GigaScience, 2024, **14**, 1–9
DOI: 10.1093/gigascience/giae121
Data Note

High-quality phenotypic and genotypic dataset of barley genebank core collection to unlock untapped genetic diversity

Zhihui Yuan [1], Maximilian Rembe[1,2], Martin Mascher [1,3], Nils Stein [1,4], Axel Himmelbach [1], Murukarthick Jayakodi [1], Andreas Börner [1], Klaus Oldach [5], Ahmed Jahoor [6], Jens Due Jensen[6], Julia Rudloff[7], Viktoria-Elisabeth Dohrendorf[8], Luisa Pauline Kuhfus[9], Emmanuelle Dyrszka[9], Matthieu Conte[9], Frederik Hinz[10], Salim Trouchaud[11], Jochen C. Reif [1], and Samira El Hanafi [1,*]

[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), OT Gatersleben, 06466 Seeland, Germany
[2]KWS SAAT SE & Co. KGaA, 37574 Einbeck, Germany
[3]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, 04103 Leipzig, Germany
[4]Crop Plant Genetics, Institute of Agricultural and Nutritional Sciences, Martin-Luther-University of Halle-Wittenberg, 06120 Halle (Saale), Germany
[5]KWS LOCHOW GmbH, 29303 Bergen, Germany
[6]Nordic Seed Germany GmbH, 31688 Nienstädt, Germany
[7]Limagrain GmbH, 31226 Peine-Rosenthal, Germany
[8]Nordsaat Saatzucht GmbH, Zuchtstation Gudow, D-23899 Gudow, Germany
[9]Syngenta France SAS, 31790, Saint-Sauveur, France
[10]Saatzucht Bauer GmbH & CO.KG, 93083 Obertraubling, Germany
[11]Secobra Saatzucht GmbH, 85368 Moosburg an der Isar, Germany
*Correspondence address. Samira El Hanafi, Breeding research departement, Leibniz Institute of Plant Genetics and Crop Plant Research, 06466 Gatersleben, Germany. E-mail: hanafi@ipk-gatersleben.de

Abstract

Background: Genebanks around the globe serve as valuable repositories of genetic diversity, offering not only access to a broad spectrum of plant material but also critical resources for enhancing crop resilience, advancing scientific research, and supporting global food security. To this end, traditional genebanks are evolving into biodigital resource centers where the integration of phenotypic and genotypic data for accessions can drive more informed decision-making, optimize resource allocation, and unlock new opportunities for plant breeding and research. However, the curation and availability of interoperable phenotypic and genotypic data for genebank accessions is still in its infancy and represents an obstacle to rapid scientific discoveries in this field. Therefore, effectively promoting FAIR (i.e., findable, accessible, interoperable, and reusable) access to these data is vital for maximizing the potential of genebanks and driving progress in agricultural innovation.

Findings: Here we provide whole genome sequencing data of 812 barley (*Hordeum vulgare* L.) plant genetic resources and 298 European elite materials released between 1949 and 2021, as well as the phenotypic data for 4 disease resistance traits and 3 agronomic traits. The robustness of the investigated traits and the interoperability of genomic and phenotypic data were assessed in the current publication, aiming to make this panel publicly available as a resource for future genetic research in barley.

Conclusions: The data showed broad phenotypic variability and high association mapping potential, offering a key resource for identifying genebank donors with untapped genes to advance barley breeding while safeguarding genetic diversity.

Keywords: Barley, plant genetic resources, elite, whole genome resequencing, disease resistance, agronomic traits

Data Description

Context

Successful plant breeding programs rely on balanced efforts between short-term goals to develop competitive cultivars and the maintenance of a broad genetic pool to guarantee long-term progress. In practice, the development of new varieties has been predominantly derived by recycling existing elite lines, leading to important genetic improvement and the reduction in the genetic diversity of elite germplasm. This could impede the breeding of potential new varieties capable of addressing and responding to constraints related to climate change, agronomical threads, and meeting the escalating social demands [1]. To overcome these limitations, leveraging genetic diversity harbored within plant genetic resources (PGRs) has been frequently suggested [2]. PGRs provide a valuable reservoir of untapped genetic potential that can be utilized to develop varieties with improved yield [3] and end-use quality, as well as enhanced resistance to both biotic and abiotic stresses, such as diseases [4], pests [5], waterlogging [6], salinity [7], and drought [6, 8].

As the most cost-effective *ex situ* conservation strategy, genebanks worldwide are committed to maintaining PGRs, which hold a diverse gene pool encompassing all the alleles of various genes, including those from wild species, landraces, and breeding stocks. However, although enormous efforts have been made

Received: September 26, 2024. **Revised:** December 5, 2024. **Accepted:** December 18, 2024

to conserve germplasm [9], it is estimated that less than 1% of the resources preserved in genebanks have been used in crop improvement [10]. The great challenge for breeders and scientists lies in finding useful barley PGRs among entire genebank collections that comprise thousands of accessions with complex patterns of genetic diversity [11]. Therefore, core collections were proposed as a strategy to streamline operational processes and mitigate costs, thereby facilitating more precise and effective research and breeding initiatives. Over the past decades, this approach has become even more attractive thanks to recent technological advancements, which have markedly reduced the costs of genotyping and led to dramatic improvements in read length, sequencing chemistry, instrumentation, and throughput [12]. As a result, generating large-scale sequencing and genotyping datasets for entire genebank collections is now feasible. This has greatly expanded the scope of genotyping efforts and underpinned the effective selection of core collections that maximize genetic diversity [13]. These advancements provide powerful tools to efficiently harness PGRs, enabling the identification of valuable and favorable genes. This has streamlined their incorporation into crop improvement efforts, ultimately speeding up the development of new and improved varieties. Coupled with extensive and high-quality phenotypic data, the systematic use of whole genome sequencing data could provide valuable insights into genetic diversity and potential breeding opportunities. Our recent findings using genome-wide association analyses highlighted the value of these data in selecting donors with potentially novel favorable genes [14].

