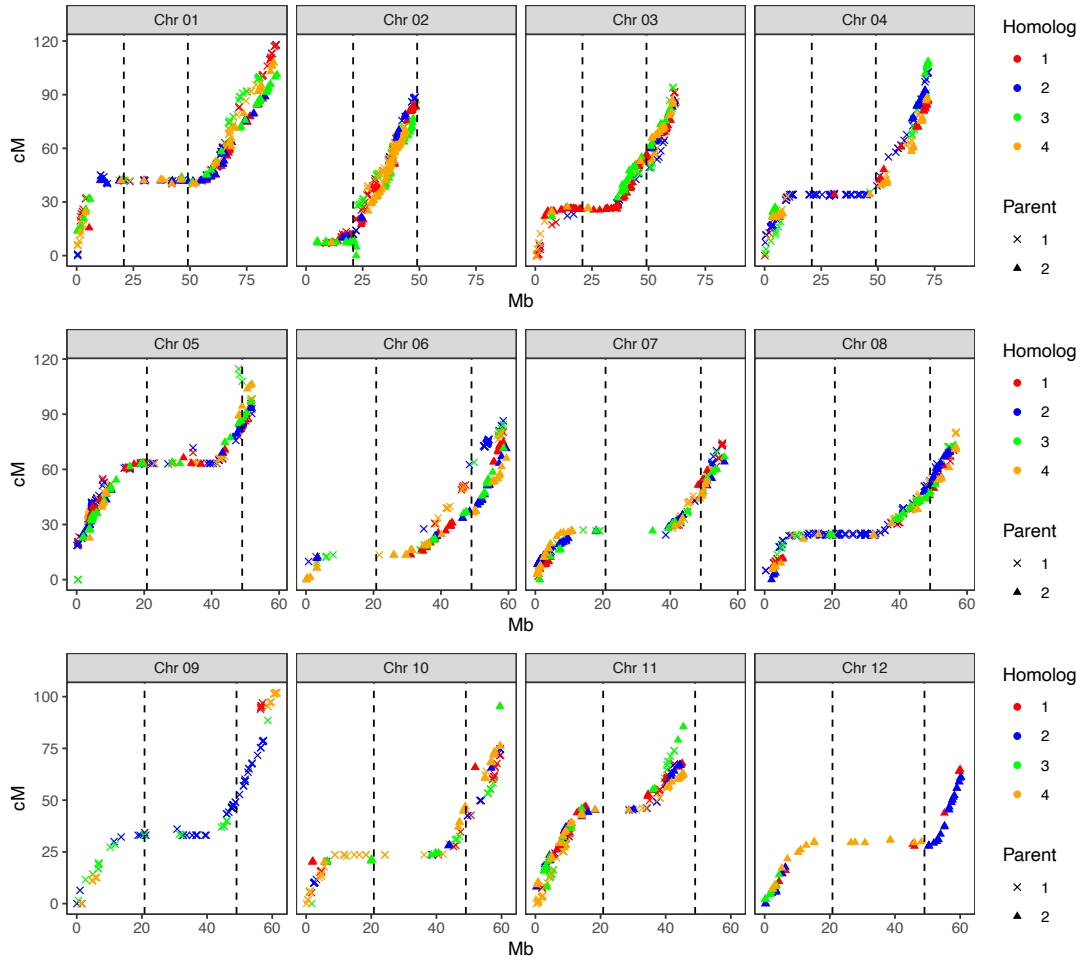

Figure S1: Marey map based on the aggregated SNP information of the four homologs and two parents for each chromosome.

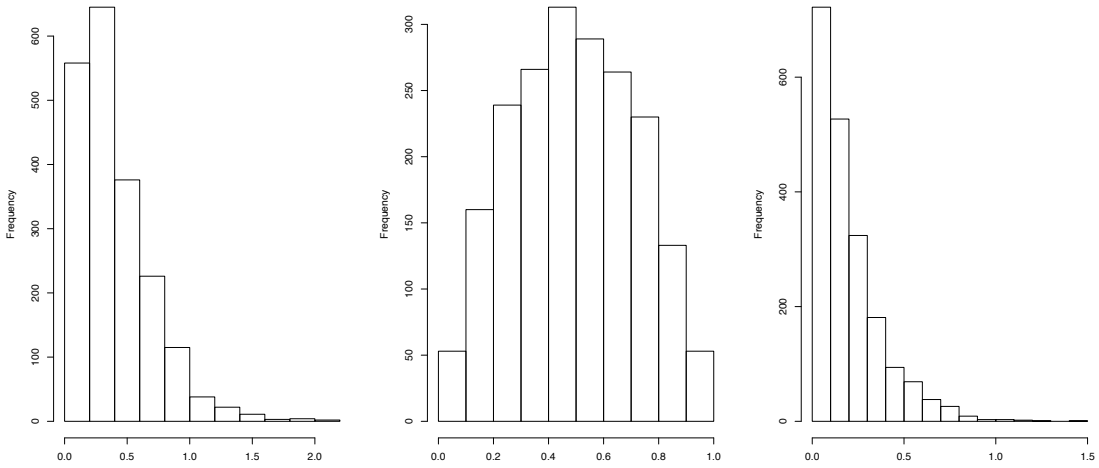

Figure S2: Histogram of additive effects (left), ratios of dominance and additive effects (middle), and dominance effects (right) for 2,000 QTL.

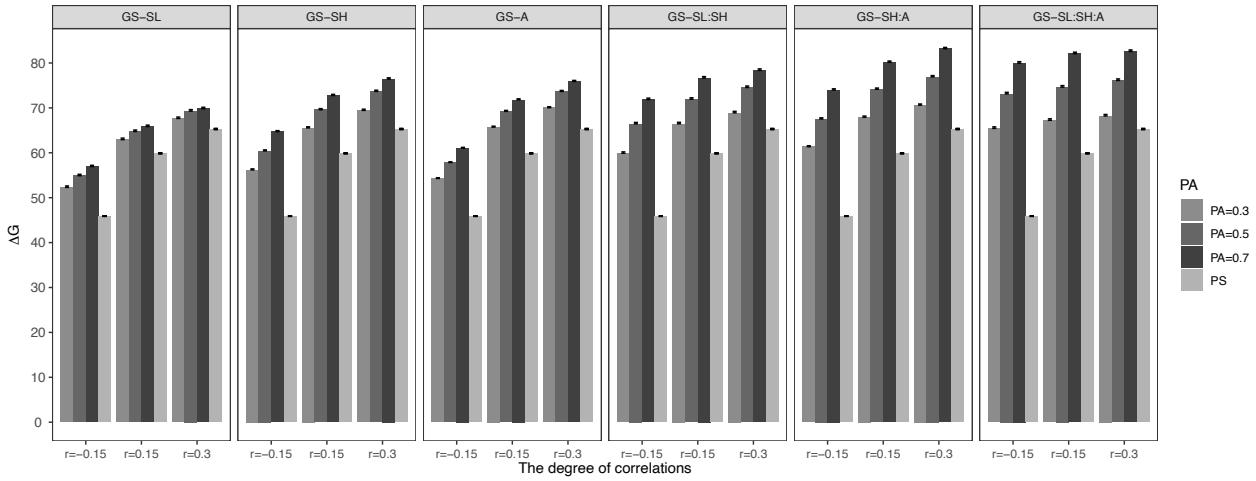

Figure S3: Genetic gain for the target trait (ΔG) on average across 1,000 simulation runs at D clone stage for different correlations between the traits (r=-0.15, 0.15, and 0.3), prediction accuracies (PA=0.3, 0.5, and 0.7), and different selection strategies, where the weight of genomic selection relative to phenotypic selection (α_k) was 0.9. Error bars represent the standard error of the genetic gain across 1,000 simulation runs. This evaluation was based on VC1 ($\sigma^2_G : \sigma^2_{G\times L} : \sigma^2_\epsilon = 1 : 1 : 0.5$).

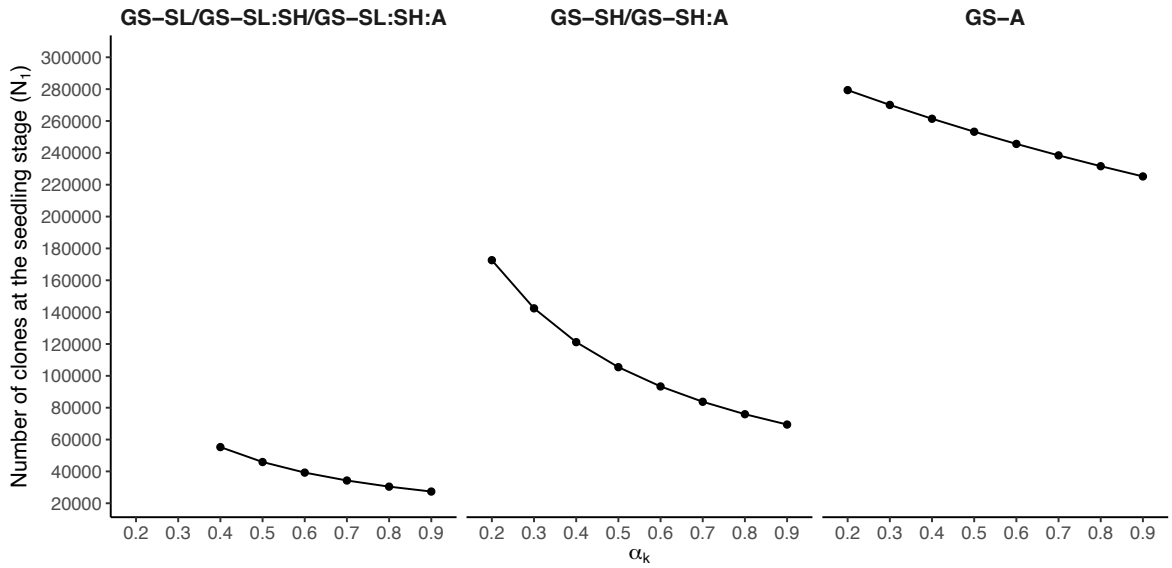

Figure S4: Number of clones in the seedling stage (N_1) across the examined weights of genomic selection relative to phenotypic selection (α_k) for three different selection strategies.

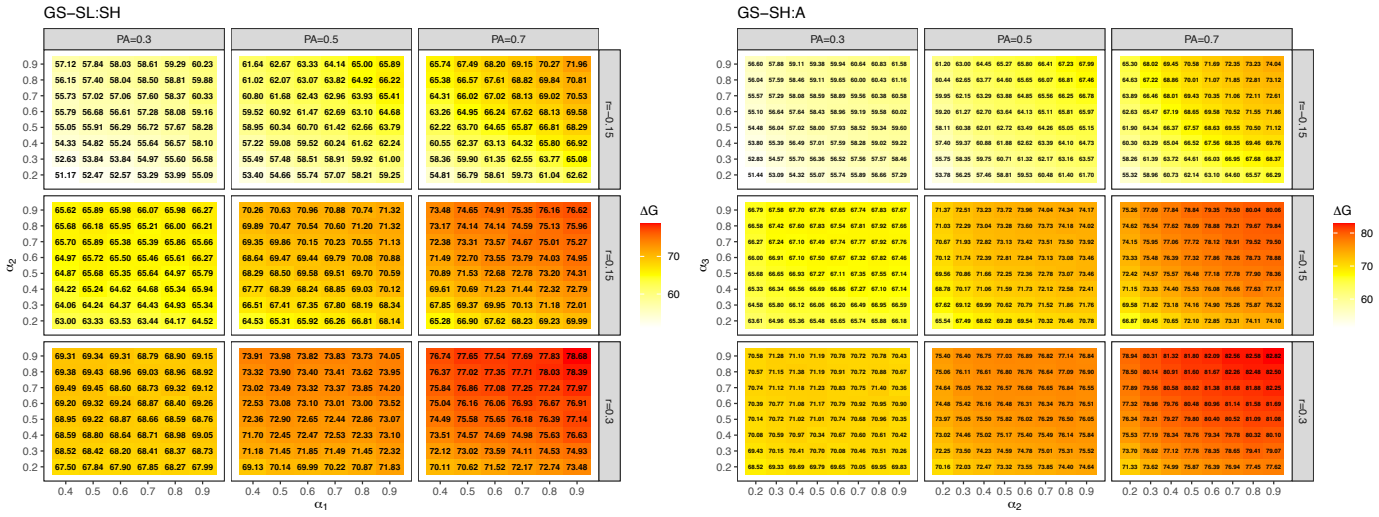

Figure S5: Genetic gain for the target trait (ΔG) on average across 1,000 simulation runs at the D clone stage for different correlations between the traits (r=-0.15, 0.15, and 0.3), and prediction accuracies (PA=0.3, 0.5, and 0.7) in the selection strategies GS-SL:SH (left) and GS-SH:A (right) for all combinations of weight of genomic selection relative to phenotypic selection (α_k). This evaluation was based on VC1 ($\sigma^2_G : \sigma^2_{G\times L} : \sigma^2_\epsilon = 1 : 1 : 0.5$).

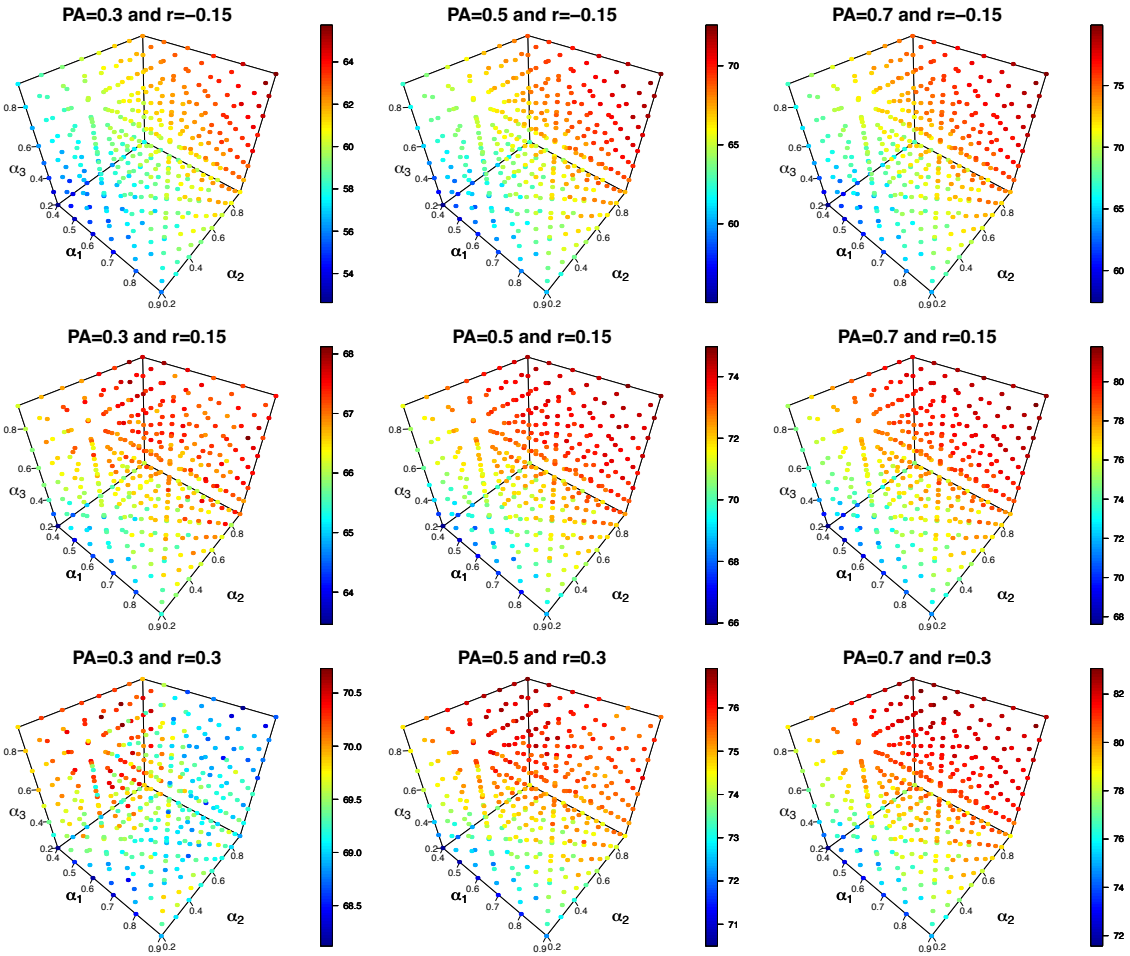

Figure S6: Genetic gain for the target trait (ΔG) across 1,000 simulation runs at the D clone stage for different correlations between the traits (r=-0.15, 0.15, and 0.3) and prediction accuracies (PA=0.3, 0.5, and 0.7) in the selection strategy GS-SL:SH:A for all combinations of weight of genomic selection relative to phenotypic selection (α_k, where $k = 1, 2, 3$).This evaluation was based on VC1 ($\sigma^2_G : \sigma^2_{G\times L} : \sigma^2_\epsilon = 1 : 1 : 0.5$).

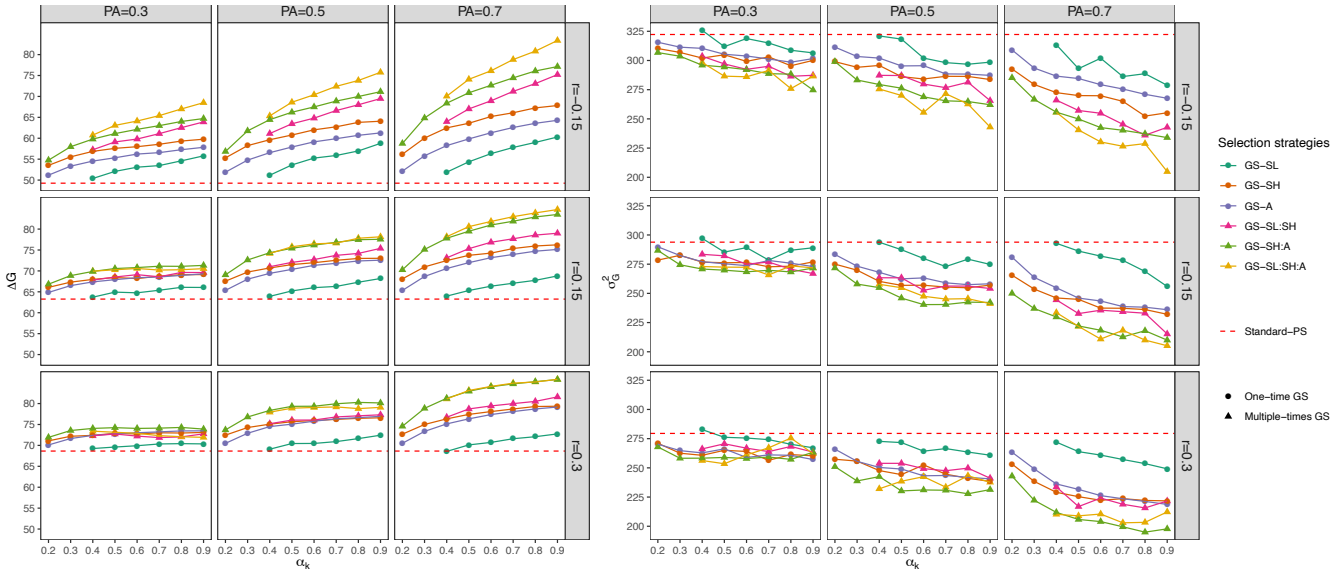

Figure S7: Genetic gain (ΔG, left) and genetic variance (σ^2_G, right) for the target trait on average across 1,000 simulation runs at D clone stage for different weights of genomic selection relative to phenotypic selection (α_k), different selection strategies, different correlations between the traits (r=-0.15, 0.15, and 0.3), prediction accuracies (PA=0.3, 0.5, and 0.7), and for the ratio of variance components VC2 ($\sigma^2_G : \sigma^2_{G\times L} : \sigma^2_\epsilon = 1 : 0.5 : 0.5$). The details regarding the selection strategies are shown in Figure 2.

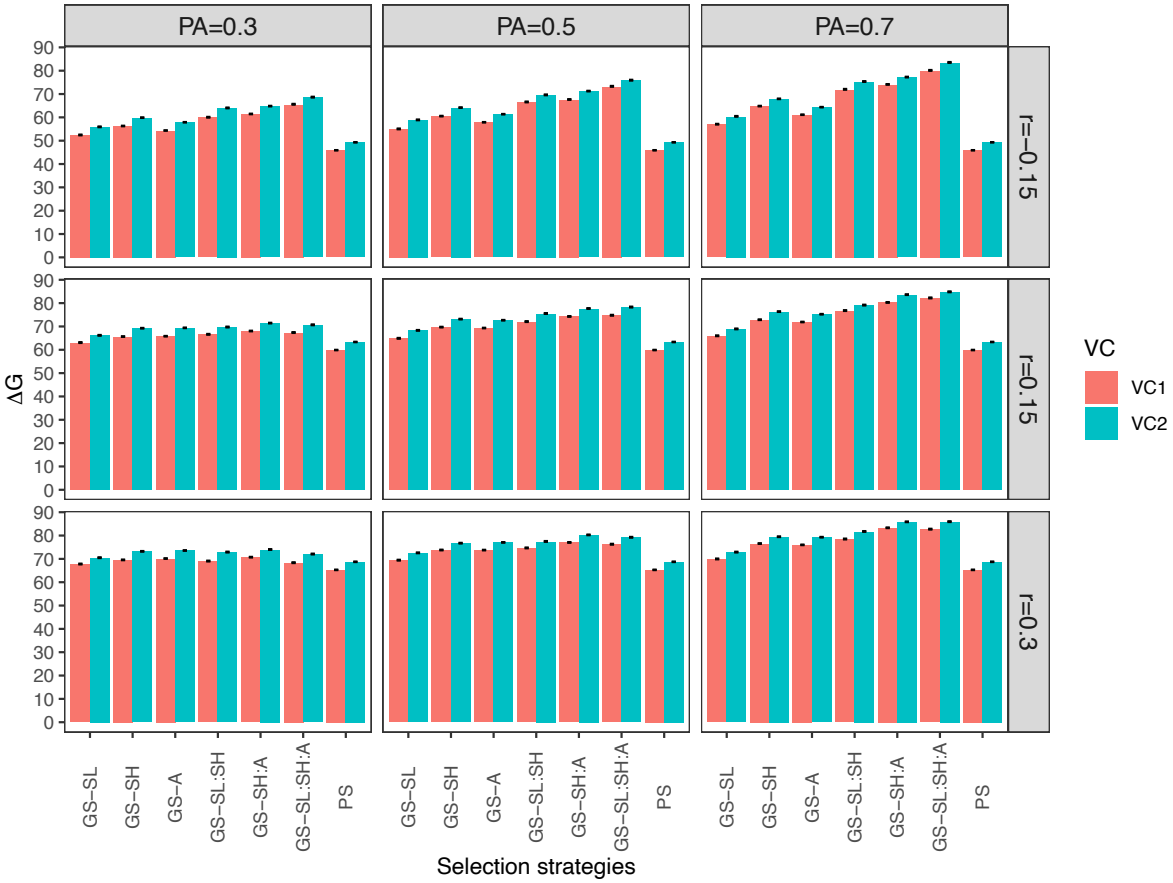

Figure S8: Genetic gain for the target trait (ΔG) on average across 1,000 simulation runs at D clone stage under different ratios variance components for the target trait ($\sigma^2_G : \sigma^2_{G\times L} : \sigma^2_\epsilon$): (1) 1 : 1 : 0.5 (VC1) and (2) 1 : 0.5 : 0.5 (VC2), different selection strategies, different correlations between the traits (r=-0.15, 0.15, and 0.3), and prediction accuracies (PA=0.3, 0.5, and 0.7) when the weight of genomic selection relative to phenotypic selection (α_k) was 0.9. Error bars represent the standard error of the genetic gain across 1,000 simulation runs.

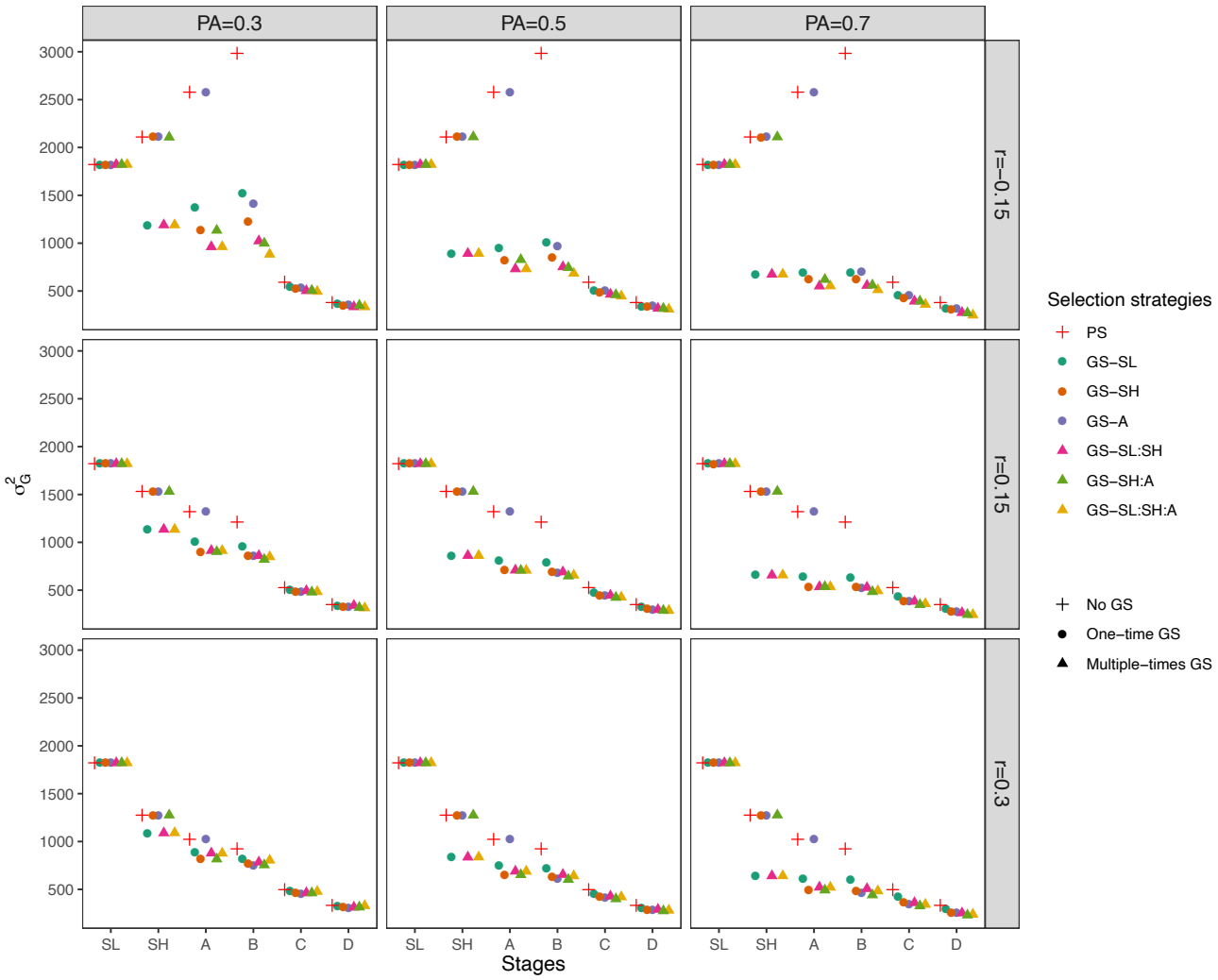

Figure S9: Genetic variance for the target trait (σ^2_G) on average across 1,000 simulation runs in the corresponding stage for different correlations between the traits (r=-0.15, 0.15, and 0.3), prediction accuracies (PA=0.3, 0.5, and 0.7) and different selection strategies, where the weight of genomic selection relative to phenotypic selection (α_k) was 0.9. This evaluation was based on VC1 ($\sigma^2_G : \sigma^2_{G\times L} : \sigma^2_\epsilon = 1 : 1 : 0.5$).

Table S1: The combination of the three different weights of genomic selection relative to phenotypic selection (α_1, α_2 and α_3) to reach the highest genetic gain for the target trait (ΔG) across 1,000 simulation runs in the strategy GS-SL:SH:A for the different correlations between the two traits (-0.15, 0.15, and 0.3) and prediction accuracies (0.3, 0.5, and 0.7).

Correlation	Prediction accuracy	α_1	α_2	α_3	ΔG	$SD_{\Delta G}$
-0.15	0.3	0.90	0.90	0.90	65.73	10.03
	0.5	0.90	0.90	0.90	72.57	10.94
	0.7	0.90	0.90	0.90	79.89	10.81
0.15	0.3	0.90	0.70	0.80	68.12	10.54
	0.5	0.80	0.90	0.90	74.97	10.23
	0.7	0.90	0.80	0.90	81.78	10.62
0.30	0.3	0.40	0.60	0.80	70.73	8.82
	0.5	0.40	0.90	0.90	76.91	9.22
	0.7	0.50	0.80	0.90	83.06	9.39

Table S2: Optimum allocation of resources to maximize genetic gain of the target trait (ΔG) for the different selection strategies and correlations between the two traits (r=-0.15, 0.15, and 0.3). The prediction accuracy was 0.3 and the phenotyping cost of early measured trait 1.4 € and genotyping cost 25 €. p_1 to p_5 , α_k, and N_1 are the selected proportion per stage, the weight of genomic selection relative to phenotypic selection, and the number of clones at the seedling stage, respectively.

Correlations	Selection strategies	ΔG [1]	$SD_{\Delta G}$ [2]	p_1	p_2	p_3	p_4	p_5	α_k	N_1
	PS	57.87 (d)	5.04	0.39	0.36	0.31	0.10	0.10	-	152,995.09
	GS-SL	57.21 (e)	5.04	0.43	0.50	0.50	0.12	0.21	0.82	23,100.40
	GS-SH	60.05 (c)	5.41	0.48	0.48	0.50	0.10	0.15	0.85	34,911.00
-0.15	GS-A	61.63 (b)	5.55	0.44	0.50	0.30	0.10	0.15	0.90	61,078.00
	GS-SL:SH	59.94 (c)	5.40	0.45	0.48	0.50	0.15	0.20	0.90	21,365.00
	GS-SH:A	62.89 (a)	5.92	0.47	0.50	0.50	0.10	0.15	0.90	34,214.00
	GS-SL:SH:A	62.59 (a)	5.81	0.48	0.49	0.49	0.14	0.21	0.90	21,157.25
	PS	67.54 (a)	6.45	0.28	0.38	0.38	0.10	0.10	-	170,906.06
	GS-SL	63.86 (c)	6.02	0.21	0.49	0.49	0.12	0.19	0.29	60,100.60
	GS-SH	66.18 (b)	6.30	0.25	0.37	0.49	0.12	0.15	0.67	88,054.48
0.15	GS-A	68.08 (a)	6.52	0.28	0.42	0.35	0.12	0.13	0.79	109,402.09
	GS-SL:SH	64.23 (c)	6.02	0.43	0.48	0.50	0.13	0.18	0.67	26,926.30
	GS-SH:A	67.60 (a)	6.50	0.19	0.48	0.47	0.12	0.17	0.84	92,069.85
	GS-SL:SH:A	65.13 (b)	6.19	0.46	0.49	0.49	0.14	0.16	0.72	24,930.00
	PS	71.42 (a)	7.05	0.23	0.39	0.42	0.10	0.10	-	178,386.46
	GS-SL	67.56 (b)	6.59	0.16	0.49	0.49	0.12	0.19	0.21	73,846.80
	GS-SH	69.74 (b)	6.83	0.17	0.33	0.49	0.13	0.17	0.59	126,127.77
0.3	GS-A	71.37 (a)	7.01	0.20	0.40	0.35	0.12	0.15	0.74	146,599.75
	GS-SL:SH	66.46 (c)	6.35	0.40	0.43	0.49	0.13	0.18	0.55	32,437.41
	GS-SH:A	70.58 (b)	6.91	0.14	0.44	0.43	0.13	0.17	0.75	118,439.54
	GS-SL:SH:A	66.84 (c)	6.37	0.44	0.46	0.45	0.14	0.17	0.59	29,865.59

[1] The letters in parentheses after ΔG represent the significance groups ($P < 0.05$) across these selection strategies within a specific correlation.

[2] $SD_{\Delta G}$ is the standard deviation of ΔG across 1,000 simulation runs.

Table S3: Optimum allocation of resources to maximize genetic gain of the target trait (ΔG) for the different selection strategies and correlations between the two traits (r=-0.15, 0.15, and 0.3). The prediction accuracy was 0.7 and the phenotyping cost of early measured trait 1.4 € and genotyping cost 25 €. p_1 to p_5 , α_k, and N_1 are the selected proportion per stage, the weight of genomic selection relative to phenotypic selection, and the number of clones at the seedling stage, respectively.

Correlations	Selection strategies	ΔG [1]	$SD_{\Delta G}$ [2]	p_1	p_2	p_3	p_4	p_5	α_k	N_1
	PS	57.87 (e)	5.04	0.39	0.36	0.31	0.10	0.10	-	152,995.09
	GS-SL	61.66 (d)	5.46	0.14	0.50	0.50	0.28	0.26	0.90	25,193.54
	GS-SH	63.81 (c)	5.69	0.45	0.15	0.50	0.20	0.20	0.85	48,549.75
-0.15	GS-A	65.73 (b)	5.95	0.46	0.47	0.16	0.10	0.25	0.90	73,034.40
	GS-SL:SH	65.66 (b)	5.97	0.24	0.37	0.50	0.26	0.25	0.90	24,744.58
	GS-SH:A	67.78 (a)	6.26	0.47	0.25	0.35	0.20	0.22	0.90	44,532.00
	GS-SL:SH:A	69.43 (a)	6.50	0.32	0.40	0.40	0.24	0.24	0.90	24,197.36
	PS	67.54 (e)	6.45	0.28	0.38	0.38	0.10	0.10	-	170,906.06
	GS-SL	66.45 (f)	6.07	0.13	0.50	0.50	0.24	0.24	0.61	36,392.38
	GS-SH	70.49 (c)	6.59	0.25	0.11	0.50	0.25	0.23	0.82	89,212.15
0.15	GS-A	73.05 (a)	6.89	0.30	0.43	0.10	0.21	0.20	0.85	124,176.68
	GS-SL:SH	68.75 (d)	6.42	0.25	0.35	0.50	0.26	0.25	0.87	25,501.81
	GS-SH:A	73.15 (a)	6.96	0.18	0.25	0.26	0.26	0.25	0.90	109,503.21
	GS-SL:SH:A	71.21 (b)	6.77	0.32	0.39	0.39	0.26	0.24	0.90	24,407.98
	PS	71.42 (c)	7.05	0.23	0.39	0.42	0.10	0.10	-	178,386.46
	GS-SL	69.26 (e)	6.44	0.11	0.49	0.49	0.22	0.23	0.46	46,171.65
	GS-SH	73.84 (b)	6.98	0.15	0.10	0.49	0.28	0.25	0.78	136,380.42
0.3	GS-A	76.29 (a)	7.23	0.22	0.39	0.10	0.23	0.22	0.85	172,683.82
	GS-SL:SH	70.22 (d)	6.59	0.22	0.30	0.50	0.29	0.27	0.76	29,922.56
	GS-SH:A	76.45 (a)	7.28	0.13	0.23	0.24	0.30	0.28	0.88	137,357.44
	GS-SL:SH:A	71.91 (c)	6.82	0.29	0.37	0.37	0.27	0.25	0.83	26,728.49

[1] The letters in parentheses after ΔG represent the significance groups ($P < 0.05$) across these selection strategies within a specific correlation.

[2] $SD_{\Delta G}$ is the standard deviation of ΔG across 1,000 simulation runs.

4 Optimal implementation of genomic selection in clone breeding programs - exemplified in potato: II. Effect of selection strategy and cross selection methods on long-term genetic gain

This manuscript is in preparation for submission.

Authors:
Po-Ya Wu, Benjamin Stich, Juliane Renner, Katja Muders, Vanessa Prigge, and Delphine van Inghelandt.

Own contribution: First author. I performed the data analyses and wrote the manuscript.

