Populating a biodigital resource center for barley (*Hordeum* sp.) using historical records and genomic prediction

Dissertation
Zur Erlangung des
Doktorgrades der Agrarwissenschaften (Dr. agr.)

Der
Naturwissenschaftlichen Fakultät III
Agrar- und Ernährungswissenschaften,
Geowissenschaften und Informatik

Der Martin-Luther-Universität Halle-Wittenberg

Vorlegt von
Frau Maria Yuli González González
Geboren am 20.02.1983 in Cauca, Colombia

1. Gutachter: Prof. Dr. Jochen Reif
2. Gutachter: Prof. Dr. Jens Léon

Eingereicht am 10.05.2021
Verteidigt am 08.11.2021

Table of Contents

1. General introduction

1.1 Barley is an important crop

Barley is the fourth main cereal crop worldwide (FAO, 2020). Cereal crops are crucial resources for global food security and provide large energy and protein components of the human diet (Lafiandra et al. 2014). Europe accounts for 61.5% of the barley global production, which amounted to 155.9 million metrics tons in 2019. The main barley producing countries in Europe are Russia, Ukraine, Spain, Germany and France (FAO, 2020). The main end-uses are animal feed, brewing malts, and human food (Nevo 2013).

Barley plays an important role as a cereal model crop in research. An important driving force has been the development of a high-quality barley genome reference sequence. In combination with the well-established protocols for genome editing, this has spurred barley research that is progressing faster than ever (Mascher et al. 2017; Beier et al.2017; Kilian and Graner, 2012; Milner et al. 2019).

1.2 Barley breeding methodologies

Barley breeding has successfully developed varieties that meet the needs of crop production such as disease resistance and specific qualities for feeding, malting and food. Currently, a major challenge for breeders is to breed high-yielding varieties to feed a growing population under conditions of global warming (Voss-Fels et al. 2019). Conventional plant breeding of barley takes at least eight years until a new variety is developed. Therefore, new strategies have been developed to accelerate the improvement rate and without increasing the costs for breeding. These improved approaches use innovations in high-throughput genotyping, genome editing, genomic selection, and speed breeding (Watson et al. 2018).

Barley breeding focused on pure lines development whereas hybrids production displayed growing potential (Verstegen et al. 2014, Mühleisen et al. 2013). Common approaches to produce pure lines are bulk and pedigree selection schemes, and doubled haploid (DH) selection methodology (Verstegen et al. 2014). Barley varieties are classified based on growth habit and morphology. For instance, spring varieties have a broad adaptation to different environments and do not require vernalization while the winter type requires it. The morphology of the ear facilitates a classification of barley in six-rowed and two-rowed types. The awn-bearing lemma firmly attached to the grain stands for hulled types, but it can be easily separated in hull-less or naked type (Taketa et al. 2004).

1.3 Important breeding goals in barley

The main target traits for selection in barley breeding are grain yield, yield stability, yield quality, straw strength, winter hardiness, and tolerance to biotic and abiotic stress (Wiegmann et al. 2019). Agronomic traits such as flowering time, plant height, and thousand grain weight are also considered as they play a role in crop yield and stability. The timing of flowering directly influences grain yield, as it needs optimization to occur during specific seasons to avoid environmental stresses such as frost, heat, and drought. Short-stature cultivars are developed to improve yield by reducing lodging and increasing the harvest index. The most common genes leading to semi-dwarf cultivars are *uzu1*, *denso*, and *sdw1* (Hellewell et al. 2000; Saisho et al. 2004). However, the strong association to low vigor and the reduction for grain yield let to avoid the use of some dwarfing genes for breeding (Verstegen et al. 2014). Breeding varieties for feeding, brewing and food implies selection for starch, protein and fiber content (Meints and Hayes 2019; Verstegen et al. 2014). Relevant features are the strong negative correlation between yield and protein content, selection based on indirect traits to achieve malt extract requirements (Bhatta et al. 2020; Kunze, 2010; Li et al. 2010)

Disease resistance is an economic and ecological mechanism to counteract pathogens causing yield reduction in crops. Major diseases in barley are powdery mildew (Blumeria graminis), speckled leaf blotch (Septoria passerinii), scald leaf blotch (Rhynchosporium secalis), net bloch (Pyrenophora teres), barley rusts like stem rust and leaf rust (Puccinia graminis resp. Puccinia hordei), and soil borne mosaic viruses (Verstegen et al. 2014). Soil borne mosaic viruses poses a serious threat to winter barley production in East Asia and Europe with yield loses up to 50% (Plumb et al. 1986). The disease is caused by different strains of barley yellow mosaic virus (BaYMV-1 and BaYMV-2) and barley mild mosaic virus (BaMMV). The symptoms typically appear in January and the bymoviruses belong to the group of winter diseases in barley. BaYMV and BaMMV are transmitted by the vector *Polymyxa graminis*, which is an obligate biotrophic parasite of plant roots and belongs to the plasmodiophorids. Bymoviruses can survive over several years in the field within the protective dormant spores of *Polymyxa graminis* (Neuhauser et al. 2010). The most effective, economically sustainable, and environmentally friendly control method is the planting resistant cultivars (Finch et al. 2014; Kanyuka et al. 2003). Genetic mapping has shown that almost 20 genes are distributed in all chromosomes of barley, conferring resistance to BaYMV and BaMMV (Perovic et al. 2019). Until now, breeding of resistant lines depended mainly on phenotypic selection and, to a lesser extent, on marker-assisted selection (MAS). Backcrossing and pyramiding are the most used breeding methods to introduce resistance against several strains into a single genotype (Ordon et al. 1999; Werner et al. 2005).

1.4 Need of broadening the genetic basis of elite breeding pools

New cultivar improvement depends of functional genetic diversity to achieve a sustainable crop production under challenging and changing environmental and pest hazards conditions (Muñoz-Amatriaín et al. 2014; Tilman et al. 2011). In this sense, genebanks have been developing and implementing the operating procedures for seed storage and plant propagation for genetic diversity preservation. *Ex situ* genebanks world-wide host 7.4 million of accessions, and the implementation of a global seed vault safe let to guard more than one million safety duplicates on the Svalbard archipelago. However, the lack of characterization data hampers the detection of useful genetic variation and its utilization to address crop breeding needs.