Moreover, the strategic deployment of core collections becomes even more compelling when combined with modern elite material [15, 16], which serves as a reference panel to define favorable alleles/genes that are absent in the elite panel. This integrated approach is essential for enhancing polygenic traits and, hence, achieving informed prebreeding decisions. To put this into practice, we selected a barley core collection [17] from the German Federal *ex situ* Genebank for Agriculture and Horticultural Crops at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) and combined it with a set of European elite material. This population was designed to (i) phenotype the whole population in multienvironmental trials for 3 agronomical traits: plant height (PLH), heading date (HD), and lodging (LOD) and 4 disease traits: *Puccinia hordei* (PUC), *Blumeria graminis hordei* (BLU), *Ramularia collo-cygni* (RAM), and *Rhynchosporium commune* (RHY); (ii) evaluate the interoperability quality for the phenotypic and genomic datasets using 5-fold cross-validation; and (iii) conduct the mantel test to check the detection power in association mapping analyses.

The data presented here can be further extended with additional PGRs and/or elite materials. They can also be integrated with alternative strategies to improve the utilization of germplasm collections by selecting untapped PGR donors, such as the development of novel association mapping methods. This will enable breeders to make more accurate predictions of trait performance, thereby enhancing the efficiency of selection processes. Furthermore, with the development of publicly accessible resources, scientists will be able to focus more on research and innovation while reducing the burden of extensive phenotyping. The insights derived from our data may significantly accelerate advancements in genomic research and breeding programs, driving improvement and fostering future collaboration and resource sharing.

Methods

Barley material and field trials

To capture a broad spectrum of geographic origins and wide genetic diversity, we selected 812 PGRs, which include 288 spring type (PGR_Spring) and 524 winter type (PGR_Winter), originating from 57 countries spanning 5 continents. Based on their performance during seed regeneration, these PGRs were thoughtfully selected from a previously described barley core 1000 collection [17], as a representative subset of the entire 21,405 barley accessions available at the IPK genebank [18], based on their performance during seed regeneration. Additionally, we incorporated 298 elite lines, including 10 local checks, which consist of 128 spring type (Elite_Spring) and 170 winter type (Elite_Winter). These elites were exclusively selected from the European registered varieties and were available through the seed market, showcasing the breeding process over time from 1949 to 2021. The study initially included 87 additional genotypes that were later excluded from certain analyses due to incomplete phenotypic or genotypic data. To maintain the integrity of the dataset and facilitate accurate adjustments for experimental design effects, we retained all relevant data, including instances of missing information.

Field trials were conducted over 3 consecutive years (2020, 2021, and 2022) across 8 locations in Germany: KWS-L/Prosselsheim (49°51′15.6″N, 10°06′04.1″E; 10.9°C average annual temperature; 565.3 mm average annual rainfall), Nordic Seed/Nienstädt (52°17′35.52″N, 9°08′57.156″E; 10.7°C average annual temperature; 638.4 mm average annual rainfall), Saatzucht Bauer/Riekofen (48°54′55.98″N, 12°21′21.744″E; 9.9°C average annual temperature; 690.3 mm average annual rainfall), Limagrain/Peine-Rosenthal (52°18′09.828″N, 10°10′28.488″E; 10.9°C average annual temperature; 607.5 mm average annual rainfall), Nordsaat/Gudow (53°33′28.0″N, 10°47′50.5″E; 10.4°C average annual temperature; 581.1 mm average annual rainfall), Syngenta/Bad Salzuflen (52°04′21.576″N, 8°41′55.86″E; 10.5°C average annual temperature; 692.8 mm average annual rainfall), Secobra-LEM/Lemgo (52°00′41.6″N, 8°52′22.7″E; 10.7°C average annual temperature; 714.3 mm average annual rainfall), and Secobra-FK/Moosburg (48°28′46.8″N, 11°54′32.6″E; 10.7°C average annual temperature; 743.1 mm average annual rainfall). The trials were sown following a generalized alpha lattice design, which organizes genotypes into incomplete blocks to minimize spatial variation. Two-row observation plots (1 m^2) with 2 replications were used, and 10 checks were included across years and locations for consistency. Each unique combination of year and location was considered a distinct environment.

Phenotyping

The whole population was phenotyped for 3 agronomy traits for their importance in barley adaptability, yield potential, and harvestability: heading date measured in days from January 1 for winter type and from the sowing date onward for the spring type, plant height measured from the soil surface to the tip of spike in centimeters (excluding awns), and lodging rated on a 1–9 scale (with a higher score indicating severe lodging). Additionally, 4 disease traits, including *P. hordei*, *B. graminis hordei*, *R. collo-cygni*, and *R. commune*, were evaluated under natural infection conditions. The disease severities were scored using an ordinal scale from 1 (fully resistant) to 9 (fully susceptible) following the guidelines of the German Federal Plant Variety Office [19].

