



OPEN ACCESS

EDITED BY

Axel Diederichsen,
Agriculture and Agri-Food Canada (AAFC),
Canada

REVIEWED BY

Photini V. Mylona,
Hellenic Agricultural Organisation (HAO),
Greece
Svein Øivind Solberg,
Inland Norway University of Applied
Sciences, Norway

*CORRESPONDENCE

Jochen C. Reif

✉ reif@ipk-gatersleben.de

RECEIVED 31 July 2023

ACCEPTED 31 October 2023

PUBLISHED 11 January 2024

CITATION

Berkner MO, Weise S, Reif JC and
Schulthess AW (2024) Genomic prediction
reveals unexplored variation in grain
protein and lysine content across a vast
winter wheat genebank collection.
Front. Plant Sci. 14:1270298.
doi: 10.3389/fpls.2023.1270298

COPYRIGHT

© 2024 Berkner, Weise, Reif and Schulthess.
This is an open-access article distributed
under the terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Genomic prediction reveals unexplored variation in grain protein and lysine content across a vast winter wheat genebank collection

Marcel O. Berkner¹, Stephan Weise², Jochen C. Reif^{1*}
and Albert W. Schulthess¹

¹Breeding Research Department, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Seeland, Germany, ²Genebank Department, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Seeland, Germany

Globally, wheat (*Triticum aestivum* L.) is a major source of proteins in human nutrition despite its unbalanced amino acid composition. The low lysine content in the protein fraction of wheat can lead to protein-energy-malnutrition prominently in developing countries. A promising strategy to overcome this problem is to breed varieties which combine high protein content with high lysine content. Nevertheless, this requires the incorporation of yet undefined donor genotypes into pre-breeding programs. Genebank collections are suspected to harbor the needed genetic diversity. In the 1970s, a large-scale screening of protein traits was conducted for the wheat genebank collection in Gatersleben; however, this data has been poorly mined so far. In the present study, a large historical dataset on protein content and lysine content of 4,971 accessions was curated, strictly corrected for outliers as well as for unreplicated data and consolidated as the corresponding adjusted entry means. Four genomic prediction approaches were compared based on the ability to accurately predict the traits of interest. High-quality phenotypic data of 558 accessions was leveraged by engaging the best performing prediction model, namely EG-BLUP. Finally, this publication incorporates predicted phenotypes of 7,651 accessions of the winter wheat collection. Five accessions were proposed as donor genotypes due to the combination of outstanding high protein content as well as lysine content. Further investigation of the passport data suggested an association of the adjusted lysine content with the elevation of the collecting site. This publicly available information can facilitate future pre-breeding activities.

KEYWORDS

genebank genomics, genomic prediction, grain quality, lysine content, protein content, wheat

1 Introduction

Worldwide, 410 Mt of consumable plant-based proteins are provided by agriculture, with soybean (*Glycine max* (L.) Merr.), maize (*Zea mays* L.), and wheat (*Triticum aestivum* L.) contributing the largest quantities (Leinonen et al., 2019). Unlike the first two crops, wheat is mostly used directly in human nutrition (OECD/FAO, 2021). Thus, it is not surprising that wheat provides on average 19% of the proteins consumed by humans, with some regional peaks reaching more than one third in North Africa as well as in West and Central Asia (Erenstein et al., 2022). Remarkable ratios were also found in some regions of South Asia: wheat consumed as flat bread accounts for about three-fifths of the daily protein consumption in Pakistani households (Hussain et al., 2004). Undoubtedly, the associated dominance in the diets are partially due to the prevalent cultivation in the respective regions but also wheat's widespread availability on a global market (Shewry and Hey, 2015). Moreover, the preference for wheat can also be assigned to the specific characteristics of the protein fraction of the wheat grain which lead to the unique baking and processing quality of wheat flour (Shewry, 2009). This is one of the reasons for wheat being processed to a diversity of breads, pastries and noodles (Shewry, 2009) and as such forms a key aspect of the cuisine in many regions.

Despite the large quantity of consumed wheat protein, the nutritional quality of this protein is rather inadequate due to the unbalanced amino acid composition. In particular, shortcomings in the lysine content are the limiting factor (Leinonen et al., 2019) which is especially problematic since the essential amino acid lysine cannot be produced by the human organism itself and thus, must be obtained from the diet (Ufaz and Galili, 2008). On the one hand, these shortcomings can be leveled out in a diverse diet which comprises lysine-rich protein sources such as legumes, meat, fish or dairy products (Ritchie et al., 2018; Leinonen et al., 2019). On the other hand, a considerable number of people, especially in developing countries, does not have the purchasing power to diversify their diet with, for example, animal-based products (Hussain et al., 2004; Pellett and Ghosh, 2004; Muleya et al., 2022). An unbalanced wheat-rich diet may result in lysine deficiency (Meybodi et al., 2019). Such a deficiency is known to cause severe physical underdevelopment in children (Batool et al., 2015). Moreover, an inadequate supply with high quality protein can affect physiological processes, the immune system as well as the cognitive development (Batool et al., 2015). Impact on elderly adults is also widely reported and for this group, deficiency results in severe impairment of health including symptoms such as anemia and fatigue (Meybodi et al., 2019). Overall, the symptoms associated with inadequate protein supply are summarized under the name protein-energy malnutrition (Meybodi et al., 2019) and affect millions of people in developing countries (Batool et al., 2015).

Some strategies have already been proposed to increase the lysine content of staple foods. For example, artificial fortification of wheat flour with ground legumes, pseudo cereals or synthesized amino acids (Hussain et al., 2004) has been shown to be effective, but may have adverse effects on the processing quality or taste of end products (Meybodi et al., 2019). Another promising strategy

could be to breed wheat varieties which combine an overall high protein content with an enrichment of lysine in the grain. In general, the potential of developing cereal crops with such characteristics has been demonstrated in maize. Naturally occurring maize mutants, such as *opaque2* and *floury2*, have been reported with a significantly elevated lysine content (Morton et al., 2016). In a case study, Muleya and collaborators (2022) concluded that the use of varieties with such a mutation reduces the risk of lysine deficiency by 21% for the poorest quintile of households in Malawi. To the best of our knowledge, an analog wheat variety has however not been developed so far. Lysine content has generally not been of interest in commercial wheat breeding programs and therefore, the potential of a breeding-based approach might be particularly high for such an orphan trait. Moreover, the naturally occurring lysine content of wheat grains is mainly influenced by the genotype and depends only to a small extent on environmental factors (Lawrence, 1976). Both arguments advocate for a breeding-based approach such as outlined: Firstly, the variation in lysine content of a large quantity of genotypes needs to be analyzed which is very laborious in the field and laboratory. The first step is followed by the identification of donor genotypes with a high lysine content in the protein fraction. Lastly, the favorable genetics of donor genotypes would be considered in pre-breeding activities and selectively transferred into the elite gene pool of modern breeding programs. While the latter step is mostly rather foreseeable, the first two are the bottlenecks for increasing the lysine content because they are time-consuming, demand resources and the result largely depends on the variation available for analysis.