Optimal implementation of genomic selection in clone breeding programs - exemplified in potato: II. Effect of selection strategy and cross-selection methods on long-term genetic gain

Po-Ya Wu[1,2], Benjamin Stich[1,2,3,4], Juliane Renner[5], Katja Muders[6], Vanessa Prigge[7], and Delphine van Inghelandt[1,*]

[1]Institute of Quantitative Genetics and Genomics of Plants, Heinrich Heine University, 40225 Düsseldorf, Germany
[2]Present address: Institute for Breeding Research on Agricultural Crops, Federal Research Centre for Cultivated Plants, 18190 Sanitz, Germany
[3]Cluster of Excellence on Plant Sciences (CEPLAS), Heinrich Heine University, 40225 Düsseldorf, Germany
[4]Max Planck Institute for Plant Breeding Research, 50829 Köln, Germany
[5]Böhm-Nordkartoffel Agrarproduktion GmbH & Co. OHG, 17111 Hohenmocker, Germany
[6]NORIKA GmbH, 18190 Sanitz, Germany
[7]SaKa Pflanzenzucht GmbH & Co. KG, 24340 Windeby, Germany
[*]Corresponding author: Delphine van Inghelandt, inghelan@hhu.de

January 25, 2024

ABSTRACT

Different cross-selection (CS) methods incorporating genomic selection (GS) have been used in diploid species to improve long-term genetic gain and simultaneously preserve diversity. However, their application to highly heterozygous and tetraploid crops, for example, potato, is lacking up to now. The objectives of our study were to (i) assess how different CS methods incorporating GS with or without consideration of genetic variability affect both short- and long-term genetic gains compared to strategies using phenotypic values (PS); and (ii) investigate the interaction effects between different genetic architectures and CS methods on long-term genetic gain. In our simulation results, implementing GS with optimal selection intensities (Optimal-GS) had a higher short- and long-term genetic gain compared to PS strategies. The CS method considering additive and dominance effects to predict progeny mean based on simulated progenies (MEGV-O) had the highest accuracy and reached the highest long-term gain among the assessed CS methods based on solely progeny mean. Compared to UC and MEGV-O, the linear combination of usefulness criteria (UC) and genome-wide diversity (called EUCD) kept the same level of genetic gain but simultaneously kept a higher diversity as well as a certain degree of genetic variance. However, the choice of the most appropriate weight to account for diversity in EUCD depends not only on the genetic architecture of the target trait but also on the breeder's breeding objectives. Therefore, these results can provide breeders with a concrete method to improve their potato breeding programs.

1 INTRODUCTION

Potato (*Solanum tuberosum* L.) is one of the most important non-cereal crops for human consumption in the world (`http://www.fao.org/faostat/en/`). In response to a growing global population, producing sufficient food becomes a big challenge for agriculture (Fróna et al., 2019). In addition, global crop production is expected to be negatively impacted by climate change due to an increase in extreme temperatures and an alternation of rainfall patterns (Abberton et al., 2016). Thus, developing methods and approaches which increase the efficiency and effectiveness to develop new varieties with high and stable yield in potato is one of the important missions of plant geneticists.

One of the necessary steps to develop varieties requires generating new genetic variability. This can be reached via (1) introducing new alleles, for instance using genetic resource collections (Sanchez et al., 2023); and (2) creating new allelic combinations. The latter is realized by meiotic recombinations that occurs after crossing parental genotypes to create new segregating populations. The next steps after crossing are to identify and select the superior varieties in the created breeding populations, as well as to determine new cross combinations to start the next breeding cycle. In a typical clonal breeding program, these steps rely until now mostly on phenotypic records, and take several years. This is especially true in potato because the target traits can only be assessed in the later stages due to low multiplication coefficient of potato (Grüneberg et al., 2009), which in turn hampers the improvement of genetic gain. With the advent of genomic selection (GS), genetic gain can be enhanced in both livestock and crop breeding (Voss-Fels et al., 2019; Xu et al., 2020). In potato, Wu et al. (2023) have shown via a simulation study that implementing GS into one

breeding cycle can improve the short-term genetic gain of the target trait compared to using phenotypic selection (PS). While incorporating GS into breeding programs has been shown to increase long-term genetic gain in diploid crops compared to PS (Gaynor et al., 2017; Gorjanc et al., 2018; Muleta et al., 2019; Lubanga et al., 2022; Sanchez et al., 2023; Werner et al., 2023), the effects of implementing GS on the long-term genetic gain in tetraploids is still unclear. Furthermore, in tetraploid and heterozygous crops, due to their complicated quantitative genetics and the importance of dominance effects, different trends of implementing GS may be expected versus diploid, which need to be assessed.

The value of new crosses is commonly estimated by the mid-parental performance based on the phenotypic records of candidate parents (Brown and Caligari, 1989). With the advent of GS, the mid-parental performance can be replaced by the estimated genetic values (EGV) from a trained GS model, which has been proven to improve genetic gain in maize compared to the one based on phenotypes (Allier et al., 2019a; Sanchez et al., 2023). However, as GS is also a truncation selection, it is accompanied by an acceleration of the fixation of favorable alleles. This in turn leads to a quick loss of genetic variation. If the candidate parents that are intermated for creating the next generation have similar genetic backgrounds, this hinders the generation of new allelic recombinations and limits the long-term improvement of genetic gain (c.f Jannink, 2010). Therefore, maintaining diversity in the breeding populations when selecting new crosses is one possibility to improve long-term genetic gain.

Several studies have proposed approaches to balance genetic gain and diversity while determining desirable new crosses. The usefulness criterion (UC) is one of the criteria used to predict the performance of a cross (Schnell and Utz, 1975). It considers the expected progeny mean (μ) and the expected response to selection in

the progenies ($ih\sigma$): UC = $\mu + ih\sigma$, where σ is the square root of progeny variance, i the selection intensity, and h the square root of heritability. Using UC to select new crosses has been shown to increase genetic gain compared to mid-parental values in simulation studies on maize of populations (Lehermeier et al., 2017; Allier et al., 2019a; Sanchez et al., 2023). Furthermore, Zhong and Jannink (2007) made a modification of the UC, called superior progeny value: S = $\mu + i\sigma$. This focuses on progeny mean and variance but ignores heritability. However, depending on the traits, both UC and S can be close to the progeny mean as (1) the progeny standard deviation becomes very small (Rembe et al., 2022) or (2) the variation in progeny mean is much higher than the variation in progeny standard deviation (Zhong and Jannink, 2007; Lado et al., 2017). These two aspects limits the advantage of cross-selection (CS) methods – like UC and S, which consider only progeny mean and variance. Therefore, investigating different weights between progeny mean and progeny variance could affect the efficiency of using progeny variance for CS on long-term genetic gain. This, however, has not yet been studied before.

The progeny mean of a bi-parental cross can be predicted by mid-parental performance based on either phenotypic records or EGV from the trained GS model. This can be assessed for inbred populations derived from inbred parents or for hybrids and outbreds in the absence of dominance effects. With the importance of dominance effects in outbred crops, it can be estimated using a formula for diploid species (Falconer and Mackay, 1996; Wolfe et al., 2021; Werner et al., 2023). However, no formula is available for tetraploid species. Thus, using the mean of parental performance could lead to inaccurate estimations of the progeny mean and no formula for predicting progney mean in tetraploid potato is available so far. Furthermore, it is not easy to obtain a reliable prediction of progeny variance (Mohammadi et al., 2015).

With the development of dense genome-wide markers and the advent of GS models, the marker effects can be well estimated (Meuwissen et al., 2001). Recently, several formulae considering linkage disequilibrium and recombination rate in parental lines have been derived to predict the progeny variance (Bonk et al., 2016; Lehermeier et al., 2017; Osthushenrich et al., 2017; Allier et al., 2019b; Wolfe et al., 2021). However, these formulae are based on diploid with either inbred or outbred parents, which cannot be applied to tetraploid potato. The simulation of virtual progenies of a cross using a genetic map and phased parental haplotype information is an alternative approach to circumvent this (Bernardo, 2014; Mohammadi et al., 2015). Softwares for that purpose are available (e.g. AlphasimR (Gaynor et al., 2021)) and can be used for such simulation in tetraploid species. The use of the average and variance of EGV among in silico progenies to estimate progeny mean and variance could improve the prediction accuracy of progeny mean compared to mid-parental values and provide an alternative to predict progeny variance for tetraploid species with heterozygous parents. This aspect, however, has not been examined earlier.

An alternative to UC and its derived methods is the optimal cross-selection (OCS) (Gorjanc et al., 2018). The basic idea of OCS is to select a group of bi-parental crosses that maximize the expected progeny mean under a certain constraint of genetic diversity or co-ancestry on the selected population of individuals who serve as parents. Through optimization algorithms (e.g. Kinghorn, 2011), this approach has proven to increase long-term genetic gain in a simulated maize breeding program with a minor penalty on the short-term genetic gain compared to using solely UC (Allier et al., 2019a; Sanchez et al., 2023). However, it is extremely more time-consuming to find the optimal parents and crosses compared to the abovementioned truncation CS methods based on ranking the performance among all possible crosses, especially

when many markers are used in the application of the tetraploid breeding program. This limits its advantage in potato.

Another possibility to quantify diversity is based on the genome-wide variation of a cross itself rather than the variation in the whole population of parents for crosses. This can be measured by the expected heterozygosity (He). Accounting for this element during the selection of new crosses may contribute to long-term genetic gain and simultaneously preserve diversity. However, to the best of our knowledge, no earlier studies have investigated the performance of a criterion including genome-wide diversity of a cross to determine new desirable crosses.

The objectives of this study were to (i) assess how different CS methods incorporating GS with or without consideration of genetic variability affect both short- and long-term genetic gains compared to strategies using phenotypic values; and (ii) investigate the interaction effects between different genetic architectures and CS methods on the long-term genetic gain.

8

2 MATERIALS AND METHODS

2.1 Potato empirical genomic dataset

A set of 80 tetraploid potato clones genotyped for 49,125 phased sequence variants across 12 chromosomes (Baig et al. in preparation) was randomly selected from a diverse panel of 100 clones and used in this simulation study. The sequence variants, including single nucleotide polymorphism and small insertion/deletion polymorphisms, have been filtered by a minor allele frequency < 0.05 and a missing rate > 0.1 and selected from all possible variants (19,649,193) to be evenly distributed every 15 kilobases. In addition, their corresponding genetic map information was estimated using a Marey map (for details see Wu et al. (2023)).

2.2 Breeding programs and selection strategies

This simulation study was based on three main selection strategies in a potato clonal breeding program: (1) Standard-PS: a scheme following a standard potato breeding program relying exclusively on PS, which serves as benchmark; (2) Optimal-PS: a scheme relying on PS but where the optimal selection intensities during the selection process were determined to maximize genetic gain; (3) Optimal-GS: a scheme based on both PS and GS where the optimal selection intensities during the selection process were determined to maximize genetic gain.

To simulate a long-term potato breeding program, 30 sequential breeding cycles were considered. Each breeding cycle of the breeding program comprised seven main stages: cross stage (X), seedling stage (SL), single hills stage (SH), A clone stage (A), B clone stage (B), C clone stage (C), and D clone stage (D). During each breeding cycle, the selection was performed following one of the above described three selection

strategies. Then, some D clones were selected as new parents for the next breeding cycle and intercrossed to create new genetic variation. The details of the approaches used to determine new crosses are described in the next section.

In order to allow a fair comparison of performance across different selection strategies and CS methods, a consistent starting point, called burn-in cycle (C_0), was required. The procedure of the potato breeding program across 30 cycles is shown in Figure 1 and its details are described here under:

- **Burn-in cycle (C_0)**
  - Step 1: 300 crosses were randomly selected from all possible crosses in the half-diallel among the 80 parents (=3,160, called candidate crosses) and served as a crossing plan. From each cross the same number of progenies, which were in the following designed as SL progenies, were generated.
  - Step 2: Selection processes from SL to D clone stage were conducted according to the chosen selection strategy (see Figure 2 in Wu et al. (2023)).
  - Step 3: The top 20 of the 60 clones at D were selected based on T_t phenotypes and were, together with the 80 parents of C_0, considered as candidate parents for cycle 1 (C_1). Therefore, the number of candidate parents in C_1 became 100 (80+20).
- **Cycle 1 (C_1)**
  - Step 1: The performance of all possible crosses in the half-diallel among the 100 parents, excluding the 300 random crosses of C_0 (= 4,650 candidate crosses), was calculated based on the chosen CS method.
  - Step 2: Based on the calculated performance from Step 1, the top 300

158 crosses were selected as the crossing plan, and from each cross the same
159 number of progenies, which were in the following designed as SL progenies,
160 were generated.

161 – Step 3: This followed Step 2 of C_0.

162 – Step 4: This followed Step 3 of C_0 except that 20 parents were randomly
163 selected from those candidate parents which were not used in the crossing
164 plan of C_1, and were removed from the candidate parents. Therefore,
165 the number of candidate parents in the next cycle (C_2) was still 100 (80-
166 20+20).

167 • **Cycle t (C_t), where t > 1**

168 – Step 1: To reduce computational time and mimic the breeder's usage,
169 only the candidate crosses consisting of those crosses between the 80 old
170 and 20 new ones and all possible crosses in the half-diallel among the 20
171 new candidate parents were considered for C_t and their performances were
172 calculated according to the CS method.

173 – Step 2 to 4: These followed steps 2 to 4 of C_1.

174 2.3 Cross-selection (CS) methods

175 Different methods were tested to select new crosses for the next cycle. The considered
176 parameters for each cross were i) the predicted progeny mean, μ; ii) the predicted
177 progeny variance, σ^2_G; and iii) the predicted progeny diversity, $He_{per-cross}$; and (iv)
178 and the linear combinations of (i), (ii) and (iii).

179 The predicted progeny mean could be evaluated in five different ways: (i) the
180 mean phenotypic values of the two parents, MPV; (ii) the mean estimated breeding

values of the two parents, MEBV-P; (iii) the mean estimated genetic values of the two parents, MEGV-P; (iv) the mean estimated breeding values among simulated offsprings, MEBV-O; and (v) the mean estimated genetic values among simulated offsprings, MEGV-O. The last two, (iv) and (v), were estimated by the mean breeding and genetic values, respectively, among 1,000 simulated progenies of an in silico cross.

To balance selection between improvement of genetic gain and maintenance of variability measured by predicted progeny variance for the selection of new crosses, the concept of UC (Schnell and Utz, 1975) was first extended by:

$$\text{EUC}: \ \mu + w_1 \cdot i \cdot \text{PA} \cdot \sigma_G \qquad [1]$$

representing an extended usefulness criterion (EUC), in which μ was the predicted progeny mean, w_1 a weight on the square root of the progeny variance (σ_G), i the selection intensity, and PA the prediction accuracy of the GS model. Here, PA replaced the square root of heritability in the response of selection when GS was implemented (Falconer and Mackay, 1996; Heffner et al., 2010). For EUC, μ was based on MEGV-O because this outperformed the other ways of progeny mean estimation in the previous assessment. σ^2_G was estimated by the variance of genetic values T_t among 1,000 simulated progenies of an in silico cross. We varied w_1 by 1, 10, 50, and 100. If $w_1 = 1$, the equation [1] is equivalent to UC. Moreover, we assumed the selected proportion per cross as 0.1 so that i corresponds to 1.755.

In addition to the EUC criterion and to keep a certain level of genomic diversity in the breeding program, $\text{He}_{per-cross}$ was incorporated into the equation [1] to create an extended usefulness criterion incorporating genomic diversity index (EUCD) by:

$$\text{EUCD}: \ \mu + i \cdot \text{PA} \cdot \sigma_G + w_2 \cdot \text{He}_{per-cross}, \qquad [2]$$

where $\mathrm{He}_{per-cross}$ was calculated by the He among 1,000 simulated progenies of an in silico cross and weighted by w_2. Due to the tetraploid nature of potato, He was determined by:

$$\mathrm{He} = \frac{1}{m}\sum_{j=1}^{m}(1 - \sum_{i=1}^{k} p_{i(j)}^4), \quad [3]$$

where m was the number of sequence variants, k the number of alleles in one sequence variant, and $p_{j(i)}$ the allele frequency of the i^{th} allele at the j^{th} sequence variant (Gallais, 2003). We only considered bi-allelic sequence variants in this study, and therefore, k was equal to 2.

The scale of σ_G and $\mathrm{He}_{per-cross}$ and their variance differed largely. To keep the same proportion for the two measurements in the equation [1] and [2], w_2 varied by 50, 500, 2500, and 5000. The resulting four different scales of weights for EUC and EUCD as well as their abbreviations are summarized in Table 1.

2.4 Simulation of genetic architectures of traits

2.4.1 Simulated true genetic and phenotypic values

Two traits, auxiliary ($\mathrm{T_a}$) and target ($\mathrm{T_t}$) trait, were considered in this study. Here, $\mathrm{T_a}$ represented the weighted sum of the auxiliary traits measured in the first three stages of the breeding program, and $\mathrm{T_t}$ the weighted sum of all market-relevant quantitative traits. The latter was controlled by 2000 quantitative trait loci (QTL). The true genetic and phenotypic values (TGV and P) for both traits were simulated following Wu et al. (2023) except for generating dominance effects. For autotetraploids, all possible genotype classes in each QTL were aaaa, Aaaa, AAaa, AAAa, and AAAA. The deviations of genetic values from their breeding values (= cumulative the number

of additive effects) for the three heterozygous classes (Aaaa, AAaa, and AAAa) were different and expressed by d_1, d_2, and d_3, respectively (Gallais, 2003) (Table 2). The simulation of the dominance effects is introduced in the next section.

The trial environments across locations and breeding cycles were assumed to be homogeneous, and therefore the variance components of trial errors for both traits were fixed. To do so, the error variance of T_a and T_t ($\sigma^2_{\epsilon_{T_a}}$ and $\sigma^2_{\epsilon_{T_t}}$) were estimated at SL of C_0 and were then both fixed and assumed for the following breeding cycles. In detail, the ratio of variance components was set for T_t as follows: $\sigma^2_{G_{T_t}} : \sigma^2_{G_{T_t}\times L} : \sigma^2_{\epsilon_{T_t}} = 1 : 1 : 0.5$, where $\sigma^2_{G_{T_t}}$ denoted the genetic variance, and $\sigma^2_{G_{T_t}\times L}$ the variance of interaction between genotype and location; and the heritability of $H^2_{T_a}$ was fixed to 0.6. At SL of C_0, $\sigma^2_{G_{T_a}}$ and $\sigma^2_{G_{T_t}}$ were estimated by the sample variance of TGV_{T_a} and TGV_{T_t}, respectively. Then, $\sigma^2_{\epsilon_{T_t}}$ was fixed to $\frac{1}{2}$ of the estimated $\sigma^2_{G_{T_t}}$. Similarly, $\sigma^2_{\epsilon_{T_a}}$ was estimated by $\frac{1-H^2_{T_a}}{H^2_{T_a}}\sigma^2_{G_{T_a}}$. However, $\sigma^2_{G_{T_t}}$ and $\sigma^2_{G_{T_t}\times L}$ varied across breeding cycles and $\sigma^2_{G_{T_t}}$ was re-estimated at SL of each cycle. Consequently, $\sigma^2_{G_{T_t}\times L}$ was controlled by the ratio of variance components.

2.4.2 Estimated breeding and genetic values

In this study, a GS model was assumed to be trained for T_t on earlier cycles of the breeding program with a prediction accuracy PA. The estimated breeding values for T_t obtained from the GS model were estimated by $EBV_{T_t} = TBV_{T_t} + \epsilon_{PA}$, where TBV_{T_t} were the true breeding values of T_t, that is, only additive effects were considered. ϵ_{PA} was the residual value following a normal distribution $N(0, \sigma^2_{\epsilon_{PA}})$, with

$$\sigma^2_{\epsilon_{\mathrm{PA}}} = \frac{1}{n'-2}\frac{1-\mathrm{PA}^2}{\mathrm{PA}^2}\sum_{i=1}^{n'}(\mathrm{TBV}_{\mathrm{T_t(i)}} - \overline{\mathrm{TBV}}_{\mathrm{T_t}})^2 \quad [4]$$

representing the error variance determined by the level of PA, where n' was the number of genotyped clones, $\mathrm{TBV}_{\mathrm{T_t(i)}}$ the $\mathrm{TBV}_{\mathrm{T_t}}$ at the i^{th} genotyped clone, and $\overline{\mathrm{TBV}}_{\mathrm{T_t}}$ the average of $\mathrm{TBV}_{\mathrm{T_t}}$ on all genotyped clones. The estimated genetic values for $\mathrm{T_t}$ ($\mathrm{EGV}_{\mathrm{T_t}}$) were obtained by replacing all TBV appearing in this section by TGV.

2.5 Economic settings and quantitative genetic parameters

The costs for phenotypic evaluation of $\mathrm{T_a}$ and $\mathrm{T_t}$ in one environment were assumed to be 1.4 and 25 €, respectively. The costs for genotypic evaluation per clone were set to 25 €. For the Standard-PS procedure, the total budget in one breeding cycle was 677,500 €. As this strategy served as benchmark, the total budget for all other selection strategies was also fixed to this amount. In this study, we chose the selection strategy GS-SH:A as Optimal-GS (see Wu et al. (2023)), and set PA and r to 0.5 and 0.15, respectively, for all selection strategies as well as CS methods. The same number of locations and number of clones at D ($\mathrm{N_6}$=60) were set as the ones in the Standard-PS (see Wu et al. (2023)). The optimal selection proportions achieving the maximum short-term genetic gain and the number of clones at SL for each selection strategy used in this study are summarized in Table S1.

In order to investigate the interaction effects between different genetic architectures and CS methods on the long-term genetic gain, we considered four different cases of dominance degree δ for $\mathrm{T_t}$: (1) No dominance effects: δ_0 was set to 0; (2) mild dominance effects: δ_1 was produced across all QTL from $N(1,1)$; (3) moderate

dominance effects: $\delta_2 = 2 \times \delta_1$; (4) strong dominance effects: $\delta_3 = 3 \times \delta_1$. The true dominance effect at each QTL was then calculated by multiplying the true additive effect by the specific δ.

2.6 Evaluations

The genetic gain and genetic variability achieved by each scenario in each breeding cycle were evaluated and ranked. The genetic gain was defined as the difference in mean $\mathrm{TGV}_{\mathrm{T_t}}$ between D clones and the 80 selected candidate parents of C_0. The level of variability was evaluated by using the genetic variance of $\mathrm{T_t}$, and the level of genomic diversity by expected heterozygosity (He) (see equation [3]) at D clone stage. To ensure statistical significance, all results in this study were based on 30 independent simulation runs.

3 RESULTS

The mean genetic gain and genetic variance of T_t, as well as the genome-wide diversity in a long-term tetraploid potato breeding program were assessed at D clone stage considering the following parameters and their interactions: (1) different selection strategies, (2) different CS methods, (3) different genetic architectures of T_t, i.e., different degree of dominance.

Regardless of the genetic architectures of T_t and under the use of the MPV method, any selection strategy based on the optimal allocation of resources (Optimal-GS and Optimal-PS) had a higher genetic gain than the Standard-PS in both short- and long-term breeding cycles (Figure 2a). Furthermore, the Optimal-GS was superior to Optimal-PS. An increase of the cycle numbers strengthened this tendency.

Regardless of the parameters: selection strategies, CS methods, and genetic architectures of T_t, an improved genetic gain was observed with increased numbers of finished breeding cycles (Figures 2a and 3a). However, the speed of increase of the genetic gain per cycle reduced at late breeding cycles compared to early ones. This trend as well as the difference in ranking among all assessed CS methods were affected by several parameters: the degree of dominance and weights (w_1 and w_2) of the modified UC. Their details are explained below.

3.1 Comparison of CS methods that only consider progeny mean

First, we observed the effects of GS implementation on genetic gain using different CS methods only focusing on the progeny mean. In general, any progeny mean predicted by in silico progenies (MEBV-O and MEGV-O) outperformed those predicted by mid-parental performance (MPV, MEBV-P, and MEGV-P) (Figure 2a). Further-

more, the MEGV-O method was superior to the MEBV-O method. The difference between the two CS methods was more obvious as both the number of breeding cycles and the degree of dominance increased. The latter had stronger influences on genetic gain compared to the former. Interestingly, the MPV (Optimal-GS) had the highest genetic gain among CS methods based on mid-parental performance.

In contrast to the genetic gain, the genetic variance of T_t decreased as the number of breeding cycles increased (Figure 2b). This tendency increased with the reduction of the degree of dominance. Furthermore, the effects of the selection strategies and the CS methods were opposite on the genetic variance in comparison to the genetic gain. As the degree of dominance increased, larger differences and fluctuations in genetic variance among these CS methods and cross cycles were observed.

Similarly, the genome-wide diversity measured by He decreased with increasing breeding cycle numbers (Figure 2c). However, a higher degree of dominance reduced this tendency. With increasing importance of dominance effects, the CS methods considering additive and dominance effects (MPV, MEGV-P, and MEGV-O) kept a higher He than those based solely on additive effects (MEBV-P and MEGV-P), especially at late cycles. Furthermore, MEGV-O method kept the highest He among the progeny mean-base CS methods, even though it had the lowest genetic variance and the highest genetic gain. Therefore, the MEGV-O was used hereafter as a measurement for the prediction of progeny mean in the weighted methods, i.e., EUC and EUCD.

3.2 Comparison of CS methods with weights on progeny variance or genome-wide diversity

Regardless of the genetic architectures of T_t, a small or no difference on genetic gain was observed at early cycles among the following CS methods: MEGV-O, EUC and EUCD with low weights (Figure 3a). As the cycle number increased, the difference was more pronounced. On average across the four levels of dominance effects, the $EUC_{w_1=1}$ (=UC) had the highest genetic gain among all EUC (731.01 at C_{30}), as well as was superior to CS methods based only on progeny mean (MEGV-O and MPV methods) (Figure 3a and Table S2). Furthermore, the EUCD with a low weight (w_2=50 or 500) yielded the highest genetic gain (734.38 at C_{30}).

Under the assumption of the same proportion for genetic variance and He with considering the four scales (A, B, C, and D) (Table 1), the effect of the different weights for EUC and EUCD was compared for the long-term genetic gain. Regardless of the genetic architectures of T_t, a small or no difference between EUCD and EUC was observed on the genetic gain when the lowest weights for w_1 and w_2 were given, that is under Scale A (Figure 3a and Table S2). Furthermore, $EUCD_{w_2=500}$ always outperformed $EUC_{w_1=10}$ (Scale B). With high dominance effects, the EUCDs were superior to the EUCs with high weights, i.e., under Scale C and D.

The ranking and the difference in genetic gain among these CS methods were influenced by the degree of dominance (Table S2). EUC and EUCD with high weights ranked better when dominance effects increased. This was especially true for EUCD. For instance, $EUCD_{w_2=5000}$ had the worst performance under no or mild dominance effects. However, under strong dominance effects, it ranked 7th and outperformed $EUC_{w_1=50\&100}$ as well as MPV. While a slow improvement of genetic gain using

EUCD$_{w_2=2500}$ was observed under the case without dominance effects, it ranked 5th under the cases with moderate and strong dominance effects. Furthermore, the difference between this CS method and the best one decreased, especially by strong dominance effects.

EUC and EUCD with low weights reached high genetic gain but were accompanied by low genetic variance (Figure 3b and Table S2). This trend was similar to the CS methods only based on mid-parental values in the previous section. In addition, with an increase in cycle numbers, the reduction of genetic variance slowed down, especially for the case with strong dominance effects. Differently, high-weighted EUC and EUCD kept relatively high genetic variance and even increased it as the cycle number increased. Under the same proportion for genetic variance and He, EUCD had a higher genetic variance than EUC under Scale C and D, except for the case with strong dominance effects under Scale C, but EUCD reached a higher genetic gain than EUC. Furthermore, with strong dominance effects, EUCD$_{w_2=5000}$ kept the highest genetic variance. However, it still performed similarly to EUC$_{w_1=10}$ regarding to genetic gain and even had a much higher genetic gain than MPV and EUC$_{w_1=50\&100}$. Furthermore, EUC dramatically decreased He along increasing cycles (Figure 3c and Table S2), which was similar to the only mean-based CS methods. This trend was not mitigated a lot as w_1 increased, except for the scenarios with low or no dominance effects. In contrast to EUC, using EUCD obviously slowed down the decline of He. A greater w_2 increased this tendency. Under the same proportion for genetic variance and He (i.e., any scale in Table 1), EUCD always kept a higher He than EUC. Furthermore, EUCD with a low w_2 reached a higher He than EUC with a high w_1, especially for a high importance of dominance effects. Overall the genetic architectures of T_t, EUCD with a low w_2 (50 or 500) achieved a high genetic gain and still kept a higher

He than the UC and the MEGV-O method. Meanwhile, its level of genetic variance remained average. Under strong dominance effects, the genetic gain reached by the EUCD with a high w_2 (e.g. $\text{EUCD}_{w_2=2500}$) had no significant difference with the highest one reached by $\text{EUCD}_{w_2=500}$. However, the He and genetic variance achieved by $\text{EUCD}_{w_2=2500}$ were higher than the one reached by $\text{EUCD}_{w_2=500}$.

4 DISCUSSION

Different CS methods accounting or not for diversity have been evaluated in diploid crops to enhance genetic gain (Gaynor et al., 2017; Allier et al., 2019a; Werner et al., 2023). However, the effects of implementing GS and different CS methods in a long-term breeding program for tetraploid crops with a highly heterozygous genome are lacking. Because of their difference in quantitative genetics compared to diploid inbred breeding, one could expect different outcomes in such an analysis. Therefore, we evaluated the efficiency of different CS methods in long-term breeding programs under different genetic architectures via a simulation study.

4.1 The effects of different selection strategies on long-term potato breeding programs

In this study, we extended an analysis (Wu et al., 2023) considering the implementation of GS in one breeding cycle, to study its impact on the long-term genetic gain. Regardless of the genetic architectures and based on MPV as CS method, a higher genetic gain (Figure 2a) in long-term breeding programs was observed with Optimal-PS compared to the benchmark Standard-PS. This follows the trend observed in the study on short-term genetic gain (Wu et al., 2023). Our observations can be explained by that Optimal-PS had lower selection proportions at B and C clone stages (i.e., higher selection intensities, Table S1) fully based on P_{T_t} selection in comparison with the benchmark procedure, leading to a higher genetic gain according to breeder's equation (Falconer and Mackay, 1996). Furthermore, the selection strategy incorporating GS reached a higher genetic gain than PS, which can be expected because the former has a higher indirect selection response than the latter at the early

stages (Wu et al., 2023). Thus, we compared the performance among the evaluated CS methods using the selection strategy GS:SH-A hereafter.

4.2 The accuracy in predicting progeny mean

Among the examined CS methods that predict progeny mean, the ranking with respect to maximum genetic gain was MEGV-O > MEBV-O > MPV > MEGV-P / MEBV-P (Figure 2a). This trend was even more pronounced as both breeding cycle numbers and dominance effects increased. One reason might be expected that the CS methods based on an in silico cross of simulated offspring can more precisely predict progeny mean compared to mid-parental performance because the former can consider the possibilities of allelic combinations for progenies of a cross. Furthermore, a high prediction accuracy of progeny mean makes it possible to identify the cross maximizing the gain from selection. Therefore, we calculated the real progeny mean as the average of all simulated SL progenies at C_0 across 30 runs to assess the accuracy to predict progeny mean using different CS methods. The prediction accuracy was estimated as correlation between real and predicted progeny mean. The results (Table S3) were in complete agreement with our finding on the ranking of CS methods for genetic gain, leading to a higher improvement of genetic gain using the CS methods based on an in silico cross of simulated offspring compared to all CS methods based on mid-parental values (Figure 2a).

Outbred crops have a highly heterozygous genome, which is accompanied by the existence of dominance effects for quantitative traits in the phenotypes of the parental genotypes. The dominance effects in outbreds can partially transmit from parents to progenies (Gallais, 2003; Endelman et al., 2018). Therefore, taking into account dominance effects to predict progeny mean can lead to a more accurate estimate

compared to additive effects only. This was clearly observed in our results based on tetraploid potato: MEGV-O had higher accuracy in predicting progeny mean compared to MEBV-O, especially as dominance effects increased. It also provided a higher long-term genetic gain, which is in accordance with a former study (Werner et al., 2023). Werner et al. (2023) showed that the genetic gain was increased when considering both additive and dominance effects to predict cross performance using a formula in a diploid crop. However, it was in discordance with our result when comparing the parental prediction of progeny mean between MEGV-P and MEBV-P, even though MEGV-P incorporated dominance effects. One explanation can be that using MEGV-P based on parental dominance effects to capture dominance effects for progenies is incorrect, leading to low accuracy in predicting progeny mean, especially with increasing dominance effects (Table S3).

One surprising aspect was that MPV had the highest genetic gain among all CS methods based on mid-parental performance. This observation was unexpected that phenotypic records outperformed estimated values from a GS model, as well as in discordance with former studies in maize breeding programs (Allier et al., 2019a; Sanchez et al., 2023), where MEBV-P reached a higher genetic gain than MPV. One explanation of the superiority of MPV compared to MEBV-P and MEGV-P is that the heritability across the four environments in our setting (0.72 at D of C_0) used in this method was higher than the assigned PA (0.5) used in MEBV-P and MEGV-P. Therefore, according to the breeder's equation, the MPV can increase more the genetic gain than other CS methods based on mid-parental performance incorporating GS model. This was also confirmed by the higher accuracy in predicting progeny mean using MPV compared to MEBV-P and MEGV-P (Table S3).

4.3 The limitation of mean-basis CS methods

Besides genetic gain, the evaluation of genetic variance and genome-wide diversity across cycles is essential because a low genetic variations in breeding materials could limit the genetic gain in the long-term (Falconer and Mackay, 1996). As expected, both the genetic variance of T_t and He decreased with cycle numbers increase (Figure 2). This was more pronounced for the genetic variance especially with the CS methods reaching the higher genetic gain. The high accuracy in predicting progeny mean that leads to the quick accumulation of favorable alleles can be one reason for this observation. Moreover, the Bulmer effect (Bulmer, 1971), which reduces the proportion of genetic variance due to linkage disequilibrium between trait-coding polymorphisms (Grevenhof et al., 2012), can further explain this result. Focusing on mean performance only to select new crosses could lead to a plateau for the genetic gain with increasing cycle numbers, which hampers the further long-term improvement of genetic gain. Therefore, CS methods considering the maintenance of diversity while maximizing long-term genetic gain are required.

4.4 The efficiency of CS methods used to balance the improvement of genetic gain and the maintenance of diversity

Besides paying attention to high progeny mean, a high variance in progenies is of fundamental importance in response to selection of gain. The UC of a cross considers these aspects and has been used to predict the mean performance of the upper fraction of its progeny, considering the genetic variance, the heritability as well as the selection intensity. Thus, this method could improve the genetic gain compared to mean-basis CS methods, which is confirmed in our study. However, while we observed

a slightly higher genetic gain using the UC compared to the MEGV-O method (Figure 3 and Table S2), the genetic variance or He were the same for UC and MEGV-O method. Furthermore, their difference in the genetic gain was not statistically significant, which is contradictory to the results of former studies (Lehermeier et al., 2017; Sanchez et al., 2023). This could be explained by lower PA (0.5) and selection intensity (1.75) used in the present study, compared to a high heritability (1) and selection intensity (2.06) in Sanchez et al. (2023). Lehermeier et al. (2017) also showed higher heritability and selection intensity lead to a higher advantage of the UC versus other methods.

On the other hand, a large ratio of mean to square root of genetic variance can be expected to weaken the merit of UC. Rembe et al. (2022) showed a comparably higher ratio (~ 3) for the trait of ear emergence compared to the other assessed traits, which could lead to a low influence of the progeny variance on UC. This could be more pronounced in our study because the ratio was around 29. Therefore, this observation could explain the very small and even no difference between progeny mean and UC in our study. Furthermore, in our study, the variance in the progeny mean was much higher (~ 90 times) than the variance in the progeny standard deviation. This is in accordance with former studies (Zhong and Jannink, 2007; Lado et al., 2017), leading to no difference between UC and progeny mean. Thus, one way to strengthen the importance of the genetic variance in the progeny could be to weight the genetic variance or to add extra measurement to the UC.

The genetic diversity of a cross can be quantified by the genetic variance of a trait, but also on a genome-wide scale by the expected heterozygosity (He). Therefore, in addition to the weight on genetic variance of T_t, i.e., EUC, one could consider weighting He can integrate another level of diversity to balance genetic gain. This is because

the latter considers the level of whole genomic variation instead of being restricted to the variation of specific loci linked to QTL of T_t like the former. In our study, on average across the four different genetic architectures, $EUCD_{w_2=50/500}$ reached the maximum genetic gain among all assessed EUCDs and a slightly higher long-term genetic gain compared to UC (Figure 3a and Table S2). Meanwhile, $EUCD_{w_2=50/500}$ kept a certain degree of genetic variance and a slightly higher He compared to UC. This confirmed our expectation, as EUCD keeps the advantage of the UC and preserves a certain genome-wide diversity by accounting for He simultaneously, helping to efficiently convert genetic variability into long-term genetic gain.