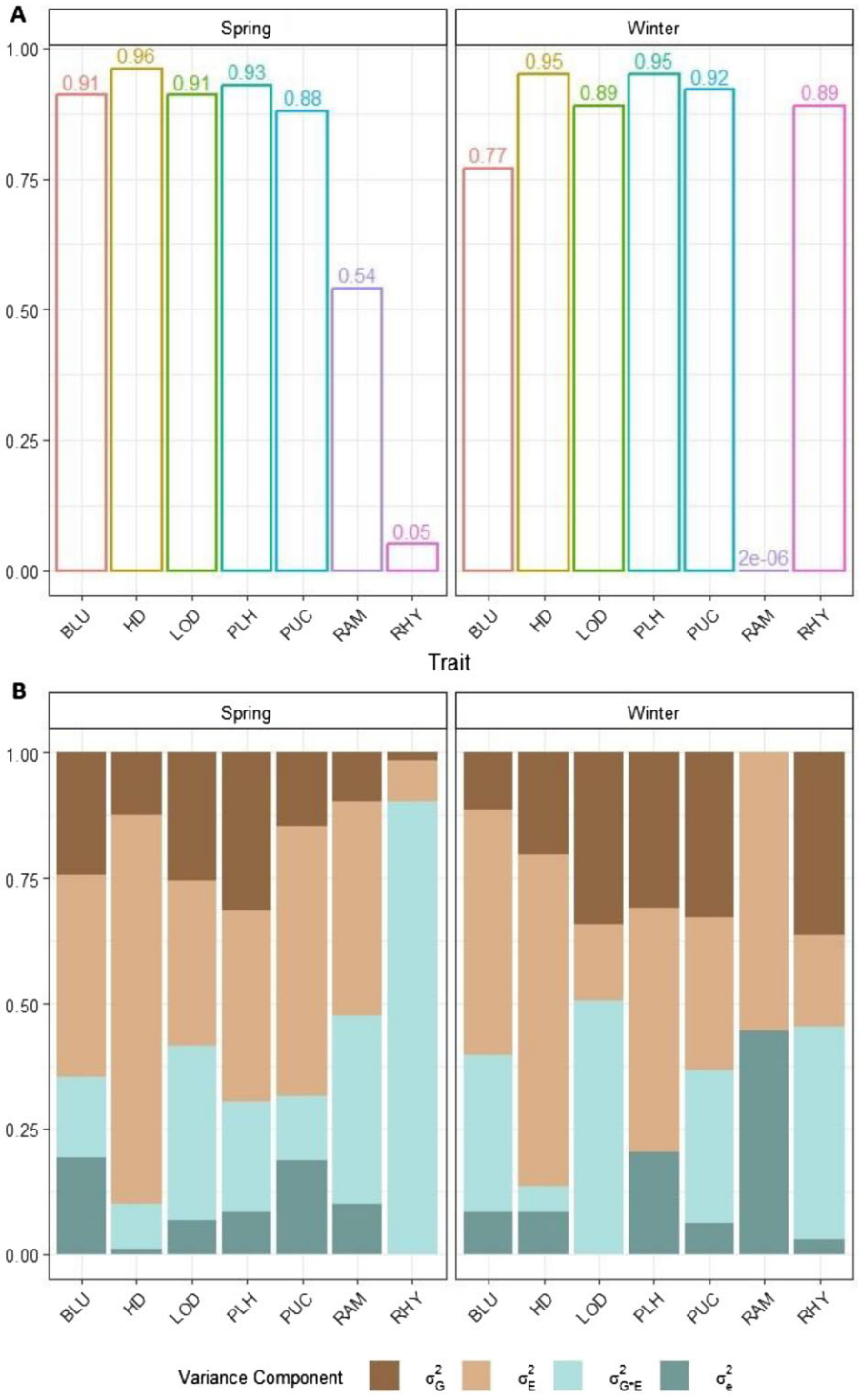

Figure 1: Heritability (A) and percentages of the different variance components (B) for the 7 traits considered in this study. BLU: *Blumeria graminis hordei*; HD: heading date; LOD: lodging; PLH: plant height; PUC: *Puccinia hordei*; RAM: *Ramularia collo-cygni*; RHY: *Rhynchosporium commune*; σ^2_G: genotypic variance; σ^2_{G*E}: variance due to genotype × environment interaction; σ^2_E: variance due to environment; σ^2_e: residual.

Phenotypic data analyses

A linear mixed model using restricted maximum likelihood (REML) method [20] was used for data analyses across environments for spring and winter barley separately. Phenotypic data were corrected for outliers following the method of Tukey and Anscombe [21]. The residuals were extracted and then normalized to flag the outliers according to a predefined significance threshold of $P < 0.01$ (Supplementary Table S2). Variance components and best linear unbiased estimations (BLUEs) of each genotype were computed from the outlier-corrected data, following model (1):

$$y_{ijkm} = \mu + E_m + g_i + g_i \times E_m + E_m : r_j \ : b_k + e_{ijkm}, \quad (1)$$

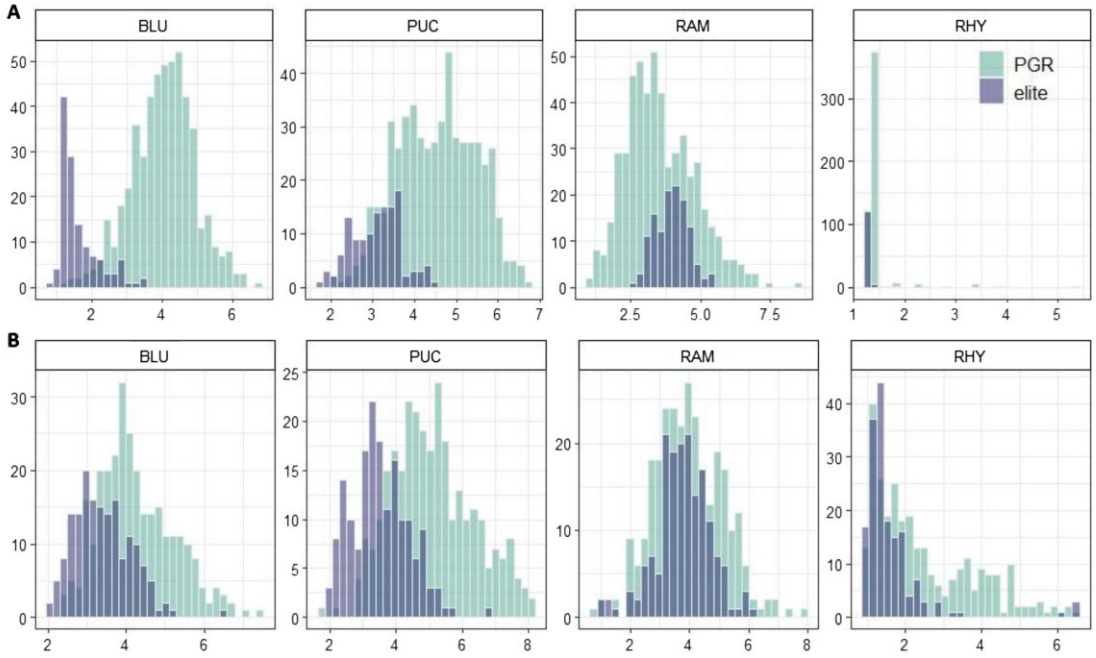

Figure 2: Histogram showing the phenotypic distribution for 4 disease traits for the spring (A) and winter (B) population. BLU: *Blumeria graminis hordei*; PUC: *Puccinia hordei*; RAM: *Ramularia collo-cygni*; RHY: *Rhynchosporium commune*.