Genebank collections for wheat are known to harbor large genetic diversity (Sansaloni et al., 2020; Schulthess et al., 2022) and phenotypic variation (Philipp et al., 2018; Schulthess et al., 2022). Diversity is however trait-specific and thus, identification of potential donor genotypes with a high content of lysine and protein rely on phenotypic investigation. Earlier attempts to screen genebank collections of wheat for both traits date back to the early 1970s. Vogel and collaborators evaluated 12,613 wheat accessions from the World Wheat Collection of the United States Department of Agriculture (USDA) (Vogel et al., 1973). In the same decade, both traits were measured for 9,706 *Triticum* accessions at the predecessor institution of the Federal *ex situ* Genebank of Agricultural and Horticultural Crops which is today hosted at the Leibniz Institute of Plant Genetics and Crop Plant Research in Gatersleben (IPK Genebank) (Lehmann et al., 1978). The aim of the aforementioned study was to screen the entire collection once and to identify accessions with a strong deviation from the population mean. The deviating accessions were re-evaluated in another year in order to account for an overestimation due to environmental effects. Until the mid-1980s, further successions were successively investigated in a structured manner (Müntz and Lehmann, 1987). Despite the sheer amount of work reflected by the work from Lehmann and collaborators (1978), this data has not been mined in depth according to today's standards and possibilities. Since then, developments in biostatistics and genomics urge the need for a re-evaluation of this historical dataset. This includes the connection of phenotypic data to genotypic data derived by next generation sequencing, which becomes more and more available for large

parts of the cereal collections at the IPK Genebank (Schulthess et al., 2022). Combining and analyzing data will undoubtedly become more important for the work of genebank curators in the future. Since the evaluation in the 1970s, the IPK Genebank has increased in size. With more than 27 thousand genebank accessions of *Triticum* species (Oppermann, 2023), the IPK Genebank preserves nowadays the 9th largest collection of plant genetic resources of wheat and its crop wild relatives (FAO, 2010). Genomic prediction could be used to characterize these new non-phenotyped parts of the collection as well as those parts without reliable phenotypic data. The power of targeted genomic prediction has recently been shown by many studies in the context of genebanks (Yu et al., 2016; Gonzalez et al., 2021; Berkner et al., 2022; Schulthess et al., 2022). Finally, informing the interested public on the newly generated information according to the FAIR (Findable, Accessible, Interoperable and Reusable) (Wilkinson et al., 2016) principles will further activate genebanks. This strategy could enable breeders to specifically select suitable donor genotypes and eventually, it may contribute to a future with less malnutrition in developing countries.

The main aim of this study was to activate historical records of the nutritional quality of wheat accessions stored at the IPK Genebank for their use in plant breeding and research. In more detail, we targeted (1) to curate the raw historical records for protein and lysine content which were generated between 1970 and 1986, (2) to analyze the data across years in order to generate outlier-corrected best linear unbiased estimates (BLUEs) for genebank accessions, (3) to apply a most suited model for genomic prediction in order to predict phenotypes for the majority of genebank accessions and (4) to suggest a set of well characterized suitable donor genotypes to breeders and the interested public.

2 Materials and methods

2.1 Curation of historical records

Historical data on protein and lysine content were compiled and curated. Some of the data originally recorded on punched tapes was unlocked; other data was recorded manually from paper files. All records were checked for accuracy and linked to the currently used accession numbers. This data originates from a large screening of the *Triticum* collection of the IPK Genebank. Between 1970 and 1986, 4,971 accessions were cultivated in 11 almost consecutive years (Figure S1), seeds were harvested and analyzed in the laboratory for protein content and lysine content. Detailed description of the procedure has been given by Lehmann and collaborators (1978). BLUEs for thousand grain weight (TGW) were used as published by Philipp and collaborators (2019).

2.2 Origin and curation of genomic data

This study relied on a genomic dataset which has been generated by Schulthess and collaborators (2022). Briefly, the authors requested 7,651 accessions from the *Triticum* collection

of the IPK Genebank and developed 7,745 isolate lines from them. From here onward, these isolates are referred to as accession samples. All 7,745 accession samples were genotyped by following a genotyping-by-sequencing approach. Reads were aligned to the first version of the reference genome var. Chinese Spring (IWGSC, 2018). After alignment, markers were rejected based on homozygosity of either the reference or alternative allele. In the next step, information of markers was omitted based on missing values (> 10%), a minimum homozygous allele count of < 10% and a maximum heterozygosity of > 1%. Later, imputation was done based on the dominant allele. Afterwards, further filtering based on a minor allele frequency of 1% led to a final matrix with 17,118 markers which was used for downstream analysis.

2.3 Outlier correction and analysis of phenotypic data

The raw data for protein and lysine content was trimmed to ensure that the data could be analyzed. Per trait, accessions were excluded from further analysis if they were represented by a single datapoint. Phenotypic values of 561 accessions remained after this trimming. Furthermore, all records of a year were omitted if no overlap with records of other years could be found. Outlier correction and calculation of BLUEs was done as described by Philipp and collaborators (2018). Briefly, the following linear mixed model was fitted to the data:

$$y_{ij} = \mu + g_i + a_j + e_{i(j)} \quad (1)$$

where y_{ij} is the protein content (or lysine content) measured on seeds of the accession i which were harvested in the year j . Accordingly, μ is the general fixed population average effect, while g_i and a_j represent the effects of the genotype and the year, respectively. The term $e_{i(j)}$ refers to the error of the model of which the variance is modelled as specific for each year. For the identification of outliers and the estimation of BLUEs, the term g_i was modelled as fixed while a_j was modelled as random. In contrast, both terms were considered as random for the calculation of heritability. Outliers were identified based on standardized residuals and with a correction for multiple testing (Holm, 1979; Nobre and Singer, 2011) as implemented by Philipp and collaborators (2018).

Heritabilities (h^2) of both traits were estimated as described by Philipp and collaborators (2018),

$$h^2 = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_e^2}{\bar{N}_Y}} \quad (2)$$

where σ_G^2 and σ_e^2 refer to the genetic variance and the average of year-specific error variances, respectively. \bar{N}_Y is the average number of years in which an accession was tested. In addition, above explained variances components were used to compute plot-based heritabilities as:

$$h_{pb}^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_e^2} \quad (3)$$

Lysine content was adjusted for protein content and TGW, because lysine content was strongly correlated with both other traits. The adjustment approach was a derivative of the approach applied by Vogel and collaborators (1975). Briefly, a multiple linear regression model was fitted on the BLUEs of lysine content in dependence on protein content and TGW. Afterwards, lysine content was adjusted genotype-wise based on the partial regression coefficients and the mean-centered protein content and TGW values as follows:

$$\begin{aligned} \text{Lysine}_{adj} = & \text{Lysine} - b_{\text{Protein}}(\text{Protein} - \overline{\text{Protein}}) \\ & - b_{\text{TGW}}(\text{TGW} - \overline{\text{TGW}}) \end{aligned} \quad (4)$$

2.4 Analysis of population structure

The relatedness of the 7,745 genotyped accession samples was studied based on a principal coordinate (PCo) analysis (Gower, 1966). For this, pair-wise Rogers' distances (Rogers, 1972) were calculated between genomic profiles of all accession samples and compiled into a distance matrix; the complexity of the distance matrix was reduced by deriving PCos (Gower, 1966). First and second PCos, which retain the highest amount of variation, were plotted against each other to graphically portray possible patterns resulting from population structure.

2.5 Genomic prediction models and their evaluation

In the present study, four different genomic prediction models, namely G-BLUP, EG-BLUP, Bayes A, and Bayesian Lasso, were compared based on their performance. The G-BLUP model (VanRaden, 2008) predicts phenotypic values based on additive genetic effects. These effects are explained by the relationship among the genotypes. The prediction model for n genotypes has the following matrix notation (Henderson, 1985):

$$y = \mathbf{1}_n \mu + \mathbf{g} + \mathbf{e} \quad (5)$$

where the phenotypic values (BLUEs), given by the vector y , are a function of the general mean (μ) and the n -dimensional vectors \mathbf{g} and \mathbf{e} , which account for the genotypic values and the model's residuals, respectively. The n -dimensional vector of ones ($\mathbf{1}_n$) assigns μ to each element of y . The vectors \mathbf{g} and \mathbf{e} follow multivariate normal distributions $\mathbf{g} \sim N(\mathbf{0}, \mathbf{G}\sigma_g^2)$ and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$ which depend on the genomic-estimated additive relationship matrix \mathbf{G} and the genetic variance (σ_g^2) or \mathbf{I} and the residual variance (σ_e^2), respectively. The $n \times n$ matrix \mathbf{G} was calculated based on the first method described in VanRaden (2008) while \mathbf{I} is an n -dimensional identity matrix.