While EUCD with a high weight kept a higher genetic variance and He along cycles, it was accompanied by a reduction of long-term genetic gain compared to EUCD with a low weight. This was not surprising because a high weight on diversity means to minimize the loss of diversity after selection. Allier et al. (2019a) had a similar approach accounting for different weights of penalty on He to balance between maximal genetic gain and minimal loss of diversity during the selection of new crosses. They also indicated that a stronger penalty on diversity reduced the improvement of genetic gain but kept higher diversity. However, this trend gradually diminished as the degree of dominance increased in our study, implying different weights should be fitted to different genetic architectures when using EUCD.

Although our proposed method EUCD cannot manage to reach a significant improvement in genetic gain compared to EUC and MEGV-O, it keeps a higher genome-wide diversity, which can balance maximal genetic gain and minimal loss of diversity in the process of selecting new crosses. Preserving diversity is very important in long-term breeding programs because it provides opportunities for breeders to promptly adjust the goals of the breeding programs in response to new requests such as changes

in climate and human usage and to develop new varieties adapted to biotic and abiotic stresses. Therefore, for the improvement of the long-term breeding program, potato breeders should choose a proper weight on He accounting to their parameters for a subsequent long-term improvement in genetic gain and nevertheless adaptability of the breeding program. In detail, to reach a high long-term genetic gain but simultaneously keep a certain diversity, $\text{EUCD}_{w_2=50/500}$ can be used for cases with no, mild, and moderate dominance effects, and $\text{EUCD}_{w_2=2500}$ for cases with strong dominance effects. However, $\text{EUCD}_{w_2=2500}$ or $\text{EUCD}_{w_2=5000}$ can be utilized if the main breeding goals are to keep maximum diversity and to reach a certain genetic gain for the cases with moderate or strong dominance effects. Therefore, the choice of the most appropriate weight on diversity in EUCD depends on not only the genetic architecture of T_t, but also the breeder's breeding objectives.

4.5 Assumptions of the present study

In this study, we assume that the parental haplotype phase is known, and therefore, the progeny variance can be predicted by in silico progenies (Bernardo, 2014; Mohammadi et al., 2015; Miller et al., 2023). However, within the current state of the art, the methodology to phase haplotype is costly (Sun et al., 2022). Thus, in current breeding programs, the possibility of estimating the progeny mean is based on mid-parent performance. In this study, MPV had a higher accuracy in predicting progeny mean compared to MEGV-P or MEBV-P because the heritability is higher than PA. However, if heritability is lower than PA, the advantage of MPV compared to MEBV-P and MEGV-P will disappear. For example, the heritability at early breeding stages is lower than the one at late breeding stages, because the former has lower experimental locations and replications than the latter. Therefore, if the

candidate parents are selected from early breeding stages, the superiority of MPV over MEBV-P or MEGV-P could diminish.

Wolfe et al. (2021) and Werner et al. (2023) predicted the progeny mean based on the formula based on allele frequencies of parents and considering additive and dominance effects from Falconer and Mackay (1996) in heterozygous diploid crops. Although Wolfe et al. (2021) showed no improvement in prediction accuracy of the progeny mean using MEGV estimated by the formula compared to MEBV estimated by mid-parental values based on the empirical breeding materials, Werner et al. (2023) indicated that the genetic gain was improved especially for the traits with the existence of dominance effects using MEGV estimated by the formula to select crosses via a simulation study. Therefore, one possibility to improve the prediction of progeny mean in future research is to develop the formula to estimate progeny mean and variance in tetraploid species. Furthermore, $\mathrm{He}_{per-cross}$ based on simulated progenies is highly correlated with $\mathrm{He}_{per-cross}$ based on parental genotypic information. Thus, the lack of information about haplotype phase does not affect the ability to quantify genome-wide diversity of a cross.

An alternative method to consider genome-wide diversity while selecting new crosses for the next breeding cycle was developed by Gorjanc et al. (2018) and Allier et al. (2019a). Their approach is called optimal cross-selection (OCS) and is based on an optimization algorithm. This approach provided a high genetic gain as well as kept a high diversity. However, this method required the optimization process to search for an optimal group of crosses, leading to extremely intensive computational calculations compared to our proposed methods–EUCD. This difference in computational time requirement is even more pronounced with an increasing number of markers and repetitions. Our study considered between 2 to 24 times more SNP and

the triple repetition numbers compared to the former studies. Thus, OCS has not been assessed yet in this study. However, the comparison of performance between the two methods still needs further research.

4.6 Conclusion

The present study demonstrated that implementing GS with optimal selection intensity per stage enhances both short- and long-term gain from selection compared to a typical tetraploid potato breeding program solely based on PS. In addition, for tetraploid and heterozygous crops, the prediction of progeny mean considering not only additive but also dominance effects (MEGV-O) is necessary. This can reach the highest prediction accuracy in progeny mean and have the highest genetic gain among all mean-based CS methods. Furthermore, combining UC and genome-wide diversity (EUCD) by a linear combination in a tetraploid potato breeding program reached the same level of long-term genetic gain. However, it simultaneously preserved a higher diversity as well as a certain degree of genetic variance compared to MEGV-O and UC. In our results, although EUCD with a low weight can reach the highest genetic gain, different genetic architectures of T_t and the breeder's breeding objectives require choosing different degree weights to achieve the high genetic gain and simultaneously preserve the diversity. These results can provide breeders with a concrete method to improve their potato breeding programs.

REFERENCES

585 Abberton, M., Batley, J., Bentley, A., Bryant, J., Cai, H., Cockram, J., Costa de
586 Oliveira, A., Cseke, L. J., Dempewolf, H., De Pace, C., Edwards, D., Gepts,
587 P., Greenland, A., Hall, A. E., Henry, R., Hori, K., Howe, G. T., Hughes, S.,
588 Humphreys, M., Lightfoot, D., Marshall, A., Mayes, S., Nguyen, H. T., Ogbon-
589 naya, F. C., Ortiz, R., Paterson, A. H., Tuberosa, R., Valliyodan, B., Varshney,
590 R. K., and Yano, M. (2016). Global agricultural intensification during climate
591 change: a role for genomics. *Plant Biotechnology Journal*, 14(4):1095–1098.

592 Allier, A., Lehermeier, C., Charcosset, A., Moreau, L., and Teyssèdre, S. (2019a). Im-
593 proving short-and long-term genetic gain by accounting for within-family variance
594 in optimal cross-selection. *Frontiers in Genetics*, 10:1006.

595 Allier, A., Moreau, L., Charcosset, A., Teyssèdre, S., and Lehermeier, C.
596 (2019b). Usefulness criterion and post-selection parental contributions in
597 multi-parental crosses: application to polygenic trait introgression. *G3:*
598 *Genes—Genomes—Genetics*, 9(5):1469–1479.

599 Bernardo, R. (2014). Genomewide selection of parental inbreds: classes of loci and
600 virtual biparental populations. *Crop Science*, 54(6):2586–2595.

601 Bonk, S., Reichelt, M., Teuscher, F., Segelke, D., and Reinsch, N. (2016). Mendelian
602 sampling covariability of marker effects and genetic values. *Genetics Selection*
603 *Evolution*, 48(1):1–11.

604 Brown, J. and Caligari, P. D. (1989). Cross prediction in a potato breeding pro-
605 gramme by evaluation of parental material. *Theoretical and Applied Genetics*,
606 77(2):246–252.

607 Bulmer, M. G. (1971). The effect of selection on genetic variability. *The American*
608 *Naturalist*, 105:201–211.

609 Endelman, J. B., Carley, C. A., Bethke, P. C., Coombs, J. J., Clough, M. E., da Silva,
610 W. L., Jong, W. S. D., Douches, D. S., Frederick, C. M., Haynes, K. G., Holm,
611 D. G., Miller, J. C., Muñoz, P. R., Navarro, F. M., Novy, R. G., Palta, J. P., Porter,
612 G. A., Rak, K. T., Sathuvalli, V. R., Thompson, A. L., and Yencho, G. C. (2018).
613 Genetic variance partitioning and genome-wide prediction with allele dosage infor-
614 mation in autotetraploid potato. *Genetics*, 209:77–87.

615 Falconer, D. S. and Mackay, T. F. C. (1996). *Introduction to quantitative genetics*.
616 Longman group, Essex, UK, 4 edition.

617 Fróna, D., Szenderák, J., and Harangi-Rákos, M. (2019). The challenge of feeding
618 the world. *Sustainability (Switzerland)*, 11(20):5816.

619 Gallais, A. (2003). *Quantitative genetics and breeding methods in autopolyploid plants*.
620 Institut national de la recherche agronomique.

621 Gaynor, R. C., Gorjanc, G., Bentley, A. R., Ober, E. S., Howell, P., Jackson, R.,
622 Mackay, I. J., and Hickey, J. M. (2017). A two-part strategy for using genomic
623 selection to develop inbred lines. *Crop Science*, 57(5):2372–2386.

624 Gaynor, R. C., Gorjanc, G., and Hickey, J. M. (2021). AlphaSimR: an R package for
625 breeding program simulations. *G3: Genes—Genomes—Genetics*, 11(2).

626 Gorjanc, G., Gaynor, R. C., and Hickey, J. M. (2018). Optimal cross selection for
627 long-term genetic gain in two-part programs with rapid recurrent genomic selection.
628 *Theoretical and Applied Genetics*, 131(9):1953–1966.

629 Grevenhof, E. M. V., Arendonk, J. A. V., and Bijma, P. (2012). Response to genomic
630 selection: the bulmer effect and the potential of genomic selection when the number
631 of phenotypic records is limiting. *Genetics Selection Evolution*, 44:1–10.

632 Grüneberg, W., Mwanga, R., Andrade, M., and Espinoza, J. (2009). Selection meth-
633 ods. Part 5: breeding clonally propagated crops. In Ceccarelli, S., Guimarães,
634 E. P., and Weltzien, E., editors, *Plant Breeding and Farmer Participation*, pages
635 275–322. Food and Agriculture Organization of the United Nations (FAO).

636 Heffner, E. L., Lorenz, A. J., Jannink, J. L., and Sorrells, M. E. (2010). Plant breeding
637 with genomic selection: gain per unit time and cost. *Crop Science*, 50(5):1681–
638 1690.

639 Jannink, J. L. (2010). Dynamics of long-term genomic selection. *Genetics Selection*
640 *Evolution*, 42:1–11.

641 Kinghorn, B. P. (2011). An algorithm for efficient constrained mate selection. *Ge-*
642 *netics Selection Evolution*, 43(1):1–9.

643 Lado, B., Battenfield, S., Guzmán, C., Quincke, M., Singh, R. P., Dreisigacker, S.,
644 Peña, R. J., Fritz, A., Silva, P., Poland, J., and Gutiérrez, L. (2017). Strategies for
645 selecting crosses using genomic prediction in two wheat breeding programs. *The*
646 *Plant Genome*, 10.

647 Lehermeier, C., Teyssèdre, S., and Schön, C. C. (2017). Genetic gain increases by
648 applying the usefulness criterion with improved variance prediction in selection of
649 crosses. *Genetics*, 207(4):1651–1661.

650 Lubanga, N., Massawe, F., Mayes, S., Gorjanc, G., and Bančič, J. (2022). Genomic

selection strategies to increase genetic gain in tea breeding programs. *The Plant Genome*, page e20282.

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

Miller, M. J., Song, Q., Fallen, B., and Li, Z. (2023). Genomic prediction of optimal cross combinations to accelerate genetic improvement of soybean (glycine max). *Frontiers in Plant Science*, 14:1–12.

Mohammadi, M., Tiede, T., and Smith, K. P. (2015). PopVar: a Genome-Wide procedure for predicting genetic variance and correlated response in biparental breeding populations. *Crop Science*, 55(5):2068–2077.

Muleta, K. T., Pressoir, G., and Morris, G. P. (2019). Optimizing genomic selection for a sorghum breeding program in haiti: a simulation study. *G3: Genes—Genomes—Genetics*, 9:391–401.

Osthushenrich, T., Frisch, M., and Herzog, E. (2017). Genomic selection of crossing partners on basis of the expected mean and variance of their derived lines. *PLOS ONE*, 12(12):e0188839.

Rembe, M., Zhao, Y., Wendler, N., Oldach, K., Korzun, V., and Reif, J. C. (2022). The potential of genome-wide prediction to support parental selection, evaluated with sata from a commercial barley breeding program. *Plants*, 11:2564.

Sanchez, D., Sadoun, S. B., Mary-Huard, T., Allier, A., Moreau, L., and Charcosset, A. (2023). Improving the use of plant genetic resources to sustain breeding programs' efficiency. *Proceedings of the National Academy of Sciences of the United States of America*, 120(14):e2205780119.

Schnell, F. and Utz, H. (1975). F1 Leistung und Elternwahl in der Zuchtung von Selbstbefruchtern. In *Ber Arbeitstag Arbeitsgem Saatzuchtleiter. Gumpenstein, Österreich*, pages 243–248.

Sun, H., Jiao, W. B., Krause, K., Campoy, J. A., Goel, M., Folz-Donahue, K., Kukat, C., Huettel, B., and Schneeberger, K. (2022). Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nature Genetics*, 54:342–348.

Voss-Fels, K. P., Cooper, M., and Hayes, B. J. (2019). Accelerating crop genetic gains with genomic selection. *Theoretical and Applied Genetics*, 132(3):669–686.

Werner, C. R., Gaynor, R. C., Sargent, D. J., Lillo, A., Gorjanc, G., and Hickey, J. M. (2023). Genomic selection strategies for clonally propagated crops. *Theoretical and Applied Genetics*, 136(4):1–17.

Wolfe, M. D., Chan, A. W., Kulakow, P., Rabbi, I., and Jannink, J. L. (2021). Genomic mating in outbred species: predicting cross usefulness with additive and total genetic covariance matrices. *Genetics*, 219(3).

Wu, P.-Y., Stich, B., Renner, J., Muders, K., Prigge, V., and van Inghelandt, D. (2023). Optimal implementation of genomic selection in clone breeding programs—Exemplified in potato: I. Effect of selection strategy, implementation stage, and selection intensity on short-term genetic gain. *The Plant Genome*, page e20327.

Xu, Y., Liu, X., Fu, J., Wang, H., Wang, J., Huang, C., Prasanna, B. M., Olsen, M. S., Wang, G., and Zhang, A. (2020). Enhancing genetic gain through genomic selection: from livestock to plants. *Plant Communications*, 1(1):100005.

696 Zhong, S. and Jannink, J. L. (2007). Using quantitative trait loci results to discrim-
697 inate among crosses on the basis of their progeny mean and variance. *Genetics*,
698 177(1):567–576.

Table 1: Overview of the different weight (w_1 and w_2) scales for the extended usefulness criterion (EUC) and extended usefulness criterion incorporating genomic diversity index (EUCD).

Criterion	Cross-selection methods (w_1, w_2)			
	Scale A	Scale B	Scale C	Scale D
EUC	$EUC_{w_1=1}$ (1,0)	$EUC_{w_1=10}$ (10,0)	$EUC_{w_1=50}$ (50,0)	$EUC_{w_1=100}$ (100,0)
EUCD	$EUCD_{w_2=50}$ (1,50)	$EUCD_{w_2=500}$ (1,500)	$EUCD_{w_2=2500}$ (1,2500)	$EUCD_{w_2=5000}$ (1,5000)

Table 2: Summary of the five genotype classes, including their coding expression, additive and dominance effects, as well as breeding and genetic values.

Genotype class	Additive effect (a)	Dominance effect			Breeding value	Genetic value
		d_1	d_2	d_3		
aaaa	0	0	0	0	0	0
Aaaa	1	1	0	0	a	a+d_1
AAaa	2	0	1	0	2a	2a+d_2
AAAa	3	0	0	1	3a	3a+d_3
AAAA	4	0	0	0	4a	4a

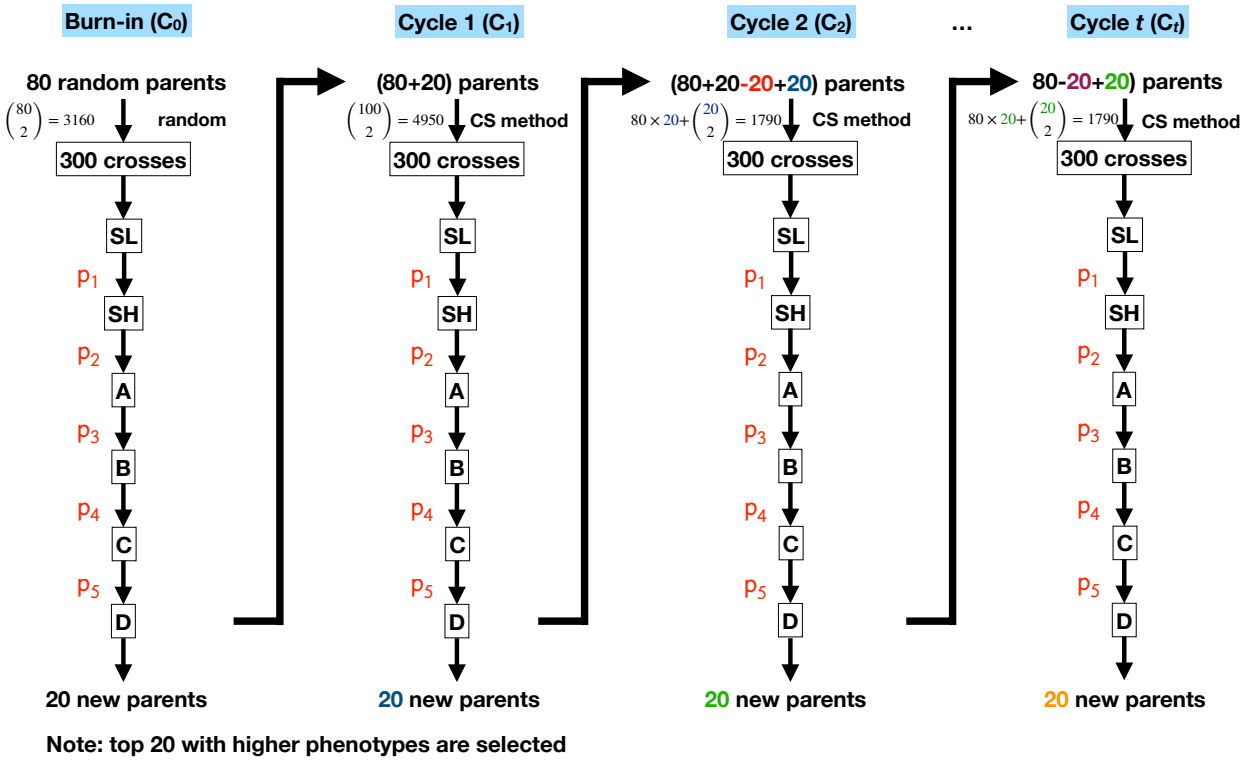

Figure 1: Graphical illustration of recurrent selection in a potato breeding program with the chosen cross-selection (CS) method to determine new crosses. Each breeding cycle of the breeding program comprised seven main stages: cross stage where 300 crosses are selected, seedling stage (SL), single hills stage (SH), A clone stage (A), B clone stage (B), C clone stage (C), and D clone stage (D). p_1 to p_2 are selection proportions at each selection stage. Their exact values for each selection strategy are shown in Table S1. The details about conducting the selection strategies in each breeding cycle are shown in Wu et al. (2023).

37

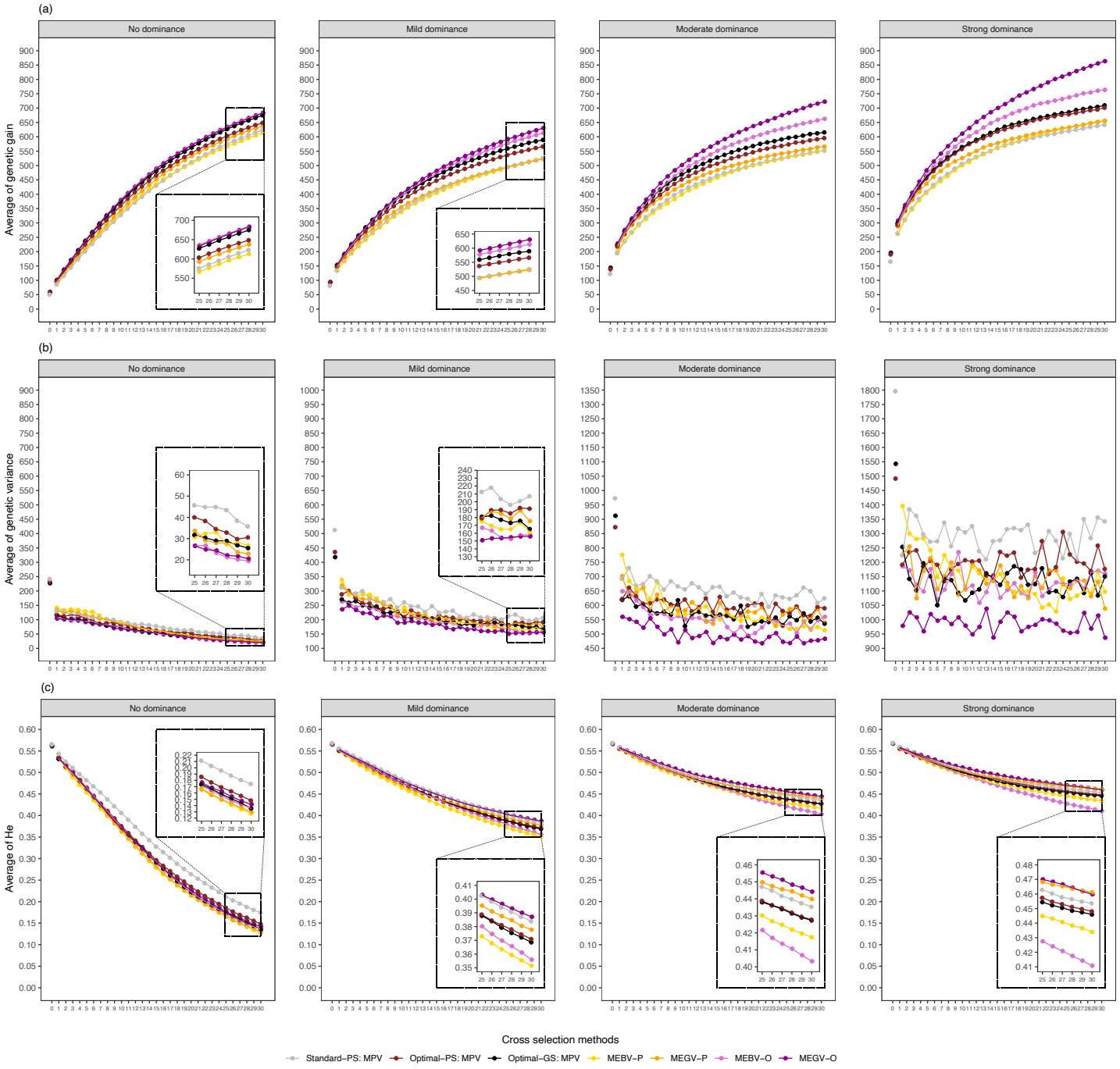

Figure 2: The evolution of genetic gain (a) and genetic variance (b) for the target trait, as well as (c) genome-wide diversity measured by expected heterozygosity (He), along the 30 breeding cycles on average across 30 simulation runs. Measures were done at D clone stage for different selection strategies (Standard-PS, Optimal-PS, and Optimal-GS), different mean-basis cross-selection methods (MPV, MEBV-P, MEGV-P, MEBV-O, and MEGV-O), and different genetic architectures of the target trait (no, mild, moderate, and strong dominance effects).

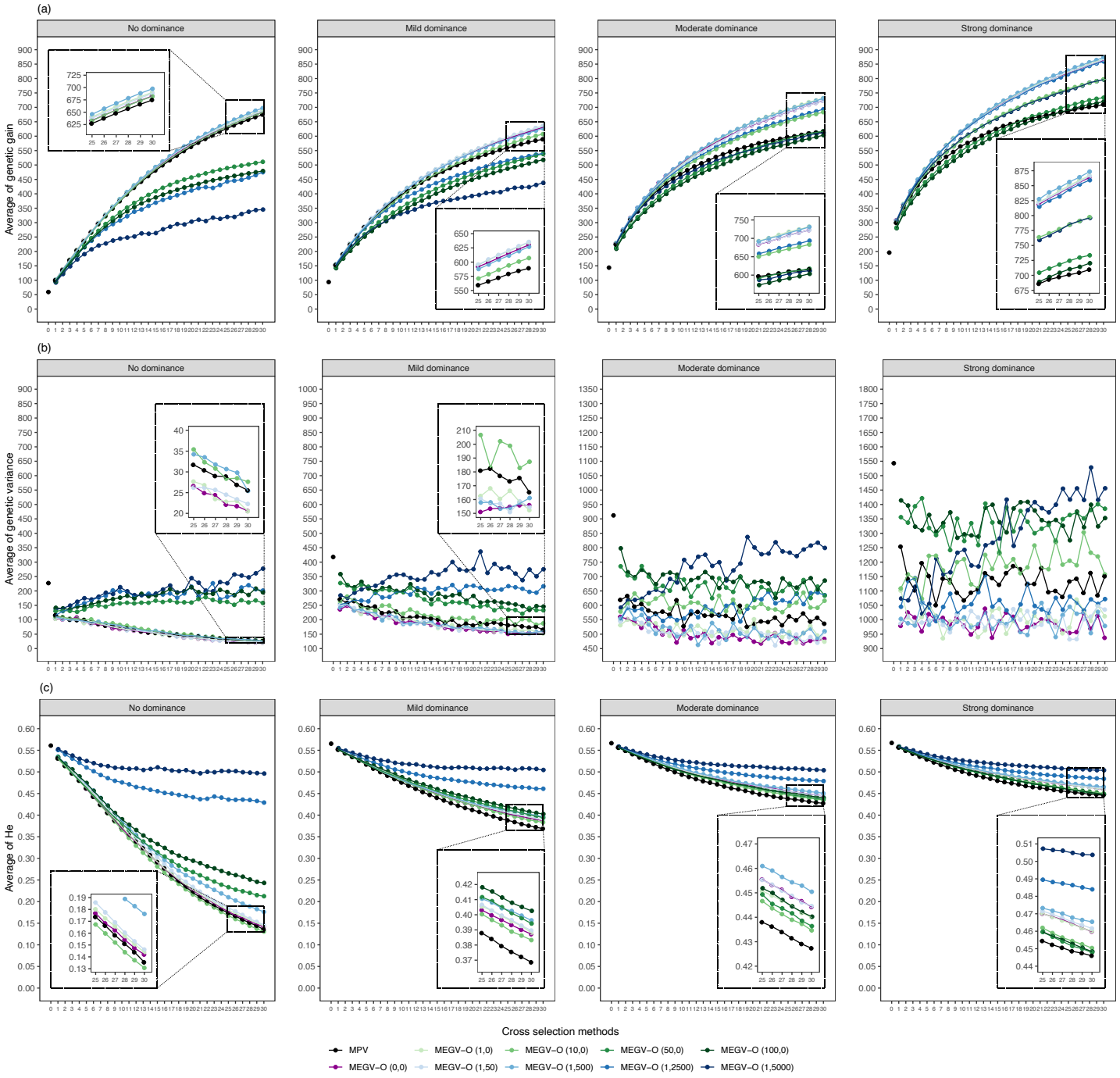

Figure 3: The evolution of genetic gain (a) and genetic variance (b) for the target trait, as well as (c) genome-wide diversity measured by expected heterozygosity (He), along the 30 breeding cycles on average across 30 simulation runs. Measures were done at D clone stage based on the Optimal-GS selection strategy for different cross-selection methods modified by usefulness criteria (EUC and EUCD), and different genetic architectures of the target trait (no, mild, moderate, and strong dominance effects). The details of EUC and EUCD are shown in Table 1.

SUPPLEMENTARY MATERIAL

Table S1: The summary of the optimal selected proportions achieving the maximum short-term genetic gain for the three selection strategies (Standard-PS, Optimal-PS, and Optimal-GS-SH:A). p_1 to p_5, α_k, and N_1 are the selected proportion per stage, the weight of genomic selection relative to phenotypic selection, and the number of clones at seedling stage (see Wu et al. (2023)), respectively.

Selection strategy	p_1	p_2	p_3	p_4	p_5	α_k	N_1
Standard-PS	∼0.33	0.1	0.15	0.2	0.2	-	300,000
Optimal-PS	∼ 0.16	0.45	0.45	0.10	0.10	-	190,252
Optimal-GS-SH:A	∼0.15	0.45	0.45	0.10	0.20	0.90	96,831

41

Table S2: The mean and standard deviation (sd) of the genetic gain and the genetic variance for the target trait, as well as genome-wide diversity measured by expected heterozygosity (He) at cycle 30 across 30 simulation runs. Simulations were based on the Optimal-GS selection strategy for different cross-selection methods modified by usefulness criteria (EUC and EUCD), and different genetic architectures of the target trait (no, mild, moderate, and strong dominance effects). The details of EUC and EUCD are shown in Table 1.

Case	Scale	Cross-selection method		Genetic gain	Genetic variance	He
				mean | sd | rank[1] | group[2]	mean | sd | rank[1]	mean | sd | rank[1]
No dominance	-	Optimal-GS: MPV	mean-base	674.86 | 45.38 | 6 | b	25.54 | 7.17 | 6	0.1354 | 0.0236 | 9
		MEGV-O		682.91 | 36.83 | 4 | ab	20.55 | 7.12 | 9	0.1419 | 0.0211 | 8
	A	MEGV-O (1,0)	EUC	690.42 | 34.25 | 2 | ab	20.50 | 6.79 | 10	0.1438 | 0.0241 | 7
		MEGV-O (1,50)	EUCD	688.26 | 37.24 | 3 | ab	22.27 | 5.59 | 8	0.1462 | 0.0241 | 6
	B	MEGV-O (10,0)	EUC	682.03 | 35.41 | 5 | ab	27.62 | 7.19 | 5	0.1307 | 0.0216 | 10
		MEGV-O (1,500)	EUCD	697.52 | 35.45 | 1 | a	25.41 | 5.29 | 7	0.1762 | 0.0193 | 5
	C	MEGV-O (50,0)	EUC	510.83 | 41.30 | 7 | c	158.38 | 37.96 | 4	0.2125 | 0.0255 | 4
		MEGV-O (1,2500)	EUCD	474.79 | 32.86 | 9 | d	201.18 | 59.78 | 2	0.4291 | 0.0106 | 2
	D	MEGV-O (100,0)	EUC	479.07 | 31.80 | 8 | d	195.13 | 45.51 | 3	0.2432 | 0.0287 | 3
		MEGV-O (1,5000)	EUCD	345.34 | 29.90 | 10 | e	277.26 | 152.84 | 1	0.4962 | 0.0107 | 1
Mild dominance	-	Optimal-GS: MPV	mean-base	589.13 | 24.70 | 6 | b	165.21 | 29.62 | 6	0.3686 | 0.0146 | 6
		MEGV-O		630.46 | 21.73 | 3 | a	155.91 | 32.38 | 8	0.3871 | 0.0120 | 8
	A	MEGV-O (1,0)	EUC	634.86 | 20.86 | 2 | a	152.32 | 29.24 | 10	0.3880 | 0.0132 | 10
		MEGV-O (1,50)	EUCD	635.53 | 18.70 | 1 | a	154.53 | 28.70 | 9	0.3896 | 0.0107 | 9
	B	MEGV-O (10,0)	EUC	607.07 | 24.09 | 5 | b	187.34 | 32.32 | 5	0.3833 | 0.0130 | 5
		MEGV-O (1,500)	EUCD	627.08 | 18.74 | 4 | a	161.12 | 36.10 | 7	0.3963 | 0.0103 | 7
	C	MEGV-O (50,0)	EUC	538.28 | 19.13 | 8 | c	233.10 | 46.12 | 4	0.3942 | 0.0122 | 4
		MEGV-O (1,2500)	EUCD	539.83 | 21.64 | 7 | c	316.13 | 86.45 | 2	0.4611 | 0.0075 | 2
	D	MEGV-O (100,0)	EUC	517.75 | 26.58 | 9 | d	245.84 | 39.13 | 3	0.4026 | 0.0147 | 3
		MEGV-O (1,5000)	EUCD	437.93 | 21.82 | 10 | e	375.27 | 101.12 | 1	0.5045 | 0.0096 | 1

Case	Scale	Cross-selection method		Genetic gain	Genetic variance	He
				mean | sd | rank[1] | group[2]	mean | sd | rank[1]	mean | sd | rank[1]
Moderate dominance	-	Optimal-GS: MPV	mean-base	615.59 | 29.91 | 8 | cd	536.13 | 110.22 | 6	0.4273 | 0.0102 | 10
		MEGV-O		722.54 | 25.61 | 3 | a	483.54 | 97.25 | 9	0.4442 | 0.0087 | 5
	A	MEGV-O (1,0)	EUC	731.97 | 24.21 | 1 | a	473.95 | 96.69 | 10	0.4400 | 0.0072 | 7
		MEGV-O (1,50)	EUCD	722.47 | 26.83 | 4 | a	508.45 | 98.52 | 8	0.4446 | 0.0088 | 4
	B	MEGV-O (10,0)	EUC	683.29 | 21.74 | 6 | b	616.05 | 115.14 | 5	0.4346 | 0.0096 | 9
		MEGV-O (1,500)	EUCD	731.00 | 25.33 | 2 | a	510.10 | 108.94 | 7	0.4504 | 0.0086 | 3
	C	MEGV-O (50,0)	EUC	617.29 | 29.37 | 7 | c	633.84 | 104.73 | 4	0.4364 | 0.0110 | 8
		MEGV-O (1,2500)	EUCD	693.48 | 28.76 | 5 | b	635.23 | 120.92 | 3	0.4788 | 0.0054 | 2
	D	MEGV-O (100,0)	EUC	603.47 | 27.35 | 10 | d	685.91 | 188.11 | 2	0.4402 | 0.0097 | 6
		MEGV-O (1,5000)	EUCD	611.82 | 27.36 | 9 | cd	799.24 | 118.21 | 1	0.5038 | 0.0060 | 1
Strong dominance	-	Optimal-GS: MPV	mean-base	709.57 | 29.15 | 10 | d	1150.66 | 223.20 | 5	0.4459 | 0.0089 | 10
		MEGV-O		863.63 | 37.04 | 4 | a	937.20 | 192.78 | 10	0.4598 | 0.0082 | 6
	A	MEGV-O (1,0)	EUC	866.76 | 34.68 | 2 | a	1012.18 | 189.86 | 8	0.4601 | 0.0074 | 5
		MEGV-O (1,50)	EUCD	866.11 | 25.14 | 3 | a	1035.21 | 240.35 | 7	0.4616 | 0.0092 | 4
	B	MEGV-O (10,0)	EUC	797.36 | 33.52 | 6 | b	1157.18 | 237.24 | 4	0.4504 | 0.0065 | 7
		MEGV-O (1,500)	EUCD	873.48 | 28.85 | 1 | a	978.41 | 199.76 | 9	0.4654 | 0.0067 | 3
	C	MEGV-O (50,0)	EUC	733.52 | 41.32 | 8 | c	1385.26 | 240.45 | 2	0.4479 | 0.0085 | 9
		MEGV-O (1,2500)	EUCD	859.17 | 36.96 | 5 | a	1071.70 | 230.82 | 6	0.4840 | 0.0054 | 2
	D	MEGV-O (100,0)	EUC	720.28 | 36.26 | 9 | cd	1352.71 | 289.42 | 3	0.4484 | 0.0107 | 8
		MEGV-O (1,5000)	EUCD	796.02 | 25.44 | 7 | b	1456.23 | 291.13 | 1	0.5037 | 0.0047 | 1

[1] The number after the sd represent the rank across these cross-selection methods within a specific genetic architecture.

[2] The letters after the rank represent the significance groups ($P < 0.05$) across these cross-selection methods within a specific genetic architecture.

Table S3: Accuracy to predict progeny mean using the different mean-basis cross-selection methods under different genetic architectures of the target trait. This was calculated as the correlation between predicted progeny mean and real progeny mean at SL of C_0 with an average across 30 simulation runs.

	No dominance	Mild dominance	Moderate dominance	Strong dominance
MPV	0.85189	0.81577	0.75472	0.70474
MEBV-P	0.49619	0.48285	0.44988	0.41854
MEGV-P	0.49617	0.47554	0.43872	0.40755
MEBV-O	0.99962	0.96148	0.88994	0.82049
MEGV-O	0.99962	0.99913	0.99840	0.99783

5 Improvement of prediction ability by integrating multi-omic datasets in barley

This manuscript was published in BMC geomics in March, 2022.

Authors:
Po-Ya Wu, Benjamin Stich, Marius Weisweiler, Asis Shrestha, Alexander Erban, Philipp Westhoff, and Delphine van Inghelandt.

Own contribution: First author. I performed the data analyses and wrote the manuscript.