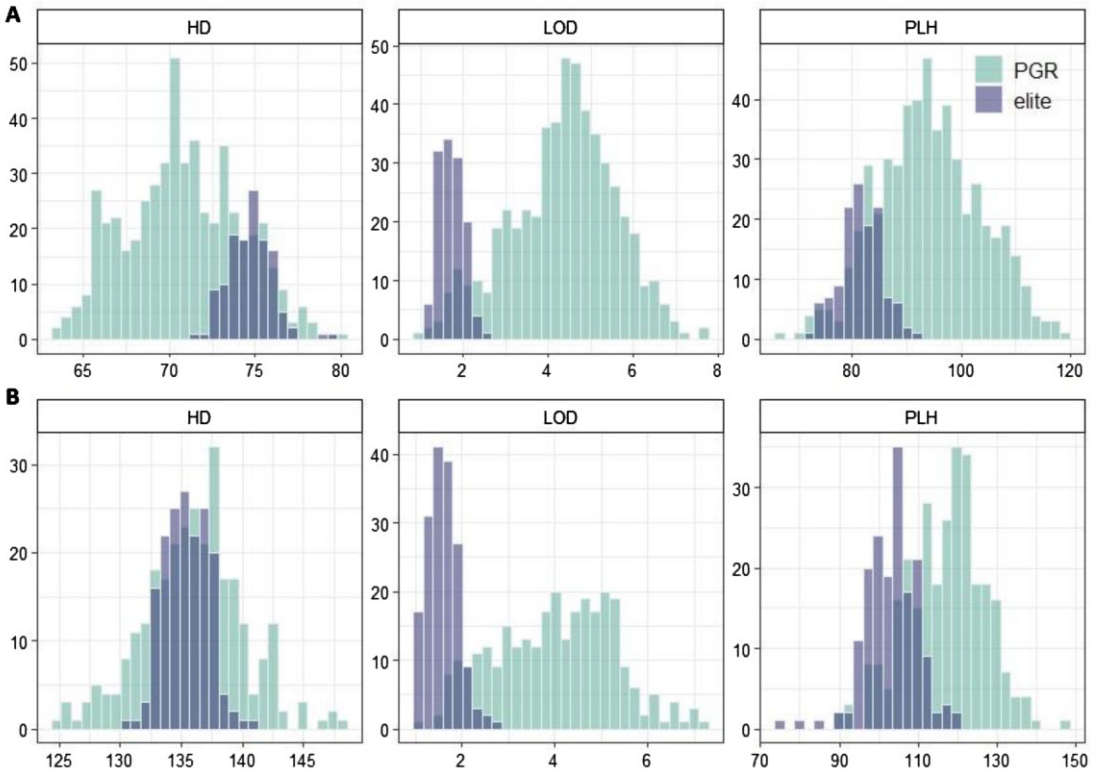

Figure 3: Histogram showing the phenotypic distribution for 3 agronomic traits for the spring (A) and winter (B) population. HD: heading date; LOD: lodging; PLH: plant height.

where y_{ijkm} denoted the vector of phenotypic values for the ith genotype (g) tested in the kth block (b) nested in the jth replication (r) in mth environment (E), μ was the common mean, and e denoted the error term of the model. We assumed that all random effects followed an independent normal distribution with different variance components. In the model (1), all terms except μ and g_i were considered random for deriving the BLUEs across environments, whereas all terms except μ were modeled as random to

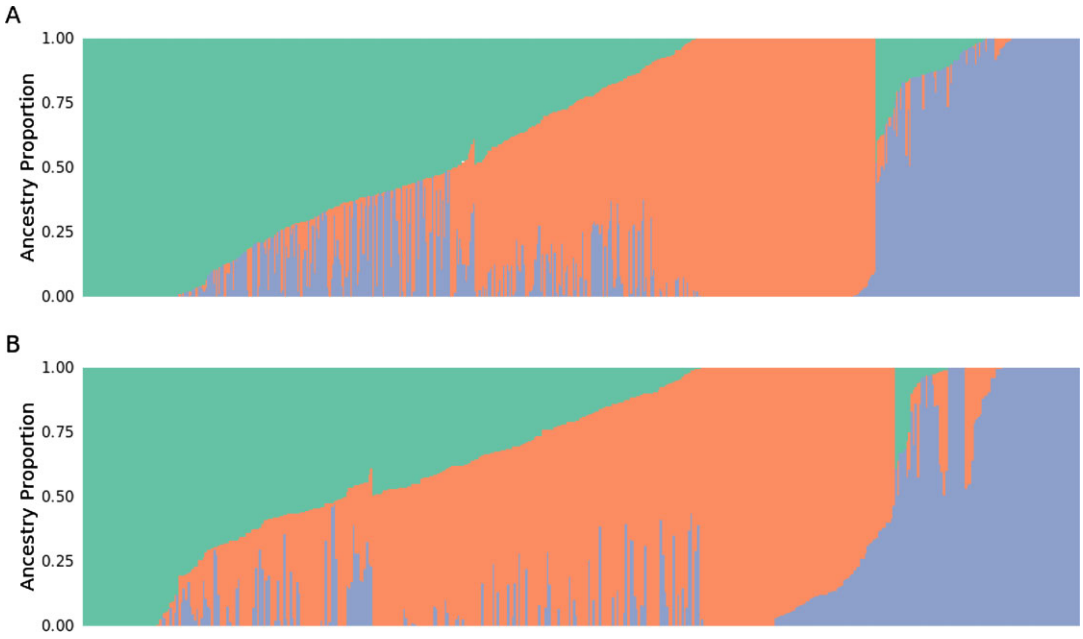

Figure 4: Admixture analysis of the spring (A) and winter (B) populations with the K = 3 admixture model. Each individual is represented as a vertical bar with color corresponding to the proportions of 3 ancestral components (K).

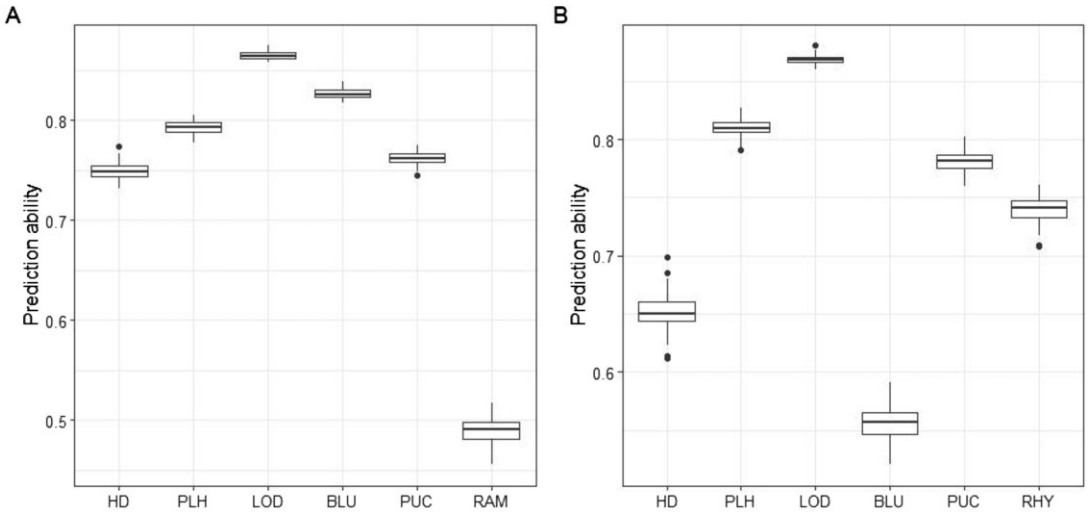

Figure 5: Fivefold cross-validation abilities of the genomic best linear unbiased prediction for heading date (HD; days), plant height (PLH; cm), lodging (LOD), *Blumeria graminis hordei* (BLU), *Puccinia hordei* (PUC), *Rhynchosporium commune* (RHY), and *Ramularia collo-cygni* (RAM), obtained in the spring (A) and winter (B) populations.