In the mid-twentieth century, the establishment of genebanks for the conservation of plant genetic resources was promoted when high-performing varieties began to replace traditional landraces. As a result, half a million of barley accessions from worldwide resources have been collected and preserved in genebanks (Verstegen et al. 2014). The Federal *ex situ* Genebank for Agricultural and Horticultural Plant Species hosted at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) in Gatersleben (Germany) maintains ~22.000 barley accessions. It is the sixth largest collection worldwide and one of the most actives in germplasm distribution. The management of the IPK barley collection includes in-field multiplication, cold storage facilities since 1976, as well as the collection, preservation, storage, and quality assessment of characterization data (Börner 2006). Moreover, the data management switched from field worksheet and manual evaluation to automatized devises such as automatic Marvin digital seed analyzer and Personal Digital Assistants (PDAs).

The need to increase the efficiency in identification of useful genetic variation has been emphasized already since 1967. Nevertheless, genebank managers, plant geneticist and breeders still point to the need of systematic evaluation of comprehensive collections. In this direction, genomic approaches for genebank management and prebreeding have been proposed to facilitate the characterization and utilization of plant genetic resources. For instance, the application of molecular markers was initially promoted for population genetics studies. But the high cost of characterization with molecular markers allowed only a reduced number of accessions to be evaluated. Consequently, curators and breeders focused on core collections (Brown 1989) for genomic characterization. Recent advances such as affordable genome-wide genotyping and low cost in genotyping technologies enables the characterization for entire genebank collections (Kilian and Graner 2012). Genotyping-by-sequencing (GBS), an application of next generation technologies (NGS), had showed good performance in detecting rare variation in barley germplasm collections (Darrier et al. 2019). In this manner, the utilization of the single nucleotide polymorphisms (SNPs) could be facilitated by GBS and the reference sequence of barley variety Morex (Mascher et al. 2017). A significant amount of barley accessions has been tested by SNP markers including a core collection with ~2,400 accessions and an entire collection comprising ~22,000 accessions (Milner et al. 2019; Muñoz-Amatriaín et al. 2014). SNPs can be used as a molecular passport data to complement, corroborate and correct traditional passport records. There are challenges for genetic resources management worldwide with whose SNPs can deal, for instance tracking the identity, identifying duplicates within and between genebanks, and monitoring the genetic integrity (Mascher et al. 2019; Milner et al. 2019).

Potential users of genetic resources such as breeders need information on the value of plant genetic resources for improving important agronomic or quality traits. The phenotypic characterization for complete collections is demanding in time and economical resources. Therefore, unlocking the historical records collected during seed regeneration and diseases screening provide the opportunity to increase the amount of information in genebanks at no extra cost (de Carvalho et al. 2013; Gonzalez et al. 2018a,b; Gonzalez et al. 2021; Milner et al. 2019;

Philipp et al. 2019). Moreover, the amount of characterization data could be increased using historical data as a training set to predict the performance of non-phenotyped individuals using the tool box of genome-wide predictions. Genomic prediction for mining global gene banks scale can make profit from historical data. Thus, data integration between different genebanks requires especial attention (Crossa et al. 2016; Jarquin et al. 2016; Yu et al. 2016). In this regard, data sharing, documentation and archiving must follow the FAIR principles, i.e. data must be findable, accessible, interoperable and reusable (Wilkinson et al. 2016).

1.5 Data availability for genebank accessions

The majority of germplasm collections have only basic passport data available, but genebank users need information to select promising candidates for their own projects. Genetic resources contain various genes and traits needed to deal with current and future challenges. However, making available characterization data to the community requires several years of intensive phenotyping and specialized equipment (Anglin et al. 2018). The low-cost alternative is mining data in public repositories such as dataverse, CGspace, PGP, and datasets linked to published articles (Anglin et al. 2018; Arend et al. 2016; Milner et al. 2019). Digital platforms are for instance the BRIDGE portal (https://bridge.ipk-gatersleben.de/), komugi (https://shigen.nig.ac.jp/wheat/komugi/), and GrainGenes (https://wheat.pw.usda.gov/GG3/). Genebanks can feed their databases with information of genetic resources evaluated by researchers outside of the genebank. Moreover, there is the possibility of make available information recorded which is still stored on paper or excel sheets.

Historic phenotypic data accumulated during seed multiplication routine is a key resource of information to leverage the untapped biodiversity of genetic resources (Keilwagen et al. 2014).

However, biases are induced by disparity in protocols of evaluation, fluctuating weather conditions and changing agronomic management across years (Krajewski et al. 2015).

The IPK genebank preserves ~151.000 accessions for ~3.000 plant species. The seed regeneration performed for about 5% of the collection each year generates a huge amount of data because it is accompanied by a routine of phenotypic characterizations. Curators score field trials to protect the genetic constitution of the accessions, reduce the effects of natural selection, and control possible seed mixtures from past management activities at the genebank (Philipp et al. 2019). The historical data collected from field trials at IPK for germplasm of barley and wheat is publicly available on e!DAL-Plant Genomics and Phenomics Research Data Repository (PGP) (Arend et al. 2016). The GBS data for ~ 20.000 barley accession is also available at PGP repository. Part of the barley germplasm data could be found at BRIDGE including phenotypic traits of spikes (König et al. 2020) and EURISCO involving taxonomic data (Kreide et al. 2019) (Fig. 1).