EG-BLUP accounts for additive-by-additive epistasis (Jiang and Reif, 2015) and can be seen as an extension to the G-BLUP model as follows:

$$y = \mathbf{1}_n \mu + \mathbf{g} + \mathbf{g}_1 + \mathbf{e} \quad (6)$$

In equation (6), the terms y , $\mathbf{1}_n$, μ , \mathbf{g} , and \mathbf{e} are as defined in equation (5). The n -dimensional vector \mathbf{g}_1 accounts for the additive-by-additive effect among genotypes. This effect follows a multivariate normal distribution $\mathbf{g}_1 \sim N(\mathbf{0}, \mathbf{H}\sigma_{g_1}^2)$, where $\mathbf{H} = \mathbf{G}\#\mathbf{G}$, with $\#$ being the Hadamard product operator (Jiang and Reif, 2015).

BayesA (Meuwissen et al., 2001) was used as the base Bayesian model for genomic prediction. In this prediction approach, priors of the regression parameters are assumed to follow a scaled-t distribution. Genomic prediction with the Bayesian Lasso was applied according to Park and Casella (2008). In this approach, the regression parameters have a double-exponential prior.

The four genomic prediction models were compared based on their ability to accurately predict phenotypes. In this comparison, the unit of quality was the prediction ability, which was defined as the correlation between the BLUEs and the predicted phenotypes. The comparison was established by means of five-fold cross-validation. All accession samples with known BLUEs were assigned to one of five equally sized groups. Four of these groups were incorporated in the prediction model as training set in order to predict the phenotypes in the remaining group, known as test set. The prediction was repeated in such a manner that each group has once been the test set and four times part of the training set. Thereafter, predicted phenotypes of all test sets were combined and the Pearson correlation coefficient with the respective BLUEs was calculated. This whole process was independently repeated 100 times. For an unbiased comparison, all models were tested based on the same training and test set. The best performing prediction model was used to predict the phenotypes of all accession samples with genomic data. In the latter case, all accession samples having available phenotypic data were used as training set.

All computational calculations, analysis as well as the creation of figures was implemented in the R environment (R v. 4.0.2). Solving the linear mixed model for data curation, BLUEs computation and the estimation of variance components from phenotypic data was done by engaging ASReml-R 4 v. 4.1.0.110 (Butler et al., 2018). Genomic prediction models were implemented with the R package BGLR v. 1.0.8 (Pérez and De Los Campos, 2014).

3 Results

3.1 Curated data with high quality

The data curation resulted in a comprehensive dataset for protein content and lysine content which comprised 11 years of experimental trials. In total, the resulting raw dataset included 5,952 records for protein content and 5,940 records for lysine content from a total of 4,971 accessions. Across years, the raw data did not only display differences in the traits' distributions; but moreover, the number of recorded data points differed strongly with a clear dominance for the year 1970 in which 3,442 records were taken per trait (Figure S1). In contrast, only six measurements were reported

for 1983; these were excluded due to the absent overlap with any other year. Despite the large amount of data, the dataset was rather incomplete with an unbalanced structure and most accessions tested only in one year (Table S1): Of all 4,971 accessions, 4,410 accessions were grown and characterized once without any replication. These records were excluded from further analysis to ensure that reliable BLUEs can be obtained for the remaining accessions. After this step, remaining accessions were evaluated in up to seven years, with an average number of 2.41 and 2.39 for protein and lysine content, respectively.

The quality of the data can be reviewed based on heritability for the two traits (Table 1). For protein content, the heritability before outlier correction reached 0.77 and could only be slightly improved due to the correction. The quality of the data for lysine content improved by 23.14% due to the removal of 17 outlier data points, leading to an increase in heritability from 0.47 to 0.58. The

estimated plot-based heritabilities behaved accordingly, as also evidenced by the negligible amount of rejected data points.

3.2 Estimated average phenotypic performance of accessions and associations of the traits

The analysis resulted in BLUEs of 558 accessions for both traits, namely protein content and lysine content. On average, accessions had a protein content of 17.61% and a lysine content of 4.17%. However, some accessions were found with a very positive deviation from the average (Figure 1): 12 accessions exhibited a lysine content of more than 5.0%. Protein content and lysine content were highly correlated ($r = 0.63$, $p < 0.01$). Out of the 558 accessions, only 319 accessions additionally had BLUEs for TGW. TGW was negatively

TABLE 1 Description of the dataset for protein content and lysine content before and after outlier correction.

Trait	Outlier correction	μ_N	σ_g^2	σ_e^2	h^2	h_{pb}^2
Protein content	before	2.410	3.894	2.808	0.77	0.58
	after	2.407	4.017	2.679	0.78	0.60
Lysine content	before	2.394	0.087	0.235	0.47	0.27
	after	2.385	0.096	0.168	0.58	0.36

Depicted are the average number of datapoints per accession (μ_N), the genetic variance (σ_g^2), the average of year-specific error variances (σ_e^2) as well as the heritability (h^2) and the plot-based heritability (h_{pb}^2). Accessions with a single datapoint per trait were disregarded here.

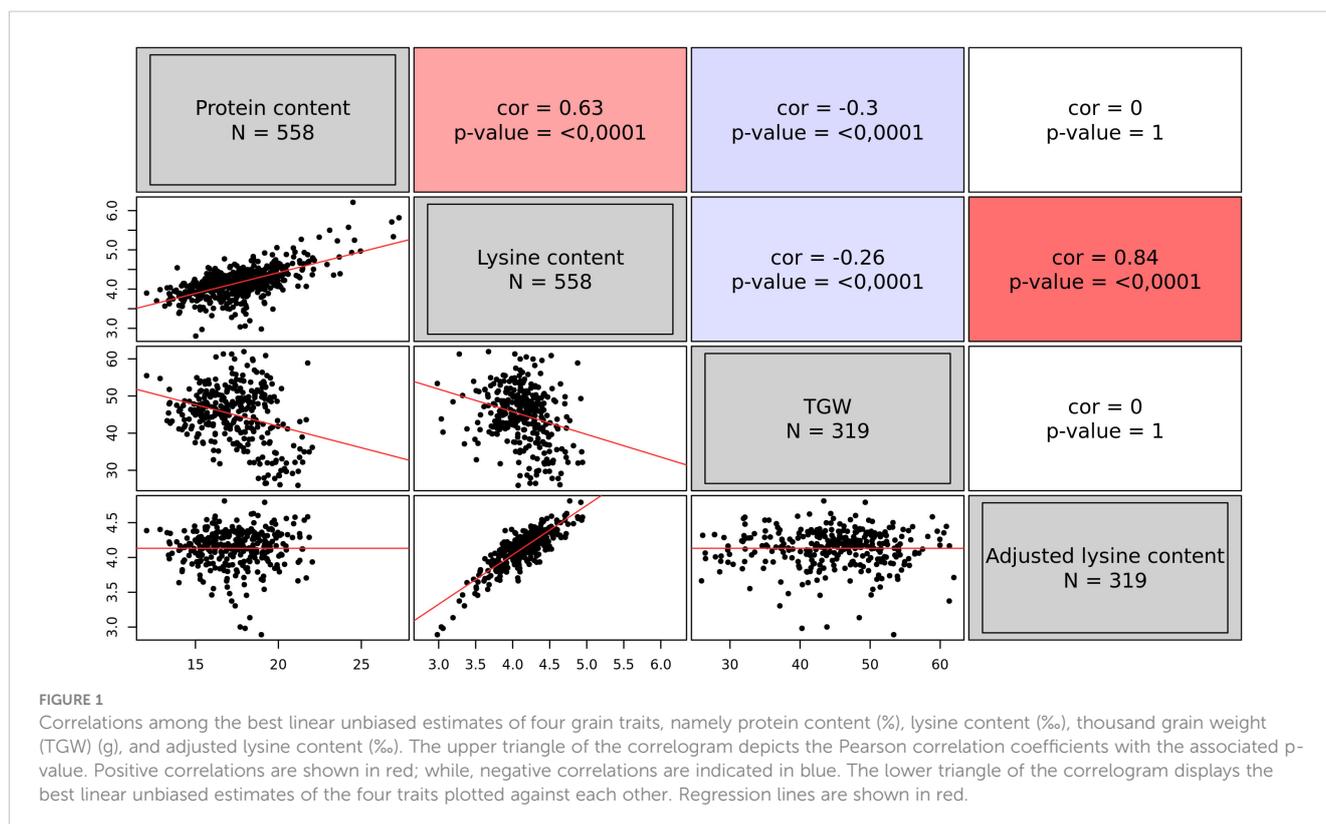


FIGURE 1

Correlations among the best linear unbiased estimates of four grain traits, namely protein content (%), lysine content (%), thousand grain weight (TGW) (g), and adjusted lysine content (%). The upper triangle of the correlogram depicts the Pearson correlation coefficients with the associated p-value. Positive correlations are shown in red; while, negative correlations are indicated in blue. The lower triangle of the correlogram displays the best linear unbiased estimates of the four traits plotted against each other. Regression lines are shown in red.

correlated with both, protein content and lysine content (Figure 1). With this information, the adjusted lysine content of these 319 accessions was calculated. The adjustment completely broke the correlations of lysine content with TGW and protein content but retained the strong correlation with lysine content *per se*.