Wu *et al. BMC Genomics* (2022) 23:200
https://doi.org/10.1186/s12864-022-08337-7

BMC Genomics

RESEARCH Open Access

Improvement of prediction ability by integrating multi-omic datasets in barley

Po-Ya Wu[1], Benjamin Stich[1,2], Marius Weisweiler[1], Asis Shrestha[1], Alexander Erban[3], Philipp Westhoff[2,4] and Delphine Van Inghelandt[1*]

Abstract

Background: Genomic prediction (GP) based on single nucleotide polymorphisms (SNP) has become a broadly used tool to increase the gain of selection in plant breeding. However, using predictors that are biologically closer to the phenotypes such as transcriptome and metabolome may increase the prediction ability in GP. The objectives of this study were to (i) assess the prediction ability for three yield-related phenotypic traits using different omic datasets as single predictors compared to a SNP array, where these omic datasets included different types of sequence variants (full-SV, deleterious-dSV, and tolerant-tSV), different types of transcriptome (expression presence/absence variation-ePAV, gene expression-GE, and transcript expression-TE) sampled from two tissues, leaf and seedling, and metabolites (M); (ii) investigate the improvement in prediction ability when combining multiple omic datasets information to predict phenotypic variation in barley breeding programs; (iii) explore the predictive performance when using SV, GE, and ePAV from simulated 3'end mRNA sequencing of different lengths as predictors.

Results: The prediction ability from genomic best linear unbiased prediction (GBLUP) for the three traits using dSV information was higher than when using tSV, all SV information, or the SNP array. Any predictors from the transcriptome (GE, TE, as well as ePAV) and metabolome provided higher prediction abilities compared to the SNP array and SV on average across the three traits. In addition, some (di)-similarity existed between different omic datasets, and therefore provided complementary biological perspectives to phenotypic variation. Optimal combining the information of dSV, TE, ePAV, as well as metabolites into GP models could improve the prediction ability over that of the single predictors alone.

Conclusions: The use of integrated omic datasets in GP model is highly recommended. Furthermore, we evaluated a cost-effective approach generating 3'end mRNA sequencing with transcriptome data extracted from seedling without losing prediction ability in comparison to the full-length mRNA sequencing, paving the path for the use of such prediction methods in commercial breeding programs.

Keywords: Barley, Deleterious SV, Transcriptome, Metabolome, Genomic prediction, Omic prediction

Background

Barley (*Hordeum vulgare* L.) is the fourth most important cereal crop in the world (FAOSTAT, http://www.fao.org/faostat/en/) and is used for human nutrition and animal feed [1]. In the context of a growing global population [2], producing sufficient food is a big challenge for agriculture [3]. In addition, climate change is expected to negatively impact global crop production by increasing extreme temperatures and altering rainfall patterns [4]. Thus, high and stable yield in barley is one of the most important breeding goals. However, in addition to directly breeding for yield, the consideration of yield-related characters during the breeding processes proved successful [5]. Leaf angle (LA) e.g. is one of the most important canopy architecture

*Correspondence: inghelan@hhu.de
[1]Institute of Quantitative Genetics and Genomics of Plants, Heinrich Heine University, 40225 Düsseldorf, Germany
Full list of author information is available at the end of the article

parameters that influence the efficiency of photosynthesis and further affect yield production [6]. In addition, the control of plant height (PH) can be used to reduce yield loss arising from lodging and adaption to variable environments through heading time (HT) alteration impacts yield [7]. Therefore, the use of approaches that help breeders to reliably select for yield and yield-related quantitative traits increases the gain of selection.

Genomic prediction (GP) has emerged as a powerful tool to increase selection gain for complex traits in both livestock and plant breeding programs [8, 9]. This method is based on the idea that the performance of individuals can be predicted from genotypic information using the GP model which was trained on those individuals with both phenotypic and genotypic information. Thus, the genotyped individuals can be preselected before their phenotypes are measured in the field to shorten the breeding cycle as well as to reduce the cost of phenotyping [10].

Typically, single nucleotide polymorphisms (SNP) serve as predictors in GP [11–13]. SNP in gene coding regions can be classified into non-synonymous (nsSNP) and synonymous SNP (sSNP), which differ in their property to change or not the amino acid sequence of a protein. Therefore, these two SNP classes may have different influence on phenotypes. In previous studies, the advantage of using these classes of SNP in comparison to randomly selected SNP for GP was explored in pig [14]. However, they observed that predictive performance of neither nsSNP nor sSNP did significantly differ from those of random SNP for most traits. In addition, Heidaritabar et al. [15] observed that nsSNP did not enhance the performance of GP in chicken. On the other hand, a protein may be able to tolerate an amino acid change due to a nsSNP and still keep its function normal [16]. Therefore, SNP can be grouped using the SIFT algorithm [17] into (1) tolerant SNP (tSNP), which can keep a protein's function normal; and (2) deleterious SNP (dSNP), which will affect a protein's function. To the best of our knowledge, the use of tSNP or dSNP as predictor of the phenotypic variation has not yet been compared.

Complex biological processes such as transcription, translation, and biochemical cascades resulting in various metabolites occur between DNA sequence and phenotypes [11], which hamper the predictive power of SNP. In addition, higher-order epistatic effects may contribute to the genetic variance of complex traits [18], which can in most of cases not directly be captured using SNP information [13, 19]. Therefore, prediction ability of phenotypic variation using SNP information for quantitative traits still leaves room for improvement. In the last years, molecular technologies were developed, which allow a cheap and high-throughput gene expression and metabolite profiling [20]. Such data can act as bridge to shorten the biological distance between genotypes and phenotypes and may even capture higher-order epistatic interactions for the prediction of phenotypic variation [21, 22].

Transcription is the first downstream processes after the DNA sequence and, thus, more likely affects the variation of traits compared to SNP. Recently, thanks to technological developments, several studies have proposed to use gene expression (GE) variation as predictor of phenotypic variation in maize [11, 21], rice [22] and barley [23]. While Schrag et al. [21] and Hu et al. [22] used GE assessed from microarray experiments for GP and showed that a considerable proportion of phenotypic variation can be explained by such information, Guo et al. [11] and Weisweiler et al. [23] used mRNA sequencing datasets to predict the performance of phenotypic traits. The advantage of mRNA sequencing compared to microarray experiment is the possibility to extract SNP and small insertions/deletions (INDEL) called sequence variants (SV hereafter), in addition to the quantification of transcript abundance. Furthermore, a single gene can often produce more than one transcript through alternative splicing, which can generate various proteins to regulate the complexity of pathways [24]. These different transcripts of the same gene can be identified using full-length mRNA sequencing. To our knowledge, transcript expression (TE) as predictor in GP has not yet been compared to GE.

Compared to the two previous levels of molecular information (DNA sequence and GE), metabolites (M) have the closest relationship to the expressed phenotype because they are the end-points of upstream biochemical processes [25], and, thus, have a high potential as predictors for GP. Previous studies on the use of metabolites to predict phenotypic traits in *Arabidopsis thaliana*, maize, wheat, and barley reported lower or higher prediction abilities compared to SNP information, depending on the traits and species [11, 21, 26–29]. Gemmer et al. [29] recommended that metabolites cannot be used alone in barley for phenotype prediction. However, the integration of expression and metabolite datasets with SNP information improved prediction abilities in comparison to the benchmark using SNP information in maize [11, 21]. Thus, the integration of several layers of omic datasets such as SV, GE, TE, and M as predictors could outperform benchmark methods and should be evaluated in GP of phenotypic traits in barley.

The objectives of our study were to (i) assess the prediction ability for three yield-related phenotypic traits (LA, PH, and HT) using different omic datasets as single predictors compared to a SNP array, where these omic datasets included different types of sequence variants (SV, dSV, and tSV), different types of transcriptome (expression presence/absence variation-ePAV, GE, and TE) sampled from two tissues, leaf and seedling, and metabolites (M); (ii) investigate the improvement in prediction ability

when combining multiple omic datasets information to predict phenotypic variation in barley breeding programs; (iii) explore the predictive performance when using SV, GE, and ePAV from simulated 3'end mRNA sequencing of different lengths as predictors.

Results

Heritability

The three phenotypic traits (LA, PH, and HT) were measured for 23 spring barley inbreds in seven environments. The adjusted entry means of the 23 inbreds ranged from 2.52 to 7.07 for LA, 48.75 to 79.75 cm for PH, and 57.31 to 82.23 days for HT (Suppl. Table S1). Heritabilities on an entry mean basis (H^2) were high and similar for LA (0.91) and HT (0.90) and with 0.83 slightly lower for PH. A total of 192 chemical entities were annotated (Suppl. Table S2) and after filtering (see methods), 144 metabolites remained for which the relative abundances were used for further analyses. A total of 101 metabolites were found in databases and, thus, it was possible to assign them according to their chemical features to 12 compound classes, while the remaining 43 metabolites were unknown (Suppl. Table S3). The heritabilities of the metabolites on an entry mean basis ranged from 0 to 0.98 with an average of 0.62 (Suppl. Fig. S1). The classification of the metabolic predictors using different degrees of heritability (0.1 to 0.8 in increments of 0.1) resulted in eight groups with 133, 128, 121, 117, 109, 93, 72 and 45 metabolites, respectively. These groups were then considered for the omic prediction described below.

Correlation and genetic dissimilarity analyses

Positive correlations between the three phenotypic traits were observed (Suppl. Fig. S2). Particularly, LA was highly and significantly correlated with HT (0.685***), where the correlation coefficients between PH and HT as well as between PH and LA were with about 0.45 considerably lower. Many metabolites were significantly ($P < 0.05$) negatively associated with the assessed phenotypic traits (Fig. 1). For instance, a cluster of some acids, amino acids, and several unknown metabolites was strongly negatively correlated with the three traits. Interestingly, we found that the same metabolites that were significantly correlated with LA were also correlated with HT. This was consistent with the phenotypic correlations between both traits (Fig. 1 and Suppl. Fig. S2).

To assess similarity/dissimilarity between these omic datasets, we performed generalized procrustes analysis (GPA) [30] on the resulting principal component analysis (PCA) obtained from each omic dataset. The dissimilarity measurements from GPA were used for principal coordinates analysis (PCoA). The first two PCo accounted for 71.86% and 20.72% of the total variability, respectively (Fig. 2). The first PCo separated the metabolites from the other features while the second PCo tended to differentiate the two tissues, leaf (*l*) and seedling (*s*). GE, TE, and ePAV datasets were similar to each other within the same tissue. This can be explained thereby that the ePAV dataset was derived from GE dataset and the GE dataset was derived from the TE dataset. $ePAV_{ls}$ was, as expected, centered between the ePAV from the individual tissues. Although SNP array, SV, dSV, and tSV clustered together, SNP array was more distant from the cluster of dSV, tSV, and SV which almost overlapped. This was due to that dSV and tSV are a subset of SV. This finding indicated that SNP, expression and metabolite features would provide different layers of biological information and might contribute differently and complementarily to the phenotypic variation.

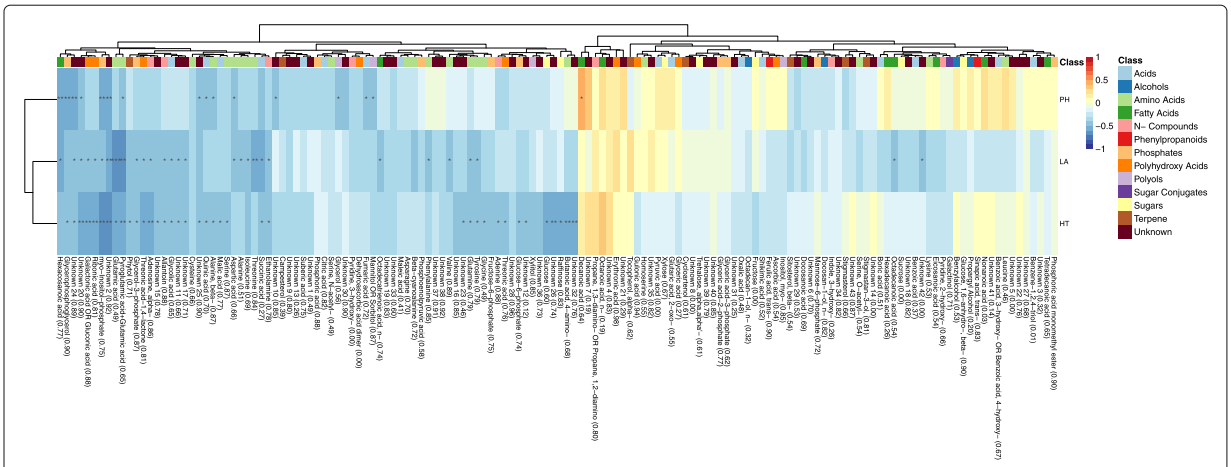

Fig. 1 Heatmap of Pearson correlation coefficients calculated between all pairs of the three phenotypic traits and the 144 metabolites. The three phenotypic traits are leaf angle (LA), plant height (PH) and heading time (HT). Correlations marked with *, **, and *** were significant at $P < 0.05$, 0.01, and 0.001, respectively. The heritability of each metabolite is given in parentheses after each metabolite's name

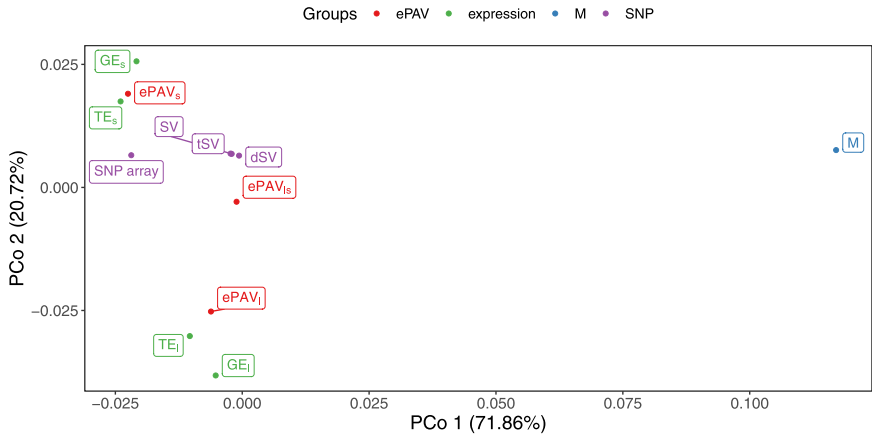

Fig. 2 Plot of the first two axes of the principal coordinate analysis for comparison of the similarity between different omic datasets based on generalized procrustes analysis. The omic datasets include SNP array, sequence variants (SV), deleterious sequence variants (dSV), tolerant sequence variants (tSV), gene expression in seedling and leaf (GE_s and GE_l), transcript expression in seedling and leaf (TE_s and TE_l), expression presence/absence variation in seedling, leaf and combining both tissues ($ePAV_s$, $ePAV_l$, and $ePAV_{ls}$), and metabolites (M). The colors show the four groups of omic datasets used in a grid search for integration of multiple predictors (Figure 5). Red represents ePAV, green expression, blue metabolite, and purple SNP and SV predictors

Omic prediction

The prediction ability of the three phenotypic traits using different single predictors was examined through five-fold cross-validation. Regardless of the predictor, the prediction abilities were higher for traits with higher heritabilities (Fig. 3). Prediction abilities based on SV, GE, TE, ePAV, and M datasets were compared to that realized with the SNP array which was used as baseline predictor. The observed median prediction ability based on the SNP array dataset ranged from 0.185 (HT) to 0.590 (LA). The prediction ability of SV extracted from mRNA sequencing dataset was slightly higher than that of SNP array dataset across the three traits. Moreover, the dSV dataset slightly outperformed the SV extracted from mRNA sequencing and the tSV dataset (Fig. 3). Even higher prediction abilities were observed for ePAV, any expression datasets from seedling (GE_s and TE_s), and metabolite datasets (Fig. 3). The prediction abilities for the ePAV dataset were significantly different among *l*, *s* and *ls*, but not consistently across the three traits (data not shown). $ePAV_{ls}$ was chosen as the best compromise across the three traits for further analyses, as it was for none of the three traits in the significance group with the lowest prediction abilities. The TE datasets slightly outperformed the GE datasets for HT and LA, and TE_s resulted in the highest prediction ability as single predictor for these traits. In contrast, no difference between TE and GE was observed for PH.

To explore whether the heritability of a metabolite affects the prediction performance, eight classes of metabolites based on different degrees of heritabilities served as predictor. The prediction ability increased when the metabolites with lower heritability (< 0.1) were not considered (Fig. 3). However, the prediction ability didn't increase significantly and consistently across the three traits with increasing heritability of the considered metabolites (data not shown). Therefore, we selected the metabolite group for which the highest prediction ability was observed across the three traits ($M_{0.6}$) for further analyses.

Pearsons correlation coefficients between pairwise predicted values of different omic datasets were calculated, and the correlation-based distance was used for PCoA analysis for each trait. Across the three examined traits, the metabolite feature was clearly separated from the other omics features (Fig. 4), and the predicted values of M were less correlated with those values of the other omic datasets than the other omic datasets among themselves (Suppl. Fig. S3). A similar result was observed between the two tissues, seedling and leaf, which were clearly separated on Fig. 4. In contrast, the predicted values from features that clustered together on Fig. 4, especially SNP array, SV, dSV, tSV, $ePAV_{ls}$, were highly correlated (Suppl. Fig. S3).

In order to evaluate whether the prediction ability can be improved by combining several predictors, a joined weighted relationship matrix of the single predictors with the highest prediction ability was established and a grid search was used to identify those combinations of dSV, $ePAV_{ls}$, TE_s, and $M_{0.6}$ resulting in the highest prediction ability. For the three examined traits, the highest median prediction ability was observed when more than one predictor was used (Fig. 5). Furthermore, the optimal weights of the four predictors to reach the maximal prediction ability differed among the three traits, but the weights of

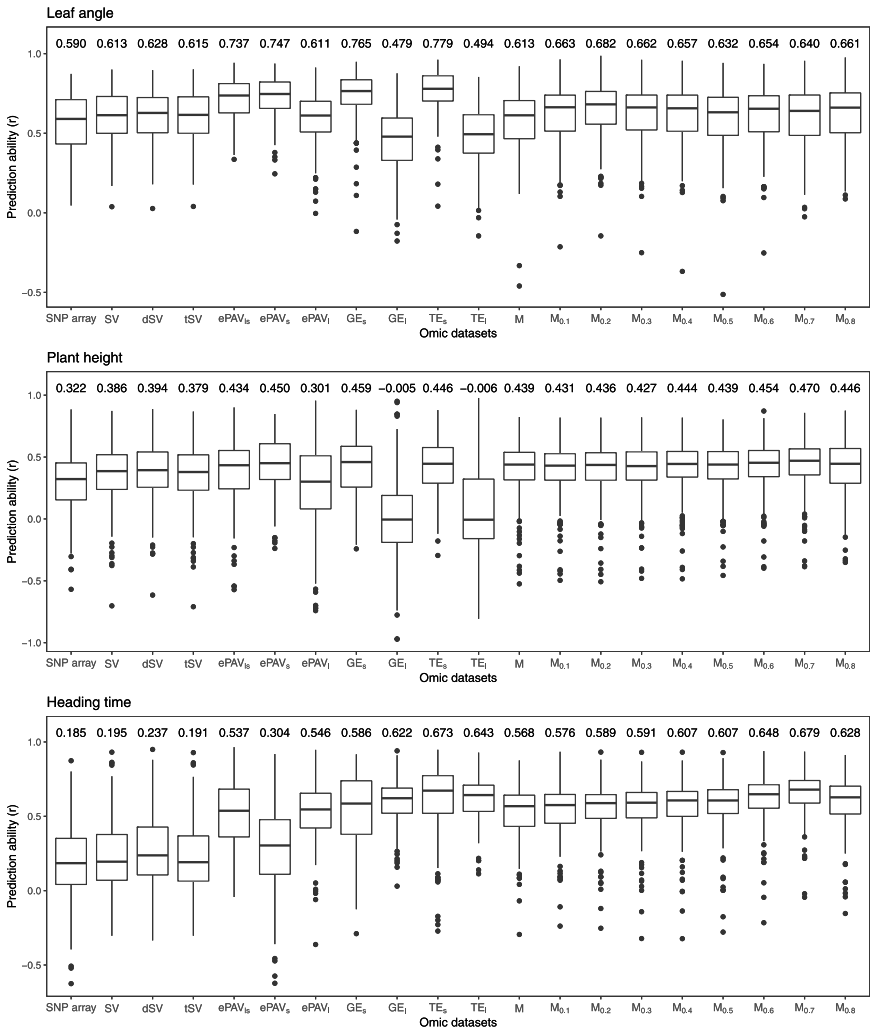

Fig. 3 Boxplot of prediction abilities for the three traits, leaf angle, plant height and heading time, based on 22 inbreds using different omic datasets as a single predictor across 200 five-fold cross-validation runs. The values given above each box represent the medians of 200 runs. The omic datasets include SNP array, sequence variants (SV), deleterious sequence variants (dSV), tolerant sequence variants (tSV), gene expression in seedling and leaf (GE_s and GE_l), transcript expression in seedling and leaf (TE_s and TE_l), expression presence/absence variation in seedling, leaf and combining both tissues ($ePAV_s$, $ePAV_l$, and $ePAV_{ls}$), metabolites filtered for their heritability (M, $M_{0.1}$, $M_{0.2}$, $M_{0.3}$, $M_{0.4}$ $M_{0.5}$, $M_{0.6}$, $M_{0.7}$, and $M_{0.8}$)

$ePAV_{ls}$ and TE_s were at least 10% and 50%, respectively. However, the optimal weight for M was, except for PH, 0, and the optimal weight for the dSV was 0 for the three traits.

We also assessed the prediction abilities of SV, GE, ePAV from 3'end mRNA sequencing that we simulated from our full-length mRNA sequencing dataset. Depending on the trait, a similar, slightly better or worse median of prediction abilities of SV, GE, ePAV were observed when considering 3'end mRNA sequencing compared to a full-length mRNA sequencing dataset as baseline (Fig. 6). Moreover, we did not observe a systematic trend on the prediction ability when increasing the length of the 3'end mRNA sequencing.

Discussion

Ability of different omic features to predict phenotypic traits

Genomic prediction has become a broadly used tool to improve the gain of selection in plant breeding [9]. The current standard procedure of genomic prediction is to use SNP markers generated from SNP array or genotyping by sequencing methods as predictors [12]. However, there are several complicated biological downstream pro-

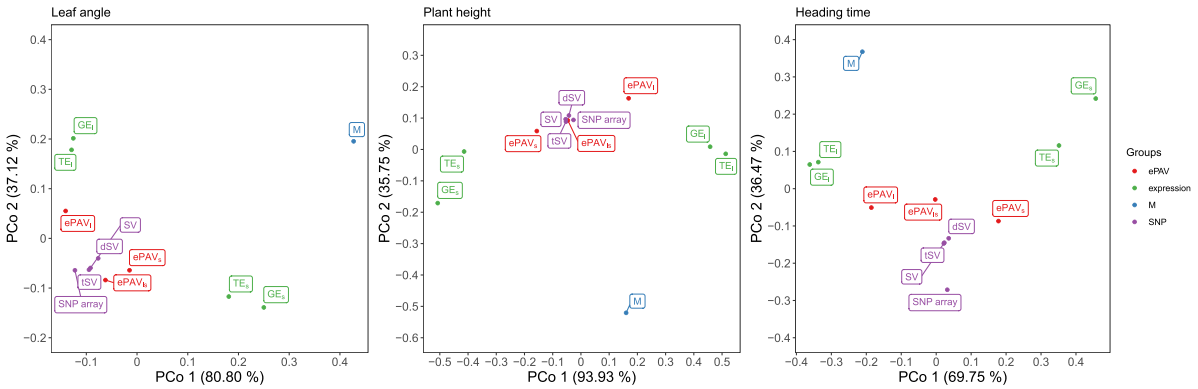

Fig. 4 Plot of the first two axes of principal coordinates analysis for comparison of predicted values from different omic datasets as a single predictor based on the median of correlation-based distance across 200 five-fold cross-validation runs for the three traits, leaf angle, plant height and heading time. The omic datasets include SNP array, sequence variants (SV), deleterious sequence variants (dSV), tolerant sequence variants (tSV), gene expression in seedling and leaf (GE_s and GE_l), transcript expression in seedling and leaf (TE_s and TE_l), expression presence/absence variation in seedling, leaf and combining both tissues ($ePAV_s$, $ePAV_l$, and $ePAV_{ls}$), and metabolites (M). The colors show the four groups of omic datasets used in a grid search for integration of multiple predictors (Figure 5). Red represents ePAV, green expression, blue metabolite, and purple SNP and SV predictors

cesses such as transcription, translation, and biochemical cascades resulting in various metabolites between DNA sequences and phenotypes [11]. Using predictors that are biologically closer to the phenotypes may increase the prediction ability in genomic predictions. With the development of high-throughput molecular technologies, the availability of such predictors from the genomic, transcriptomic, or metabolomic level is ensured [20]. In this pilot study, we aim to compare different types of omic datasets for their predictive performance in order to prioritize them for their later evaluation in large-scale experiments. We hold that this is true also with only 23 inbreds of our study, especially as these inbreds are representative of and cover most of the genotypic diversity of barley [23].

For the three examined traits, any of the SV information generated from mRNA sequencing (SV, dSV, as well as tSV) resulted in a higher prediction ability compared to the SNP data produced with the 50K SNP array (Fig. 3). This might be explained by the higher number of SV features, as increasing the number of predictors can increase the extent of linkage disequilibrium between SNP and quantitative trait loci (QTL) [23, 31]. In addition, INDEL information was included in the SV, which was not the case in the SNP array. INDEL are one type of genetic variation in living organisms that involve larger DNA fragments than single variants and have been identified in known genes (c.f. [32, 33]). Therefore, they are very usefull for the developpment of functional markers [34] and

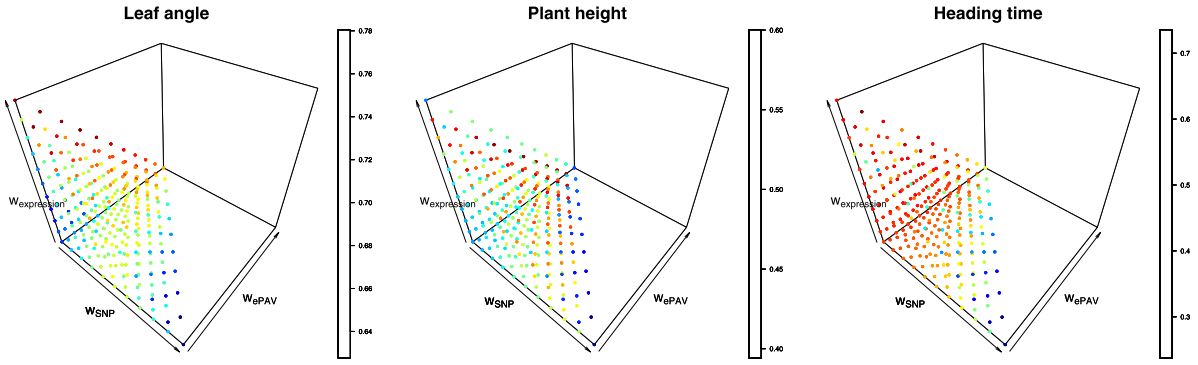

Fig. 5 Prediction ability for the three traits, leaf angle, plant height, and heading time, from 22 inbreds for 286 combinations of the joined weighted matrix which differ in the weights of four predictors, deleterious sequence variants (dSV), expression presence/absence variation in combined leaf and seedling ($ePAV_{ls}$), transcript expression in seedling (TE_s), and metabolite with a heritability on an entry mean basis >0.6 ($M_{0.6}$). Plotted values represent medians across 200 cross-validation runs

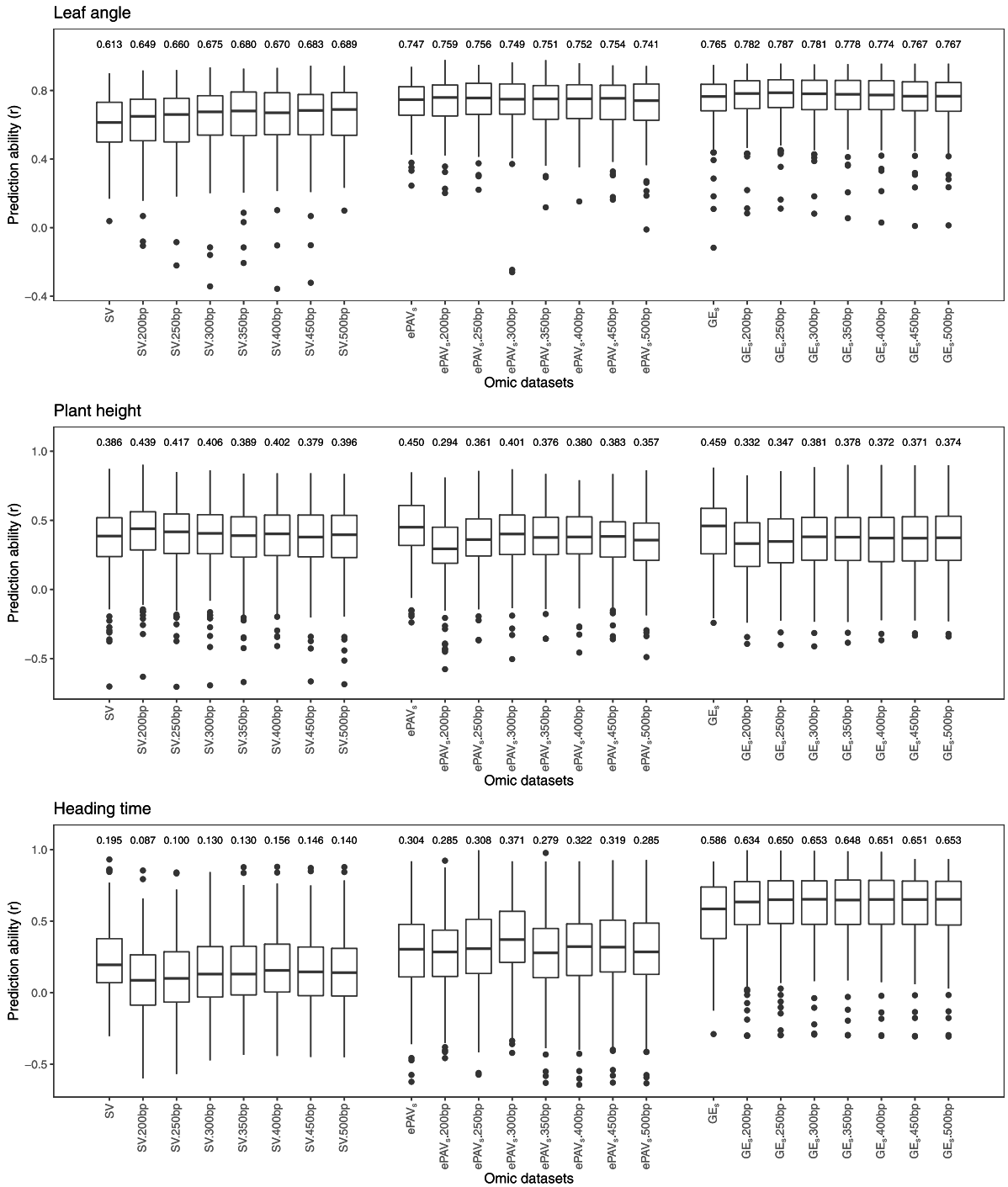

Fig. 6 Boxplot of prediction ability for the three traits, leaf angle, plant height and heading time, from 22 inbreds using different omic datasets from simulated 3'end mRNA sequencing with seven length categories (200, 250, 300, 350, 400,450, and 500 bp) as a single predictor across 200 cross-validation runs. The omic datasets from full-length mRNA sequencing are used as a baseline. The values given above each box represent medians of 200 prediction abilities. The omic datasets include sequence variants (SV), gene expression in seedling (GE_s), and expression presence/absence variation in seedling ($ePAV_s$)

are expected to cause extreme change in the phenotypes. This could be a further explaination why SV had better predictive performance than SNP array. Our observation is in agreement with the finding that the PCo 1 resulting from the GPA separated clearly SV and SNP array (Fig. 2), which indicates that SV and SNP array provide different information.

SV in gene coding regions can be classified into nsSV and sSV, where the former can change the amino acid sequence of proteins, but not the latter. However, not all amino acid changes lead to significant changes of the protein. This can be explored by the SIFT algorithm in classifying SV into dSV and tSV based on the conversion of amino acid sequences [16], where the former cause a loss of protein function but not the latter. Kono et al. [35] showed that known phenotype-altering variants were more frequently inferred as deleterious than the genome-wide average, and have a higher probability to contribute to phenotypic variation. Thus, we compared the prediction ability of dSV and tSV compared to that of SV across the three traits.

The predicted phenotypic values based on the three different classes of SV were highly correlated with each other (Suppl. Fig. S3), which can be expected because dSV and tSV are a subset of SV and clustered together in the GPA (Fig. 2). However, the prediction ability for the three phenotypic traits using dSV information was slightly higher than using tSV and all SV information, despite the fact that the number of dSV features was far smaller (15,868) than the number of tSV features (117,698) and the total number of SV. This trend of a higher prediction ability for dSV was even more pronounced when adjusting for differences in the number of features by resampling simulations (data not shown). Our finding is in discordance with the results of Do et al. [14] and Heidaritabar et al. [15], who observed no difference between the prediction performance of nsSNP and randomly sampled SNPs. A first explanation for our different findings could be that the former cited studies classified the SNP based on whether they may induce amino acid change or not, whereas our study distinguished tolerant/deleterious SNP. Secondly, the SNP used for GP by Heidaritabar et al. [15] were imputed for all genotypes from a 60K SNP array. This might have hampered the improvement of prediction ability in comparison to our study, which is based on real variant data for all inbreds (except few missing data that were mean-imputed). Our finding indicated that the preselection of variants based on their theoreticaly predicted protein function could improve prediction performance of traits, which can be of considerable importance for breeders.

The features derived from the transcriptome datasets (GE, TE, as well as ePAV) led to increased prediction abilities by 62.81% compared to SNP array and even SV on average across the three traits and two tissues. This finding was inconsistent with the results of previous studies [11, 21], who observed that the prediction abilities based on transcriptomic datasets were a little lower (5.30% and 0.03%) than those based on genomic information averaged across the examined traits. This difference might be caused by the complex genetic architectures of traits evaluated and tissue sampled in the studies cited above. However, the use of transcriptomic datasets as predictors still had reasonable prediction abilities in the former studies, which is in accordance with our results and can be explained by the fact that with such datasets expression levels can be quantified and physiological epistasis even captured.

A single gene can encode multiple distinct transcripts through alternative splicing, which allows organisms to increase the protein diversity based on the same set of genes [36], and therefore could lead to more phenotypic variation. As a consequence, a higher prediction ability could be expected for phenotypic traits predicted from TE compared to GE information. This was confirmed by our findings (Fig. 3), and suggests that TE information might be more efficient than GE information in predicting the performance of traits when the full-length mRNA sequencing has been performed.

All the datasets generated by mRNA sequencing from seedling were well separated from those from leaf (Fig. 2). Similarly, the correlation between predicted patterns based on the transcriptomic dataset of the two tissues was low (Fig. 4 and Suppl. Fig. S3) , which indicated that different types of tissue offer dissimilar information concerning the phenotypic variation and influence the prediction ability. In general, the prediction ability was considerably higher for the datasets from seedling in comparison with the datasets from leaf on average across the three traits (Fig. 3). This might be explained by the fact that more diverse genes are expressed in seedling than in leaf.

Only for HT, expression information from leaf (GE_l, TE_l) achieved the same level of prediction ability as that from seedling. One explanation for this finding might be that HT is triggered by environmental factors in later developmental stages and therefore the causal expression features for this trait are more likely to be revealed in leaf than in early developmental stages like seedling.

A total of 53 of the 144 metabolites quantified in our study were significantly correlated with at least one of the three phenotypic traits (Fig. 1). This suggests that the metabolites can be used for selection for phenotypes. In addition, the metabolite feature was clearly separated from the other features in the similarity/dissimilarity analysis (Fig. 2). More importantly, the correlations between the predicted values based on metabolic feature and other omic datasets were low, and lower than the correlation between different other omic datasets (Suppl. Fig. S3).

This finding suggested that the metabolites can provide another biological layer of information to capture the phenotypic variation. We observed across the three traits that prediction abilities based on metabolites were considerably higher compared to SNP or SV information (Fig. 3). This finding is in contradiction to results of previous studies [11, 29] who revealed considerably lower prediction ability using metabolites as predictor. This might be caused by the high accuracy of the metabolite assessment used in our study. The average heritability on an entry mean basis across 144 metabolites was with about 0.62 considerably higher than that observed by Guo et al. [11] with 0.49 and Gemmer et al. [29] with 0.26. This aspect was studied further by leaving out those metabolites with heritabilities < 0.1. This resulted in an increased prediction ability for all traits, which suggested that higher accuracy of metabolites can bring stable information in the prediction of phenotypes.