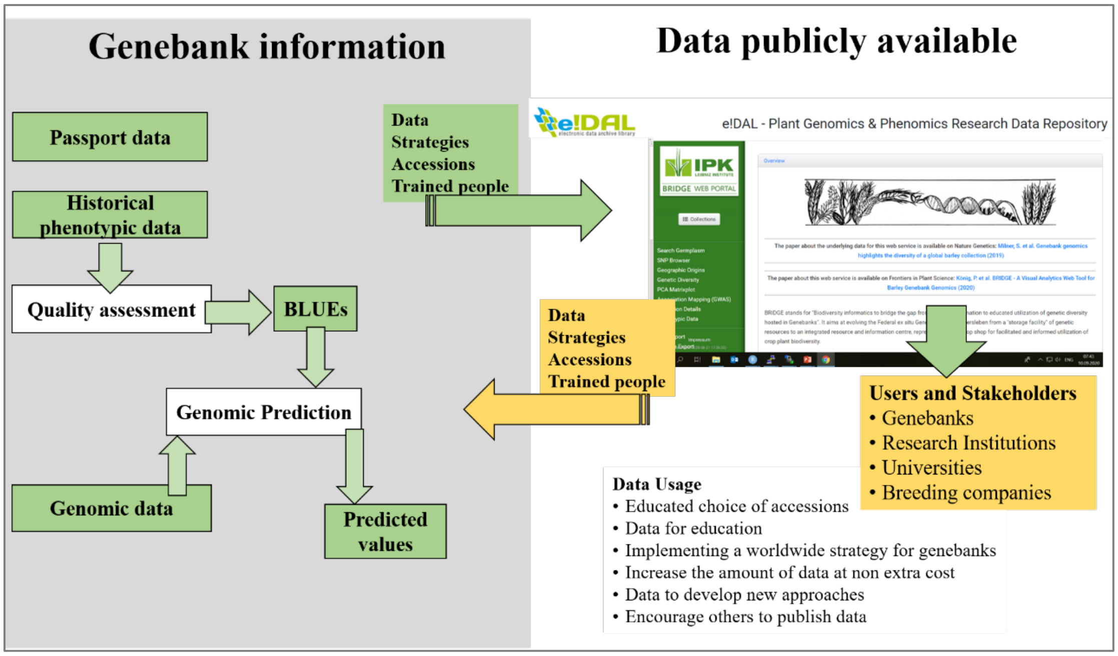

Figure 1. Publishing information of genebanks following the FAIR principles, i.e. data must be findable, accessible, interoperable and reusable, as strategy to make an efficient use of plant genetic resources.

Historical phenotypic data from seed regeneration and disease screenings for barley accessions are non-orthogonal across traits and years (Gonzalez et al. 2018b; Gonzalez et al. 2021; Philipp et al. 2019). In this respect, for barley, 12 accessions were tested for thousand grain weight in 1984, while 4,789 accessions were tested for plant height in 1970. Moreover, some accessions were tested more than once in a year before they were classified as winter or spring type. Over the period of seven decades, the accessions were evaluated up to 22 years. These huge datasets with information for up to ~13,000 accessions are from long periods of evaluation. The data from seed regeneration are very unbalanced, as the trials were not installed for gathering evaluation data. They were rather conducted in order to (i) preserve the genetic diversity stored under a size sample and quality thresholds pre-established, (ii) preserve new genotypes, (iii) research, and (iv) seed distribution (Börner 2006).

Determining precise estimates of the performance of accessions based on historical information requires an assessment of data quality. The evaluation strategy may include methods for outlier detection on unbalanced sets (Bernal-Vasquez et al. 2016; Estaghvirou et al. 2014) and bias assessment on first- and second-degree statistics (González et al. 2018a; Piepho and Möhring 2006). The possible bias results from accessions that were regenerated in blocks, depending on the year when they entered the genebank, especially before the establishment of cold storage facilities. The performance estimation must therefore deal with limitations such as biased estimations and unbalanced data sets. In this sense, the restricted maximum likelihood algorithm (REML), which is used to solve mixed model equations, becomes a suitable alternative (Patterson and Thompson 1971). Furthermore, working with historical data sets entails a management for huge amount of records and interdisciplinary knowledge for plausibility checks. Then, both, automated standard operating procedures and manual quality assessment, are relevant (González et al. 2018a; Mascher et al. 2019).

1.6 Exploitation of germplasm with the aid of genomics

The potential of genomic prediction and genome-wide association mapping encourages their utilization to improve efficiency in germplasm characterization. In particular, genomic prediction is suitable for complex traits given the estimation of all marker effects across the entire genome to calculate genomic estimate breeding values (GEBVs) (Heffner et al. 2009; Jiang et al. 2021; Waugh R. 2014). The promising performance of genomic prediction has been observed on a wide range of studies. For animal breeding, the selection in layer chickens based on genome-wide prediction outperformed the conventional breeding scheme for most of 16 traits (Wolc et al. 2015). Studies in dairy cattle showed increments in rates of genetic gain per year from 50% to 100% for yield traits, and between threefold to fourfold for lowly heritable traits since the implementation of genome-wide prediction (García-Ruiz et al. 2016).

Respecting plants, the assessment of genome-wide prediction started on populations from breeding programs. In this respect, correlations up to 0.79 among observed and predictive values indicated that genomic selection in plant breeding can be suitable for selecting among lines that are not phenotyped (Crossa et al. 2010). Genomic prediction has been also tested on plant genetic resources revealing promising results. For instance, the assessment of wheat landraces and genebank accessions of cauliflower showed prediction accuracies ranging from 0.16 to 0.67 (Crossa et al. 2016; Thorwarth et al. 2018). Moreover, using soybean germplasm accessions were achieved correlations between predicted and observed values up to 0.92 based on independent validation trials (Jarquin et al. 2016). For barley, moderate to high prediction accuracies have been reported (Nielsen et al. 2016; Philipp et al. 2016; Sallam et al. 2015; Thorwarth et al. 2017). Thus, genome-wide prediction is a promising tool to add information to plant genetic resources hosted in genebanks.

1.7 Objectives

The main goal of the present work was to examine strategies to advance the IPK barley collection into a bio-digital resource center facilitating an educated choice of genetic resources for research and breeding. In particular, the objectives were to:

(1) Develop and evaluate strategies to assess the quality of historic phenotypic data for plant height, flowering time, and thousand grain weight;

(2) Compare the phenotypic diversity of the collection with the phenotypic variation currently exploited in barley breeding in Central Europe;

(3) Illustrate the potential use of the phenotypic data in order to unlock the valuable diversity for research and breeding;

(4) Provide all research data and the presented results in a FAIR-way to be easily re-used; and

(5) Investigate the potential of genomic prediction based on historical screening data of plant responses against to *Barley yellow mosaic viruses* for populating the IPK bio-digital resource center of barley;

Peer-reviewed scientific articles

2. Unlocking historical phenotypic data from an *ex situ* collection to enhance the informed utilization of genetic resources of barley (*Hordeum* sp.)