3.3 Comparison of different genomic prediction approaches

Both genotypic data and phenotypic data, which were available for 337 accession samples, were used for further genomic analysis and prediction of protein and lysine content. The number of genotyped accession samples with phenotypic information was only slightly lower for adjusted lysine content. In addition, no

clear relationship pattern was observed between the distribution of the 337 accession samples along the PCos of the genomic distances and the phenotypic variation. Thus, no subpopulations were found with substantially higher or worse performing accessions compared to the population average. All in all, the available training set with reliable phenotypes corresponds to a representative sample of the whole winter wheat collection (Figure 2). Therefore, despite its limited size and provided high cross-validated prediction abilities, reliable predictions should be expected for both, phenotyped and non-phenotyped accessions.

Four different genome-wide prediction approaches were implemented and compared based on the correlation between BLUEs and predicted phenotypes. EG-BLUP outperformed G-BLUP and both Bayesian methods for the prediction of protein content, lysine content, and adjusted lysine content (Figure 3). In

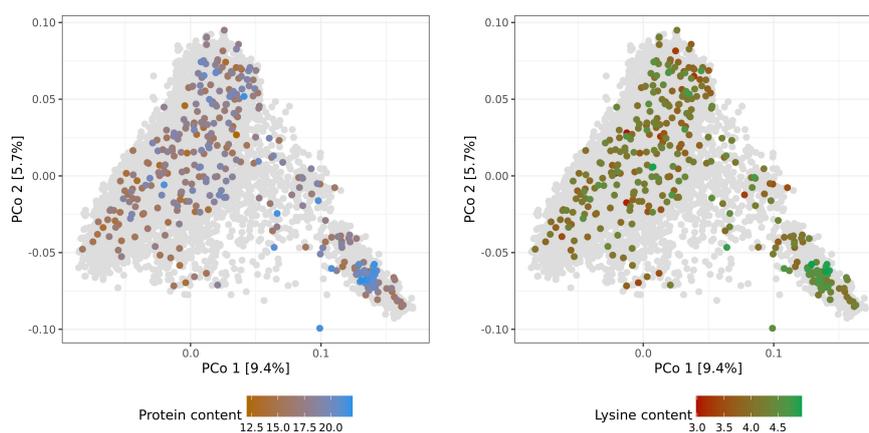


FIGURE 2

Molecular diversity of the IPK winter wheat collection covered by accession samples with best linear unbiased estimates of protein (%) and lysine (%) content. Distributions of the 337 phenotypic values are depicted via colorcoding and shown separately per trait. Biplots are based on the first and second principal coordinates (PCo) from the Rogers' distances between 7,745 accession samples characterized with genotyping-by-sequencing. Gray dots represent genotyped accession samples lacking best linear unbiased estimates.

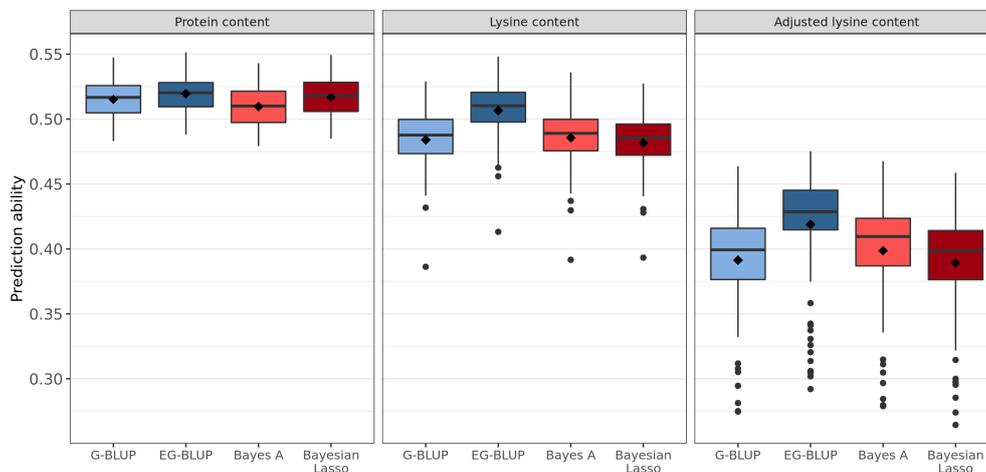


FIGURE 3

Distribution of genomic prediction abilities estimated in 100 five-fold cross-validation runs for protein, lysine, and adjusted lysine content. Four genomic prediction models were considered: G-BLUP, EG-BLUP, Bayes A, and Bayesian Lasso. Boxes enclose 50% of the central data, including median (horizontal black bold line) and mean (black diamonds), while whiskers are $\pm 1.5 \times$ interquartile range and dots represent extreme values.

terms of average cross-validated prediction abilities, EG-BLUP was 5.05% superior than the best of the three other alternative approaches for the prediction of adjusted lysine content. In addition, two different approaches were compared for the prediction of adjusted lysine content (Figure S2). The separate prediction of lysine content, protein content and TGW in order to calculate the derived trait based on these predictions was marginally less accurate than using the derived trait for the genomic prediction.

3.4 Predicted phenotypes of 7,745 accession samples

Protein content, lysine content, and adjusted lysine content were predicted for 7,745 accession samples by applying EG-BLUP - the most-accurate prediction model in cross-validations. For all accession samples, predicted protein content, lysine content, and adjusted lysine content averaged 16.85%, 4.08‰, and 4.13‰, respectively, with associated standard deviations of 0.74%, 0.11‰, and 0.09‰, correspondingly (Figure 4). Some accessions had outstanding values for the three traits with highest predicted values of protein content, lysine content, and adjusted lysine content amounting to 20.92%, 4.8‰, and 4.64‰, respectively. Interestingly, we found few accessions that had high values for both protein content and adjusted lysine content.

For all three traits, the size of the training set was rather small compared with the test set. For example, information of only 329 accession samples was used in order to predict the adjusted lysine content for 7,416 accession samples (Figure 4). Interestingly, the mean and median were both lower in the test set compared with the

training set. This deviation was most dominant for the predicted protein content where the means were 17.31% and 16.82% in training set and test set, respectively.

3.5 Definition of promising donor genotypes

The newly explored information can be used to select germplasm for pre-breeding programs with yet unexploited genetic diversity. To motivate future germplasm usage, favorable accessions were preliminary selected based on culling levels for predicted protein content and predicted adjusted lysine content in parallel. For both traits, the more stringent threshold was set to 99.9% of the normal distribution and five accessions could preliminary be selected (Figure 5). The respective accessions had not only favorable predicted phenotypes; moreover, these accessions were also recorded with particularly high BLUEs for both traits. Thus, the prediction can be seen as a confirmation of the genetic superiority of these accessions. Additional 19 accessions were identified with a more relaxed threshold for the culling levels selection ($z = 0.99$) (Table S2).