estimate the variance component for deriving heritability, following model (2):

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_{g\times E}^2}{\overline{n_E}} + \frac{\sigma_e^2}{\overline{n_R}}}, \quad (2)$$

where σ_g^2 denoted the genotypic variance, $\sigma_{g\times E}^2$ denoted the interaction between genotype and environment, σ_e^2 denoted the residual variance, $\overline{n_R}$ denoted the average number of replications per genotype, and $\overline{n_E}$ denoted the average number of environments in which the genotypes were evaluated. ASReml-R [22] was employed for all mixed linear models that were applied in the phenotypic analysis.

Whole genome shotgun sequencing

Whole genome sequencing (WGS) of the 1,110 genotypes (812 PGRs and 298 elite lines) was performed at IPK Gatersleben. High molecular weight (HMW) DNA was extracted from the leaves (8 g) of greenhouse-grown (21°C/18°C day/night temperature) 7-day-old seedlings following a previously established protocol [23]. The Illumina Nextera libraries were prepared and sequenced using the Illumina NovaSeq 6000 platform [24]. Raw sequencing reads were trimmed using cutadapt [version 3.3; 25] and aligned to the MorexV3 reference genome [26] using Minimap2 [version 2.20; 27]. The resultant alignment records were sorted with Novosort (V3.09.01; http://www.novocraft.com). Finally, a total of 149,380,812 single nucleotide polymorphisms (SNPs) for the

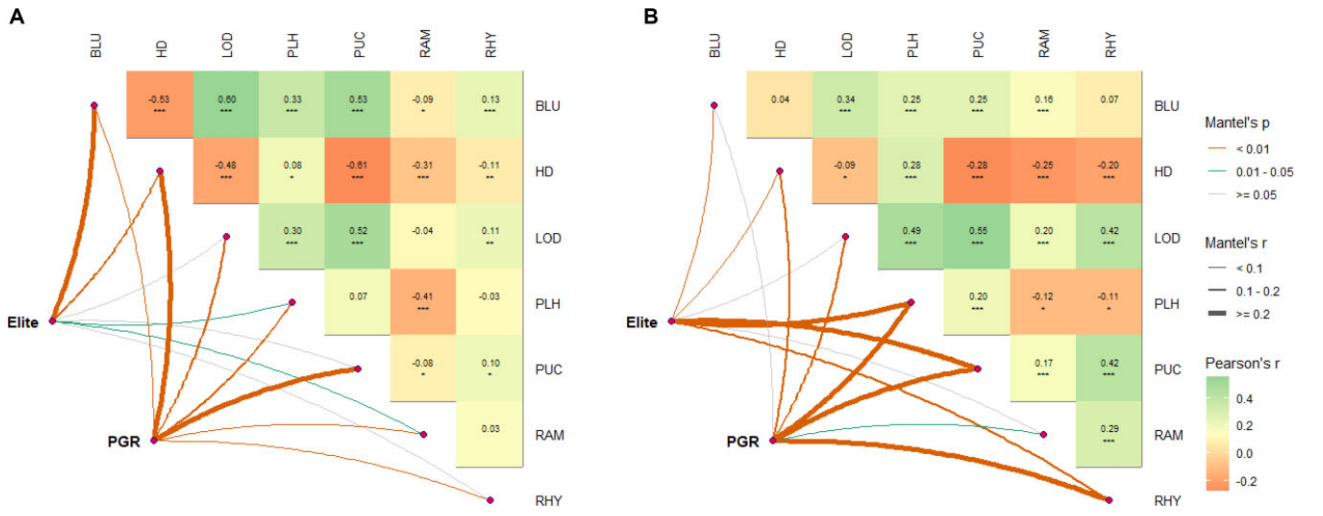

Figure 6: Pairwise correlations for the recorded traits and the Mantel tests between tested traits versus elite materials and plant genetic resources (PGR) for spring barley (A) and winter barley (B). The lines represent significant relationships, where the width of the line represents the Mantel *r* statistic value and the different colors of the lines represent different degrees of significance. The Pearson correlation coefficient between different traits is shown in the heatmap matrix. BLU: *Blumeria graminis hordei*; HD: heading date; LOD: lodging; PLH: plant height; PUC: *Puccinia hordei*; RAM: *Ramularia collo-cygni*; RHY: *Rhynchosporium commune*. *$P < 0.05$. **$P < 0.01$. *** $P < 0.001$.

1,110 genotypes were initially outputted by BCFtools [version 1.9; 28].

Quality control for SNP data

The resulting raw genotypic data were used to extract the corresponding datasets of the 4 subgroups. Only biallelic SNPs with a minor allele frequency >0.05 and missing rate <0.1 were retained by PLINK [version 1.9; 29] for each of the 4 subgroups. These meticulous steps yielded datasets comprising 17,759,260 SNPs for Elite_Spring, 26,903,811 for Elite_Winter, 54,934,336 for PGR_Spring, and 46,434,685 for PGR_Winter.

The resulting filtered genotypic data were used as input to phase and impute missing values using Beagle [version 5.2; 30], leveraging linkage disequilibrium to infer missing data accurately. Subsequently, an r^2 cutoff of 0.2 was set to prune markers by PLINK (version 1.9) with a sliding window size of 50 kb and a step size of 10 kb. The final number of SNPs available differed in the 4 subgroups due to the aforementioned process: 710,855 of Elite_Spring, 945,074 of Elite_Winter, 2,321,327 of PGR_Spring, and 1,775,972 of PGR_Winter. For each tested SNP, homozygous for the most frequent allele, heterozygous, and homozygous for the alternative allele were coded as 0, 1, and 2 by PLINK (version 1.9), respectively.