Published in:

Theoretical and Applied Genetics (2018) 131:2009-2019.

DOI: 10.1007/s00122-018-3129-z

Authors: Maria Y. González, Norman Philipp, Albert W. Schulthess, Stephan Weise, Yusheng Zhao, Andreas Börner, Markus Oppermann, Andreas Graner, Jochen C. Reif

The original paper has been published and available online:

https://doi.org/10.1007/s00122-018-3129-z

Peer-reviewed scientific articles

3. Unbalanced historical phenotypic data from seed regeneration of a barley *ex situ* collection

Published in:

Scientific Data 5:180278

DOI: 10.1038/sdata.2018.278

Authors: Maria Y. González, Stephan Weise, Yusheng Zhao, Norman Philipp, Daniel Arend, Andreas Börner, Markus Oppermann, Andreas Graner, Jochen C. Reif & Albert W. Schulthess

The original paper has been published and available online:

https://www.nature.com/articles/sdata2018278.pdf

Peer-reviewed scientific articles

4. Genebank genomics highlights the diversity of a global barley collection

Published in:

Nature Genetics 51:319-326(2019)

DOI: 10.1038/s41588-018-0266-x

Authors: Sara G Milner, Matthias Jost, Shin Taketa , Elena Rey Mazón, Axel Himmelbach, Markus Oppermann, Stephan Weise, Helmut Knüpffer, Martín Basterrechea, Patrick Köni, Danuta Schüler, Rajiv Sharma, Raj K Pasam, Twan Rutten, Ganggang Guo, Dongdong Xu, Jing Zhang, Gerhard Herren, Thomas Müller, Simon G Krattinger, Beat Keller, Yong Jiang, Maria Y González , Yusheng Zhao , Antje Habekuß , Sandra Färber, Frank Ordon, Matthias Lange, Andreas Börner, Andreas Graner, Jochen C Reif, Uwe Scholz, Martin Mascher, Nils Stein

The original paper has been published and available online:

https://www.nature.com/articles/s41588-018-0266-x.pdf

Peer-reviewed scientific articles

5. 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 barley soilborne mosaic viruses

Published in:

Theoretical and Applied Genetics (2021)

DOI: 10.1007/s00122-021-03815-0

Authors: Maria Y. Gonzalez, Yusheng Zhao, Yong Jiang, Nils Stein, Antje Habekuss, Jochen C. Reif, Albert W. Schulthess

The original paper has been published and available online:

https://link.springer.com/content/pdf/10.1007/s00122-021-03815-0.pdf

6. General discussion

Plant genetic resources have the potential to contribute significantly to crop adaptation to a changing climate. However, the actual use of plant genetic resources in plant breeding is limited and contrasts with their potential. An important building block to exploit plant genetic resources in plant breeding is to make genomic and phenotypic information for plant genetic resources available for both breeding research and applied plant breeding. This can be done in combination with the development of a suitable bioinformatic toolbox that facilitates data mining in order to implement a targeted selection of promising plant genetic resources in plant breeding.

Internationally, there are some extensive projects to increase the information density of plant genetic resources. For example, in the Seeds of Discovery project (https://seedsofdiscovery.org/tag/cimmyt/), CIMMYT has systematically genotyped and phenotyped large parts of its extensive maize and wheat collections. The IPK in Gatersleben has also genotypically characterized its entire *ex situ* barley collection in the frame of the BRIDGE project. The activities in BRIDGE focused on the description and placing of the molecular diversity of the IPK barley collection and collection-related research questions such as redundancy within and between genebanks. For example, one result of these analyses was that the IPK barley collection covers a wide diversity when compared to the Barley Core Collection (BCC), which was established to represent the global barley diversity. However, the IPK collection has also collection gaps and accessions from Turkey and Central and Eastern Asian are underrepresented (Milner et al. 2019). An important component that was missing from the BRIDGE project activities was the generation and use of comprehensive phenotypic data. This was a major gap that we were able to fill in the context of this PhD by opening up historical data.

6.1 Large historical datasets of a diverse collection

The barley collection of the IPK comprises ~20,000 accessions. About 5% of the collection is regenerated annually in field plots of up to 3.75 m². During the regeneration process, phenotypic information is collected for each accession for traits such as plant height, flowering time, or thousand grain weight. In addition, sporadic attempts have been made in the past to characterize portions of the collection. For example, the Federal Research Center for Cultivated Plants (Julius Kühn-Institut, JKI) and its predecessor organizations have annually screened a small sample of barley accessions in the IPK collection for soil-borne diseases, a limiting factor for winter barley production for which the effective control method is breeding resistant varieties. In this way, extensive historical data has been accumulated over the years. Considering plant height, flowering date, thousand grain weight, and resistance to barley mosaic viruses BaYMV and BaMMV, up to 2,968 data points for winter barley and up to 9,898 data points for spring barley accessions are available in the IPK genebank documentation system (Fig. 2).