Generally, (di)-similarity between (1) different omic datasets (Fig. 2) and also between (2) the correlation between predicted phenotypic traits based on different omic datasets (Fig. 4 and Suppl. Fig. S3) was observed in our study. This suggested complementation between different biological perspectives to the phenotypic variation. Therefore, combining predictors covering different layers of biological information in an integrative model could have an advantage over the GP model based on single predictors, and was examined in our study.

Increasing prediction abilities by combining multiple predictors

In this study, a grid search was used to identify those combinations of dSV, $ePAV_{ls}$, TE_s, and $M_{0.6}$ in the joined weighted relationship matrix of GBLUP model maximizing the prediction ability. The highest prediction ability across the three examined traits was observed when more than one predictor was used and, for each of the three traits, without the contribution of the dSV (Fig. 5). This finding might be explained by the fact that transcriptome and metabolome information are closer to phenotypes than gene information according to the central dogma of molecular biology, and can capture together more genetic variation and physiological epistasis caused by complicated networks and interactions between genes than when using only one single predictor [11].

On the other hand, even if a higher prediction ability for all three examined traits was observed if more than one predictor was used (Fig. 5), the optimal weight of each component in the joined weighted relationship matrix depended highly on the traits. For instance, metabolite information was needed to obtain the highest prediction ability for PH, but not for the other traits. Transcriptome was the most important component, but the weight ranged from 0.5 to 0.9 across the three traits. From the physiological point of view, this might be explained by the different genetic architectures of the different traits and their exposure to different environments at different developmental stages and tissues. We observed the tendency that for traits with a lower heritability more different omic predictors were needed to result in the highest prediction ability. Further research on traits with high genetic complexity and low heritability such as yield is needed to test this hypothesis.

Summary: application in breeding programs

The results of our study suggested that combining the information of SV, expression, as well as metabolite dataset into genomic prediction models can improve the prediction ability of phenotypic traits. Especially, the expression datasets were the most important components for this improvement (Fig. 5). To be implemented in breeding programs, such datasets have to be created approximately at the costs of one traditional phenotyping unit (c.f. [37]). This implies that the datasets of SV, gene expression, and metabolite are sampled from one tissue, to avoid the cost of multiple sampling at several stages. The goal of this study was to compare predictors for their ability to predict phenotypic traits. The results of our study indicate that the higher and more stable predictive performance across traits can be achieved from gene and transcript expression gained on seedling samples. Seedling samples combine both aptitude in reaching a high prediction ability but can be also generated in a cost-effective and high-throughput manner. Thus, they are recommended as the best tissue to predict the variation of phenotypes in barley populations. However, for other crops such as tuber crops, different approaches and tissues might be needed, which requires further research.

The limited budget available in practical breeding programs for full-length mRNA sequencing hampers the use of such approaches. Instead, 3'end mRNA sequencing could be a cost-effective alternative method to obtain transcriptome information. For 3'end mRNA sequencing, only 50-800bp at the 3'end of the genes are sequenced. Interestingly, we observed that the prediction abilities of SV, GE, ePAV from simulated 3'end mRNA sequencing were on average across the three traits similar to those from the full-length mRNA sequencing (Fig. 6). Therefore, our finding suggested that transcriptome data can be generated from the 3'end mRNA sequencing without losing prediction ability in comparison to the full-length mRNA sequencing, paving the path for the use of such prediction methods in commercial breeding programs.

Although this study is based on a limited number of barley inbreds, it can be considered as a pilot research showing how different omic datasets can improve prediction of phenotypic variation and will open the path to

perform such analysis on a bigger scale, e.g. on segregating populations derived from the 23 inbreds [38].

Materials and methods

Plant materials and phenotypic data collection

This study was based on 23 spring barley inbreds which were selected from a worldwide collection [39] to maximize phenotypic and genotypic diversity [23]. The 23 inbreds were planted as replicated checks in a field experiment laid out as an augmented row-column design. The experiment was performed in seven agro-ecologically diverse environments (Cologne from 2017 to 2019, Mechernich and Quedlinburg from 2018 to 2019) in Germany in which the checks were replicated 10 to 21 times per environment. At each environment, three yield-related phenotypic traits were assessed. The leaf angle (LA) was scored on a scale from 1 (erect) to 9 (very flat) on four-week-old plants. The heading time (HT) was recorded as days after planting. Furthermore, the plant height (PH, cm) was measured after heading (only assessed in Cologne and Mechernich).

Omic datasets

Metabolite profiling

The metabolite profiling of our study was based on leaf samples collected for the 23 barley inbreds with quadruplicates in a greenhouse experiment, where no phenotypic traits were assessed. Seeds of the 23 spring barley inbreds were sown in controlled conditions with 16 hours light and eight hours dark at 22 °C. Plantlets were cultivated for two weeks and then moved to vernalisation in a growth chamber. After five weeks of vernalisation, the plants were repotted and returned to the greenhouse. After one week, one 3 x 1cm piece of the central part of the youngest fully developed leaf was harvested from two plants of the same inbred, pooled, and immediately flash frozen in liquid nitrogen. The collection of all samples was done within one hour to minimize the variation due to circadian rhythms. Each of the 92 samples was analyzed one time via gas chromatography-mass spectrometry (GC-MS) using an adapted protocol from Lisec et al. [40]. Metabolites were extracted from 45-55 mg frozen mortared samples with 1.5 ml of a 1:2.5:1 H_2O:methanol:chloroform (v:v:v) mixture pre-cooled to -20 °C, then mixed on a rotator for 10 min and centrifuged at 20,000 g for 2 min (both at 4 °C). A total of 30 μl of the supernatant were dried completely in a vacuum concentrator and derivatized in two steps via an MPS-Dual-head autosampler (Gerstel): (1) with 10 μl methoxyamine hydrochloride (Acros organics; freshly prepared at 20 mg/ml in pure pyridine (Sigma-Aldrich)) and shaking at 37 °C for 90 min, (2) adding 90 μl N-Methyl-N-(trimethylsilyl)trifluoroacetamide (MSTFA; Macherey-Nagel) and shaking at 37 °C for 30 min. After incubation for 2 hours at room temperature, 1 μl of derivatized compounds was injected at a flow of 1 ml/min with an automatic liner exchange system in conjunction with a cold injection system (Gerstel) in splitless mode (ramping from 50 °C to 250 °C at 12 °C/s) into the GC. Chromatography was performed using a 7890B GC system (Agilent Technologies) with a 30 m long, 0.25 mm internal diameter, HP-5MS column with 5% phenyl methyl siloxane film (Agilent 19091S-433). The oven temperature was held constant at 70°C for 2 min and then ramped at 12.5°C/min to 320°C at which it was held constant for 5 min; resulting in a total run time of 27 minutes.

Metabolites were ionized with an electron impact source at 70V and 200 °C source temperature and recorded in a mass range of m/z 60 to m/z 800 at 20 scans per second with a 7200 GC-QTOF (Agilent Technologies). Raw data files exported from MassHunter Qualitative (v b07, Agilent Technologies) in the mzData format (*mzdata.xml) were converted to the NetCDF format (*.cdf) and baseline-corrected via MetAlign (v 041012, [41]) using default parameters. Baseline-correction was visually inspected using OpenChrom (v 1.3.0, [42]). Quantitative analysis of GC-MS-based metabolite profiling experiments was then performed using TagFinder (v 4.1, [43]). After evaluating the uniqueness and linearity of each fragment, the aggregated fragment intensity was calculated as the average of the maximum scaled fragment intensity. For relative quantification, aggregated fragment intensities of the compounds were normalized to those of the internal standard ribitol (Sigma-Aldrich) which was added to the extraction buffer. Mass spectral annotation was manually supervised using the Golm Metabolome Database mass-spectral library (http://gmd.mpimp-golm.mpg.de/download/) after conversion of absolute time in retention indices [44]. The raw data, details of the quantification and annotation steps, and the processed metabolite profiles are available (https://www.ebi.ac.uk/metabolights/MTBLS1561). The compounds corresponding to contaminations, siloxane, ribitol, and dimethylphenylalanine were removed. Furthermore, if several compounds were identified as the same metabolite, the one with the greatest heritabilty, for which the calculation is described below, was retained.

SNP genotyping, RNA extraction, sequencing, and quantification of gene expression

The Illumina 50K barley SNP array [45] was used to genotype the 23 inbreds of our study [23]. This dataset is designated in the following as SNP array.

mRNA was extracted from leaf and seedling samples of the 23 inbreds as described earlier by Weisweiler et al. [23]. 46 polyA enriched RNA libraries were prepared at the Max Planck Genome Centre Cologne (https://mpgc.mpipz.mpg.de/home/). In addition, two tissue sam-

ples of one of the inbreds and one tissue sample of two other inbreds had to be removed during the data cleaning process. Reads were trimmed, adapter and low quality regions were removed. Afterwards, reads were mapped using HISAT2 (version 2.0.5) [46] to the Morex referene sequence version 1 [47]. Transcript calling was performed with StringTie (version 2.1.3) [48]. Newly identified and annotated genes were included to the dataset as described by Weisweiler et al. [23]. The expression data for the 23 inbreds was separated into gene expression and transcript expression data. The expression quantified as fragments per kilobase of exon model per million fragments mapped (FPKM) was measured for every transcript of a gene, resulting in one FPKM-value per gene and the corresponding FPKM-value for each transcript of a gene. The FPKM-values of genes and transcripts are designated in the following as GE and TE, where the indexes l and s were used to separate the leaf (GE_l, TE_l) and seedling (GE_s, TE_s) samples. For further details see Weisweiler et al. [23].

Determination of ePAV

For each tissue separately, a presence call was made for each inbred-gene combination in the matrix of presence/absence calls, if GE >0 and an absence call if GE = 0. No presence/absence call ("NA") was made for the inbreds with $0< GE <10\%$ of the maximum value of GE for a gene-tissue combination (cf. [49]). Tissue specific ePAV calls were combined to an across tissue ePAV call as described in detail by Weisweiler et al. [23]. The ePAV detection procedure resulted in three ePAV data sets, namely ePAV leaf ($ePAV_l$), ePAV seedling ($ePAV_s$), and one across both tissues ($ePAV_{ls}$).

Sequence variant calling

Variant calling of SNP and small INDEL and their filtering was performed with samtools (version 1.11) and bcftools (version 1.10.2) as described by Weisweiler et al. [23], and the dataset is designed in the following as SV. SIFT4G (version 2.4) was used to annotate and predict tolerant and deleterious variants. The prediction was done based on the conversion of amino acid sequences [16]. Amino acid substitutions were classified according to their effect on the protein functions and were predicted as tolerant if the score was >0.05 and as deleterious if the score was <= 0.05. The SIFT4G database was build based on the uniref 90 database (downloaded 2020/04/29) and the Morex reference sequence version 1 [47] with the tool SIFT4_Create_Genomic_DB.

Simulation of 3'end mRNA sequencing

For the simulation of 3'end mRNA sequencing, GE_s was only measured based on the last 200, 250, 300, 350, 400, 450, and 500 bp at the 3'end of each gene. To the same reduced set of sequence data, the ePAV detection procedure and the SV calling procedure has been applied resulting in seven different GE, ePAV, and SV datasets.

Statistical analyses

Adjusted entry means, variance components, and heritability

Based on visual inspections of quantile-quantile (Q-Q) plots of residuals as well as residuals vs. fitted values plots, phenotypic outliers were removed. Each of the phenotypic traits was then analysed across the environments using the following mixed model:

$$y_{ijk} = \mu + E_j + G_i + (G \times E)_{ij} + \varepsilon_{ijk}, \quad (1)$$

where y_{ijk} was the observed phenotypic value for the i^{th} genotype at the j^{th} environment within the k^{th} replication, μ the general mean, G_i the effect of the i^{th} inbred, E_j the effect of the j^{th} environment, $(G \times E)_{ij}$ the interaction between the i^{th} inbred and the j^{th} environment, and ε_{ijk} the random error. To estimate adjusted entry means for all inbreds, G_i was treated as fixed and the other effects as random. As the samples for metabolites were collected from one environment, the model [1] was reduced to:

$$y_{ik} = \mu + G_i + \varepsilon_{ik}, \quad (2)$$

where y_{ik} was the observed metabolite for the i^{th} inbred within the k^{th} replication, and ε_{ik} the random error. The resulting adjusted entry means of phenotypic traits and metabolites for each inbred were used in further analyses, where the adjusted entry means of metabolites were designated as M.

To estimate the genetic variance (σ^2_G), model (1) and (2) were used but considering G_i as random. The heritability on an entry mean basis for the phenotypic traits and metabolites was then calculated as $H^2 = \sigma^2_G/(\sigma^2_G + \bar{v}/2)$, where $\bar{v}$ was the mean variance of difference between two adjusted entry means [50].

Prediction of phenotypic traits from multi-omic datasets

The performance to predict phenotypic variation of different types of predictors: (1) SNP array, (2) sequence variants (SV), (3) deleterious sequence variants (dSV), (4) tolerant sequence variants (tSV), (5) $ePAV_s$, (6) $ePAV_l$, (7) $ePAV_{ls}$, (8) gene expression in seedling (GE_s), (9) gene expression in leaf (GE_l), (10) transcript expression in seedling (TE_s), (11) transcript expression in leaf (TE_l), (12) metabolite (M), was compared based on the most stable and widely used model in GP, genomic best linear unbiased prediction (GBLUP) model [51], which can be described as

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \boldsymbol{\varepsilon} \quad (3)$$

where $\mathbf{y}$ is the vector of the adjusted entry means of the examined trait, $\mathbf{1}$ the unit vector, μ the general mean, $\mathbf{Z}$ the incidence matrix of genotypic effects, and $\mathbf{u}$ the vector of genotypic effects that are assumed be normal

distributed with $N(0, \mathbf{G}\sigma_u^2)$, in which $\mathbf{G}$ denotes the relationship matrix between inbreds and σ_u^2 the genetic variance. In addition, $\boldsymbol{\varepsilon}$ is the vector of residuals following a normal distribution $N(0, \mathbf{I}\sigma_e^2)$. In this study, only additive effects were modeled.

For each of the above mentioned omic dataset, the monomorphic features and the features with missing rates > 0.2 have been filtered out. $\mathbf{W}$ was defined as a matrix of feature measurements for the respective omic dataset that is designated in the following as predictor. The dimensions of $\mathbf{W}$ were the number of barley inbreds (n) times the number of features in the corresponding predictor (m) (Table 1). Because of genotyping problems for one of the inbreds, 22 inbred lines were used for further analyses ($n = 22$).

For each predictor, the additive relationship matrix $\mathbf{G}$ was defined as $\mathbf{G} = \frac{\mathbf{W}^*\mathbf{W}^{*T}}{m}$, where $\mathbf{W}^*$ is a matrix of feature measurement for the respective predictor, whose columns are centered and standardized to unit variance of $\mathbf{W}$, and $\mathbf{W}^{*\mathbf{T}}$ is the transpose of $\mathbf{W}^*$. In addition, to assess the impact of the heritability of a metabolite on the prediction performance, only those metabolites with a heritability on an entry mean basis higher than t, where t varied from 0.1 to 0.8 in increments of 0.1, were considered, and the datasets were designated as $M_{0.1}$, $M_{0.2}$, $M_{0.3}$, $M_{0.4}$ $M_{0.5}$, $M_{0.6}$, $M_{0.7}$ and $M_{0.8}$.

In order to understand whether the different omic datasets can capture similar genetic information, Pearsons correlation coefficients between pairwise predicted values of different omic datasets were calculated. Subsequently, 1 − the correlation coefficients among all pairs of predictors was used as the correlation-based distance in a PCoA. Furthermore, to investigate the performance of a joined weighted relationship matrix [21] to predict phenotypic variation, the matrices $\mathbf{G}$ in model (3) of four predictors were weighted and summed up to one joined weighted relationship matrix, where we varied:

1. the weight of SNP (w_{SNP}): the weight of the most representative SNP datasets was determined as the one from the SNP array, SV, tSV, or dSV which has the most stable prediction performance across the three traits (dSV).
2. the weight of ePAV (w_{ePAV}): the weight of the most representative ePAV datasets was determined as the one from $ePAV_{ls}$, $ePAV_s$, or $ePAV_l$ which has most stable prediction performance across the three traits ($ePAV_{ls}$).
3. the weight of expression ($w_{expression}$): the weight of the most representative of the expression datasets was determined as the one from GE_s, GE_l, TE_s, or TE_l which has most stable prediction performance across the three traits (TE_s).
4. the weight of metabolite (w_M, $1 - w_{SNP} - w_{ePAV} - w_{expression}$): the weight of the most representative metabolite datasets was determined as the one from M, $M_{0.1}$, $M_{0.2}$, $M_{0.3}$, $M_{0.4}$ $M_{0.5}$, $M_{0.6}$, $M_{0.7}$, or $M_{0.8}$ which has most stable prediction performance across the three traits ($M_{0.6}$).

A grid search, varying any weight (w) from 0 to 1 in increments of 0.1, resulted in 286 different combinations of joined weighted relationship matrix, where the summation of four weights in each combination must be equal to 1. In addition, the performance of SV, GE_s, and $ePAV_s$ from simulated 3'end mRNA sequencing of different length as described above was explored.

Five-fold cross-validation was used to assess the model performance. Prediction abilities were obtained by calculating Pearson correlations between observed (y) and predicted ($\hat{y}$) adjusted entry means in the validation set of each fold. The median prediction ability across the five folds within each replicate was calculated and the median of the median across the 200 replicates was used for further analyses.

Table 1 The number of features and the abbreviations for each omic dataset used in this study

Omic dataset	Abbreviation	Number of features
50K SNP array	SNP array	38,285
Sequence variants	SV	133,566
Deleterious sequence variants	dSV	15,868
Tolerant sequence variants	tSV	117,698
Expression presence/absence variation in seedling	$ePAV_s$	27,445
Expression presence/absence variation in leaf	$ePAV_l$	26,653
Expression presence/absence variation in combining leaf and seedling	$ePAV_{ls}$	36,235
Gene expression in seedling	GE_s	67,844
Gene expression in leaf	GE_l	60,888
Transcript expression in seedling	TE_s	250,490
Transcript expression in leaf	TE_l	220,749
Metabolites	M	144

Correlation and genetic similarity analyses

Correlations among the three phenotypic traits, and between the three phenotypic traits and the individual metabolites were measured as Pearson correlation coefficient. Principal component analysis (PCA) was performed on each omic dataset (SNP array, SV, dSV, tSV, $ePAV_s$, $ePAV_l$, $ePAV_{ls}$, GE_l, GE_s, TE_s, TE_l, and M). To evaluate similarity/dissimilarity among the various datasets, generalized procrustes analysis (GPA) [30] was performed

based on the PCA results. Subsequently, 1 − the procrustes similarity indexes among all pairs of omic datasets was used as dissimilarity measurements in a principal coordinates analysis (PCoA).

All analyses have been performed using the statistical software R [52].

Abbreviations

GP: Genomic prediction; GBLUP: Genomic best linear unbiased prediction; SNP: Single nucleotide polymorphisms; INDEL: Insertions/deletions; nsSNP: Non-synonymous single nucleotide polymorphisms; sSNP: Synonymous single nucleotide polymorphisms; tSNP: Tolerant single nucleotide polymorphisms; dSNP: Deleterious single nucleotide polymorphisms; SV: Sequence variants; dSV: Deleterious sequence variants; tSV: Tolerant sequence variants; *l*: Leaf; *s*: Seedling; ePAV: Expression presence/absence variation; $ePAV_s$: Expression presence/absence variation in seedling; $ePAV_l$: Expression presence/absence variation in leaf; $ePAV_{ls}$: Expression presence/absence variation in combining leaf and seedling; GE: Gene expression; GE_s: Gene expression in seedling; GE_l: Gene expression in leaf; TE: Transcript expression; TE_s: Transcript expression in seedling; TE_l: Transcript expression in leaf; M: Metabolites; QTL: Quantitative trait loci; GPA: Generalized procrustes analysis; PCA: Principal component analysis; PCoA: Principal coordinates analysis; LA: Leaf angle; PH: Plant height; HT: Heading time; GC-MS: Gas chromatography-mass spectrometry; FPKM: Fragments per kilobase of exon model per million fragments mapped; w_{SNP}: Weight of single nucleotide polymorphisms; w_{ePAV}: Weight of expression presence/absence variation; $w_{expression}$: Weight of expression; w_M: Weight of metabolite

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-08337-7.

Additional file 1: Supplemental Materials.

Additional file 2: Supplementary Table S1.

Additional file 3: Supplementary Table S2.

Acknowledgements

Computational infrastructure and support were provided by the Centre for Information and Media Technology at Heinrich Heine University Düsseldorf. We acknowledge the comments of Joachim Kopka (Department of Molecular Physiology, Max-Planck-Institute of Molecular Plant Physiology, 14476 Potsdam-Golm, Germany) and Dominik Brilhaus (Cluster of Excellence on Plant Sciences (CEPLAS), Heinrich Heine University, 40225 Düsseldorf, Germany) on an earlier version of this manuscript. The authors thank Florian Esser, and George Alskief for technical assistance with performing the field experiments at Cologne and Mechernich as well as Dr. Frankziska Wespel (Saatzucht Breun) and her team for realizing the field experiment at Quedlinburg. We acknowledge excellent technical assistance of Elisabeth Klemp, Katrin Weber, and Maria Graf for GC-MS measurements. We are grateful to Amaury de Montaigu for collecting and processing the metabolite samples.

Authors' contributions

DVI and BS designed and coordinated the project, MW processed genotypic and transcriptomic datasets, PW conducted GS-MS measurements, AE did datamining of GC-MS dataset, AS contributed to data interpretation, PYW performed the data analyses, and DVI, PYW, and BS wrote the manuscript. All authors read and approved the final manuscript.

Funding

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 2048/1, Project ID: 390686111). The funders had no influence on study design, the collection, analysis and interpretation of data, the writing of the manuscript, and the decision to submit the manuscript for publication. Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

The sequencing datasets have been deposited in the NCBI Sequence Read Archive (SRA) under accession PRJNA534414. The metabolite dataset have been deposited in the MetaboLights (https://www.ebi.ac.uk/metabolights/MTBLS1561). The phenotypic dataset of the adjusted entry means for the three traits can be found in Supplementary Table S1. The annotation and abundance of metabolites can be found in Supplementary Table S2.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Institute of Quantitative Genetics and Genomics of Plants, Heinrich Heine University, 40225 Düsseldorf, Germany. [2]Cluster of Excellence on Plant Sciences (CEPLAS), Heinrich Heine University, 40225 Düsseldorf, Germany. [3]Department of Molecular Physiology, Max-Planck-Institute of Molecular Plant Physiology, 14476 Potsdam-Golm, Germany. [4]Institute of Plant Biochemistry, Heinrich Heine University, 40225 Düsseldorf, Germany.

Received: 17 September 2021 Accepted: 20 January 2022
Published online: 12 March 2022

References

1. Newton AC, Flavell AJ, George TS, Leat P, Mullholland B, Ramsay L, Revoredo-Giha C, Russell J, Steffenson BJ, Swanston JS, Thomas WTB, Waugh R, White PJ, Bingham IJ. Crops that feed the world 4. Barley: a resilient crop? Strengths and weaknesses in the context of food security. Food Secur. 2011;3(2):141–78. https://doi.org/10.1007/s12571-011-0126-3.
2. FAO. The Future of Food and Agriculture – Trends and Challenges. Rome. 2017. http://www.fao.org/3/i6583e/i6583e.pdf. Accessed on 7 May 2021.
3. Fróna D, Szenderák J, Harangi-Rákos M. The challenge of feeding the world. Sustainability (Switzerland). 2019;11(20):5816. https://doi.org/10.3390/su11205816.
4. Abberton M, Batley J, Bentley A, Bryant J, Cai H, Cockram J, Costa de Oliveira A, Cseke LJ, Dempewolf H, De Pace C, Edwards D, Gepts P, Greenland A, Hall AE, Henry R, Hori K, Howe GT, Hughes S, Humphreys M, Lightfoot D, Marshall A, Mayes S, Nguyen HT, Ogbonnaya FC, Ortiz R, Paterson AH, Tuberosa R, Valliyodan B, Varshney RK, Yano M. Global agricultural intensification during climate change: a role for genomics. Plant Biotechnol J. 2016;14(4):1095–8. https://doi.org/10.1111/pbi.12467.
5. Sreenivasulu N, Schnurbusch T. A genetic playground for enhancing grain number in cereals. Trends Plant Sci. 2012;17(2):91–101. https://doi.org/10.1016/J.TPLANTS.2011.11.003.
6. Mantilla-Perez MB, Salas Fernandez MG. Differential manipulation of leaf angle throughout the canopy: current status and prospects. J Exp Bot. 2017;68(21-22):5699–717. https://doi.org/10.1093/JXB/ERX378.
7. Bezant J, Laurie D, Pratchett N, Chojecki J, Kearsey M. Marker regression mapping of QTL controlling flowering time and plant height in a spring barley (Hordeum vulgare L.) cross. Heredity. 1996;77(1):64–73. https://doi.org/10.1038/hdy.1996.109.
8. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157(4):1819–29.
9. Desta ZA, Ortiz R. Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci. 2014;19(9):592–601. https://doi.org/10.1016/j.tplants.2014.05.006.
10. Xu Y, Liu X, Fu J, Wang H, Wang J, Huang C, Prasanna BM, Olsen MS, Wang G, Zhang A. Enhancing genetic gain through genomic selection: from livestock to plants. Plant Commun. 2020;1(1):100005. https://doi.org/10.1016/j.xplc.2019.100005.
11. Guo Z, Magwire MM, Basten CJ, Xu Z, Wang D. Evaluation of the utility of gene expression and metabolic information for genomic prediction in

maize. Theor Appl Genet. 2016;129(12):2413–27. https://doi.org/10.1007/s00122-016-2780-5.
12. Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y, Dreisigacker S, Singh R, Zhang X, Gowda M, Roorkiwal M, Rutkoski J, Varshney RK. Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 2017;22(11):961–75. https://doi.org/10.1016/j.tplants.2017.08.011.
13. Li Z, Gao N, Martini JWR, Simianer H. Integrating gene expression data Into genomic prediction. Front Genet. 2019;10(FEB):126. https://doi.org/10.3389/fgene.2019.00126.
14. Do DN, Janss LLG, Jensen J, Kadarmideen HN. SNP annotation-based whole genomic prediction and selection: an application to feed efficiency and its component traits in pigs. J Anim Sci. 2015;93(5):2056–63. https://doi.org/10.2527/jas.2014-8640.
15. Heidaritabar M, Calus MPL, Megens H-J, Vereijken A, Groenen MAM, Bastiaansen JWM. Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. J Anim Breeding Genet. 2016;133(3):167–79. https://doi.org/10.1111/jbg.12199.
16. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. Nat Protoc. 2016;11(1):1–9. https://doi.org/10.1038/nprot.2015.123.
17. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001;11(5):863–74. https://doi.org/10.1101/gr.176601.
18. Taylor MB, Ehrenreich IM. Higher-order genetic interactions and their contribution to complex traits. Trends Genet. 2015;31(1):34–40. https://doi.org/10.1016/j.tig.2014.09.001.
19. Wang X, Xu Y, Hu Z, Xu C. Genomic selection methods for crop improvement: current status and prospects. Crop J. 2018;6(4):330–40. https://doi.org/10.1016/j.cj.2018.03.001.
20. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinforma Biol Insights. 2020;14. https://doi.org/10.1177/1177932219899051.
21. Schrag TA, Westhues M, Schipprack W, Seifert F, Thiemann A, Scholten S, Melchinger AE. Beyond genomic prediction: combining different types of omics data can improve prediction of hybrid performance in maize. Genetics. 2018;208(4):1373–85. https://doi.org/10.1534/genetics.117.300374.
22. Hu X, Xie W, Wu C, Xu S. A directed learning strategy integrating multiple omic data improves genomic prediction. Plant Biotechnol J. 2019;17(10):2011–20. https://doi.org/10.1111/pbi.13117.
23. Weisweiler M, de Montaigu A, Ries D, Pfeifer M, Stich B. Transcriptomic and presence/absence variation in the barley genome assessed from multi-tissue mRNA sequencing and their power to predict phenotypic traits. BMC Genomics. 2019;20(1):787. https://doi.org/10.1186/s12864-019-6174-3.
24. Swarup R, Crespi M, Bennett MJ. One gene, many proteins: mapping cell-specific alternative splicing in plants. Dev Cell. 2016;39(4):383–5. https://doi.org/10.1016/j.devcel.2016.11.002.
25. Rattray NJW, Deziel NC, Wallach JD, Khan SA, Vasiliou V, Ioannidis JPA, Johnson CH. Beyond genomics: understanding exposotypes through metabolomics. Human Genomics. 2018;12(1):1–14. https://doi.org/10.1186/s40246-018-0134-x.
26. Meyer RC, Steinfath M, Lisec J, Becher M, Witucka-Wall H, Törjék O, Fiehn O, Eckardt Ä, Willmitzer L, Selbig J, Altmann T. The metabolic signature related to high plant growth rate in Arabidopsis thaliana. Proc Natl Acad Sci U S A. 2007;104(11):4759–64. https://doi.org/10.1073/pnas.0609709104.
27. Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, Sulpice R, Altmann T, Stitt M, Willmitzer L, Melchinger AE. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet. 2012;44(2):217–20. https://doi.org/10.1038/ng.1033.
28. Longin F, Beck H, Gütler H, Heilig W, Kleinert M, Rapp M, Philipp N, Erban A, Brilhaus D, Mettler-Altmann T, Stich B. Aroma and quality of breads baked from old and modern wheat varieties and their prediction from genomic and flour-based metabolite profiles. Food Res Int. 2020;129. https://doi.org/10.1016/j.foodres.2019.108748.
29. Gemmer MR, Richter C, Jiang Y, Schmutzer T, Raorane ML, Junker B, Pillen K, Maurer A. Can metabolic prediction be an alternative to genomic prediction in barley?,. PLoS ONE. 2020;15(6):0234052. https://doi.org/10.1371/journal.pone.0234052.
30. Gower JC. Generalized procrustes analysis. Psychometrika. 1975;40(1):33–51. https://doi.org/10.1007/BF02291478.
31. Goddard ME, Hayes BJ. Genomic selection. J Anim Breeding Genet. 2007;124(6):323–30. https://doi.org/10.1111/j.1439-0388.2007.00702.x.
32. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. 2006;16(9):1182–90. https://doi.org/10.1101/GR.4565806.
33. Clark TG, Andrew T, Cooper GM, Margulies EH, Mullikin JC, Balding DJ. Functional constraint and small insertions and deletions in the ENCODE regions of the human genome. Genome Biol. 2007;8(9):1–14. https://doi.org/10.1186/GB-2007-8-9-R180.
34. Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES. Dwarf8 polymorphisms associate with variation in flowering time. Nat Genet. 2001;28(3):286–9. https://doi.org/10.1038/90135.
35. Kono TJY, Fu F, Mohammadi M, Hoffman PJ, Liu C, Stupar RM, Smith KP, Tiffin P, Fay JC, Morrell PL. The role of deleterious substitutions in crop genomes. Mol Biol Evol. 2016;33(9):2307–17. https://doi.org/10.1093/molbev/msw102.
36. Black DL. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. Cell. 2000;103(3):367–70. https://doi.org/10.1016/S0092-8674(00)00128-8.
37. Cobb JN, DeClerck G, Greenberg A, Clark R, McCouch S. Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. TAG Theor Appl Genet Theor Angew Genet. 2013;126(4):867. https://doi.org/10.1007/S00122-013-2066-0.
38. Casale F, Van Inghelandt D, Weisweiler M, Li J, Stich B. Genomic prediction of the recombination rate variation in barley - A route to highly recombinogenic genotypes. Plant Biotechnol J. 2021. https://doi.org/10.1111/PBI.13746.
39. Haseneyer G, Stracke S, Paul C, Einfeldt C, Broda A, Piepho H-P, Graner A, Geiger HH. Population structure and phenotypic variation of a spring barley world collection set up for association studies. Plant Breed. 2009;129(3):271–9. https://doi.org/10.1111/j.1439-0523.2009.01725.x.
40. Lisec J, Schauer N, Kopka J, Willmitzer L, Fernie AR. Gas chromatography mass spectrometry-based metabolite profiling in plants. Nat Protoc. 2006;1(1):387–96. https://doi.org/10.1038/nprot.2006.59.
41. Lommen A. Metalign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. Anal Chem. 2009;81(8):3079–86. https://doi.org/10.1021/ac900036d.
42. Wenig P, Odermatt J. OpenChrom: a cross-platform open source software for the mass spectrometric analysis of chromatographic data. BMC Bioinformatics. 2010;11. https://doi.org/10.1186/1471-2105-11-405.
43. Luedemann A, Strassburg K, Erban A, Kopka J. Data and text mining TagFinder for the quantitative analysis of gas chromatography-mass spectrometry (GC-MS)-based metabolite profiling experiments. Bioinformatics. 2008;24(5):732–7. https://doi.org/10.1093/bioinformatics/btn023.
44. Strehmel N, Hummel J, Erban A, Strassburg K, Kopka J. Retention index thresholds for compound matching in GC-MS metabolite profiling. J Chromatogr B Anal Technol Biomed Life Sci. 2008;871(2):182–90. https://doi.org/10.1016/j.jchromb.2008.04.042.
45. Bayer MM, Rapazote-Flores P, Ganal M, Hedley PE, Macaulay M, Plieske J, Ramsay L, Russell J, Shaw PD, Thomas W, Waugh R. Development and evaluation of a barley 50k iSelect SNP array. Front Plant Sci. 2017;8:1792. https://doi.org/10.3389/fpls.2017.01792.
46. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60. https://doi.org/10.1038/nmeth.3317.
47. Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, Bayer M, Ramsay L, Liu H, Haberer G, Zhang XQ, Zhang Q, Barrero RA, Li L, Taudien S, Groth M, Felder M, Hastie A, Šimková H, Stanková H, Vrána J, Chan S, Munõz-Amatriaín M, Ounit R, Wanamaker S, Bolser D, Colmsee C, Schmutzer T, Aliyeva-Schnorr L, Grasso S, Tanskanen J, Chailyan A, Sampath D, Heavens D, Clissold L, Cao S, Chapman B, Dai F, Han Y, Li H, Li X, Lin C, McCooke JK, Tan C, Wang P, Wang S, Yin S, Zhou G, Poland JA, Bellgard MI, Borisjuk L, Houben A, Doleael J, Ayling S, Lonardi S, Kersey P, Langridge P, Muehlbauer GJ, Clark MD, Caccamo M, Schulman

AH, Mayer KFX, Platzer M, Close TJ, Scholz U, Hansson M, Zhang G, Braumann I, Spannagl M, Li C, Waugh R, Stein N. A chromosome conformation capture ordered sequence of the barley genome. Nature. 2017;544(7651):427–33. https://doi.org/10.1038/nature22043.
48. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–5. https://doi.org/10.1038/nbt.3122.
49. Jin M, Liu H, He C, Fu J, Xiao Y, Wang Y, Xie W, Wang G, Yan J. Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. Sci Rep. 2016;6(1):1–12. https://doi.org/10.1038/srep18936.
50. Piepho HP, Möhring J. Computing heritability and selection response from unbalanced plant breeding trials. Genetics. 2007;177(3):1881–8. https://doi.org/10.1534/genetics.107.074229.
51. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91(11):4414–23. https://doi.org/10.3168/jds.2007-0980.
52. R Core Team. R: A Language and Environment for Statistical Computing. 2019. https://www.r-project.org/. Accessed on 2 Sept 2019.

Publisher's Note

SUPPLEMENTAL MATERIALS

List of Supplemental Tables

- Table S1: The adjusted entry means of the 23 inbreds for the three traits, leaf angle (LA), plant height (PH) and heading time (HT).

- Table S2: The information of 192 chemical entries and their relative abundance for each inbred.

- Table S3: The classification of the 144 metabolites based on their chemical properties.

List of Supplemental Figures

- Figure S1: Distribution of heritabilities (H^2) for the 144 metabolites. The average (0.62) is indicated as red vertical line.

- Figure S2: Pearson correlation coefficients calculated between all pairs of adjusted entry means of the three phenotypic traits.