3.6 Association of lysine with altitude

The five accessions which were preliminary selected based on the stringent culling levels selection originated from Nepal and Afghanistan (Table S2). A characteristic of these accessions is the high altitude of their collecting sites. The altitudes of the collecting sites ranged between 1,925 m and 2,975 m above sea level, as

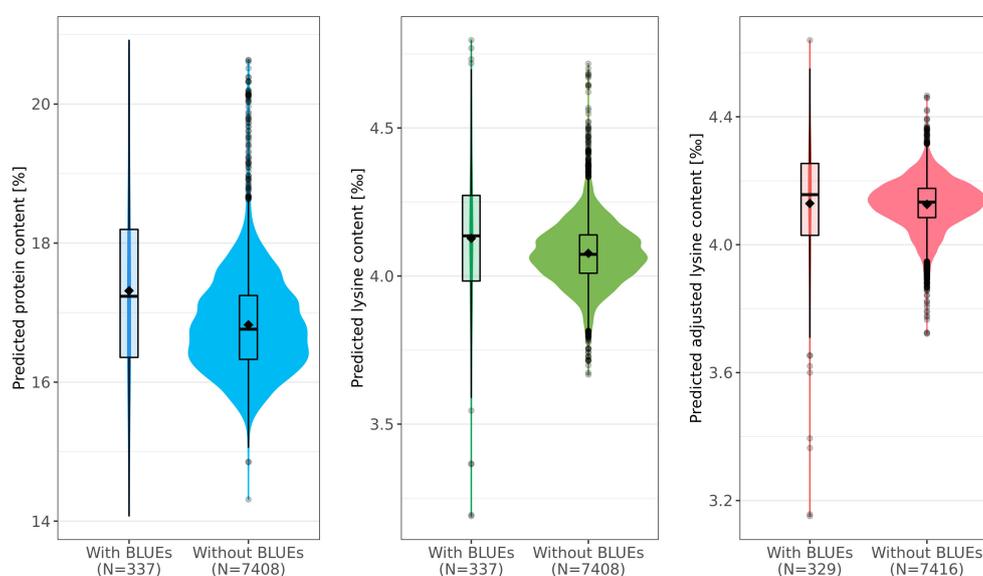


FIGURE 4

Predicted phenotypes of 7,745 wheat accession samples for protein (%), lysine (‰), and adjusted lysine (‰) content. For each trait, distributions of the prediction are shown separately for those accessions with best linear unbiased estimates (BLUEs, left) and those for which only genotypic data was present (right). Boxes enclose 50% of the central data, including median (horizontal black bold line) and mean (black diamonds), while whiskers are $\pm 1.5 \times$ interquartile range and dots represent extreme values.

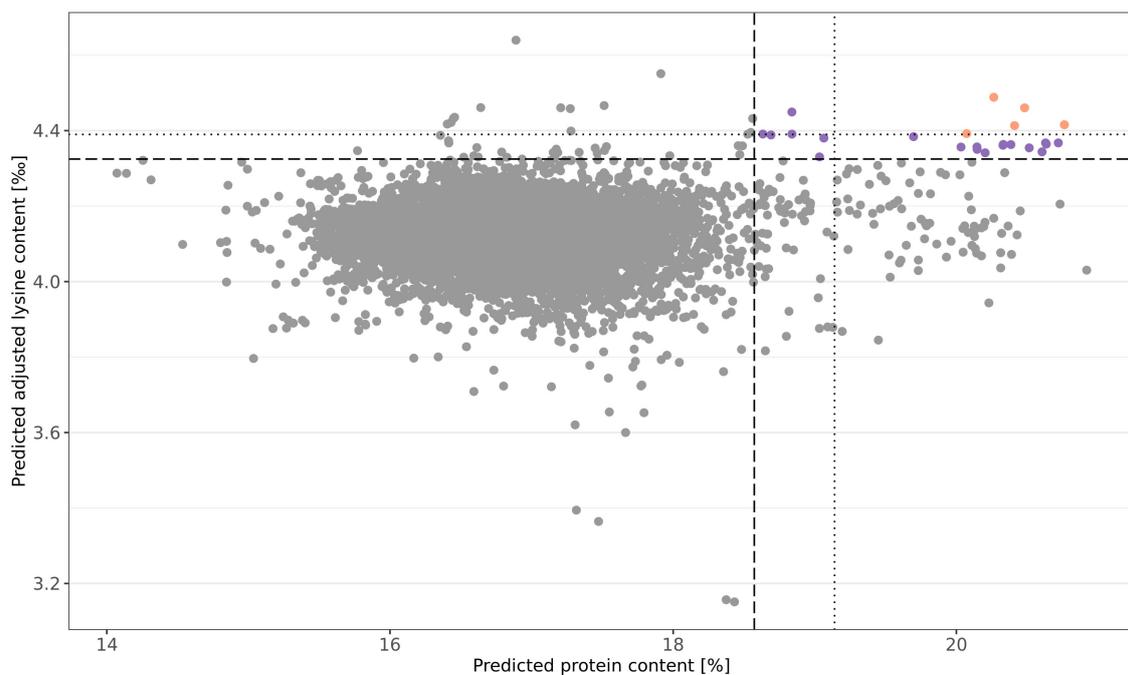


FIGURE 5

Culling levels selection based on genomic predicted protein (%) and adjusted lysine (‰) content. Shown are the predicted phenotypes of 7,745 wheat accession samples. Selection was performed with two intensities: a stringent threshold was defined by 0.999 of the normal distribution (dotted line) while a more relaxed selection threshold resulted from 0.99 of the normal distribution (dashed line). Orange and purple dots represent accessions samples which were selected based on the stringent and relaxed threshold, respectively.

indicated by the genebank catalog (Oppermann, 2023) and the description of the collecting expedition (Witcombe, 1975). The altitude of the collecting site is only known for 927 accession samples which are 12.12% of the examined winter wheat collection. The correlation between the altitude and the predicted traits was analyzed despite the incomplete data. Predicted protein and lysine contents were positively correlated with the altitude of the collecting site; the correlation coefficient amounted to 0.23 and 0.50, respectively ($p < 0.01$). The predicted adjusted lysine content was also positively correlated with the altitude of the collecting site ($r = 0.45$, $p < 0.01$) (Figure S3).

4 Discussion

The analysis of the present historical data was affected by their non-orthogonal structure. Most of the data trace back to the year 1970, with the vast majority of the accessions tested without replication. Lehmann and collaborators (1978) planned to screen all accessions once with the aim of identifying accessions with a strong positive deviation from the mean of the total collection. In order to reduce cost and workload, a repetition of the analysis was conducted only for accessions with high protein content ($> 18.5\%$) and lysine content ($> 4.5\%$). This strategy has two major disadvantages: First, the potential of some accessions may have been underestimated in a single testing year which would then have resulted in an erroneous rejection of this accession. Second, the total phenotypic variation cannot be estimated correctly without reliable

data from the underperforming accessions. However, precise estimates of variation are a prerequisite for optimized allocation of resources in breeding. On the one hand, this bias in the database for prediction should lower the quality of the prediction especially for the underperforming accessions (Zhao et al., 2012). On the other hand, this should be less relevant in our case because the selection decision in the first stage was based on data collected in one year coupled with the moderate plot-based heritabilities (Table 1). To account for the described shortcomings, we excluded accessions which were represented by only a single datapoint, as unreplicated data would provoke large uncertainty of the phenotypic data.

The exclusion of accessions with unreplicated data had strong consequences for the generated BLUEs. Trimming of the dataset reduced the number of accessions in the dataset to 11% and due to the multiple-stage testing strategy, resulted in higher mean values of protein content (15.92% to 17.66%) and lysine content (3.72‰ to 4.10‰). During the time of data collection, the multiple-stage evaluation resulted in optimized selection gain. From today's perspective, the shortcomings in the dataset however highlight the need to systematically plan screenings in a way which already consider the proper statistical evaluation. Especially with limited resources, repeated phenotyping of a well-chosen subset of accessions should be favored since missing phenotypic information can be determined by genomic predictions that rely on cheap genotyping of whole genebank collections (Yu et al., 2016). Given the selection strategy elaborated above, it was important to investigate whether selection decisions stood in the way of a representative training population. Inspection of the PCos

and distribution of phenotypic values suggests that we found this to be the case to a limited extent (Figure 2).