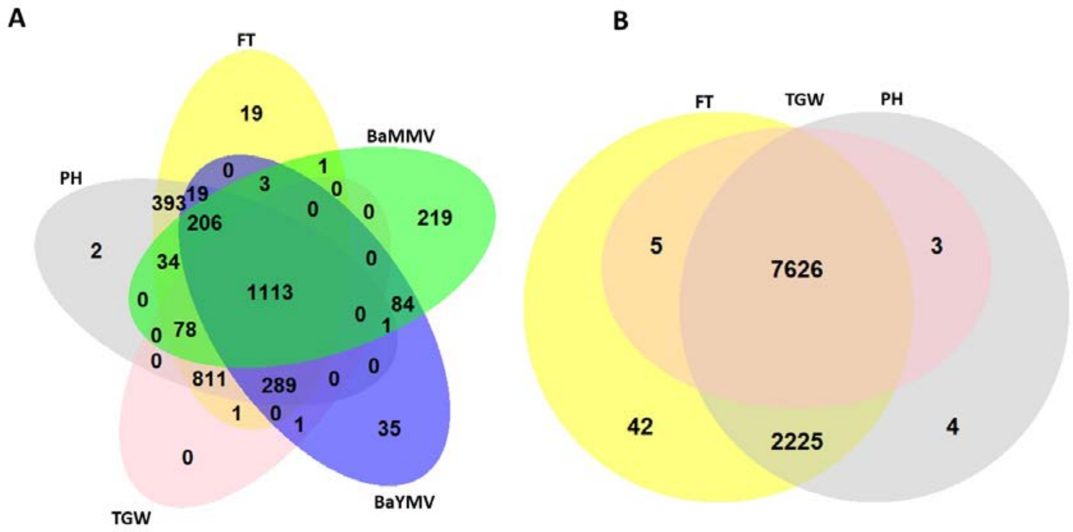

Figure 2. Venn diagrams showing the number of A) winter and B) spring barley accessions phenotyped for flowering time (FT), plant height (PH), thousand grain weight (TGW), and resistance against the barley mosaic viruses BaYMV and BaMMV.

The historical data stored at the IPK genebank documentation systems was an untapped treasure, reflecting a value of approximately 1.8 million euros, assuming a cost per plot of 30 euros and ignoring seed logistics costs. The challenge in using the unbalanced historical data was to implement an appropriate procedure for data curation and analysis. Mixed linear models using the restricted maximum likelihood algorithm (REML) (Patterson and Thompson 1971) are a powerful tool for analyzing such nonorthogonal data. The general form of a linear mixed model is $\mathbf{Y} = \mathbf{X}\hat{a} + \mathbf{Zu} + \mathbf{e},$ where $\mathbf{Y}$ corresponds to the response vector, $\mathbf{X}$ and $\mathbf{Z}$ are known design matrices, â is a vector of fixed effects, **u** corresponds to random effects and **e** stands for error terms. We used linear mixed models and implemented outlier detection to assess the quality of the historical data (Gonzalez et al. 2018a), which is a strategy successfully used to curate phenotypic data generated in plant breeding programs (Galiano-Carneiro et al. 2020; Neuweiler et al. 2020; Trini et al. 2020).

For the historical barley data, outlier detection using the rescaled mean absolute deviation of the standardized residuals and the Bonferroni-Holm test (Bernal-Vasquez et al. 2016) showed excellent results. A maximum of 2.5% of the data points in the historical data were detected as outliers and removed, which increased heritability by up to ~17% for flowering time, plant height, and thousand grain weight for both winter and spring barley (Gonzalez et al. 2018a). Consequently, the implemented data curation pipeline is suitable as a blueprint for further studies and is already being used in other international projects such as the 'Activated GEnebank NeTwork (AGENT)' (https://www.agent-project.eu/).

Timely curation of data has the advantage of incorporating further information for the evaluation of results in the case of outlier tests. Thus, the great experience of the curators in the genebank can be better used to differentiate technical or agronomic causes from actual biological outliers. Image data on field trials, for example by drone overflight, provide a simple initial way to document field conditions and should be used as standard for propagation crops in the future.

Missing data are common in designed experiments. However, in our case, the data from the seed regenerations have a missing value structure that deviates from a random scenario in the early years (1946-1976) before the introduction of seed cold storage. During this period, propagation occurred in blocks. Blocks of accessions often correspond to the year in which the accessions were added to the genebank, and they were generally assigned to the same geographic region. However, in our comprehensive resampling study, we clearly demonstrated that the bias in the variance of genotypes and residuals was, on average, negligible for the scenarios tested, and the genotypic values could also be estimated without bias. Consequently, the historical barley accession data are suitable to populate a bio-digital resource center for the IPK barley collection.

The phenotypic data demonstrated the impressive diversity in the barley collection at IPK Gatersleben. Spring barley showed ranges of 96 days for flowering time, 111.2 cm for plant height, and 68.7 g for thousand grain weight (Table 1). These ranges are substantially larger than those reported for other data sets. For example, Maurer et al (2015) observed a range of 50 days for flowering time in a population that included backcrosses with 25 exotic donors. In addition, the accessions evaluated for BaYMV and BaMMV also included highly resistant ones. These results demonstrate that the barley collection of IPK Gatersleben, is a promising resource to unlock useful diversity for plant breeding.

Table 1. Number of barley accessions, number of phenotypic records (total and average per accession), mean ± standard deviation, range, coefficient of variation (CV) of the best linear unbiased estimations (BLUEs), and heritability for Flowering time, Plant height, thousand grain weight, and BaYMV and BaMMV susceptibilities. * Stands for genomic heritability (Gonzalez et al. 2018a, b; Gonzalez et al. 2021).

Data Ressource	Trait	Type	No. of accessions	Phenotypic records	Records per accession	Mean±S.D.	Range	CV (%)	Heritability	Data Published on
Seed regeneration trials	Flowering time (days)	Spring	9.898	43.264	4.4	78.3±6.3	49.2-145.1	8.0	0.88	Original and processed data published on Plant Genomics and Phenomics Research Data Repository (PGP)
		Winter	2.968	10.101	3.5	146.8±3.3	128.3-178.2	3.3	0.83	
	Plant height (cm)	Spring	9.858	41.933	4.2	96.5±6.3	37.4-148.6	6.5	0.86	
		Winter	2.947	10.238	3.5	103.7±15.9	25.5-176.5	15.9	0.87	
	Thousand grain weight (g)	Spring	7.634	33.854	4.4	44.8±7.3	3.2-71.9	16.3	0.92	
		Winter	2.293	7.748	3.4	43.2±7.4	15.6-68.4	17.1	0.92	
Disease screening	BaYMV	Winter	1.751	4.145	2.4	3.7±2.5	-0.7-11.5	67.0	0.48*	Original data included as supplementary files for a paper DOI:10.1038/s41588-018-0266-x
	BaMMV	Winter	1.739	2.444	1.4	4.8±3.4	0.99-9.0	70.0	0.63*	

6.2 Association mapping as an entry point for allele mining and potential of marker-assisted selection

Association mapping is a promising approach for identifying candidate genes. This was impressively demonstrated with the comprehensive historical information in combination with genotyping-by-sequencing data in this work (Milner et al. 2019). Candidate gene information is an important first step for targeted allele mining, however, validation of candidate genes appears to be another important intermediate step.