- Figure S3: Heatmap of correlation coefficients calculated between all pairs of the predicted values of omic datasets for the three traits, leaf angle, plant height and heading time, across 200 five-fold cross-validation runs. The values given in each cell represent the medians of 200 runs. The omic datasets include SNP array, sequence variants (SV), deleterious sequence variants (dSV), tolerant sequence variants (tSV), gene expression in seedling and leaf (GE_l

and GE_s), transcript expression in seedling and leaf (TE_l and TE_s), expression presence/absence variation in seedling, leaf and combining both tissues ($ePAV_s$, $ePAV_l$, and $ePAV_{ls}$), and metabolites (M)

Table S1: The adjusted entry means of the 23 inbreds for the three traits, leaf angle (LA), plant height (PH) and heading time (HT).

`Suppl_Table_S1_AEM_3traits_23_inbreds.csv`

Table S2: The information of 192 chemical entries and their relative abundance for each inbred.

`Suppl_Table_S2_192_analytes_information.csv`

Table S3: The classification of the 144 metabolites based on their chemical properties.

Metabolites	Number
Amino Acids	22
Acids	16
Phosphates	11
Fatty Acids	10
Polyhydroxy Acids	9
N-Compounds	8
Alcohols	4
Sugars	7
Terpene	7
Polyols	4
Phenylpropanoids	2
Sugar Conjugates	1
Unknown	43

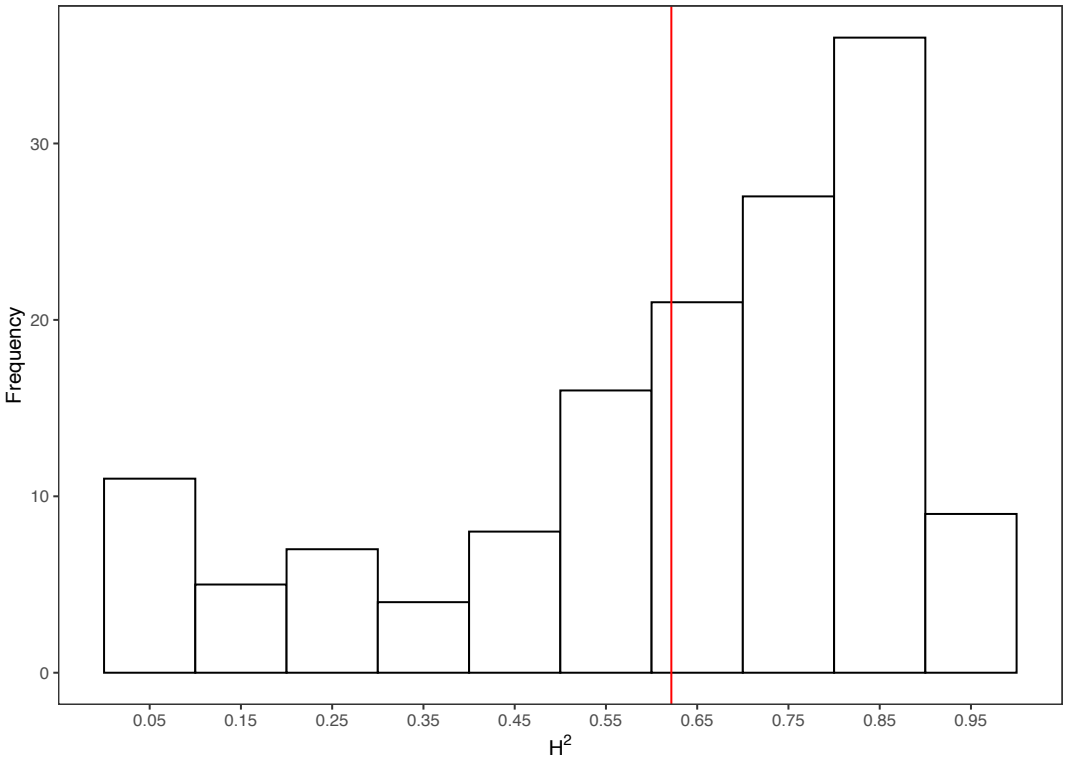

Figure S1: Distribution of heritabilities (H^2) for the 144 metabolites. The average (0.62) is indicated as red vertical line.

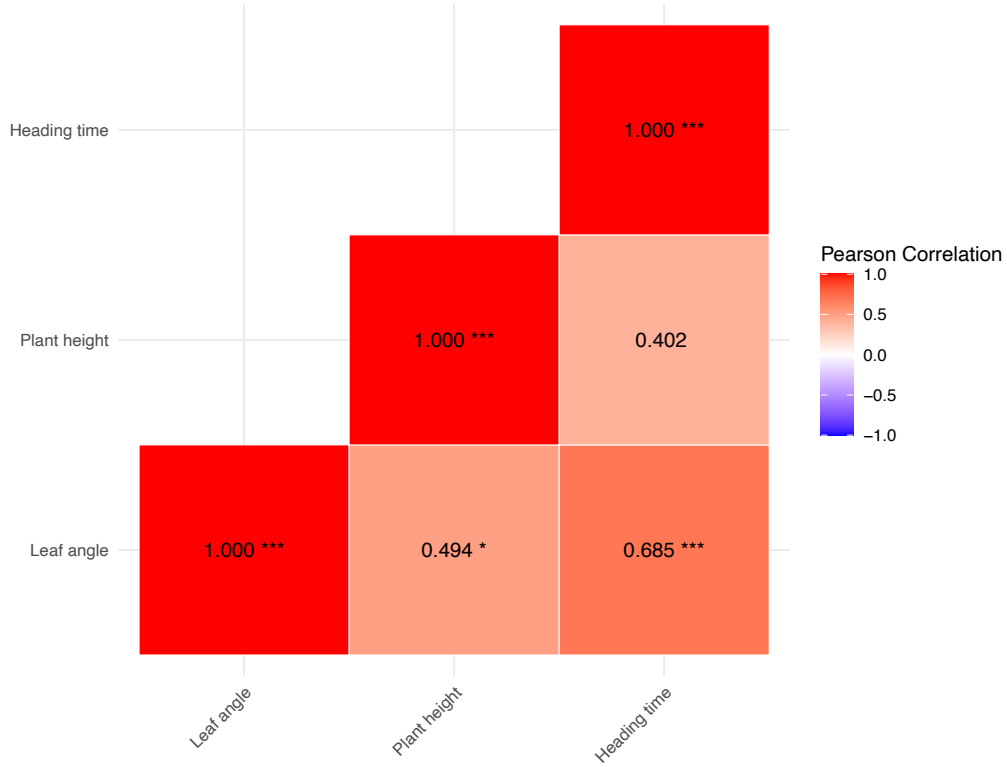

Figure S2: Pearson correlation coefficients calculated between all pairs of adjusted entry means of the three phenotypic traits.

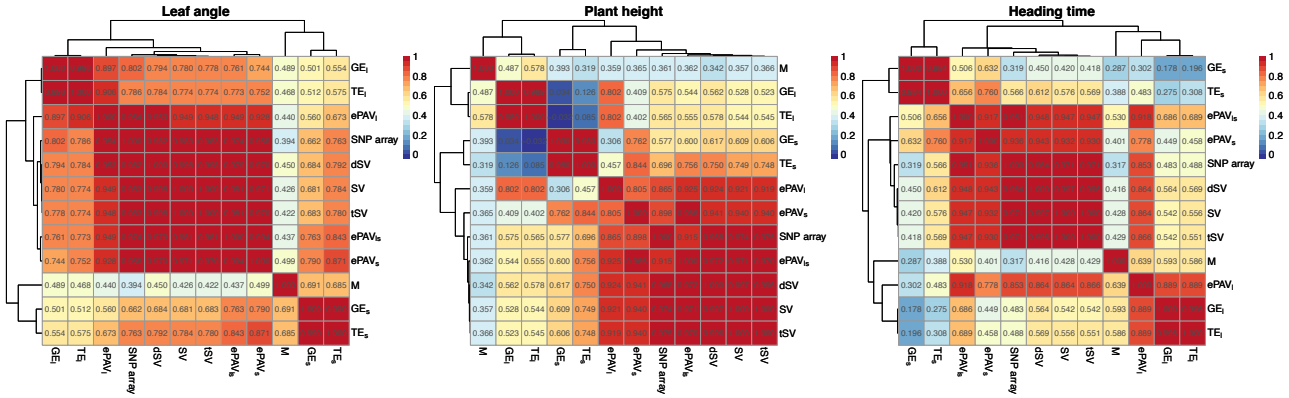

Figure S3: Heatmap of correlation coefficients calculated between all pairs of the predicted values of omic datasets for the three traits, leaf angle, plant height and heading time, across 200 five-fold cross-validation runs. The values given in each cell represent the medians of 200 runs. The omic datasets include SNP array, sequence variants (SV), deleterious sequence variants (dSV), tolerant sequence variants (tSV), gene expression in seedling and leaf (GE_l and GE_s), transcript expression in seedling and leaf (TE_l and TE_s), expression presence/absence variation in seedling, leaf and combining both tissues ($ePAV_s$, $ePAV_l$, and $ePAV_{ls}$), and metabolites (M).

6 Structural variants in the barley gene pool: precision and sensitivity to detect them using short-read sequencing and their association with gene expression and phenotypic variation

This manuscript was published in Theoretical and Applied Genetics in August, 2022.

Authors:
Marius Weisweiler, Christopher Arlt*, **Po-Ya Wu***, Delphine van Inghelandt, Thomas Hartwig, and Benjamin Stich.
*: These authors contributed equally

Own contribution: Co-second author. I performed the data analyses.

Theoretical and Applied Genetics (2022) 135:3511–3529
https://doi.org/10.1007/s00122-022-04197-7

ORIGINAL ARTICLE

Structural variants in the barley gene pool: precision and sensitivity to detect them using short-read sequencing and their association with gene expression and phenotypic variation

Marius Weisweiler[1] · Christopher Arlt[1] · Po-Ya Wu[1] · Delphine Van Inghelandt[1] · Thomas Hartwig[2] · Benjamin Stich[1,3]

Received: 11 May 2022 / Accepted: 3 August 2022 / Published online: 27 August 2022

Abstract

Key message **Structural variants (SV) of 23 barley inbreds, detected by the best combination of SV callers based on short-read sequencing, were associated with genome-wide and gene-specific gene expression and, thus, were evaluated to predict agronomic traits.**

Abstract In human genetics, several studies have shown that phenotypic variation is more likely to be caused by structural variants (SV) than by single nucleotide variants. However, accurate while cost-efficient discovery of SV in complex genomes remains challenging. The objectives of our study were to (i) facilitate SV discovery studies by benchmarking SV callers and their combinations with respect to their sensitivity and precision to detect SV in the barley genome, (ii) characterize the occurrence and distribution of SV clusters in the genomes of 23 barley inbreds that are the parents of a unique resource for mapping quantitative traits, the double round robin population, (iii) quantify the association of SV clusters with transcript abundance, and (iv) evaluate the use of SV clusters for the prediction of phenotypic traits. In our computer simulations based on a sequencing coverage of 25x, a sensitivity > 70% and precision > 95% was observed for all combinations of SV types and SV length categories if the best combination of SV callers was used. We observed a significant ($P < 0.05$) association of gene-associated SV clusters with global gene-specific gene expression. Furthermore, about 9% of all SV clusters that were within 5 kb of a gene were significantly ($P < 0.05$) associated with the gene expression of the corresponding gene. The prediction ability of SV clusters was higher compared to that of single-nucleotide polymorphisms from an array across the seven studied phenotypic traits. These findings suggest the usefulness of exploiting SV information when fine mapping and cloning the causal genes underlying quantitative traits as well as the high potential of using SV clusters for the prediction of phenotypes in diverse germplasm sets.

Introduction

Researchers began to study genomic rearrangements and structural variants (SV) about 60 years ago. These studies investigated somatic chromosomes, biopsies, and cell cultures from lymphomas to understand the role of abnormal chromosome numbers as well as SV for the development of cancer (Jacobs and Strong 1959; Nowell and Hungerford 1960; Manolov and Manolov 1972; Craig-Holmes et al. 1973; Mitelman et al. 1979).

The development of sequencing by synthesis pioneered by Frederick Sanger (Sanger et al. 1977) enabled in the following years the first sequenced genomes of prokaryotes (e.g., *Escherichia coli*) and eukaryotes (e.g., yeast) (Goffeau et al. 1996; Blattner et al. 1997). Next milestones of sequencing by synthesis were the sequenced genomes of

Communicated by Takao Komatsuda.

Christopher Arlt and Po-Ya Wu authors contributed equally.

✉ Benjamin Stich
benjamin.stich@hhu.de

[1] Institute for Quantitative Genetics and Genomics of Plants, Universitätsstraße 1, 40225 Düsseldorf, Germany

[2] Institute for Molecular Physiology, Universitätsstraße 1, 40225 Düsseldorf, Germany

[3] Cluster of Excellence on Plant Sciences, From Complex Traits towards Synthetic Modules, Universitätsstraße 1, 40225 Düsseldorf, Germany

Arabidopsis thaliana as first plant species (The Arabidopsis Genome Iniative 2000) and of human (Craig Venter et al. 2001). Due to the development of next-generation sequencing (NGS) platforms such as 454 and Illumina, studies aiming for genome-wide variant detection in 100s or 1000s of samples as in the 1000 genome project (Altshuler et al. 2012) became possible.

Three different approaches have been proposed to detect SV based on NGS data: assembling, long-read sequencing, and short-read sequencing (Mahmoud et al. 2019). For crop and especially for cereal species, the assembly approach is a tough challenge because of the large genome size and the high proportion of repetitive elements in the genomes (Neale et al. 2014; Mascher et al. 2017). Long-read mapping requires Pacific Biosciences or Nanopore sequencing data which results in high costs if many accessions should be sequenced and, thus, is not affordable for many research groups. In contrast, short-read sequencing is well-established for SV detection in the human genome (Chaisson et al. 2019; Ebert et al. 2021). Various software tools have been developed to detect SV from short-read sequencing data and were benchmarked based on human genomes (Cameron et al. 2019; Kosugi et al. 2019).

More recently there is also an increased interest in using such approaches for SV detection in plant genomes (Fuentes et al. 2019; Zhou et al. 2019; Guan et al. 2021). Fuentes et al. (2019) evaluated several SV callers to detect SV in the rice genome. However, no study evaluated the performance of SV callers for transposon-rich complex cereal genomes.

Several studies have examined the distribution and frequency of SV in the genomes of rice and maize (Wang et al. 2018; Yang et al. 2019; Kou et al. 2020). Despite the importance of cereals for human nutrition, only Jayakodi et al. (2020) performed a genome-wide study on SV in barley, with a focus on large SV in 20 barley accessions.

In humans, SV have been described to have an up to ~ 50fold stronger influence on gene expression than single nucleotide variants (SNV) (Chiang et al. 2017). SV also have been associated with changes in transcript abundance in plants such as in cucumber (Zhang et al. 2015), maize (Yang et al. 2019), tomato (Alonge et al. 2020), and soybean (Liu et al. 2020a). However, the role and frequency of SV in gene regulatory mechanisms in small grain cereals is widely unexplored.

In humans, several studies have shown that phenotypic variation is more likely to be caused by SV than by SNV (Alkan et al. 2011; Baker 2012; Sudmant et al. 2015; Schüle et al. 2017; McColgan and Tabrizi 2018). In plants, individual SV have been associated with traits such as aluminum tolerance in maize (Maron et al. 2013), disease resistance and domestication in rice (Xu et al. 2012), or plant height (Li et al. 2012) and heading date (Nishida et al. 2013) in wheat. In barley, individual SV have been associated with traits such as Boron toxicity tolerance (Sutton et al. 2007) and disease resistance (Muñoz-Amatriaín et al. 2013). In grapevine and rice, it has been shown that SV have a low variant frequency due to purifying selection (Zhou et al. 2019; Kou et al. 2020). However, few studies have examined the ability to predict quantitatively inherited phenotypic traits using SV in comparison to SNV.

The objectives of our study were to (i) facilitate SV discovery studies by benchmarking SV callers and their combinations with respect to their sensitivity and precision to detect SV in the barley genome, (ii) characterize the occurrence and distribution of SV clusters in the genomes of 23 barley inbreds that are the parents of a unique resource for mapping quantitative traits, the double round robin population (Casale et al. 2022), (iii) quantify the association of SV clusters with transcript abundance, and (iv) evaluate the use of SV clusters for the prediction of phenotypic traits.

Methods

Benchmarking of variant callers for detecting SV and INDELs in the barley genome

Computer simulations

We used Mutation-Simulator (version 2.0.3) (Kühl et al. 2021) to simulate INDELs, deletions, duplications, inversions, insertions, and translocations in the first chromosome of the Morex reference sequence v2 (Monat et al. 2019) as this was the genome sequence available when our study was performed. Furthermore, it is not expected that the reference version impacts the results of the simulations. In accordance with Fuentes et al. 2019, we considered five SV length categories for each of the above mentioned SV types (except translocations) (A: 50–300 bp; B: 0.3–5 kb; C: 5–50 kb; D: 50–250 kb; E: 0.25–1 Mb) plus INDELs (2-49bp). Translocations were simulated for 50 bp–1 Mb (ABCDE). We simulated SV with a mutation rate of $1.9x10^{-6}$ for the SV length categories A-C and INDELs, whereas mutation rates of $3.8x10^{-6}$ and $1.9x10^{-7}$ were assumed for SV length categories D and E, respectively. For each type of SV, we used BBMap's randomreads.sh (BBMap - Bushnell B. - http://sourceforge.net/projects/bbmap/) to simulate 2x150 bp Illumina reads with a sequencing coverage of 1.5x, 3x, 6x, 12.5x, 25x, and 65x as well as LRSim (version 1.0) (Luo et al. 2017) to simulate linked-reads with a sequencing coverage of 14x and 25x. Illumina- and linked-reads were simulated with a minimum, average, and maximum base quality of 25, 35, and 40, respectively.

Theoretical and Applied Genetics (2022) 135:3511–3529

SV detection

The simulated Illumina reads were mapped to the first chromosome of the Morex reference sequence v2 using BWA-MEM (version 0.7.15) whereas LongRanger align (version 2.2.2) was used for the simulated linked-reads. The SV callers Pindel (version 0.2.5b9) (Ye et al. 2009), Delly (version 0.8.1) (Rausch et al. 2012), GRIDSS (version 2.8.3) (Cameron et al. 2017), Manta (version 1.6.0) (Chen et al. 2016), Lumpy (smoove version 0.2.5) (Layer et al. 2014), and NGSEP (version 3.3.2) (Duitama et al. 2014) were used to identify SV based on the mapped reads. GATK's HaplotypeCaller (4.1.6.0) (Poplin et al. 2017), Pindel, and GRIDSS were used to detect INDELs. The workflow was implemented in Snakemake (version 5.10.0) (Köster et al. 2021). A SV call was only kept if it passed the built-in filter of the corresponding SV caller. For INDELs and all SV types and length categories, only homozygous, alternative variant calls were considered. Deletions annotated as "replacement" (RPL) by Pindel were removed. We calculated the sensitivity (1), precision (2), and the F1-score (3) as

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) \quad (1)$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (2)$$

$$F1\text{ - core} = 2 * (\text{Precision*Sensitivity}/\text{Precision} + \text{Sensitivity}) \quad (3)$$

for all combinations of SV types*SV callers, where TP was the number of true positives, FP the number of false positives, and FN the number of false negatives. For INDELs, a TP INDEL had break points that did differ ≤ 2 bp from those of the simulated INDEL and the length did differ by ≤ 5bp. For SV length category A, a TP SV had break points that did differ ≤ 10 bp from those of the simulated SV and the SV length did differ by ≤ 20 bp. For the other SV length categories, a TP SV had break points and length differences compared to the simulated SV of ≤ 50 bp. For insertions where no SV length was detected, the start of a TP insertion had a break point that did differ ≤ 10 bp from this of the simulated insertion. For translocations, a TP translocation had break points that did differ ≤ 50 bp from those of the simulated translocation.

We also evaluated combinations of SV callers for their precision and sensitivity to detect SV. The following procedure was used to decide for the combinations that were examined: First, for those SV callers, which have shown a precision ≥ 95% for all SV length categories for a particular SV type, SV calls were combined via logical or ("|"). Second, for those SV callers with a precision ≤ 95% in at least one SV length category, SV calls were combined with a logical and ("&"). If the precision of the combination of the second step increased to ≥ 95% in all SV length categories, SV calls of this combination were kept for the particular SV type and were combined with a logical or with those of the first step. The threshold of ≥ 95% precision was used to reduce the number of FP SV calls to a reasonable level.

Detection of SV, SNV, and INDELs in the barley genome

Genetic material and sequencing

Our study was based on 23 spring barley inbreds (Weisweiler et al. 2019) that were selected out of a worldwide collection of 224 inbreds (Haseneyer et al. 2010) (Supplementary Table S6) using the MSTRAT algorithm (Gouesnard 2001). These inbreds are the parents of the double round robin population (Casale et al. 2022). Paired-end sequencing libraries with an insert size of 425 bp were sequenced (2x150 bp) to a ~25x coverage on the Illumina HiSeqX platform by Novogene Corporation Inc. (Sacramento, USA).

SV, INDELs, and SNV detection

The quality of the raw reads was checked by fastqc. Reads were adapter- and quality-trimmed using Trimmomatic (version 0.39) (Bolger et al. 2014). The trimmed reads were mapped to the Morex reference sequence v3 (Mascher et al. 2021) using BWA-MEM. PCR-duplicates were removed using PICARD (version 2.22.0).

Based on the results of the benchmarking of different SV callers using simulated data, the results of specific SV callers were combined as explained above. The final set of deletions for each inbred were those that were identified by Manta | GRIDSS | Pindel | Delly | (Lumpy & NGSEP) where homozygous-reference (0/0) and heterozygous variant (0/1) calls were removed. Additionally, deletions annotated by Pindel as RPL were removed. In analogy, the duplications were identified by Manta | GRIDSS | Pindel | (Delly & Lumpy). Insertions of the SV length category A were identified by Manta | GRIDSS | Delly, where insertions of the SV length categories B-E were called using Manta. Inversions were identified by Manta | GRIDSS | Pindel. Translocations were called from pairs of break points identified by Manta | GRIDSS | (Delly & Lumpy). INDELs were detected by GATK's HaplotypeCaller | GRIDSS | Pindel where homozygous-reference (0/0) and heterozygous variant (0/1) calls were discarded. SV which were located in a region of the reference sequence, where the sequence only consists of N's, were excluded. For genome regions, where break points of different SV overlapped or were inconsistent in the same inbred, only the smallest SV was considered.

The number of false positives could be increased by detecting large SV clusters; therefore, SV clusters larger than 1 Mb were not considered in our study. The SV of the 23 inbreds were grouped together to SV clusters based on the similarity of sizes and the position in the genome according to the following procedure. The distance from a SV to the next SV in such a SV cluster had to be smaller than 20 bp for the SV length category A and 50 bp for the SV length category B - E and the difference of the two break points had to be smaller than 10 or 50 bp as described above. SV with a larger difference between break points were kept as separate SV and SV clustering was pursuing. Each SV cluster was genotyped across the examined 23 barley inbreds.

SNV and INDELs were called using GATK. First, GATK's HaplotypeCaller was used in single sample GVCF mode, afterward GATK's CombineGVCFs was used to combine the SNV across the 23 inbreds. Combined SNV were genotyped using GATK's GenotypeGVCFs. SNV were filtered using GATK's VariantFiltration where variants below the following filtering thresholds were removed: QD < 2.0; QUAL < 30.0; SOR > 3.0; FS > 60.0; MQ < 40.0; MQRankSum < −12.5; ReadPosRankSum < −8.0. Heterozygosity of SNV for each genotype was low (1.0–1.7%) and therefore such SNV were not discarded to avoid removing true positives.

PCR validation of SV

A total of 25 of the detected SV were targeted for validation by PCR amplification of genome regions of and around the SV in Morex and Unumli-Arpa. This included six SV length category A deletions, five SV length category A insertions, six SV length category B deletions and eight SV length category C-E deletions. In order to determine the SV allele, we required the amplification of two differently sized fragments in the two inbreds. For each SV, a regular primer pair was created with the position defined by the validation strategy (Supplementary Fig. S1). If needed, a second right primer was added to the PCR reaction. The primers were designed using Primer3 (Untergasser et al. 2012) and Blast+ (Camacho et al. 2009).

Plant material was sampled for the PCR validation from adult plants and seedlings grown under controlled conditions. DNA was extracted from 100 mg frozen plant material using the DNeasy Plant Mini Kit (Qiagen, Germany) according to the manufacturer's instructions. The PCR reaction mixture contained in a final volume of 20 µL: 0.2 mM dNTP, Fw/Rev Primer 0.5 µM, 50 ng DNA, 1.5 U/µL DreamTaq DNA Polymerase (Thermo Fischer Scientific, USA), Polymerase-Buffer 1X and water. Amplified fragments were separated by gel electrophoresis and the validation success was determined by comparing the PCR product sizes with the calculated values based on the SV detection.

Location of SV clusters

SV clusters were classified and annotated based on their location in the genome, their distance relative to genes, or other genomic features. SV clusters were grouped into four gene-associated and one intergenic SV cluster categories: 5 kb upstream/downstream gene-associated SV clusters were located in the 5 kb region from the 3′- or 5′- end of a gene. Intron and exon gene-associated SV clusters were located in the gene sequence, where the genic sequence was separated into intronic and exonic sequences. SV clusters which were not located in the four gene-associated SV cluster categories were determined as intergenic SV clusters. A gene-associated SV cluster could be classified in more than one category if its sequence covers several genomic features.

To check if the detected SV clusters were transposable elements, the genomic positions of SV clusters were compared to the transposable elements annotation file of the Morex reference sequence v3 (Mascher et al. 2021). Deletions, duplications, inversions, INDELs, and insertions with known length were annotated as transposable elements if the reciprocal overlap was ≥ 80% (Fuentes et al. 2019). Insertions with unknown length were classified as transposable elements if the detected break point of the insertion was inside the transposable element sequence. Translocations were classified as transposable element, if at least one of the two break points was located inside a transposable element sequence.

SV hotspots were identified using the following procedure: The average number of SV clusters in non-overlapping 1 Mb windows across each of the seven chromosomes was determined. Using this number, we calculated for each window based on the Poisson distribution the expected number of SV clusters. Windows with more SV clusters than the Q_{99} of the expected Poisson distribution were designated as SV hotspots (Guan et al. 2021).

Population genetic analyses

Linkage disequilibrium (LD) measured as r^2 (Hill and Robertson 1968) was calculated between each SV type and linked SNV. Nucleotide diversity (π) was calculated in 100 kb windows along the seven chromosomes separately for SV clusters (deletions, insertions, duplications, inversions) and SNV using vcftools (version 0.1.17) (Danecek et al. 2011).

SV clusters and gene expression

SV clusters which were assigned into one of the gene-associated SV categories, namely 5 kb up- or downstream, introns, and exons, were associated with the genome-wide gene expression of the 23 barley inbreds. Gene expression for the

seedling tissue measured as fragments per kilobase of exon model per million fragments mapped was available for all inbreds from an earlier study (Weisweiler et al. 2019). This information was the basis of a principal component analysis. For all gene-associated SV clusters with a minor allele frequency (MAF) > 0.15, Pearson's correlation coefficient with the first three principal components was estimated, where the presence and absence of SV clusters were used as metric character. This analysis was performed to examine the association between SV clusters and genome-wide gene expression (Liu et al. 2020b). A permutation procedure with 1,000 iterations was used to test the mean absolute values of the correlations for their significance. In addition to this evaluation of the effect of SV clusters on the genome-wide gene expression level, we also examined the significance of the effect of gene-associated SV clusters with a MAF > 0.15 on the expression of individual genes. In order to do so, the mixed linear model with population structure and kinship matrix (PK model) (Stich et al. 2008) was used. The population structure matrix consisted of the first two principal components calculated from 133,566 SNV and INDELs derived from mRNA sequencing (Weisweiler et al. 2019). From the same information, the kinship matrix was calculated as described by Endelman and Jannink (2012).

Assessment of phenotypic traits

For the assessment of phenotypic traits under field conditions, the 23 inbreds were planted as replicated checks in an experiment laid out as an augmented row-column design. The experiment was performed in seven environments (Cologne from 2017 to 2019, Mechernich and Quedlinburg from 2018 to 2019) in Germany in which the checks were replicated multiple times per environment. For each environment, seven phenotypic traits were assessed. Heading time (HT) was recorded as days after planting, leaf angle (LA) was scored on a scale from 1 (erect) to 9 (very flat) on four-week-old plants, and plant height (PH, cm) was measured after heading in Cologne and Mechernich. Seed area (SA, mm^2), seed length (SL, mm), seed width (SW, mm), and thousand grain weight (TGW, g) were measured based on full-filled grains from Cologne (2017–2019) and Quedlinburg (2018) by using MARVIN seed analyzer (GTA Sensorik, Neubrandenburg, Germany).

Prediction of phenotypes

Each of the phenotypic traits was analyzed across the environments using the following mixed model:

$$y_{ijk} = \mu + E_j + G_i + (G \times E)_{ij} + \varepsilon_{ijk}, \quad (4)$$

where y_{ijk} was the observed phenotypic value for the i^{th} genotype at the j^{th} environment within the k^{th} replication; μ the general mean, G_i the effect of the i^{th} inbred, E_j the effect of the j^{th} environment, $(G \times E)_{ij}$ the interaction between the i^{th} inbred and the j^{th} environment, and ε_{ijk} the random error. This allowed to estimate adjusted entry means for all inbreds.

The performance to predict the adjusted entry means of each barley inbred for each trait using different types of predictors: (1) single nucleotide polymorphism (SNP) array, which was generated by genotyping the 23 inbreds using the Illumina 50K barley SNP array (Bayer et al. 2017), (2) gene expression (3) SNV & INDELs, (3a) SNV, (3b) INDELs, (4) SV clusters, (4a) deletions, (4b) duplications, (4c) insertions, (4d) inversions, (4e) translocations, was compared based on genomic best linear unbiased prediction (GBLUP) (VanRaden 2008).

For each predictor, the monomorphic features and the features with missing rates > 0.2 and identical information were discarded. **W** was defined as a matrix of feature measurement for the respective predictor. The dimensions of **W** were the number of barley inbreds ($n = 23$) times the number of features in the corresponding predictor (m) ($m_{SNP\ \text{array}} = 38,025$, $m_{\text{gene expression}} = 67,844$, $m_{\text{SNV\&INDELs}} = 3,025,217$, $m_{\text{SNV}} = 2,338,565$, $m_{\text{INDELs}} = 686,918$, $m_{\text{SVclusters}} = 458,330$, $m_{\text{deletions}} = 183,219$, $m_{\text{duplications}} = 93,073$, $m_{\text{insertions}} = 70,143$, $m_{\text{inversions}} = 6,582$, $m_{\text{translocations}} = 105,313$). The additive relationship matrix **G** was defined as $\mathbf{G} = \frac{\boldsymbol{W}^*\boldsymbol{W}^{*T}}{m}$, where $\boldsymbol{W}^*$ was a matrix of feature measurement for the respective predictor, whose columns are centered and standardized to unit variance of $\boldsymbol{W}$, and $\boldsymbol{W}^{*T}$ was the transpose of $\boldsymbol{W}^*$.

Furthermore, to investigate the performance of a joined weighted relationship matrix (Schrag et al. 2018) to predict phenotypic variation, the three **G** matrices in GBLUP model of the three predictors, SNV &INDELs, gene expression, and SV clusters, were weighted and summed up to one joined weighted relationship matrix. A grid search, varying any weight (w) from 0 to 1 in increments of 0.1, resulted in 66 different combinations of joined weighted relationship matrix, where the summation of three weights in each combination must be equal to 1.

Fivefold cross-validation was used to assess the model performance. Prediction abilities were obtained by calculating Pearson's correlations between observed (y) and predicted ($\hat{y}$) adjusted entry means in the validation set of each fold. The median prediction ability across the five folds within each replicate was calculated and the median of the median across the 200 replicates was used for further analyses.

Results

Precision and sensitivity of SV callers

Six tools (Table 1) which call SV based on short-read sequencing data were evaluated with respect to their precision and sensitivity to detect five different SV types (deletions, insertions, duplications, inversions, and translocations) in five SV length categories (A: 50–300 bp; B: 0.3–5 kb; C: 5–50 kb; D: 50–250 kb; E: 0.25–1 Mb) using computer simulations. The precision of Delly, Manta, GRIDSS, and Pindel to detect deletions of all five SV length categories based on 25x sequencing coverage ranged from 97.8–100.0%, whereas the precision of Lumpy and NGSEP was lower with values between 75.0 and 89.8% (Table 2). The sensitivity of NGSEP was with 78.6–87.5% the highest but that of Manta was with 79.7–81.1% only slightly lower. We evaluated various combinations of SV callers and observed for the combination of Manta | GRIDSS | Pindel | Delly | (Lumpy & NGSEP) an increase of the sensitivity to detect deletions compared to the single SV callers up to a final of 89.0% without decreasing the precision considerably (99.1%).

Manta was the only SV caller which allowed the detection of insertions of all SV length categories with precision values as high as 99.8–100.0%. The combination of Manta | GRIDSS | Delly for the SV length category A has shown a high sensitivity (88.4%) and precision (99.8%). This combination was therefore used for the detection of insertions of SV length category A in further analyses.

The sensitivity of the SV callers Delly, Manta, Lumpy, and GRIDSS to detect duplications of the SV length category A was with values from 28.2 to 39.4% very low. In contrast, Pindel could detect these duplications with a sensitivity of 75.7%. For the other SV length categories, the combination of Manta | GRIDSS | Pindel could increase the sensitivity to detect duplications by 2–7% compared to using a single SV caller while the precision ranged between 97.6 and 99.3%.

The performance of Lumpy and NGSEP to detect inversions reached precision values of 81.5–98.5% and sensitivity values of 66.1–80.0% that were on the same low level as for deletions. Delly performed well for detecting inversions in SV length categories B to D, but for E and especially for A, the performance was lower compared to that of the other SV callers. Overall, Pindel was the only SV caller with a combination of both, high precision and sensitivity to detect inversions. These precision and sensitivity values could be further improved across all SV length categories by combining the calls of Pindel with that of Manta | GRIDSS (Table 2).

The combination of GRIDSS | Pindel | GATK increased the sensitivity to detect INDELs (2–49 bp) by 3% compared to using the single callers (Supplementary Table S1). With 6%, an even higher difference for the sensitivity to detect translocations was observed between the combination of Manta | GRIDSS | (Delly & Lumpy) and single callers.

In a next step, different sequencing coverages from 1.5x to 65x were simulated and the performance of the best combination of SV callers for each of the SV types was compared to their performance with 25x sequencing coverage (Supplementary Fig. S1). For deletions, the F1-score, which is harmonic mean of the precision and sensitivity, for 65x sequencing coverage was ~2% higher than for 25x sequencing coverage. Only marginal differences were observed between the F1-score of 65x or 25x sequencing coverage for calling duplications and inversions. Interestingly, the F1-score for calling translocations and insertions was with 2% and 9%, respectively, higher in the scenario with 25x than with 65x sequencing coverage. For 12.5x sequencing coverage, the F1-score was still on an high level with values > 80% for each SV type (Supplementary Fig. S2). With a further reduced sequencing coverage, the F1-score also decreased. Finally, the performance of our pipeline to detect SV was evaluated based on 14x and 25x linked-read sequencing data. For all SV types and SV length categories, with the exception of deletions and duplications in SV length category D and A, respectively, the F1-score was 2–7% higher based on Illumina sequencing data than based on linked-read sequencing data.