Population structure

Subsequently, the aforementioned post–quality control markers were used to investigate the population structure within and across spring and winter barley accessions using principal coordinate analysis (PCoA) based on pairwise Rogers's distance [31]. PCoA was performed using the R package ape [version v5.7–1; 32]. Additionally, the population structure was tested using ADMIXTURE [version 1.3.0; 33]. The optimal number of population components was determined based on cross-validation function (–cv).

Moreover, linkage disequilibrium (LD) analyses of the 4 subgroups was carried out separately by determining the pairwise squared allele-frequency correlations (r^2) between markers [34] and then combined to estimate LD decay across the entire genome. A decay curve was fitted for each subgroup using nonlinear regression of pairwise r^2 against the distance (Mb) between the markers. LD within a specific physical distance of 2 Mb was calculated and visualized using PopLDdecay [version 3.40; 34].

Genomic–phenotypic data interoperability

To evaluate the interoperability for the phenotypic and genomic datasets, we calculated the accuracy of the genomic best linear unbiased prediction (GBLUP) [35]. First, the mixed-model equations for genomic prediction were computed using REML in the rrBLUP R package [v4.6.1; 36]. Prediction accuracies were then estimated through 5-fold cross-validation. In this process, both phenotypic and genomic datasets were randomly subdivided into 5 groups. The first 4 groups served together as the training set, whereas the fifth group corresponded to the prediction set. The random sampling was repeated 100 times, giving a total of 500 cross-validation runs. Genomic prediction ability was thereafter defined as the correlation between BLUEs across environments for a trait and the corresponding predicted values.

Mantel correlation

Following the imputation process, we used PLINK (version 1.9) to construct a genetic relationship matrix. To further explore the association between phenotypic variation and population structure, the correlation between the genetic relationship matrix and the absolute trait differences (Euclidean distance matrix) in each subgroup was tested using a Mantel test [37] implemented in the R package vegan [v2.6–4; 36] and visualized by linkET R package [v0.0.7.4; 38]; 999 permutations were used to evaluate the significance of the test.

Data Validation and Quality Control

High heritability estimation highlights the robustness of the phenotypic data

The quality and reliability of the phenotypic data were rigorously assessed by estimating the heritability of the evaluated traits. After outlier correction, the heritability estimates for most traits were generally high, exceeding 0.5 (Fig. 1A). Notable exceptions

included RHY in the spring population ($h^2 = 0.05$) and RAM ($h^2 = 2E\text{-}06$) in the winter population. Variance components analysis revealed that environment (σ_e^2) accounts for the largest proportion of the total variance, while genotype and genotype × environment interaction were less pronounced, with the exception of LOD and RHY in both the spring and winter population, as well as PUC in the winter population (Fig. 1B). This suggests that factors such as temperature fluctuations, varying levels of precipitation, and humidity across different climate zones may have influenced the observed phenotypic performance. These environmental conditions likely influenced growth patterns and trait expression, leading to larger phenotypic variability in traits with high heritability and restricted variability in traits with low heritability.

The resulting BLUEs showed a normal distribution for most disease traits (Fig. 2). However, RHY showed left skew in both spring and winter population, while BLU and PUC displayed left skew in the elite population for both spring and winter type. The left skew of RHY suggests low disease pressure across 3 years, hence resulting in a small proportion of susceptible genotypes. The left skew of elite population of BLU and PUC also suggests that PGRs tend to be more susceptible than the elite materials for the 2 diseases. For agronomic traits (Fig. 3), the PGR population showed a normal distribution, while elite lines showed a normal distribution in HD and PLH only in the winter population.

Furthermore, several significant correlations were observed between the evaluated traits (Fig. 4). For pairing of agronomic and disease traits, it was observed that HD was negatively correlated with all the disease traits, except for BLU in the winter barley population. Those observations suggest a strategic plant response given that delayed heading allows plants to evade disease infection through spatial or temporal adjustments. Moreover, LOD was positively correlated with all the disease traits, except for RAM in the spring barley population. PLH was positively correlated with BLU and PUC while negatively correlated with RAM and RHY.

Whole genome sequencing data show high genetic diversity and high marker densities

WGS data of the 1,110 genotypes showed an average coverage of 4.7×, with a range spanning from 0.5× to 22.6× across all samples with a mapping rate from 94% to 99%, providing a solid foundation for downstream genetic analyses and ensuring a comprehensive representation of the genomic information across the diverse set of genotypes.

Building on this comprehensive genomic dataset, we performed PCoA to assess the genetic diversity among the spring and winter barley population, as reported in our companion study [14]. The first 2 coordinates together explained 11.66% and 11.25% of the spring and winter population, respectively. As anticipated, the inclusion of PGRs significantly broadened the genetic diversity compared to the elite materials. Notably, the elite spring population formed a tight, cohesive cluster, indicating less genetic diversity, while the elite winter population exhibited a more dispersed pattern, reflecting greater genetic variability.

To further complement the population structure analyses, the optimal number of genetic components (K = 3) was determined based on cross-validation results. The admixture analysis revealed distinct population structures within the spring and winter populations (Fig. 4), with individuals showing varying proportions of the 3 inferred components. These results highlight the contrasting levels of genetic diversity and population structure within each spring and winter barley genotype.

For the intrachromosomal decay of LD (r^2), PGR was faster in both the spring and winter population as compared to elite materials. The slower LD decay in the elite population may be due to genetic bottlenecks and/or high selection pressures that produce specific linkage between alleles that control specific phenotypes.

High genomic prediction accuracies support the interoperability of genomic and phenotypic data

Systematic errors can occur during field trials, which will systematically disrupt the connectivity between genotype and phenotype data and, in turn, decrease the value of the data for subsequent integrated analyses. To assess potential data imbalances, we used the cross-validated accuracy of genomic prediction as a quality measure for genomic–phenotypic data interoperability.

Integrating phenotypic data with WGS data resulted in 652 spring and 458 winter barley genotypes. Overall, the genomic–phenotypic data interoperability was in general high (Fig. 5), with a maximum prediction accuracy observed for lodging in both spring and winter populations. Disease-resistant traits showed moderate to high prediction abilities, suggesting that genomic data can be reliably used, thereby potentially accelerate breeding efforts for resistant varieties. In parallel, this robust result ensures reliable data quality, enabling comprehensive analyses to explore genotype–phenotype relationships and lay a solid foundation for future studies aimed at finding marker-trait associations and understanding the genetic mechanisms underlying key traits in barley.