Candidate gene information can be used to predict the performance of the many genotyped but not phenotyped barley accessions using the previously identified diagnostic molecular markers. This procedure is commonly referred to as marker-assisted selection (MAS). We tested the potential of MAS using virus resistance data and observed low to moderate prediction accuracies ranging from 0.24 to 0.42 for BaYMV and from 0.26 to 0.40 for BaMMV (Gonzalez et al. 2021). One explanation for this is a possibly complex genetic architecture of BaYMV and BaMMV in the diverse collection of barley accessions. This led us to investigate prediction models that might be better suited for predicting complex traits.

6.3 Filling the gaps of IPK barley collection by genomic prediction

Genomic prediction is a powerful tool to fill the gaps in genebank information for non-phenotyped accessions. Trait performance information is important for identifying promising accessions for research and breeding. The potential of genomic prediction has been investigated for soilborne viruses on winter barley (Gonzalez et al. 2021). Genomic prediction in winter barley included a training population with both phenotypic and genotypic data to predict the performance of non-phenotyped individuals that have genotyping-by-sequencing profiles. This model achieved high predictive abilities of 0.62 for BaYMV and 0.64 for BaMMV susceptibilities, substantially higher than the predictive accuracies in MAS. Our results are consistent with other studies that had successfully tested the potential of genomic prediction was in genetic resources of sorghum (Yu et al. 2016), wheat (Crossa et al. 2016), and cauliflower (Thorwarth et al. 2018). As a result, genebanks are encouraged to take advantage of the potential and limitations of genomic prediction to collect comprehensive information for their accessions. Consequently, genomic prediction has now been used for additional traits to estimate missing values from the traits flowering time, thousand grain weight, and plant height for the IPK barley collection (Jiang et al. 2021).

Previous studies showed the potential to increase the accuracy of the prediction via a proper model choice by accounting for fixed effect for known genes (Bernardo, 2014; Zhao et al 2014). We also tested such a model, the weighted genome-wide best linear unbiased prediction (W-BLUP), using historical screening data for soilborne diseases. The W-BLUP approach increased predictive ability by 3.0% for BaYMV and 5.0% for BaMMV compared with standard method. Therefore, W-BLUP is the appropriate approach to fill the gaps of BaYMV and BaMMV susceptibilities in winter barley.

In summary, with the previous studies, we were able to predict the performance of full winter and spring barley collection for flowering time, thousand grain weight, and plant height. For BaYMV and BaMMV susceptibilities, this was only successful for winter barley, as we felt that prediction from winter barley to spring barley was not promising due to massive differences between both populations, and no data were available for validation. Therefore, there is still a need to implement a method for predicting spring barley performance that can consider winter barley as a training set (Fig. 3). In this regard, Jiang et al. (2021) demonstrated for the traits flowering time, thousand grain weight, and plant height that spring barley performance can be predicted using a winter barley training set. However, the prediction abilities were significantly lower compared to those within the winter barley or spring barley group. Thus, the prediction has to be handled with caution and should be flanked by standard errors of the predictions.

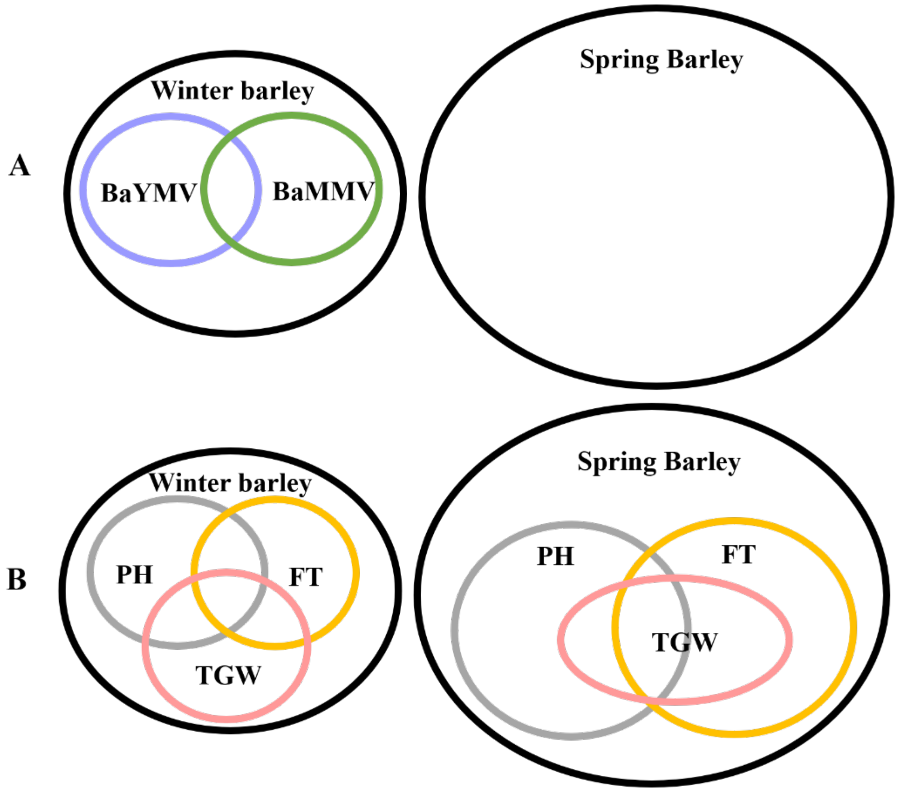

Figure 3. Venn diagrams showing the genotyped groups of winter and spring barley, and the phenotyped groups for (A) barley yellow mosaic virus (BaYMV), and barley mild mosaic virus (BaMMV), and B) the phenotyped groups on seed regeneration trials for flowering time (FT), plant height (PH) and thousand grain weight (TGW).