Table 1 Properties of structural variant (SV) callers for short-read sequencing that were compared in our study, where split reads (SR), paired-end reads (PE), read depth (RD), and local alignments (LA) are the underlying detection principles

SV caller	Detection principle				Deletion	Insertion		Inversion	Duplication	Translocation
	SR	PE	RD	LA		≤500bp	>500bp			
Pindel[1]	x				x	x	x	x	x	
Delly[2]	x	x			x	x		x	x	x
Lumpy[3]	x	x	x		x			x	x	x
Manta[4]	x	x		x	x	x	x	x	x	x
GRIDSS[5]	x	x		x	x	x		x	x	x
NGSEP[6]			x		x	x	x	x		

[1]Ye et al. (2009), [2]Rausch et al. (2012), [3]Layer et al. (2014), [4]Chen et al. (2016), [5]Cameron et al. (2017), [6] Duitama et al. (2014)

Table 2 Sensitivity/precision of structural variant (SV) callers and combinations of them (for details see Material & Methods) to detect deletions, insertions, duplications, and inversions of the SV length categories A (50–300 bp), B (0.3–5 kb), C (5–50 kb), D (50–250 kb), and E (0.25 – 1 Mb)

SV caller	SV length category				
	A	B	C	D	E
	Deletions				
Delly	58.1/97.8	76.2/99.4	72.5/99.3	72.4/100.0	75.0/100.0
Manta	79.7/100.0	81.1/99.8	79.9/99.6	79.7/99.4	81.0/100.0
Lumpy	60.0/78.1	70.5/86.5	66.8/85.6	62.5/79.0	64.3/80.6
GRIDSS	79.0/99.5	80.7/99.9	77.8/99.9	78.1/100.0	77.4/100.0
Pindel	87.4/99.9	68.4/99.7	83.6/99.4	80.2/100.0	67.9/100.0
NGSEP	84.1/87.3	83.1/83.4	83.5/82.2	87.5/89.8	78.6/75.0
Combination	89.0/99.1	86.9/99.4	86.7/99.2	86.5/99.4	86.9/100.0
	Insertions				
Delly	3.4/100.0				
Manta	88.4/99.8	74.1/100.0	72.1/100.0	72.5/100.0	75.0/100.0
GRIDSS	45.5/100.0				
Pindel	6.6/93.0				
NGSEP	64.1/59.2	26.8/29.6	35.5/40.5	30.5/32.1	26.0/26.5
Combination	88.4/99.8	74.1/100.0	72.1/100.0	72.5/100.0	75.0/100.0
	Duplications				
Delly	28.2/99.0	75.1/96.8	74.7/95.4	75.3/97.2	71.7/91.7
Manta	39.0/99.5	80.5/99.8	82.7/99.8	83.9/98.7	82.6/97.4
Lumpy	31.5/98.4	67.9/84.8	67.7/82.6	68.3/81.9	65.2/80.0
GRIDSS	39.4/99.8	80.0/100.0	80.0/100.0	83.3/100.0	79.4/100.0
Pindel	75.7/98.1	57.8/99.0	88.1/99.8	83.9/99.4	73.9/100.0
Combination	75.8/98.1	87.3/99.1	90.8/99.3	89.8/98.2	89.1/97.6
	Inversions				
Delly	49.7/70.4	84.6/99.2	85.5/99.4	82.6/99.4	78.2/98.6
Manta	77.0/99.0	87.0/99.9	87.3/99.9	90.0/100.0	82.8/100.0
Lumpy	66.1/88.5	76.8/96.2	75.3/97.4	77.4/94.8	74.7/98.5
GRIDSS	76.9/99.1	86.9/99.8	85.2/99.9	87.9/100.0	82.8/100.0
Pindel	83.5/99.2	90.7/99.9	90.2/99.9	89.0/100.0	77.0/100.0
NGSEP	0.0/0.0	75.7/87.9	75.3/81.5	80.0/85.4	77.0/88.2
Combination	88.4/98.1	91.5/99.8	90.9/99.8	93.2/100.0	85.1/100.0

SV clusters across the 23 parental inbreds of the double round robin population

Across the 23 barley inbreds that are the parents of a new resource for mapping natural phenotypic variation, the double round robin population, we detected 458,671 SV clusters using the best combination of SV callers (Table 3). These comprised 183,489 deletions, 70,197 insertions, 93,079 duplications, 6,583 inversions, and 105,323 translocations. Additionally, 6,381,352 INDELs were detected

Table 3 Summary of detected structural variants (SV) and small insertions and deletions (2–49 bp, INDELs) across 23 diverse barley inbreds, where MAF was the minor allele frequency, and TE were SV clusters which were annotated as transposable elements in the Morex reference sequence v3

SV type	Number of SV calls	Number of SV clusters		
			MAF > 0.05	TE
Deletions	714,867	183,489	78,823	16,846
Insertions	241,522	70,197	29,672	279 (17,718)[1]
Duplications	195,710	93,079	58,793	6,608
Inversions	14,961	6,583	4,116	92
Translocations	251,956	105,323	61,572	0 (54,258)[1]
INDELs	29,637,520	6,381,352	4,134,064	21

[1]Because of missing endpoint information no reciprocal overlap criterion applied

across the seven chromosomes. The proportion of SV clusters which were annotated as transposable elements varied from 1.4% for inversions to 51.5% for translocations.

We performed a PCR-based validation for detected deletions and insertions (Supplementary Table S2, Supplementary Fig. S3). Six out of six deletions and five out of five insertions up to 0.3 kb could be validated (Supplementary Fig. S4). Additionally, we could validate eight out of eleven deletions between 0.3 and 460 kb (Supplementary Fig. S5), where for the three not validated deletions, the expected fragments were not observed in the non-reference parental inbred.

The number of SV clusters present per inbred ranged from less than 40,000 to more than 80,000 (Fig. 1A). We observed no significant ($P > 0.05$) correlation between the sequencing coverage, calculated based on raw, trimmed, and mapped reads, of each inbred as well as the number of detected SV clusters in the corresponding inbred. A two-sided t-test resulted in no significant ($P > 0.05$) association between the number of SV clusters of an inbred and the spike morphology as well as the landrace versus variety status of the inbreds. In contrast, principal component analyses based on the presence/absence matrices of the SV clusters revealed a clustering of inbreds by spike morphology, geographical origin, and landrace vs. variety status (Supplementary Fig. S6).

Out of the 458,671 SV clusters, 50.6% (232,071) appeared in only one of the 23 inbreds, whereas 19.7% (90,256) were detected in at least five inbreds (Fig. 1B, Supplementary Fig. S7). Additional analyses revealed a significant although weak negative correlation ($r = -0.06681$, $P = 2.07\text{x}10^{-314}$) between the length of a SV cluster and its MAF. The average MAF of SV clusters with a length of 250 kb to 1 Mb and of 50–250 kb was 0.08, respectively, while that of SV clusters with a length of 50 bp–50 kb was 0.13 (Supplementary Fig. S8). SV clusters annotated as transposable elements had a shorter average length of 5,853 bp and a higher MAF of 0.16 compared to SV clusters that were not annotated as transposable elements (10,605 bp, 0.12). Deletions and insertions of the SV length category A were the most common detected SV clusters with a fraction of 41.7 and 48.4%, respectively (Supplementary Table S3). In contrast, for duplications, the

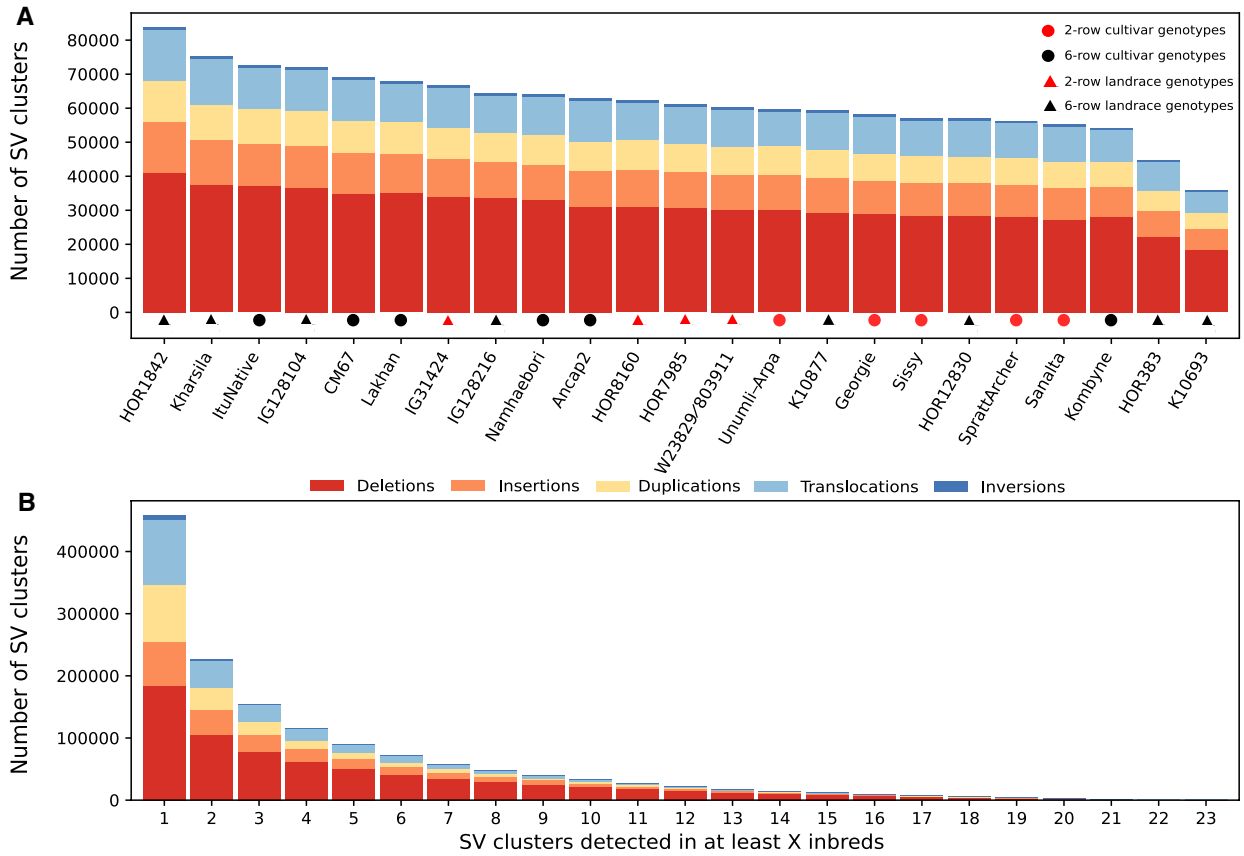

Fig. 1 Stacked bar graph of the number of different types of structural variant (SV) clusters detected in the 23 inbreds (**A**) and SV clusters which were detected in at least the given number of the inbreds (**B**)

largest fraction were that for SV clusters of the SV length category C (55.9%). The average MAF of the individual SV types was the highest for insertions with 0.17, followed by deletions, inversions, translocations, and duplications with values of 0.14, 0.11, 0.10, and 0.10, respectively.

Characterization of the SV clusters

After examining the length of the detected SV clusters and their presence in the 23 barley inbreds, we investigated the distribution of the SV clusters across the barley genome. We observed a significant correlation ($r = 0.5653$, $P < 0.01$) of nucleotide diversity (π) of SV clusters and SNV, measured in 100 kb windows along the seven chromosomes (Supplementary Fig. S9). The SV clusters were predominantly present distal of pericentromeric regions. In contrast to SNV, the frequency of all SV types, and especially that of duplications, increased in centromeric regions (Fig. 2). For all centromeres, a significantly ($P < 0.01$) higher number of SV clusters was observed compared to what is expected based on a Poisson distribution and, thus, were designated as SV hotspots. The proportion of SV clusters in pericentromeric regions was with 14.5% considerably lower compared to what is expected based on the physical length of these regions (25.7%). Only 4.5% of all detected SV hotspots were observed in pericentromeric regions.

We also examined if SV clusters provide additional genetic information compared to that of closely linked SNV. To do so, we determined the extent of LD between each SV cluster and SNV located within 1 kb and compared this with the extent of LD between the closest SNV to the SV cluster and the SNV within 1 kb. Across the different SV types, 33.7–74.3% have at least one SNV within 1 kb that showed an $r^2 \geq 0.6$ (Supplementary Table S4). In contrast, 89.2–89.9% of SNV that are closest to the SV cluster showed an $r^2 \geq 0.6$ to another SNV within 1 kb.

In the next step, we examined the presence of SV clusters relative to the position of genes. The highest proportion of SV clusters (~60%) was located in intergenic regions of the genome (Fig. 3). The second largest fraction (~30%) of SV clusters was present in the 5 kb up- or downstream regions of genes, which is considerably higher compared to that of INDELs (~17%) and SNV (~16%). Within the group of SV clusters that were 5 kb up- or downstream to genes, a particularly high fraction were inversions. On average across all SV types, about 10% of SV clusters were located in introns and exons, with inversions being the exception again, showing a considerably higher rate.

The enrichment of SV clusters proximal to genes lead us to assess their physical distance relative to the transcription start site (TSS) of the closest genes and compare this to SNV. The number of SV clusters at the TSS was approximately 10% lower than 5kb upstream of the TSS (Fig. 4). A similar trend was observed for the 5kb downstream regions (~7%). In comparison, the absolute number of SNV around the TSS was more than ten times lower than the number of SV clusters. With the exception of a distinct peak at position two downstream of the TSS, the number of SNV around the TSS followed the same trends as described for the SV clusters above.

Association of SV clusters with gene expression

We tested if the SV clusters could be associated with the genome-wide gene expression differences of the 23 inbreds. As a first step, a principal component analysis of the gene expression matrix, which included all genes and inbreds, was performed. The loadings of all 23 inbreds on principal component (PC) 1 explained 19.7% of the gene expression variation and were correlated with the presence/absence status of all inbreds for each gene-associated SV cluster. The average absolute correlation coefficient of gene-associated SV clusters and the PC1 of gene expression was 0.17 and higher than the Q_{95} of the coefficient observed for randomized presence/absence pattern and the PC1 (Supplementary Fig. S10, Supplementary Fig. S11). Similar observations were made for the association of gene-associated SV clusters with PC2 and PC3 of 0.17 and 0.19, respectively, for the above-mentioned gene expression matrix (Supplementary Fig. S12). In addition, we investigated a possible association between SV clusters and gene expression on the basis of individual genes. For a total of 1,976 out of 21,140 gene-associated SV clusters a significant ($P < 0.05$) association with the gene expression of the associated gene was observed (Fig. 5).

Prediction of phenotypic variation from SV clusters

The prediction ability of seven quantitative phenotypic traits using SV clusters as well as SNV from a SNP array, genome-wide gene expression information, SNV and INDELs (SNV & INDELs) were examined as predictors through five-fold cross-validation. The median prediction ability across all traits ranged from 0.509 to 0.648. The SV clusters had the highest prediction power, followed by SNV & INDELs, SNP array, and gene expression in decreasing order (Fig. 6). Compared to these differences, those among the median prediction abilities of the different SV types were small. The highest prediction ability was observed for insertions and the lowest for inversions. We also evaluated the possibility to combine SNV and INDELs with gene expression and SV cluster information using different weights to increase the prediction ability (Supplementary Fig. S13). The mean of the optimal weight across the seven traits was highest for gene expression (0.41) and lowest for SV clusters (0.23) (Supplementary Table S5).

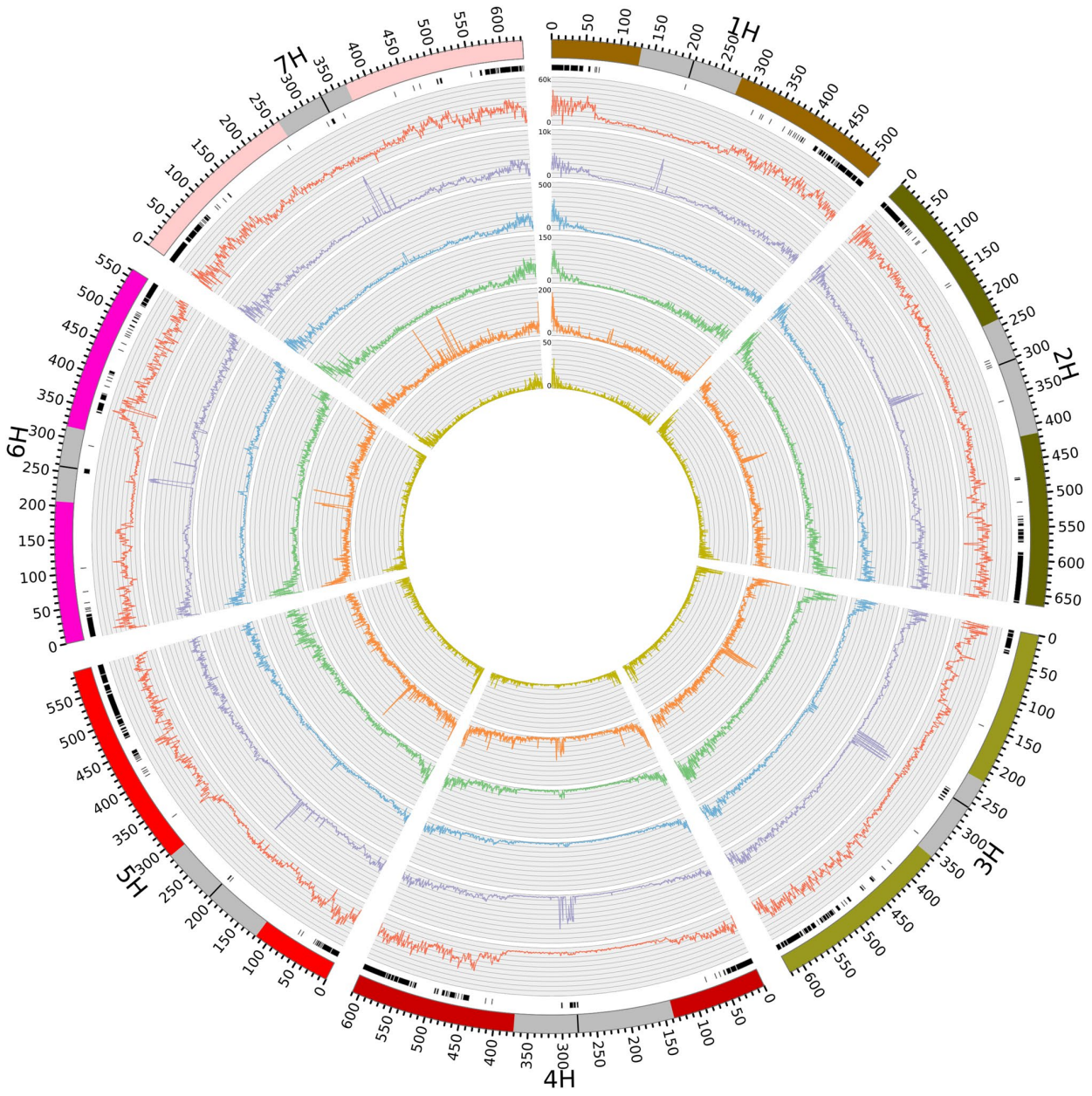

Fig. 2 Distribution of genomic variants among 23 barley inbreds across the seven chromosomes. The outermost circle denotes the chromosome number, the physical position, and as gray bar the pericentromeric regions (Casale et al. 2022) plus the centromeres (black) according to the Morex reference sequence v3. The next inner circles report the SV cluster hotspots (black bars), frequencies of single-nucleotide variants (red), small insertions and deletions (2–49 bp, INDELs, purple), deletions (blue), insertions (green), duplications (orange), and inversions (yellow) which were detected among the 23 inbreds (color figure online)

Discussion

The improvements to sequencing technologies made SV detection in large genomes possible (Della Coletta et al. 2021). Despite these advances, the relative high cost of third compared to second generation sequencing makes the former less affordable and scalable for many research groups. This fact is particularly strong if genotypes have to be analyzed. We therefore used computer simulations to study the precision and sensitivity of SV detection based on different sequencing coverages of short-read sequencing data in the model cereal barley. We also evaluated

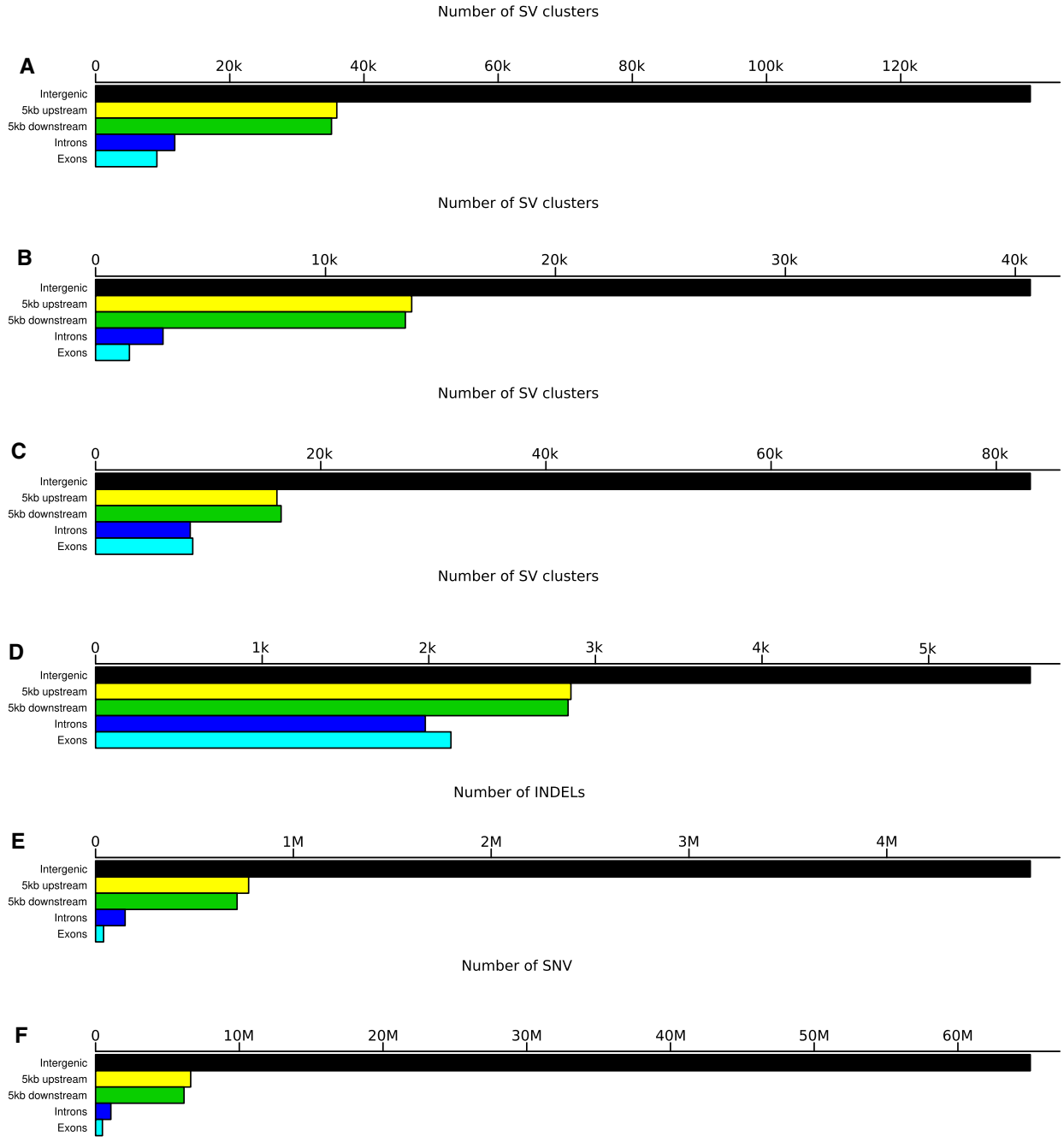

Fig. 3 The occurrence of deletions (**A**), insertions (**B**), duplications (**C**), inversions (**D**), small insertions and deletions (2 –0 49 bp, INDELs, **E**), and single-nucleotide variants (SNV) (**F**) in five genomic regions

whether linked-read sequencing offered by BGI (Wang et al. 2019) or formerly 10x Genomics (Weisenfeld et al. 2017) is advantageous for SV detection compared to classical Illumina sequencing.

Limitations of our study

In our study, the different SV types were always determined in comparison against one reference sequence. The number of insertions present in this reference inbred determines the number of detected deletions and vice versa. However, this

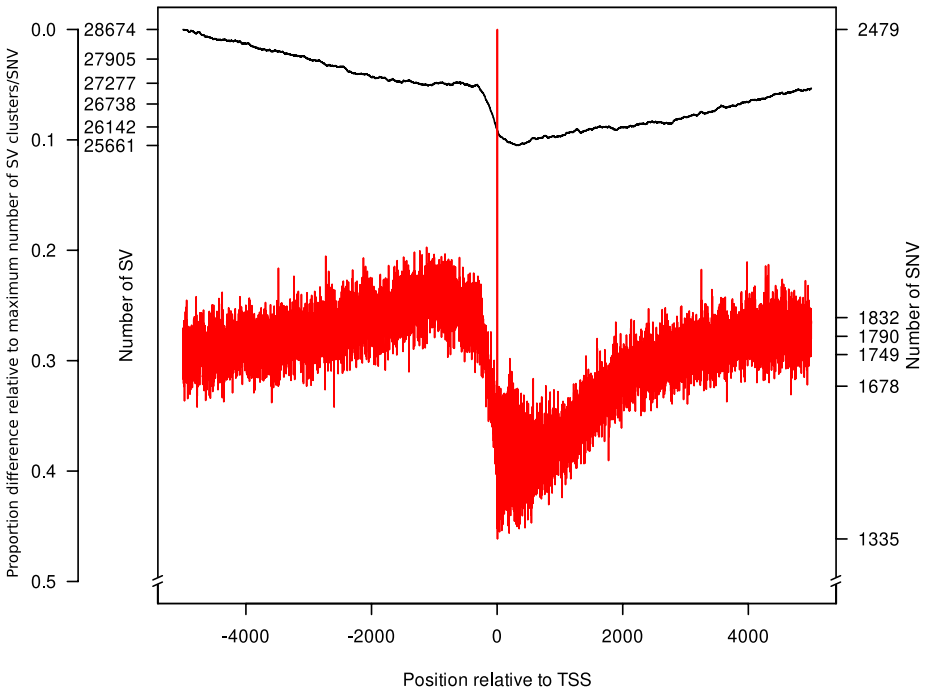

Fig. 4 Distribution of structural variant (SV) clusters (black) and single-nucleotide variants (SNV, red) among 23 barley inbreds relative to the transcription start site (TSS) of a gene (x-axis). SV clusters and SNV were counted for every position from 5kb up- and downstream around the TSS of all genes (y-axes). As third y-axis, the proportion difference relative to the maximum number of SV clusters/SNV is illustrated (color figure online)

is just a matter of nomenclature. Additionally, the usage of short-read sequencing data and only one reference sequence could lead to detect false positive SV calls, due to differences in the mapping efficiency of the evaluated inbreds due to differences in relatedness. In our study, however, the average mapping quality for the 23 inbreds was high and varied only moderately between 41 and 46. Therefore, the influence of the relatedness should be weak. However, this aspect should be considered when interpreting the SV data set.

Precision and sensitivity to detect SV in complex cereal genomes using short-read sequencing data are high

The costs for creating linked-read sequencing libraries is considerably higher compared to that of classical Illumina libraries. Taking this cost difference into account, a fair comparison of precision and sensitivity to detect SV is between 25x Illumina and 14x linked-reads. However, even when directly compared at equal (25x) sequencing coverage, the F1-score, which is the harmonic mean of the precision and sensitivity, on average across all SV types and SV length categories was higher for Illumina compared to linked-reads (Supplementary Fig. S1). One reason might be that the SV callers used in our study do not fully exploit linked-read data. In our study, linked-read information was only used to improve the mapping against the reference genome (Marks et al. 2019). More recently, SV callers have been described that exploit linked information of linked-read data as VALOR2 (Karaoğlanoğlu et al. 2020) or LEVIATHAN (Morisse et al. 2021). However, the SV callers that were available at the time the simulations were performed had a very limited spectrum of SV types and SV length categories they could detect, e.g., LongRanger wgs (Zheng et al. 2016) and NAIBR (Elyanow et al. 2018). In addition, we have observed for these SV callers in first pilot simulations considerably lower values for precision and sensitivity to detect SV compared to the classical short-read SV callers. Therefore, only short-read SV callers were evaluated in detail.

One further aspect that we examined was the influence of the sequencing coverage on sensitivity and precision of SV detection. Only a marginal difference between the F1-scores of the best combination of SV callers for a 25x vs. 65x Illumina sequencing coverage was observed (Supplementary Fig. S1). In addition, for some SV length categories, the F1-score for 25x compared to 65x sequencing coverage was actually higher. A possible explanation for this observation may be that a higher sequencing coverage can lead to an increased number of spuriously aligned reads (Kosugi et al. 2019). These reads can lead to an increased rate of false positive SV detection (Gong et al. 2021). Our result suggests that for homozygous genomes, Illumina short-read sequencing coverage of 25x is sufficient to detect SV with a high precision and sensitivity. We therefore made use of this sequencing coverage not only for further simulations but also to re-sequence the 23 barley inbreds of our study.

In addition, we also tested if a lower sequencing coverage could be used for SV detection to reduce the cost for

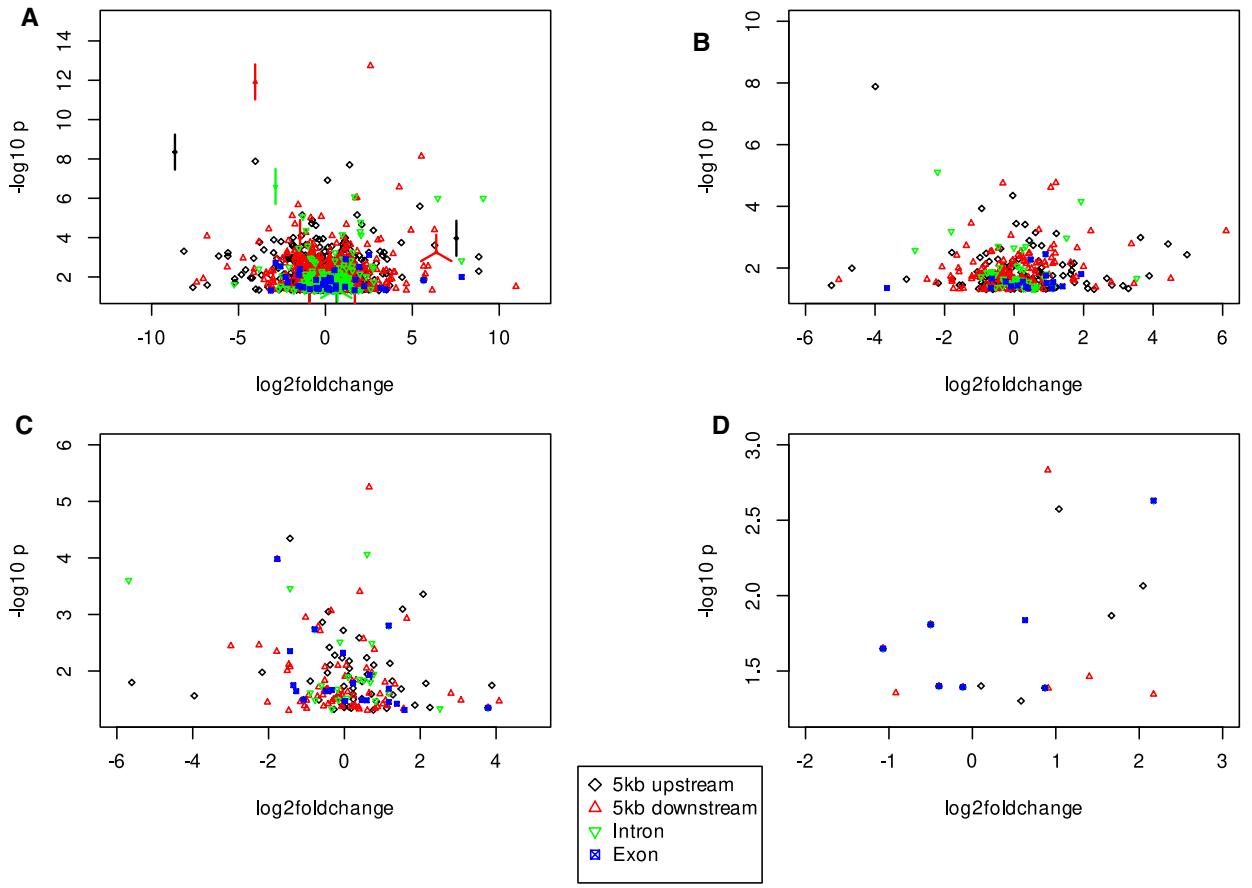

Fig. 5 Association of gene-associated (for details see Material & Methods) deletions (**A**), insertions (**B**), duplications (**C**), and inversions (**D**) with a minor allele frequency > 0.15 with the expression of individual genes assessed using the PK mixed linear model. The gene-associated structural variant (SV) clusters were classified based on their occurrence relative to genes in 5kb up- or downstream, introns, and exons. Values of SV clusters with the same coordinates are illustrated as points with edges, where each edge represents one SV cluster

sequencing further. We observed lower F-scores for all SV types using a sequencing coverage of 12.5x than for 25x (Supplementary Fig. S2). However, the F1-score was still > 80% for all SV types suggesting that even a sequencing coverage of 12.5x would have been sufficient for SV detection in barley. When decreasing the sequencing coverage further, the precision and sensitivity to detect SV decreased considerably. Therefore, a sequencing coverage of 12.5x could be used to detect SV clusters in a small discovery panel as it was performed in our study. In a next step, a larger panel of hundreds of accessions could be used for genotyping the detected SV clusters based on a lower sequencing coverage. However, the performance of such two-step approaches needs first to be evaluated based on computer simulations.

The SV callers evaluated here were chosen based on former benchmarking studies in human (Cameron et al. 2019; Chaisson et al. 2019; Kosugi et al. 2019) as well as rice (Fuentes et al. 2019) and pear (Liu et al. 2020b). Across all SV types and SV length categories, we observed the highest precision and sensitivity for Manta and GRIDSS followed by Pindel with only marginally lower values (Table 2). This finding is in accordance with results of Cameron et al. (2019) for humans. In comparison with the results of Fuentes et al. (2019), we observed a considerably lower sensitivity and precision for Lumpy and NGSEP (Table 2). This difference in performance of the SV callers in rice and barley might be explained by the difference in genome length as well as the high proportion of repetitive elements in the barley genome (Mascher et al. 2017).

Despite the high sensitivity and precision observed for some SV callers, we observed even higher values when using them in combination (Table 2). This can be explained by the different detection principles such as paired-end reads, split reads, read depth, and local assembling that are underlying the different SV callers. Our observation indicates that a combined use of different short-read SV callers is highly

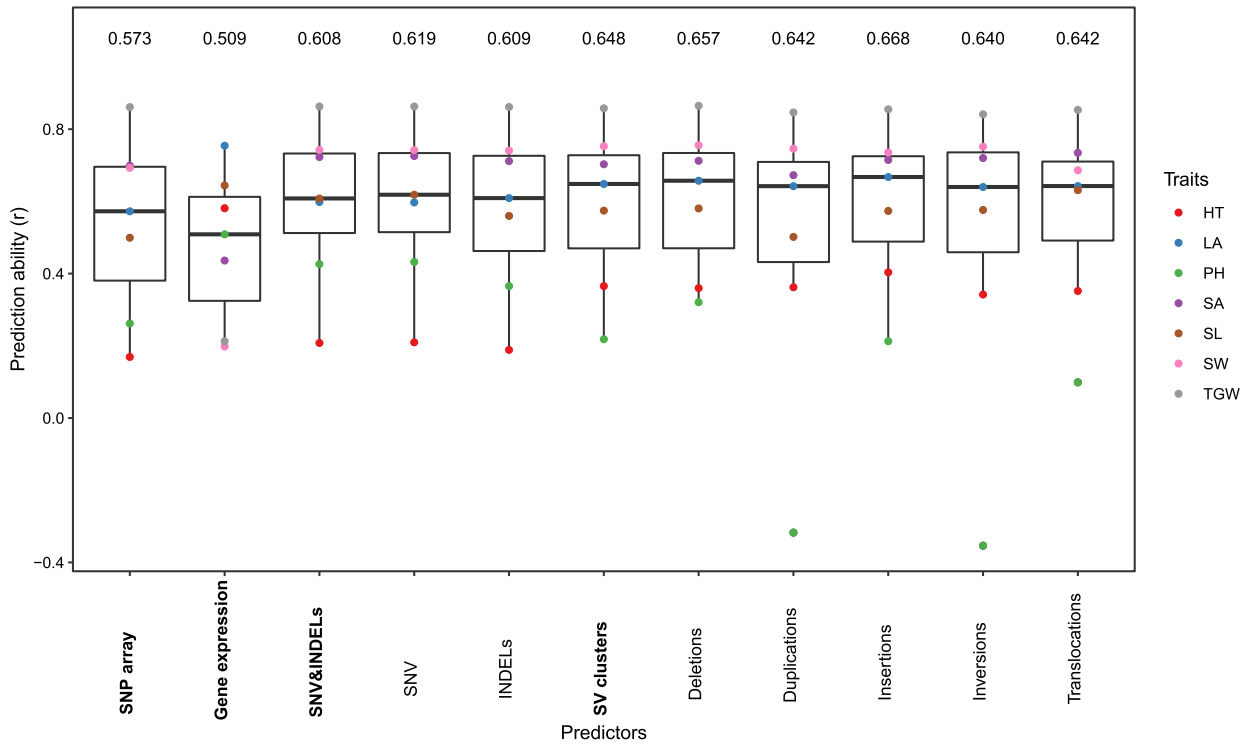

Fig. 6 Boxplot of the median prediction abilities across the seven traits heading time (HT), leaf angle (LA), plant height (PH), seed area (SA), seed length (SL), seed width (SW), thousand grain weight (TGW) based on 23 inbreds using different predictors. The points in each box represent the medians of 200 five-fold cross-validation runs for each trait. The predictors were: features from SNP array, gene expression, single nucleotide variants (SNV) and small insertions and deletions (2–49 bp, INDELs), as well as structural variant (SV) clusters individually as well as combined together

recommended. This approach was then used for SV detection in the set of 23 spring barley inbreds.