4.1 Strong associations between seed traits

The results showed a strong positive correlation between the BLUEs for protein and lysine content (Figure 1). An association of these traits has already been reported based on a large screening of the USDA World Wheat Collection (Vogel et al., 1975). The authors reported an even stronger correlation of 0.804 and 0.871 for the years 1972 and 1973, respectively. Furthermore, these authors reported a slightly negative correlation of TGW with protein content ($r = -0.278$) and lysine content ($r = -0.266$), respectively, in the year 1972. Thus, these correlations are in the same order of magnitude as the correlations found in the present study. In conclusion, these correlations indicate that accessions with a higher protein content do also have a higher overall lysine content. This is hardly surprising, because lysine is part of many groups of proteins even though not in equal abundance. Furthermore, lighter grains were identified to have an overall higher protein and lysine content. Arguably, this is due to the heterogeneous distribution of both components in the wheat grain. The storage proteins in the endosperm have a significantly lower lysine content than the embryo and bran (Vogel et al., 1976). In line with this, the relative lysine content of wheat grains decreases during the grain filling and maturation of the seed, thus, when the endosperm increases in size (Molino et al., 1988). Arguably, the fraction of the endosperm on the whole grain is larger in accessions with a high TGW. This suggests that the proportion of tissues with low lysine content increases in heavier grains. On the other hand, Vogel and collaborators (1976) also found a strong correlation of 0.91 between the lysine content of the endosperm and of the whole grain and concluded that the whole grain trait values are sufficiently reliable for selection. The data set analyzed in the present study was generated based on whole grain samples (Lehmann et al., 1978), which thus represent a mixture of embryo, bran, and endosperm. Unfortunately, modern milling processes white flour which contains exclusively endosperm tissue (Yu and Tian, 2018). Therefore, accessions with beneficial characteristics could hypothetically rely on an elevation of the lysine content in seed tissues rarely used in human nutrition. In this regard, the distribution of amino acids should be further investigated in the future, especially in outperforming accessions. In the present study, the intention was to identify genotypes that have a high proportion of lysine in the protein fraction independently of the seed size; thus, the adjustment of lysine content was important to account for the described associations.

4.2 EG-BLUP with high potential for genomic prediction

The comparison of genomic prediction models has shown that EG-BLUP outperforms the other models in terms of prediction ability (Figure 3). This is consistent with the findings of previous

genomic prediction studies in wheat. EG-BLUP resulted in more accurate predictions compared with G-BLUP for the prediction of TGW, plant height, and yellow rust resistance (Berkner et al., 2022). The particular advantage of this model is its ability to account for additive effect but also for additive-by-additive epistasis (Jiang and Reif, 2015). These results highlight therefore the importance of additive-by-additive epistasis and are thus in line with previous finding in wheat (Jiang et al., 2017; Raffo et al., 2022). The superiority across many different traits, demonstrates the robustness of the EG-BLUP model when confronted with different genetic architectures.

According to our cross-validated comparison, genomic prediction of derived traits such as adjusted lysine content can be performed accurately (Figure S2), but requires however some careful attention. In the present case, the implemented adjustment method relied on the availability of phenotypic values for three traits per accession. This restriction reduces the size of the training set and thus, the information which can be used for genomic prediction in a multiple-trait context (Schulthess et al., 2016). Moreover, the adjustment method relies on associations between lysine content and the two associated traits. These associations are, however, only valid for the examined set of genotypes and can differ between subsets of accessions such as for region-specific subpopulations. Even though the prediction of the derived trait based on the adjusted lysine content itself was most accurate (Figure S2), it might be more appropriate to predict the basis traits separately if subsampling is planned later, if the traits are biased by subpopulations, or if the availability of data is very unbalanced across different traits.

4.3 Enrichment of the genebank catalog facilitates new strategies for breeding programs

The study presents three types of data, namely, curated raw data, BLUEs, and predicted phenotypes for interested stakeholders in breeding and research. Without any doubt, the estimated (BLUEs) and genomic predicted phenotypic performance can be used for targeted selection of accessions. Although all the above-mentioned data has now become publicly available, we wanted to examine specifically accessions with high protein content and adjusted lysine content. With high selection intensities (culling levels of $z = 0.999$), we selected five promising accessions: While one of the preliminary selected accessions came from Afghanistan, four accessions originate in the Arun valley in Nepal (Oppermann, 2023). The latter ones derived from a collecting expedition in 1971. Considering that the collecting sites of all four accessions were located in neighboring villages (Witcombe, 1975), they arguably share one common mechanism of upregulated synthesis and storage of protein and lysine. At reduced selection intensity with a culling level of $z = 0.99$, 19 additional accession were identified and 14 of these originate from the very same expedition to Nepal. Witcombe and Rao (1976) evaluated the accessions of that collecting journey based on 39 traits and clustered plant material based on phenotypic characteristics but ignoring the geographic

proximity of the collecting sites. Interestingly, 14 accessions from the preliminary selected 18 accessions with Nepalese origin derived from the same phenotypic cluster. [Witcombe and Rao \(1976\)](#) identified the high altitude as one factor which leads to the common characteristics of this cluster.

4.4 Potential rationales for lysine's association with high altitudes

Adjusted lysine content was associated with the altitude of the collecting sites of accessions. On the one hand, it could be argued that the clustering of accessions with high adjusted lysine content collected at high altitudes in Nepal was caused by a spontaneous and rare mutation unrelated to selective advantages. On the other hand, high adjusted lysine content was significantly associated ($r = 0.45$; $p < 0.01$) ([Figure S3](#)) with the altitude of the collecting site in our study for a larger sample of 927 accessions for which altitude information of the collecting site was available. We thus advance the hypothesis that lysine content could play a role in the adaptation to the altitude at which these landraces have been grown continuously for many cropping seasons. High-altitude environments share common features, such as low temperatures, strong exposure to wind, drought due to lower humidity, high ultraviolet radiation, and hypoxia ([Tranquillini, 1963](#); [Jinjiu et al., 2021](#)). All of these characteristics can cause abiotic stress to plants, but only the latter two are specific to high altitudes. Therefore, high lysine content could be relevant for the adaptation to high ultraviolet radiation or hypoxia.

The involvement of lysine in the tolerance to various abiotic stresses, such as drought and salinity, has been summarized by [Kishor and collaborators \(2020\)](#). [Yadav and collaborators \(2019\)](#) investigated the impact of drought on the metabolite profile of wheat plants in glasshouse experiments. Drought-resilient wheat lines showed an increase in the lysine content of the vegetative tissue as well as impacts on other amino acids such as serin and asparagin. Additionally, [Ding and collaborators \(2016\)](#) reported a more than twofold increase in lysine content under hypoxic conditions in seedlings of rice (*Oryza sativa* L.). In contrast to this finding, the high adjusted lysine content in the present study is however not just a reaction of the wheat plant to abiotic stress. The high adjusted lysine content reflects a permanent adaptation which also results in higher lysine contents when these accessions are not facing the stresses of high altitudes. The present study relies on field trials conducted at 110 m above sea level. Moreover, the current data reflects the lysine content of mature seeds but not of vegetative plant tissue such as seedling. For seed tissue, we can only speculate about a possible interplay of lysine with stresses such as ultraviolet radiation or hypoxia.

Abiotic stress resistance of seeds is thought to be partially mediated via proteins of the late embryogenesis abundant (LEA) protein families ([Zan et al., 2020](#)). For instance, drought, extreme temperatures and ultraviolet radiation are associated with the expression of members of this protein family ([Wang et al., 2008](#); [Zan et al., 2020](#)). In wheat, [Liu and collaborators \(2019\)](#) identified 179 genes encoding such proteins in the genome of var. Chinese

Spring. These proteins cluster into eight groups with distinct characteristics. Dehydrins are one of these groups and their protective characteristics relies on the K-segment which is specifically enriched in lysine ([Yang et al., 2015](#); [Zan et al., 2020](#)). In line with this, [Bhattacharya and collaborators \(2019\)](#) found that the amino acid composition of LEA proteins in wheat can largely rely on lysine. In the case of one analyzed protein, lysine accounted for more than one quarter of all amino acids. If abundance of such proteins is causal for high lysine contents remain however speculation and this urges the need for further investigation.