6.4 Populating the IPK barley bio-digital resource center

A necessary strategy to activate genebank collections, provide easy access to plant diversity, and facilitate the selection of useful genetic variations, is to further develop the collections into bio-digital resource centers. In these, relevant information on accessions should be available and searchable. Furthermore, it is necessary to provide and integrate tools for data mining for different questions such as allele mining. This approach entails interdisciplinary efforts including agronomy, plant genetics, seed biology, and computer skills for data management.

The IPK genebank is actively working to make data publicly available to users. Recently, the BRIDGE web tool was developed to serve as a data warehouse and exploratory data analysis for the IPK barley collection. The information housed corresponds to passport, genotypic and phenotypic data, and ear images. Genotyping-by-sequencing profiles for 22,621 accessions and phenotypic records for 9,527 accessions with at least one observed phenotype are included. Functionalities included are data export, SNP browser, interactive world map, interactive scatter plots to examine population structure using Principal Component Analysis, and Manhattan plots of genome-wide association studies. In addition, BRIDGE is linked to the IPK's Genebank Information System (GBIS) for germplasm ordering services (König et al. 2020). BRIDGE follows FAIR principles (Wilkinson et al. 2016) and minimum information requirements for a plant phenotyping experiment (MIAPPE), which are essential for exploring genomic strategies among genebanks worldwide (Yu et al. 2016). BRIDGE, as IPK's barley bio-digital resource center, can now be extended for the data curated in this work. The developed blueprint is not limited to barley but also urges for other species.

6.5 Global view

There are approximately half a million barley accessions hosted at genebanks worldwide and for parts of them genotypic and phenotypic information have been generated that could be used in a global virtual genebank. In this regard core collections representing the genetic diversity for complete collections, have been tested for several traits, focusing in those which are difficult or expensive to score. For instance, powdery mildew was screened for 223 accessions from Czech, 159 accessions from Spain, and 93 accessions from Serbia (Silvar et al., 2011, Dreiseitl and Zavřelová, 2018, Šurlan-Momirović et al., 2016). Net blotch resistance was evaluated on 336 accessions from ICARDA’S germplasm (Amezrou et al., 2018). Moreover, feed quality traits was assessed on 1,480 accessions of the USDA barley core collection (Bowman et al. 2001). However, most of the times the evaluation data is produced for a reduced number of accessions. In contrast, some extensive collections have been generated huge data sets. Such as, the unbalanced historical data sets from seed regeneration and disease screening routine performed at IPK genebank with up to ~ 13.000 (60% of the total collection) accessions tested (Gonzalez et al. 2018a; Milner et al. 2019). Moreover, datasets on evaluation data of USDA barley genetic resources tested for a period of 20 years for diseases (Barley yellow dwarf, spot blotch, net blotch, stripe rust) and insect resistance (Russian wheat aphid) for up to 24,800 (75% of the total collection) (Bonman et al. 2005).

The phenotypic data generated for the above-mentioned core collections are a valuable resource that can be used as training data for genomic prediction models. A prerequisite, of course, is that genomic data can be integrated across genebanks. This is certainly possible for similar marker systems such as genotyping-by-sequencing or DarTSeq technology and represents the nucleus for designing bio-digital resource centers on a global scale.

7. Summary

Plant genetic resources contain the genetic diversity needed in plant breeding to achieve a sustainable crop production. However, the lack of trait performance estimates limits the selection of promising candidates of genebanks collections. This study explored potential strategies to populate a bio-digital resource center of barley by mining historical data of seed regeneration trials, and using genomic prediction based on historical information of plant responses against to *Barley yellow mosaic viruses*.

The historical data of seed regeneration trials involves records collected for seven decades for flowering time, plant height and thousand grain weight for up to 12.872 accessions of spring and winter barley. This unbalanced data was analyzed using a quality assessment routine and linear mixed models. Outlier removal lead to increase up to 17% on heritability estimates. The resampling study showed on average a negligible bias for variance of genotypes and residuals, and best linear unbiased estimations (BLUEs) could also be obtained without bias. The BLUEs showed a broad phenotypic variation for all tested traits, which also revealed high heritability estimates ranging from 0.83 to 0.92. These findings highlight the suitability of phenotypic values to be used as training set for genomic prediction. The original and processed data is available under the FAIR principles (findable, accessible, interoperable and reusable).

The potential of genomic prediction was tested using a training population with both phenotypic and genotypic data to predict the performance of non-phenotyped individuals that have genotyping-by-sequencing profiles. The study included information for barley yellow mosaic virus (BaYMV) susceptibility for 1,751 accessions, barley mild mosaic virus (BaMMV) susceptibility for 1,771 accessions, and single nucleotide polymorphism profiles (SNP) for 3,838 winter barley accessions. The prediction abilities were computed as correlations between the predicted and observed phenotypes. The marker assisted selection method (MAS) showed low to moderate prediction abilities amounting to 0.42. The genomic best linear unbiased

prediction (GBLUP) revealed higher values than MAS, with prediction abilities of 0.62 for BaYMV and 0.64 for BaMMV. When markers with significant major effects got more weight as in weighted genome best linear unbiased prediction (W-BLUP), the prediction abilities increased up to 5% respecting those of GBLUP. Thus, W-BLUP is the appropriate approach to predict the performance of non-phenotyped accessions for BaYMV and BaMMV susceptibilities in winter barley.

Genebanks are encouraged to unlock historical phenotypic data and test genomic prediction to fill the gaps of phenotypic information. The developed blueprint allowed to leverage a large data set, and could be adapted to other collections to promote the utilization plant genetic resources for crop improvement.

8. Zusammenfassung

Pflanzengenetische Ressourcen enthalten die genetische Vielfalt, die in der Pflanzenzüchtung benötigt wird, um eine nachhaltige Pflanzenproduktion zu erreichen. Das Fehlen von Schätzungen der Merkmalsleistung schränkt jedoch die Selektion vielversprechender Kandidaten aus den Sammlungen von Genbanken ein. Diese Studie untersuchte mögliche Strategien zur Bestückung eines bio-digitalen Ressourcenzentrums für Gerste, indem historische Daten von Saatgut-Regenerationsversuchen ausgewertet wurden und genomische Vorhersagen auf der Grundlage historischer Informationen über Resistenzen gegen Gersten-Gelbmosaikviren getroffen wurden.