Mantel test results indicate a high detection power in association mapping

Accurate mapping requires addressing the complexities inherent in genetic relatedness among individuals. In such way, especially when dealing with panels comprising both elite lines and PGRs, the intricate patterns of genetic relationship can pose significant challenges. Specially, when phenotype variation is influenced by genetic relatedness, it becomes crucial to differentiate between genuine associations and those resulting from shared genetic backgrounds. This complexity underscores the importance of robust methods to effectively uncover meaningful correlations and enhance the reliability of association mapping. Therefore, by minimizing genotype–phenotype covariance, we can reduce the risk of spurious associations [39]. The Mantel test is a widely used approach to examine the association between 2 matrices. The results revealed a moderate to low correlation between genetic distance and Euclidean phenotypic distance matrix, indicating a lack of strong association between phenotypic variation and genome-wide genetic differences (Fig. 6; Mantel's r ranged from −0.02 to 0.29 in spring barley and from 0 to 0.32 in winter barley), which in turn is expected to increase the detection power in association mapping.

Additional Files

Supplementary Table S1. List of 1,110 genotypes in this dataset.
Supplementary Table S2. The number and proportion of outliers identified for each trait.

Abbreviations

BLU: *Blumeria graminis hordei*; BLUE: best linear unbiased estimations; HD: heading date; IPK: Institute of Plant Genetics and Crop Plant Research; LD: linkage disequilibrium; LOD: lodging; PCoA: principal coordinate analysis; PGR: plant genetic resources; PLH: plant height; PUC: *Puccinia hordei*; RAM: *Ramularia collo-cygni*; RHY:

Rhynchosporium commune; SNP: single nucleotide polymorphism; WGS: whole genome sequencing.

Acknowledgments

We are grateful for the technical assistance of Mary Ziems and Annette Marlow for providing seeds of plant material, Susanne König and Ines Walde for technical assistance during sequencing data production, and Anne Fiebig, Daniel Arend, and Matthias Lange for support with data management and submission to repositories.

Author Contributions

K.O., A.J., J.D.J., J.R., V.D., L.P.K., E.D., M.C., F.H., and S.T.: cultivation and provision of phenotypic data of all spring and winter barleys over 3 years in 1 to 2 locations. Z.Y.: genotypic data analyses and curation. M.R. and S.E.H.: phenotypic data analyses. M.M., M.J., A.H., and N.S.: generated and processed the genomic data. N.S., M.M., A.B., S.E.H., and J.C.R.: edited and revised the manuscript. A.B.: developed the core 1000 population. J.C.R., N.S., S.E.H., and Z.Y.: designed the study. Z.Y. and S.E.H.: wrote the paper. All authors read and approved the final manuscript.

Funding

This research work is funded by German Ministry of Food and Agriculture under the project Structural genome variation, haplotype diversity and the barley pan-genome—Exploring structural genome diversity for barley breeding (SHAPE) phases 1 and 2 (BMBF FKZ 031B0190A; 031B0884A).

Data Availability

Phenotypic records: The raw phenotypic data described here as well as the ready-to-use phenotypic values (BLUEs) and the R script to import and curate the raw phenotypic data to compute heritability and BLUEs are available in the e!DAL-PGP Repository [40] and can be directly accessed here [41].

Raw sequencing reads: FASTQ files containing raw reads for 1,110 genotypes were submitted by [24] and deposited at the European Nucleotide Archive [42] under BioProjects PRJEB53924 (Illumina resequencing data). Sequenced genotypes are findable through their "SAMEA" IDs. The integrated Elite and PGR "SAMEA" BioSample IDs connected with plant material passports, passport data sources, SSD, and IPK genebank DOIs are listed in Supplementary Table S1.

SNP markers: Variant calling results based on read mapping against the reference sequence of MorexV3 were stored as Variant Call Format (VCF). All the VCF files are located at the European Nucleotide Archive under the project number PRJEB80159.

The script for filtering VCF files, imputation, admixture process, mantel test, and cross-validation is accessible at https://github.com/yzh1023/data-publication.git [43]. Other data further supporting this work are openly available in the *GigaScience* repository, GigaDB [44].

Competing Interests

The authors declare that they have no competing interests.

References

1. Ellegren H, Galtier N. Determinants of genetic diversity. Nat Rev Genet 2016;17:422–33. https://doi.org/10.1038/nrg.2016.58.
2. Halewood M, Chiurugwi T, Sackville Hamilton R, et al. Plant genetic resources for food and agriculture: opportunities and challenges emerging from the science and information technology revolution. New Phytol 2018;217:1407–19. https://doi.org/10.1111/nph.14993.
3. Dillon SL, Shapter FM, Henry RJ, et al. Domestication to crop improvement: genetic resources for *Sorghum* and *Saccharum* (Andropogoneae). Ann Bot 2007;100:975–89. https://doi.org/10.1093/aob/mcm192.
4. Deng Y, Ning Y, Yang D, et al. Molecular basis of disease resistance and perspectives on breeding strategies for resistance improvement in crops. Mol Plant 2020;13:1402–19. https://doi.org/10.1016/j.molp.2020.09.018.
5. Radchenko EE, Abdullaev RA, Anisimova IN. Genetic resources of cereal crops for aphid resistance. Plants 2022;11:1490. https://doi.org/10.3390/plants11111490.
6. Valliyodan B, Ye H, Song L, et al. Genetic diversity and genomic strategies for improving drought and waterlogging tolerance in soybeans. J Exp Bot 2016;68:1835–49. https://doi.org/10.1093/jxb/erw433.
7. Razzaq A, Saleem F, Wani SH, et al. De-novo domestication for improving salt tolerance in crops. Front Plant Sci 2021;12:681367. https://doi.org/10.3389/fpls.2021.681367.
8. Missanga JS, Venkataramana PB, Ndakidemi PA. Recent developments in *Lablab purpureus* genomics: a focus on drought stress tolerance and use of genomic resources to develop stress-resilient varieties. Legume Sci 2021;3:e99. https://doi.org/10.1002/leg3.99.
9. Wambugu PW, Ndjiondjop M-N, Henry RJ. Role of genomics in promoting the utilization of plant genetic resources in genebanks. Brief Functional Genomics 2018;17:198–206. https://doi.org/10.1093/bfgp/ely014.
10. Sharma S, Upadhyaya HD, Varshney RK, et al. Pre-breeding for diversification of primary gene pool and genetic enhancement of grain legumes. Front Plant Sci 2013;4:309. https://doi.org/10.3389/fpls.2013.00309.
11. Odong TL, Jansen J, van Eeuwijk FA, et al. Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. Theor Appl Genet 2013;126:289–305. https://doi.org/10.1007/s00122-012-1971-y.
12. Salgotra RK, Chauhan BS. Genetic diversity, conservation, and utilization of plant genetic resources. Genes 2023;14:174. https://doi.org/10.3390/genes14010174.
13. El Hanafi S, Jiang Y, Kehel Z, et al. Genomic predictions to leverage phenotypic data across genebanks. Front Plant Sci 2023;14:1227656. https://doi.org/10.3389/fpls.2023.1227656.
14. Yuan ZH, Rembe M, Mascher M, et al. Capitalizing genebank core collections for rare and novel disease resistance loci to enhance barley resilience. J Exp Bot 2024;75(18):5940–54. https://doi.org/10.1093/jxb/erae283.
15. Cazenave X, Petit B, Lateur M, et al. Combining genetic resources and elite material populations to improve the accuracy of genomic prediction in apple. G3 (Bethesda) 2022;12:jkab420. https://doi.org/10.1093/g3journal/jkab420.
16. Sehgal D, Vikram P, Sansaloni CP, et al. Exploring and mobilizing the gene bank biodiversity for wheat improvement. PLoS One 2015;10:e0132112. https://doi.org/10.1371/journal.pone.0132112.