The present study has not only outlined a strategy to mine historical data but also to leverage the data by genomic prediction. Moreover, this study equipped breeders and researchers with data for protein, lysine, and adjusted lysine content of in total 7,651 accession which can serve breeders to select suitable accessions for their pre-breeding programs. This might build the starting point of varieties which are not just high in protein but which further have a more favorable composition of amino acids and might help to overcome protein-energy malnutrition in future.

Data availability statement

This study comprises the publication of three different types of information. These are namely the raw data in ISA-Tab format, the R code for the calculation of BLUEs and genomic prediction with all input files, and the most important output files of the analysis. The output files include BLUEs of protein and lysine content as well as the predictions of protein content, lysine content and adjusted lysine content. The aforementioned information is available via the e!DAL ([Arend et al., 2014](#)) online repository (<https://dx.doi.org/10.5447/ipk/2023/20>).

Author contributions

MB: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – original draft. SW: Data curation, Writing – review & editing. JR: Conceptualization, Methodology, Supervision, Writing – review & editing. AS: Conceptualization, Methodology, Supervision, Validation, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the German Federal Ministry of Education and Research as part of the Project GeneBank2.0 [grant no. FKZ031B0184A to AS] and by the AGENT project that is financed by the European Union's Horizon 2020 research and innovation program [grant agreement no. 862613 to MB]. Open access publishing received financial support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) [grant no. 491250510].

Acknowledgments

This paper is dedicated to Andreas Graner, who recognized very early the value of the historical quality data of the IPK wheat collection according to the motto “the value of a collection increases with the associated information density” and thus initiated a pillar to transform the IPK Genebank into a bio-digital resource center. This research is available as a preprint on bioRxiv (Berkner et al., 2023).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Arend, D., Lange, M., Chen, J., Colmsee, C., Flemming, S., Hecht, D., et al. (2014). e!DAL - a framework to store, share and publish research data. *BMC Bioinform.* 15, 214. doi: 10.1186/1471-2105-15-214
- Batool, R., Butt, M. S., Sultan, M. T., Saeed, F., and Naz, R. (2015). Protein–energy malnutrition: A risk factor for various ailments. *Crit. Rev. Food Sci. Nutr.* 55, 242–253. doi: 10.1080/10408398.2011.651543
- Berkner, M. O., Schulthess, A. W., Zhao, Y., Jiang, Y., Oppermann, M., and Reif, J. C. (2022). Choosing the right tool: Leveraging of plant genetic resources in wheat (*Triticum aestivum* L.) benefits from selection of a suitable genomic prediction model. *Theor. Appl. Genet.* 135, 4391–4407. doi: 10.1007/s00122-022-04227-4
- Berkner, M. O., Weise, S., Reif, J. C., and Schulthess, A. W. (2023). Genomic unveiling of the diversity in grain protein and lysine content throughout a genebank collection of winter wheat. *bioRxiv*. [Preprint]. doi: 10.1101/2023.07.05.547805
- Bhattacharya, S., Dhar, S., Banerjee, A., and Ray, S. (2019). Structural, functional, and evolutionary analysis of late embryogenesis abundant proteins (LEA) in *Triticum aestivum*: A detailed molecular level biochemistry using in silico approach. *Comput. Biol. Chem.* 82, 9–24. doi: 10.1016/j.compbiolchem.2019.06.005
- Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. J., and Thompson, R. (2018). “ASReml-R reference manual version 4,” (Hemel Hempstead: VSN International Ltd).
- Ding, J., Yang, T., Feng, H., Dong, M., Slavina, M., Xiong, S., et al. (2016). Enhancing contents of γ -aminobutyric acid (GABA) and other micronutrients in dehulled rice during germination under normoxic and hypoxic conditions. *J. Agric. Food Chem.* 64, 1094–1102. doi: 10.1016/j.jultsonch.2017.08.029
- Erenstein, O., Jaleta, M., Mottaleb, K. A., Sonder, K., Donovan, J., and Braun, H.-J. (2022). “Chapter 4 global trends in wheat production, consumption and trade” in *Wheat Improvement*. Eds. M. P. Reynolds and H.-J. Braun (Cham: Springer), 47–66.
- FAO (2010). “Chapter 3 The state of ex situ conservation,” in *The second report on the state of the world's plant genetic resources for food and agriculture* (Rome: FAO), 53–90.
- Gonzalez, M. Y., Zhao, Y., Jiang, Y., Stein, N., Habekuss, A., Reif, J. C., et al. (2021). Genomic prediction models trained with historical records enable populating the German ex situ genebank bio-digital resource center of barley (*Hordeum* sp.) with information on resistances to soilborne barley mosaic viruses. *Theor. Appl. Genet.* 134, 2181–2196. doi: 10.1007/s00122-021-03815-0
- Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53, 325. doi: 10.2307/2333639
- Henderson, C. R. (1985). Best linear unbiased prediction using relationship matrices derived from selected base populations. *J. Dairy Sci.* 68, 443–448. doi: 10.3168/jds.S0022-0302(85)80843-2
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6, 65–70.
- Hussain, T., Abbas, S., Khan, M. A., and Scrimshaw, N. S. (2004). Lysine fortification of wheat flour improves selected indices of the nutritional status of predominantly cereal-eating families in Pakistan. *Food Nutr. Bull.* 25, 114–122. doi: 10.1177/156482650402500202
- IWGSC (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361, 1–13. doi: 10.1126/science.aar7191
- Jiang, Y., and Reif, J. C. (2015). Modeling epistasis in genomic selection. *Genetics* 201, 759–768. doi: 10.1534/genetics.115.177907
- Jiang, Y., Schmidt, R. H., Zhao, Y., and Reif, J. C. (2017). A quantitative genetic framework highlights the role of epistatic effects for grain-yield heterosis in bread wheat. *Nat. Genet.* 49, 1741–1746. doi: 10.1038/ng.3974

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpls.2023.1270298/full#supplementary-material>

- Jinqiu, Y., Bing, L., Tingting, S., Jinglei, H., Zelai, K. L., Lu, L., et al. (2021). Integrated physiological and transcriptomic analyses responses to altitude stress in oat (*Avena sativa* L.). *Front. Genet.* 12. doi: 10.3389/fgenet.2021.638683
- Kishor, P. B. K., Suravajhala, R., Rajasheker, G., Marka, N., Shridhar, K. K., Dhulala, D., et al. (2020). Lysine, lysine-rich, serine, and serine-rich proteins: link between metabolism, development, and abiotic stress tolerance and the role of ncRNAs in their regulation. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.546213
- Lawrence, J. M. (1976). Environmental influences on wheat lysine content. *J. Agric. Food Chem.* 24, 356–358. doi: 10.1021/jf60204a064
- Lehmann, C. O., Rudolph, A., Hammer, K., Meister, A., Müntz, K., and Scholz, F. (1978). Eiweißuntersuchungen am Getreide- und Leguminosen-Sortiment Gatersleben - Teil 1: Gehalt an Rohprotein und Lysin von Weizen sowie von Weizen-Art- und -Gattungsbastarden. *Die Kulturpflanze* 26, 133–161. doi: 10.1007/BF02146158
- Leinonen, I., Iannetta, P. P. M., Rees, R. M., Russell, W., Watson, C., and Barnes, A. P. (2019). Lysine supply is a critical factor in achieving sustainable global protein economy. *Front. Sustain. Food Syst.* 3. doi: 10.3389/fsufs.2019.00027
- Liu, H., Xing, M., Yang, W., Mu, X., Wang, X., Lu, F., et al. (2019). Genome-wide identification of and functional insights into the late embryogenesis abundant (LEA) gene family in bread wheat (*Triticum aestivum*). *Sci. Rep.* 9, 1–11. doi: 10.1038/s41598-019-49759-w
- Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genet.* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819
- Molino, I. M. M., Rojo, B., Martinez-Carrasco, R., and Pérez, P. (1988). Amino acid composition of wheat grain. 1: changes during development. *J. Sci. Food Agric.* 42, 29–37. doi: 10.1002/jsfa.2740420105
- Meybodi, N. M., Mirmoghaddae, L., Sheidaei, Z., and Mortazavian, A. M. (2019). Wheat bread: potential approach to fortify its lysine content. *Curr. Nutr. Food Sci.* 15, 1–8. doi: 10.2174/1573401315666190228125241
- Morton, K. J., Jia, S., Zhang, C., and Holding, D. R. (2016). Proteomic profiling of maize opaque endosperm mutants reveals selective accumulation of lysine-enriched proteins. *J. Exp. Bot.* 67, 1381–1396. doi: 10.1093/jxb/erv532
- Muleya, M., Tang, K., Broadley, M. R., Salter, A. M., and Joy, E. J. M. (2022). Limited supply of protein and lysine is prevalent among the poorest households in Malawi and exacerbated by low protein quality. *Nutrients* 14, 1–12. doi: 10.3390/nu14122430
- Müntz, K., and Lehmann, C. O. (1987). Reserveproteinforschung und genbank. *Die Kulturpflanze* 35, 25–52. doi: 10.1007/BF02163328
- Nobre, J. S., and Singer, J. M. (2011). Leverage analysis for linear mixed models. *J. Appl. Stat.* 38, 1063–1072. doi: 10.1080/02664761003759016
- OECD/FAO (2021). *Data about the production and use of wheat, maize, and soybean*. Available at: https://stats.oecd.org/Index.aspx?DataSetCode=HIGH_AGLINK_2021 (Accessed January 18, 2023).
- Oppermann, M. (2023). *Data from: IPK genebank accessions passport data snapshot 2023-02-17* (e!DAL - Plant Genomics & Phenomics Research Data Repository). doi: 10.5447/ipk/2023/6
- Park, T., and Casella, G. (2008). The bayesian lasso. *J. Am. Stat. Assoc.* 103, 681–686. doi: 10.1198/016214508000000337
- Pellett, P. L., and Ghosh, S. (2004). Lysine fortification: Past, present, and future. *Food Nutr. Bull.* 25, 107–113. doi: 10.1177/156482650402500201