Validation of SV in the barley genome

A PCR-based approach was used to validate a small subset of all detected SV. In accordance with earlier studies (Zhang et al. 2015; Yang et al. 2019; Guan et al. 2021), we evaluated the agreement between the detected SV and PCR results (Supplementary Fig. S3) for deletions and insertions up to 0.3 kb (Supplementary Fig. S4). For eleven out of the eleven SV, we observed a perfect correspondence.

Our PCR results further suggested that the SV callers were able to detect eight out of eleven deletions between 0.3 and 460 kb (Supplementary Fig. S5) based on the short-read sequencing of the non-reference parental inbred Unumli-Arpa. In four of the eleven PCR reactions, however, more than one band was observed. This was true three times for the non-reference genotype Unumli-Arpa and one time for Morex (Supplementary Fig. S5B). In two of the four cases, PCR indicated the presence of both SV states in one genome. This was true for Morex as well as Unumli-Arpa and might be due to the complexity of the barley genome which increases the potential for off-target amplification.

In conclusion, for 19 of the 22 tested SV (Supplementary Table S2), the SV detected in the non-reference parental inbred by the SV callers was also validated by PCR. This high validation rate implies in addition to the high precision and sensitivity values observed for SV detection in the computer simulations that the SV detected in the experimental data of the 23 barley inbreds can be interpreted.

Characteristics of SV clusters in the barley gene pool

Across the 23 spring barley inbreds that have been selected out of a world-wide diversity set to maximize phenotypic and genotypic diversity (Weisweiler et al. 2019), we have identified 458,671 SV clusters (Table 3). This corresponds to 1 SV cluster every 9,149 bp and corresponds to what was observed by Jayakodi et al. (2020). This number is in agreement with the number of SV clusters detected for cucumber (9,788 bp^{-1}) (Zhang et al. 2015) or peach (8,621 bp^{-1}) (Guan et al. 2021). Other studies have revealed a higher number of SV clusters than observed in our study. This might be due to the considerably higher number of re-sequenced accessions

in rice (214 bp^{-1}) (Fuentes et al. 2019), tomato (3,291 bp^{-1}) (Alonge et al. 2020), and grapevine (1,260 bp^{-1}) (Zhou et al. 2019).

The highest proportion of SV clusters detected in our study were deletions, followed in decreasing order by translocations, duplications, insertions, and inversions (Table 3). This is in disagreement with earlier studies where the frequency of duplications was considerably lower compared to that of insertions (Zhang et al. 2015; Zhou et al. 2019; Guan et al. 2021). Barley's high proportion of duplications compared to other crops may be due to its high extent of repetitive elements (Mascher et al. 2017).

In contrast to earlier studies in grapevine and peach (e.g., Zhou et al. 2019; Guan et al. 2021) we observed a strong non-uniform distribution of SV clusters across the genome. Only 14.5% of the SV clusters were located in pericentromeric regions, which make up 25.7% of the genome, whereas the rest was located distal of the pericentromeric regions (Fig. 2). This pattern was even more pronounced for SV hotspots, i.e., regions with a significantly ($P < 0.05$) higher amount of SV clusters than expected based on the average genome-wide distribution. Almost all SV hotspots (95.5%) were located distal of the pericentromeric regions (74.3% of the genome) where higher recombination rates are observed. Our observation indicates that the majority of SV clusters in barley might be caused by mutational mechanisms related to DNA recombination-, replication-, and/or repair-associated processes and might be only to a lower extent due to the activity of transposable elements. This is supported by the observation that, with the exception of translocations, only 1.4 to 25.2% of SV clusters were located in genome regions annotated as transposable elements (Table 3).

To complement our genome-wide analysis of barley SV clusters, we also examined their occurrence relative to genes and their association with gene expression.

Association of SV clusters with transcript abundance

About 60% of the SV clusters were detected in the intergenic space (Fig. 3). The remaining SV clusters were gene-associated and detected in regions either 5kb up- or downstream of genes (~30%) while ~10% were detected in introns and exons (Fig. 3). These values are in the range of those previously reported for rice (~75%, NA, exons: ~6%) (Fuentes et al. 2019), potato (~37%, ~37%, ~26%) (Freire et al. 2021), and peach (~52%, ~27%, ~21%) (Guan et al. 2021). The higher proportion of SV clusters in genic regions in potato and peach compared to the cereal genomes might suggest that SV clusters are more frequently associated with gene expression in clonally than in sexually propagated species. A possible explanation for this observation could be the degree of heterozygosity in clonal species, which is considerably higher compared to that in selfing species such as rice and barley. Hence, it is plausible that they better tolerate SV clusters close to genes.

Our study was based on 23 barley inbreds which confer a limited statistical power to detect SV cluster-gene expression associations. However, this leads not to an increased proportion of false positive associations. Therefore, the findings are discussed here.

We observed that the average absolute correlation coefficient of gene-associated SV clusters and global gene expression measured as loadings on the principal components was with 0.17 significantly ($P < 0.05$) different from 0 (Supplementary Fig. S12). In addition, 700 gene-associated SV clusters were individually associated ($P < 0.05$) with genome-wide gene expression. A further 1,976 alleles of gene-associated SV clusters were significantly ($P < 0.05$) associated with the expression of the corresponding 1,594 genes (Fig. 5). Additional support is given by the observation that despite SV clusters have a similar distribution across the genome as SNV, SV clusters covered more positions (in bp) of promoter regions than SNV (Fig. 4). These figures of significantly gene-associated SV clusters are in agreement with earlier figures for tomato (Alonge et al. 2020) and soybean (Liu et al. 2020a) and highlight the high potential of SV clusters to be associated with phenotypic traits.

Genomic prediction

Because of the limited number of inbreds included in this study, the power to identify causal links between SV clusters and phenotypes is low when considering only the 23 inbreds. However, instead of examining the association of individual SV clusters with phenotypic traits, we evaluated their potential to predict seven phenotypic traits in comparison with various other molecular features which is expected to provide reasonable information also with a limited sample size (Weisweiler et al. 2019).

We observed that the ability to predict these seven traits was higher for SV clusters compared to the benchmark data from a SNP array (Fig. 6). This might be explained by the considerably higher number of SV clusters than variants included in the SNP array. However, we observed the same trend when comparing the prediction ability of SV clusters to that of the much more abundant SNV & INDELs. This indicates that the SV clusters comprise genetic information that is not comprised by SNV & INDELs. Our result is supported by the observation that when examining the combination of SNV and INDELs with gene expression and SV clusters to predict phenotypic traits, an increase of the prediction ability was observed compared to the ability observed for the individual predictors (Supplementary Table S5). Furthermore, our observation of a different prediction ability

between SV clusters and SNV & INDELs can be explained by a lower extent of LD between SV clusters and linked SNV compared to that between SNV and linked SNV (Supplementary Table S4). These findings together illustrate the high potential of using SV clusters for the prediction of phenotypes in diverse germplasm sets. Such type of applications might be used also in commercial plant breeding programs. From a cost perspective such approaches will be realistic if SV detection is possible from low coverage sequencing. This might be possible when comprehensive reference sets of SV per species are available as was, for example, generated in our study for barley. However, this requires further research.

Usefulness of SV information for QTL fine mapping and cloning

The inbred lines included in our study are the parents of a new resource for joint linkage and association mapping in barley, the double round robin population (HvDRR, Casale et al. 2022). This population consists of 45 biparental segregating populations with a to total of about 4,000 recombinant inbred lines and is available from the authors upon reasonable request. The detailed characterization of the SV pattern of the parental inbreds, presented in this study, will therefore be an extremely valuable information for the ongoing and future QTL fine mapping and cloning projects exploiting one or multiple of the HvDRR sub-populations.

To illustrate this, we have mapped the naked grain phenotype in six HvDRR sub-populations (HvDRR03, HvDRR04, HvDRR20, HvDRR23, HvDRR44, HvDRR46) to chromosome 7H (7H:525,620,758-525,637,446). Taketa et al. (2008) discovered a 17 kb deletion harboring an ethylene response factor gene on chromosome 7H that caused naked caryopses in barley. In our study, two parental inbreds, namely Kharsila and IG128104, are naked barley. For both inbreds, the SV calls revealed the same 17 kb deletion on chromosome 7H. In contrast, the deletion was absent in the 21 other parental inbreds. This illustrates the potential of exploiting SV information of parental inbreds for gene QTL and gene cloning.

Furthermore, four indels which occur in the 5kb up-/downstream and genic regions of the VRS1 gene were significantly ($P < 0.01$) associated with the rowtype of the parental inbreds.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00122-022-04197-7.

Acknowledgements Computational infrastructure and support were provided by the Center for Information and Media Technology (ZIM) at Heinrich Heine University Düsseldorf.

Author Contribution Statement MW and BS designed and coordinated the project; TH extracted DNA and prepared the libraries; DVI contributed phenotypic data; MW, CA, and PW performed the analyses; MW and BS wrote the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 2048/1, Project ID: 390686111). The funders had no influence on study design, the collection, analysis and interpretation of data, the writing of the manuscript, and the decision to submit the manuscript for publication.

Data Availability Raw DNA sequencing data of the 23 barley inbreds have been deposited into the NCBI Sequence Read Archive (SRA) under the accession PRJNA77700. Raw mRNA sequencing data are available under the accession PRJNA534414. Data of gene expression, SNP array, adjusted entry means of phenotypes, INDELs, and SV clusters are available via figshare (https://doi.org/10.6084/m9.figshare.16802473). SNV data are available via zenodo (https://doi.org/10.5281/zenodo.6451025). Snakemake workflows are available via github (https://github.com/mw-qggp/SV_barley). Further scripts are available from the authors upon request.

Declarations

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

Ethics approval and consent to participate The authors declare that the experimental research on plants described in this paper complied with institutional and national guidelines.

Consent for publication All authors read and approved the final manuscript.

References

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12:363–376

Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, Levy Y, Harel TH, Shalev-Schlosser G, Amsellem Z, Razifard H, Caicedo AL, Tieman DM, Klee H, Kirsche M, Aganezov S, Ranallo-Benavidez TR et al (2020) Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell 182:145-161. e23

Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander

ES, Lee C, Lehrach H, Mardis ER, Marth GT, McVean GA et al (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65

Baker M (2012) Structural variation: the genome's hidden architecture. Nat Methods 9:133–137

Bayer MM, Rapazote-Flores P, Ganal M, Hedley PE, Macaulay M, Plieske J, Ramsay L, Russell J, Shaw PD, Thomas W, Waugh R (2017) Development and evaluation of a barley 50k iSelect SNP array. Front Plant Sci 8:1792

Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y (1997) The complete genome sequence of Escherichia coli K-12. Science 277:1453–1462

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinform 10:421

Cameron DL, Schröder J, Penington JS, Do H, Molania R, Dobrovic A, Speed TP, Papenfuss AT (2017) GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res 27:1–11

Cameron DL, Di Stefano L, Papenfuss AT (2019) Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nat Commun 10:3240

Casale F, Van Inghelandt D, Weisweiler M, Li J, Stich B (2022) Genomic prediction of the recombination rate variation in barley—a route to highly recombinogenic genotypes. Plant Biotechnol J 20:676–690

Chaisson MJ, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, Fan X, Wen J, Handsaker RE, Fairley S, Kronenberg ZN, Kong X, Hormozdiari F, Lee D, Wenger AM, Hastie AR, Antaki D et al (2019) Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun 10:1784

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32:1220–1222

Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, Montgomery SB, Battle A, Conrad DF, Hall IM (2017) The impact of structural variation on human gene expression. Nat Genet 49:692–699

Craig Venter J, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Yuan Wang Z, Wang A, Wang X, Wang J, Wei MH, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu SC, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Lai Cheng M, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Ni Tint N, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Deslattes Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X (2001) The sequence of the human genome. Science 291:1304–1351

Craig-Holmes AP, Moore FB, Shaw MW (1973) Polymorphism of human C-band heterochromatin. I. Frequency of variants. Am J Hum Genet 25:181–192

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158

Della Coletta R, Qiu Y, Ou S, Hufford MB, Hirsch CN (2021) How the pan-genome is changing crop genomics and improvement. Genome Biol 22:3

Duitama J, Quintero JC, Cruz DF, Quintero C, Hubmann G, Foulquié-Moreno MR, Verstrepen KJ, Thevelein JM, Tohme J (2014) An integrated framework for discovery and genotyping of genomic variants from high-throughput sequencing experiments. Nucleic Acids Res 42:e44

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Mari RS, Yilmaz F, Zhao X, Hsieh P, Lee J, Kumar S, Lin J, Rausch T, Chen Y, Ren J, Santamarina M, Höps W, Ashraf H et al (2021) Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science 372:eabf7117

Elyanow R, Wu HT, Raphael BJ (2018) Identifying structural variants using linked-read sequencing data. Bioinformatics 34:353–360

Endelman JB, Jannink JL (2012) Shrinkage estimation of the realized relationship matrix. G3 Genes|Genomes|Genetics 211:1405

Freire R, Weisweiler M, Guerreiro R, Baig N, Hüttel B, Obeng-Hinneh E, Renner J, Hartje S, Muders K, Truberg B, Rosen A, Prigge V, Bruckmüller J, Lübeck J, Stich B (2021) Chromosome-scale reference genome assembly of a diploid potato clone derived from an elite variety. G3 Genes|Genomes|Genetics 11:jkab330

Fuentes RR, Chebotarov D, Duitama J, Smith S, De la Hoz JF, Mohiyuddin M, Wing RA, McNally KL, Tatarinova T, Grigoriev A, Mauleon R, Alexandrov N (2019) Structural variants in 3000 rice genomes. Genome Res 29:870–880

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Life with 6000 genes. Science 274:546–567

Gong T, Hayes VM, Chan EK (2021) Detection of somatic structural variants from short-read next-generation sequencing data. Brief bioinform 22:1–15

Gouesnard B (2001) MSTRAT: an algorithm for building germ plasm core collections by maximizing allelic or phenotypic richness. J Hered 92:93–94

Guan J, Xu Y, Yu Y, Fu J, Ren F, Guo J, Zhao J, Jiang Q (2021) Genome structure variation analyses of peach reveal population dynamics and a 1.67 Mb causal inversion for fruit shape. Genome Biology 22:13

Haseneyer G, Stracke S, Paul C, Einfeldt C, Broda A, Piepho HP, Graner A, Geiger HH (2010) Population structure and phenotypic variation of a spring barley world collection set up for association studies. Plant Breed 129:271–279

Hill WG, Robertson A (1968) Linkage disequilibrium among neutral genes in finite populations. Theor Appl Genet 38:226–231

Jacobs PA, Strong JA (1959) A case of human intersexuality having a possible XXY sex-determining mechanism. Nature 183:302–303

Jayakodi M, Padmarasu S, Haberer G, Bonthala VS, Gundlach H, Monat C, Lux T, Kamal N, Lang D, Himmelbach A, Ens J, Zhang XQ, Angessa TT, Zhou G, Tan C, Hill C, Wang P, Schreiber M, Fiebig A, Budak H, Xu D et al (2020) The barley pan-genome reveals the hidden legacy of mutation breeding. Nature 588:284–289

Karaoğlanoğlu F, Ricketts C, Ebren E, Rasekh ME, Hajirasouliha I, Alkan C (2020) VALOR2: characterization of large-scale structural variants using linked-reads. Genome Biol 21:72

Köster J, Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, Wilm A, Holtgrewe M, Rahmann S, Nahnsen S (2021) Sustainable data analysis with Snakemake. F1000Research 10:33

Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y (2019) Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol 20:117

Kou Y, Liao Y, Toivainen T, Lv Y, Tian X, Emerson JJ, Gaut BS, Zhou Y (2020) Evolutionary genomics of structural variation in asian rice (Oryza sativa) domestication. Mol Biol Evol 37:3507–3524

Kühl MA, Stich B, Ries DC (2021) Mutation-simulator: fine-grained simulation of random mutations in any genome. Bioinformatics 37:568–569

Layer RM, Chiang C, Quinlan AR, Hall IM (2014) LUMPY: a probabilistic framework for structural variant discovery. Genome Biol 15:R84

Li Y, Xiao J, Wu J, Duan J, Liu Y, Ye X, Zhang X, Guo X, Gu Y, Zhang L, Jia J, Kong X (2012) A tandem segmental duplication (TSD) in green revolution gene Rht-D1b region underlies plant height variation. New Phytol 196:282–291

Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou GA, Zhang H, Liu Z, Shi M, Huang X, Li Y, Zhang M, Wang Z, Zhu B, Han B, Liang C, Tian Z (2020) Pan-genome of wild and cultivated soybeans. Cell 182:162–176

Liu Y, Zhang M, Sun J, Chang W, Sun M, Zhang S, Wu J (2020) Comparison of multiple algorithms to reliably detect structural variants in pears. BMC Genomics 21:61

Luo R, Sedlazeck FJ, Darby CA, Kelly SM, Schatz MC (2017) LRSim: a linked-reads simulator generating insights for better genome partitioning. Comput Struct Biotechnol J 15:478–484

Mahmoud M, Gobet N, Cruz-dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ (2019) Structural variant calling: the long and the short of it. Genome Biol 20:246

Manolov G, Manolov Y (1972) Marker band in one chromosome 14 from Burkitt lymphomas. Nature 237:33–34

Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, Bjornson K, Catalanotti C, Delaney J, Fehr A, Fiddes IT, Galvin B, Heaton H, Herschleb J, Hindson C, Holt E, Jabara CB, Jett S, Keivanfar N, Kyriazopoulou-Panagiotopoulou S, Lek M et al (2019) Resolving the full spectrum of human genome variation using linked-reads. Genome Res 29:635–645

Maron LG, Guimarães CT, Kirst M, Albert PS, Birchler JA, Bradbury PJ, Buckler ES, Coluccio AE, Danilova TV, Kudrna D, Magalhaes JV, Piñeros MA, Schatz MC, Wing RA, Kochian LV (2013) Aluminum tolerance in maize is associated with higher MATE1 gene copy number. Proc Natl Acad Sci U S A 110:5241–5246

Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, Bayer M, Ramsay L, Liu H, Haberer G, Zhang XQ, Zhang Q, Barrero RA, Li L, Taudien S, Groth M, Felder M, Hastie A, Šimková H, Staňková H, Vrána J, Chan S, Muñoz-Amatriaín M, Ounit R, Wanamaker S, Bolser D, Colmsee C, Schmutzer T, Aliyeva-Schnorr L, Grasso S, Tanskanen J, Chailyan A, Sampath D, Heavens D, Clissold L, Cao S, Chapman B, Dai F, Han Y, Li H, Li X, Lin C, Mccooke JK, Tan C, Wang P, Wang S, Yin S, Zhou G, Poland JA, Bellgard MI, Borisjuk L, Houben A, Doležel J, Ayling S, Lonardi S, Kersey P, Langridge P, Muehlbauer GJ, Clark MD, Caccamo M, Schulman AH, Mayer KFX, Platzer M, Close TJ, Scholz U, Hansson M, Zhang G, Braumann I, Spannagl M, Li C, Waugh R, Stein N (2017) A chromosome conformation capture ordered sequence of the barley genome. Nature 544:427–433

Mascher M, Wicker T, Jenkins J, Plott C, Lux T, Koh CS, Ens J, Gundlach H, Boston LB, Tulpová Z, Holden S, Hernández-Pinzón I, Scholz U, Mayer KF, Spannagl M, Pozniak CJ, Sharpe AG, Simková H, Moscou MJ, Grimwood J, Schmutz J, Stein N (2021) Long-read sequence assembly: a technical evaluation in barley. Plant Cell 33:1888–1906

McColgan P, Tabrizi SJ (2018) Huntington's disease: a clinical review. Eur J Neurol 25:24–34

Mitelman F, Catovsky D, Manolova Y (1979) Reciprocal 8;14 translocation in EBV-negative B-cell acute lymphocytic leukemia with Burkitt-type cells. Int J Cancer 24:27–33

Monat C, Padmarasu S, Lux T, Wicker T, Gundlach H, Himmelbach A, Ens J, Li C, Muehlbauer GJ, Schulman AH, Waugh R, Braumann I, Pozniak C, Scholz U, Mayer KF, Spannagl M, Stein N, Mascher M (2019) TRITEX: chromosome-scale sequence assembly of Triticeae genomes with open-source tools. Genome Biol 20:284

Morisse P, Legeai F, Lemaitre C (2021) LEVIATHAN: efficient discovery of large structural variants by leveraging long-range information from Linked-Reads data. bioRxiv. https://doi.org/10.1101/2021.03.25.437002

Muñoz-Amatriaín M, Eichten SR, Wicker T, Richmond TA, Mascher M, Steuernagel B, Scholz U, Ariyadasa R, Spannagl M, Nussbaumer T, Mayer KFX, Taudien S, Platzer M, Jeddeloh JA, Springer NM, Muehlbauer GJ, Stein N (2013) Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. Genome Biol 14:R58

Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD, Martínez-García PJ, Vasquez-Gross HA, Lin BY, Zieve JJ, Dougherty WM, Fuentes-Soriano S, Wu LS, Gilbert D, Marçais G, Roberts M, Holt C, Yandell M, Davis JM, Smith KE, Dean JF, Lorenz WW, Whetten RW, Sederoff R, Wheeler N, McGuire PE, Main D, Loopstra CA, Mockaitis K, DeJong PJ, Yorke JA, Salzberg SL, Langley CH (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol 15:R59

Nishida H, Yoshida T, Kawakami K, Fujita M, Long B, Akashi Y, Laurie DA, Kato K (2013) Structural variation in the 5' upstream region of photoperiod-insensitive alleles Ppd-A1a and Ppd-B1a identified in hexaploid wheat (Triticum aestivum L.), and their effect on heading time. Mol Breed 31:27–37

Nowell P, Hungerford D (1960) Chromosome studies on normal and leukemic human leukocytes. J Natl Cancer Inst 25:85–109

Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, Shakir K, Thibault J, Chandran S, Whelan C, Lek M, Gabriel S, Daly MJ, Neale B, MacArthur DG, Banks E (2017) Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. https://doi.org/10.1101/201178

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28:333–339

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci 74:5463–5467

Schrag TA, Westhues M, Schipprack W, Seifert F, Thiemann A, Scholten S, Melchinger AE (2018) Beyond genomic prediction: combining different types of omics data can improve prediction of hybrid performance in maize. Genetics 208:1373–1385

Schüle B, McFarland KN, Lee K, Tsai YC, Nguyen KD, Sun C, Liu M, Byrne C, Gopi R, Huang N, Langston JW, Clark T, Gil FJJ, Ashizawa T (2017) Parkinson's disease associated with pure ATXN10 repeat expansion. npj Parkinson's Dis 3:27

Stich B, Möhring J, Piepho HP, Heckenberger M, Buckler ES, Melchinger AE (2008) Comparison of mixed-model approaches for association mapping. Genetics 1783:1745–1754

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MHY, Konkel MK, Malhotra A, Stütz AM, Shi X, Casale FP, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJ, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HY, Mu XJ, Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal E, Ding L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lameijer EW, McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalin AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO (2015) An integrated map of structural variation in 2,504 human genomes. Nature 526:75–81

Sutton T, Baumann U, Hayes J, Collins NC, Shi BJ, Schnurbusch T, Hay A, Mayo G, Pallotta M, Tester M, Langridge P (2007) Boron-toxicity tolerance in barley arising from efflux transporter amplification. Science 318:1446–1449

Taketa S, Amano S, Tsujino Y, Sato T, Saisho D, Kakeda K, Nomura M, Suzuki T, Matsumoto T, Sato K, Kanamori H, Kawasaki S, Takeda K (2008) Barley grain with adhering hulls is controlled by an ERF family transcription factor gene regulating a lipid biosynthesis pathway. Proc Natl Acad Sci U S A 105:4062–4067

The Arabidopsis Genome Iniative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 48:796–815

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3-new capabilities and interfaces. Nucleic Acids Res 40:e115

VanRaden P (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, Mansueto L, Copetti D, Sanciangco M, Palis KC, Xu J, Sun C, Fu B, Zhang H, Gao Y, Zhao X, Shen F, Cui X, Yu H, Li Z, Chen M, Detras J, Zhou Y, Zhang X, Zhao Y, Kudrna D, Wang C, Li R, Jia B, Lu J, He X, Dong Z, Xu J, Li Y, Wang M, Shi J, Li J, Zhang D, Lee S, Hu W, Poliakov A, Dubchak I, Ulat VJ, Borja FN, Mendoza JR, Ali J, Gao Q, Niu Y, Yue Z, Naredo MEB, Talag J, Wang X, Li J, Fang X, Yin Y, Glaszmann JC, Zhang J, Li J, Hamilton RS, Wing RA, Ruan J, Zhang G, Wei C, Alexandrov N, McNally KL, Li Z, Leung H (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557:43–49

Wang O, Chin R, Cheng X, Yan Wu MK, Mao Q, Tang J, Sun Y, Anderson E, Lam HK, Chen D, Zhou Y, Wang L, Fan F, Zou Y, Xie Y, Zhang RY, Drmanac S, Nguyen D, Xu C, Villarosa C, Gablenz S, Barua N, Nguyen S, Tian W, Liu JS, Wang J, Liu X, Qi X, Chen A, Wang H, Dong Y, Zhang W, Alexeev A, Yang H, Wang J, Kristiansen K, Xu X, Drmanac R, Peters BA (2019) Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. Genome Res 29:798–808

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB (2017) Direct determination of diploid genome sequences. Genome Res 27:757–767

Weisweiler M, Montaigu AD, Ries D, Pfeifer M, Stich B (2019) Transcriptomic and presence/absence variation in the barley genome assessed from multi-tissue RNA sequencing and their power to predict phenotypic traits. BMC Genomics 20:787

Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, Li J, He W, Zhang G, Zheng X, Zhang F, Li Y, Yu C, Kristiansen K, Zhang X, Wang J, Wright M, McCouch S, Nielsen R, Wang J, Wang W (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol 30:105–111

Yang N, Liu J, Gao Q, Gui S, Chen L, Yang L, Huang J, Deng T, Luo J, He L, Wang Y, Xu P, Peng Y, Shi Z, Lan L, Ma Z, Yang X, Zhang Q, Bai M, Li S, Li W, Liu L, Jackson D, Yan J (2019) Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nat Genet 51:1052–1059

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25:2865–2871

Zhang Z, Mao L, Chen H, Bu F, Li G, Sun J, Li S, Sun H, Jiao C, Blakely R, Pan J, Cai R, Luo R, Van de Peer Y, Jacobsen E, Fei Z, Huang S (2015) Genome-wide mapping of structural variations reveals a copy number variant that determines reproductive morphology in cucumber. Plant Cell 27:1595–1604

Zheng X, Medsker B, Forno E, Simhan H, Juan C, Sciences R (2016) Haplotyping germline and cancer genomes using high-throughput linked-read sequencing. Nat Biotechnol 34:303–311

Zhou Y, Minio A, Massonnet M, Solares E, Lv Y, Beridze T, Cantu D, Gaut BS (2019) The population genetics of structural variants in grapevine domestication. Nat Plants 5:965–979

SUPPLEMENTARY INFORMATION

Only the results related to genomic selection of the traits using different structural variants have been put here.

Table S5: The optimal weights of the three predictors single nucleotide variants (SNV) and Indel (SNV&Indel), structural variants (SV) and gene expression that resulted in the highest prediction abilities for the seven traits heading time (HT), leaf angle (LA), plant height (PH), seed area (SA), seed length (SL), seed width (SW), and thousand grain weight (TGW).

Traits	SNV&INDELs	SV clusters	Gene expression	Prediction ability
HT	0.0	0.1	0.9	0.63
LA	0.0	0.4	0.6	0.79
PH	0.0	0.1	0.9	0.54
SA	0.9	0.0	0.1	0.74
SL	0.6	0.0	0.4	0.70
SW	0.0	1.0	0.0	0.75
TGW	1.0	0.0	0.0	0.86
Mean (median)	0.36 (0)	0.23 (0.1)	0.41 (0.4)	

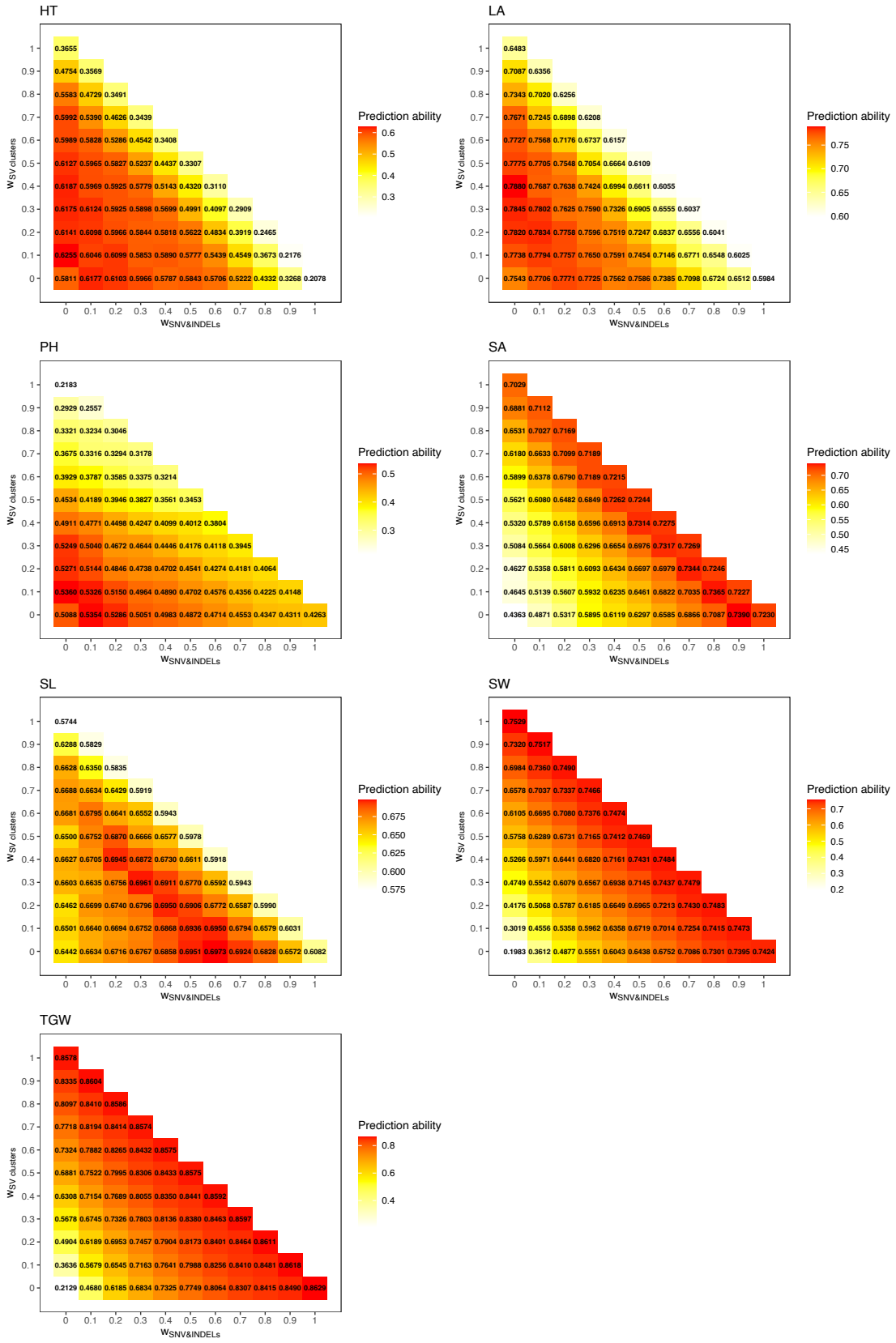

Fig. S13: Prediction ability for the seven phenotypic traits heading time (HT), leaf angle (LA), plant height (PH), seed area (SA), seed length (SL), seed width (SW), and thousand grain weight (TGW) from 23 inbreds for 66 combinations of the joined weighted matrices which differ in the weights of three predictors single nucleotide variants (SNV) and small insertions and deletions (2 - 49bp, INDELs, SNV&INDELs, x-axis), structural variant (SV) clusters (y-axis), and gene expression. Plotted values represent medians across 200 cross-validation runs.

7 List of publications

1. **Wu, P.-Y.**, Stich, B., Weisweiler, M., Shrestha, A., Erban, A., Westhoff, P., and van Inghelandt, D. (2022). Improvement of prediction ability by integrating multi-omic datasets in barley. BMC Genomics, 23:1–15.

2. Weisweiler, M., Arlt, C.*, **Wu, P.-Y.***, van Inghelandt, D., Hartwig, T., and Stich, B. (2022). Structural variants in the barley gene pool: precision and sensitivity to detect them using short-read sequencing and their association with gene expression and phenotypic variation. Theoretical and Applied Genetics, 135:3511–3529.

3. Shrestha, A., Cosenza, F., van Inghelandt, D., **Wu, P.-Y.**, Li, J., Casale, F. A., Weisweiler, M., and Stich, B. (2022). The double round-robin population unravels the genetic architecture of grain size in barley. Journal of Experimental Botany, 73:7344–7361.

4. **Wu, P.-Y.**, Stich, B., Renner, J., Muders, K., Prigge, V., and van Inghelandt, D. (2023). Optimal implementation of genomic selection in clone breeding programs — exemplified in potato: I. Effect of selection strategy, implementation stage, and selection intensity on short-term genetic gain. The Plant Genome, e20327.

5. Cosenza, F., Shrestha, A., van Inghelandt, D., Casale, F. A., **Wu, P.-Y.**, Weisweiler, M., Li, J., Wespel, F., and Stich, B. (2024). Genetic mapping reveals new loci and alleles for flowering time and plant height using the double round-robin population of barley. Journal of Experimental Botany. Accepted.

6. **Wu, P.-Y.**, Stich, B., Renner, J., Muders, K., Prigge, V., and van Inghelandt, D. (2024). Optimal implementation of genomic selection in clone breeding programs - exemplified in potato: II. Effect of selection strategy and cross selection methods on long-term genetic gain. In preparation.

*Contributed equally

8 Acknowledgements

I am very grateful to my two academic supervisors: Prof. Dr. Benjamin Stich and Dr. Delphine van Inghelandt for their advice, suggestions, and support during this thesis work, especially to Dr. Delphine van Inghelandt for many discussions and her never-ending patience in revising my work.

Sincere thanks to Prof. Dr. Friedrich Longin for agreeing to act as my reviewer for my thesis.

Many thanks to Dr. Delphine Van Inghelandt, Prof. Dr. Benjamin Stich, Dr. Juliane Renner, Dr. Katja Muders, Dr. Vanessa Prigge, Dr. Marius Weisweiler, Dr. Asis Shrestha, Alexander Erban, Dr. Philipp Westhoff for being my co-authors of my first author's publications.

Thanks to Prof. Dr. Benjamin Stich, Ines Sigge, Dr. Delphine van Inghelandt, Dr. Asis Shrestha, Dr. Michael Schneider, Dr. Suresh Bontha, Dr. Marius Weisweiler, Nadia Baig, Francesco Cosenza, Yanrong Gao, Federico Casale, Ricardo Guerreiro, Maria Schmidt, Christopher Arlt, Marius Kühl, Alessio Maggiorelli, Anjali Walpola Mudalige Dona, Twinkal Lapasiya, Aashu, Kathrin Thelen, Amelie Kok, Stephanie Krey, Konstantin Shek, and all unmentioned members and former members of the Institute for Quantitative Genetics and Genomics of Plants for creating a great and pleasant work atmosphere.

The financial support from the Federal Ministry of Food and Agriculture (Fachagentur Nachwachsende Rohstoffe) is gratefully acknowledged.

I would like to thank the breeding companies: Böhm-Nordkartoffel Agrarproduktion GmbH & Co. OHG, NORIKA GmbH, and SaKa Pflanzenzucht GmbH & Co. KG for the good collaborations.

I also sincerely thank Dr. Sigrun Wegener–Feldbrügge and her colleagues at the Junior Scientist and International Researcher Center (JUNO) for supporting the organizational and official matters, making my life in Germany easier and smoother.

I want to thank my dear friends: Meng-Ying Lin, Wei-Yun Lai, Chih-Yun Chen, Pei-Tzu Kao, Jen-Hsiang Ou, Tien-Cheng Wang, Sharry Huang, Shao-Pu Tsai, Wei-Jiun Lin, Yi-Chieh Chen, Tianyu Lan, Allegra Grappadelli, Shanny Fan, and all the other friends who I do not have space to name for always encouraging, supporting, and believing in me. Without them, my life would have been much dimmer. Lastly, I am grateful to my family for their unconditional love and support. Thank you all for always being there for me.