17. Milner SG, Jost M, Taketa S, et al. Genebank genomics highlights the diversity of a global barley collection. Nat Genet 2019;51:319–26. https://doi.org/10.1038/s41588-018-0266-x.
18. Oppermann M, Weise S, Dittmann C, et al. GBIS: the information system of the German Genebank. Database 2015;2015:bav021. https://doi.org/10.1093/database/bav021.
19. Bundessortenamt. Richtlinien für die durchführung von landwirtschaftlichen wertprüfungen und sortenversuchen. 2000; www.bundessortenamt.de/bsa/sorten/sortenzulassung/richtlinien-fuer-die-durchfuehrung-von-landwirtschaftlichen-wertpruefungen-und-sortenversuchen.
20. Patterson HD, Thompson R. Recovery of inter-block information when block sizes are unequal. Biometrika 1971;58:545–54. https://doi.org/10.1093/biomet/58.3.545.
21. Anscombe FJ, Tukey JW. The examination and analysis of residuals. Technometrics 1963;5:141–60. https://doi.org/10.1080/00401706.1963.10490071.
22. Butler DG, Cullis BR, Gilmour AR, et al. ASReml estimates variance components under a general linear. 2023; https://asreml.kb.vsni.co.uk/wp-content/uploads/sites/3/2018/07/ASReml-Package.pdf.
23. Dvorak J, McGuire PE, Cassidy B. Apparent sources of the A genomes of wheats inferred from polymorphism in abundance and restriction fragment length of repeated nucleotide sequences. Genome 1988;30:680–89. https://doi.org/10.1139/g88-115.
24. Jayakodi M, Lu Q, Pidon H, et al. Structural variation in the pangenome of wild and domesticated barley. Nature 2024;636:654–62. https://doi.org/10.1038/s41586-024-08187-1.
25. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal 2011;17:10. https://doi.org/10.14806/ej.17.1.200.
26. Mascher M, Wicker T, Jenkins J, et al. Long-read sequence assembly: a technical evaluation in barley. Plant Cell 2021;33:1888–906. https://doi.org/10.1093/plcell/koab077.
27. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018;34:3094–100. https://doi.org/10.1093/bioinformatics/bty191.
28. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011;27:2987–93. https://doi.org/10.1093/bioinformatics/btr509.
29. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am Hum Genet 2007;81:559–75. https://doi.org/10.1086/519795.
30. Browning BL, Zhou Y, Browning SR. A one-penny imputed genome from next-generation reference panels. Am Hum Genet 2018;103:338–48. https://doi.org/10.1016/j.ajhg.2018.07.015.
31. Rogers JS. Measures of genetic similarity and genetic distance. In: Wheleer M, ed. Studies in genetics VII. Austin, TX: University of Texas Publication; 1972:145–53.
32. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 2019;35:526–28. https://doi.org/10.1093/bioinformatics/bty633.
33. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 2009;19:1655–64. https://doi.org/10.1101/gr.094052.109.
34. Hill WG, Robertson A. Linkage disequilibrium in finite populations. Theor Appl Genet 1968;38:226–31. https://doi.org/10.1007/BF01245622.
35. Zhang C, Dong SS, Xu JY, et al. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics 2019;35:1786–88. https://doi.org/10.1093/bioinformatics/bty875.
36. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci 2008;91:4414–23. https://doi.org/10.3168/jds.2007-0980.
37. Mantel N. The detection of disease clustering and a generalized regression approach. Cancer Res 1967;27:209–20.
38. Endelman JB. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 2011;4:250–55. https://doi.org/10.3835/plantgenome2011.08.0024.
39. Myles S, Peiffer J, Brown PJ, et al. Association mapping: critical considerations shift from genotyping to experimental design. Plant Cell 2009;21:2194–202. https://doi.org/10.1105/tpc.109.068437.
40. Arend D, Junker A, Scholz U, et al. PGP repository: a plant phenomics and genomics data publication infrastructure. Database 2016;2016:baw033. https://doi.org/10.1093/database/baw033.
41. Yuan ZH, El Hanafi S, Reif J. Diseases resistance and agronomic traits of 853 plant genetic resources and 344 European elite genotypes in multi-environments. e!DAL - Plant Genomics & Phenomics Research Data Repository. 2024. https://doi.org/10.5447/IPK/2024/7. Accessed 12 August 2024.
42. Li W, Cowley A, Uludag M, et al. The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Res 2015;43:W580–84. https://doi.org/10.1093/nar/gkv279.
43. Scripts for "High-quality phenotypic and genotypic dataset of barley genebank core-collection to unlock untapped genetic diversity." 2024; https://github.com/yzh1023/data-publication. Accessed 19 December 2024.
44. Yuan Z, Rembe M, Mascher M, et al. Supporting data for "High-Quality Phenotypic and Genotypic Dataset of Barley Genebank Core Collection to Unlock Untapped Genetic Diversity." GigaScience Database. 2024. https://doi.org/10.5524/102638.

Received: September 26, 2024. **Revised:** December 5, 2024. **Accepted:** December 18, 2024