- Pérez, P., and De Los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198, 483–495. doi: 10.1534/genetics.114.164442
- Philipp, N., Weise, S., Oppermann, M., Börner, A., Graner, A., Keilwagen, J., et al. (2018). Leveraging the use of historical data gathered during seed regeneration of an ex situ genebank collection of wheat. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00609
- Philipp, N., Weise, S., Oppermann, M., Börner, A., Keilwagen, J., Kilian, B., et al. (2019). Historical phenotypic data from seven decades of seed regeneration in a wheat ex situ collection. *Sci. Data* 6, 1–9. doi: 10.1038/s41597-019-0146-y
- Raffo, M. A., Sarup, P., Guo, X., Liu, H., Andersen, J. R., Orabi, J., et al. (2022). Improvement of genomic prediction in advanced wheat breeding lines by including additive-by-additive epistasis. *Theor. Appl. Genet.* 135, 965–978. doi: 10.1007/s00122-021-04009-4
- Ritchie, H., Reay, D. S., and Higgins, P. (2018). Beyond calories: A holistic assessment of the global food system. *Front. Sustain. Food Syst.* 2. doi: 10.3389/fsufs.2018.00057
- Rogers, J. S. (1972). “Measures of genetic similarity and genetic distance,” in *Studies in Genetics VII*, ed. Wheeler, M. (Austin, TX: University of Texas), 145–153.
- Sansaloni, C., Franco, J., Santos, B., Percival-Alwyn, L., Singh, S., Petroli, C., et al. (2020). Diversity analysis of 80,000 wheat accessions reveals consequences and opportunities of selection footprints. *Nat. Commun.* 11, 1–12. doi: 10.1038/s41467-020-18404-w
- Schulthess, A. W., Kale, S. M., Liu, F., Zhao, Y., Philipp, N., Rembe, M., et al. (2022). Genomics-informed prebreeding unlocks the diversity in genebanks for wheat improvement. *Nat. Genet.* 54, 1544–1552. doi: 10.1038/s41588-022-01189-7
- Schulthess, A. W., Wang, Y., Miedaner, T., Wilde, P., Reif, J. C., and Zhao, Y. (2016). Multiple-trait- and selection indices-genomic predictions for grain yield and protein content in rye for feeding purposes. *Theor. Appl. Genet.* 129, 273–287. doi: 10.1007/s00122-015-2626-6
- Shewry, P. R. (2009). Wheat. *J. Exp. Bot.* 60, 1537–1553. doi: 10.1093/jxb/erp058
- Shewry, P. R., and Hey, S. J. (2015). The contribution of wheat to human diet and health. *Food Energy Secur.* 4, 178–202. doi: 10.1002/fes3.64
- Tranquillini, W. (1963). The physiology of plants at high altitudes. *Annu. Rev. Plant Physiol.* 15, 345–362. doi: 10.1146/annurev.pp.15.060164.002021
- Ufaz, S., and Galili, G. (2008). Improving the content of essential amino acids in crop plants: goals and opportunities. *Plant Physiol.* 147, 954–961. doi: 10.1104/pp.108.118091
- VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980
- Vogel, K. P., Johnson, V. A., and Mattern, P. J. (1973). Results of systematic analyses for protein and lysine composition of common wheats (*Triticum aestivum* L.) in the USDA World Collection. *Historical Res. Bulletins Nebraska Agric. Experiment Station* 258, 1–27.
- Vogel, K. P., Johnson, V. A., and Mattern, P. J. (1975). Reevaluation of common wheat from the USDA world wheat collection for protein and lysine content. *Historical Res. Bulletins Nebraska Agric. Experiment Station* 272, 1–36.
- Vogel, K. P., Johnson, V. A., and Mattern, P. J. (1976). Protein and lysine content of grain, endosperm, and bran of wheats from the USDA world wheat collection. *Crop Sci.* 16, 655–660. doi: 10.2135/cropsci1976.0011183X001600050014x
- Wang, B., Wang, Y., Zhang, D., Li, H., and Yang, C. (2008). Verification of the resistance of a LEA gene from *Tamarix* expression in *Saccharomyces cerevisiae* to abiotic stresses. *J. For. Res.* 19, 58–62. doi: 10.1007/s11676-008-0010-y
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 1–9. doi: 10.1038/sdata.2016.18
- Witcombe, J. R. (1975). *University College Bangor NEPAL Expedition 1971* (Bangor: University College of North Wales).
- Witcombe, J. R., and Rao, A. R. (1976). The genecology of wheat in a Nepalese centre of diversity. *J. Appl. Ecol.* 13, 915–924. doi: 10.2307/2402266
- Yadav, A. K., Carroll, A. J., Estavillo, G. M., Rebetzke, G. J., and Pogson, B. J. (2019). Wheat drought tolerance in the field is predicted by amino acid responses to glasshouse-imposed drought. *J. Exp. Bot.* 70, 4931–4948. doi: 10.1093/jxb/erz224
- Yang, W., Zhang, L., Lv, H., Li, H., Zhang, Y., Xu, Y., et al. (2015). The K-segments of wheat dehydrin WZY2 are essential for its protective functions under temperature stress. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.00406
- Yu, X., Li, X., Guo, T., Zhu, C., Wu, Y., Mitchell, S. E., et al. (2016). Genomic prediction contributing to a promising global strategy to turbocharge gene banks. *Nat. Plants* 2, 1–7. doi: 10.1038/nplants.2016.150
- Yu, S., and Tian, L. (2018). Breeding major cereal grains through the lens of nutrition sensitivity. *Mol. Plant* 11, 23–30. doi: 10.1016/j.molp.2017.08.006
- Zan, T., Li, L., Li, J., Zhang, L., and Li, X. (2020). Genome-wide identification and characterization of late embryogenesis abundant protein-encoding gene family in wheat: Evolution and expression profiles during development and stress. *Gene* 736, 1–14. doi: 10.1016/j.gene.2020.144422
- Zhao, Y., Gowda, M., Longin, F. H., Würschum, T., Ranc, N., and Reif, J. C. (2012). Impact of selective genotyping in the training population on accuracy and bias of genomic selection. *Theor. Appl. Genet.* 125, 707–713. doi: 10.1007/s00122-012-1862-2