Die historischen Daten von Saatgut-Regenerationsversuchen umfassen Aufzeichnungen, die über sieben Jahrzehnte für Blütezeit, Pflanzenhöhe und Tausendkorngewicht für bis zu 12.872 Akzessionen von Sommer- und Wintergerste gesammelt wurden. Diese unausgewogenen Daten wurden mit einer Qualitätsbewertungsroutine und linearen gemischten Modellen analysiert. Die Entfernung von Ausreißern führte zu einer Erhöhung der Heritabilitätsschätzungen um bis zu 17%. Die Resampling-Studie zeigte im Durchschnitt eine vernachlässigbare Verzerrung der Schätzung der Varianz der Genotypen und Residuen, und die Best Linear Unbiased Estimations (BLUEs) konnten ebenfalls ohne Verzerrung ermittelt werden. Die BLUEs zeigten eine breite phänotypische Variation für alle getesteten Merkmale, die auch hohe Heritabilitätsschätzungen im Bereich von 0,83 bis 0,92 ergaben. Diese Ergebnisse unterstreichen die Eignung der phänotypischen Werte als Trainingsset für die genomische Vorhersage. Die ursprünglichen und verarbeiteten Daten sind unter Einhaltung der FAIR-Prinzipien (findable, accessible, interoperable and reusable) verfügbar.

Das Potenzial der genomischen Vorhersage wurde unter Verwendung einer Trainingspopulation mit sowohl phänotypischen als auch genotypischen Daten getestet, um die

Leistung von nicht phänotypisierten Individuen vorherzusagen, die Genotyping-by-Sequencing-Profile aufweisen. Die Studie umfasste Informationen zur Anfälligkeit für Gerstengelbmosaikvirus (BaYMV) für 1.751 Akzessionen, Anfälligkeit für Gerstenmildmosaikvirus (BaMMV) für 1.771 Akzessionen und Einzelnukleotid-Polymorphismus-Profile (SNP) für 3.838 Wintergersten-Akzessionen. Die Vorhersagegenauigkeiten wurden als Korrelationen zwischen den vorhergesagten und beobachteten Phänotypen berechnet. Die Marker-gestützte Selektion (MAS) zeigte eine geringe bis mittlere Vorhersagegenauigkeit in Höhe von 0,42. Die Genomic Best Linear Unbiased Predictions (GBLUP) zeigten höhere Werte als die MAS, mit Vorhersagefähigkeiten von 0,62 für BaYMV und 0,64 für BaMMV. Wenn Marker mit signifikanten Haupteffekten mehr Gewicht bekamen, wie bei der gewichteten Genomic Best Linear Unbiased Predictions (W-BLUP), stiegen die Vorhersagefähigkeiten um bis zu 5 % im Vergleich zu denen von GBLUP. Somit ist W-BLUP der geeignete Ansatz, um die Leistung von nicht phänotypisierten Akzessionen für BaYMV- und BaMMV-Anfälligkeiten in Wintergerste vorherzusagen.

Genbanken werden ermutigt, historische phänotypische Daten zu erschließen und genomische Vorhersagen zu testen, um die Lücken der phänotypischen Informationen zu füllen. Die entwickelte Blaupause ermöglichte die Nutzung eines großen Datensatzes und kann für andere Sammlungen angepasst werden, um die Nutzung pflanzengenetischer Ressourcen für die Verbesserung von Kulturpflanzen zu fördern.

9. General references

- Amezrou R, Verma RPS, Chao S, Brueggeman RS, Belqadi L, Arbaoui M, Rehman S and Gyawali S (2018) Genome-wide association studies of net form of net blotch resistance at seedling and adult plant stages in spring barley collection. Mol Breeding 38, 58
- Anglin NL, Amri A, Kehel Z, Ellis D (2018) A Case of Need: Linking Traits to Genebank Accessions. Biopreserv Biobank 16:337-349
- Arend D, Junker A, Scholz U, Schüler D, Wylie J, Lange M (2016) PGP repository: a plant phenomics and genomics data publication infrastructure. Database: the journal of biological databases and curation 2016
- Bhatta M, Gutierrez L, Cammarota L, Cardozo F, Germán S, Gómez-Guerrero B, Pardo MA, Lanaro V, Sayas M, Castro AJ (2020). Multi-trait genomic prediction model increased the predictive ability for agronomic and malting quality traits in barley (*Hordeum vulgare* L.). G3-Genes Genom Genet 10(3), 1113-1124.
- Beier S, Himmelbach A, Colmsee C, Zhang X-Q, Barrero RA, Zhang Q, Li L, Bayer M, Bolser D, Taudien S, Groth M, Felder M, Hastie A, Šimková H, Staňková H, Vrána J, Chan S, Muñoz-Amatriaín M, Ounit R, Wanamaker S, Schmutzer T, Aliyeva-Schnorr L, Grasso S, Tanskanen J, Sampath D, Heavens D, Cao S, Chapman B, Dai F, Han Y, Li H, Li X, Lin C, McCooke JK, Tan C, Wang S, Yin S, Zhou G, Poland JA, Bellgard MI, Houben A, Doležel J, Ayling S, Lonardi S, Langridge P, Muehlbauer GJ, Kersey P, Clark MD, Caccamo M, Schulman AH, Platzer M, Close TJ, Hansson M, Zhang G, Braumann I, Li C, Waugh R, Scholz U, Stein N, Mascher M (2017) Construction of a map-based reference genome sequence for barley, *Hordeum vulgare* L. Scientific Data 4:170044
- Bernal-Vasquez A-M, Utz H-F, Piepho H-P (2016) Outlier detection methods for generalized lattices: a case study on the transition from ANOVA to REML. Theoretical and Applied Genetics 129:787-804
- Bernardo R (2014) Genomewide selection when major genes are known. Crop Sci 54:68-75
- Böner A (2006) Preservation of plant genetic resources in the biotechnology era. Biotechnol J 1:1393-1404
- Bonman J, MBockelman HE, Jackson F and Steffenson B J (2005) Disease and insect resistance in cultivated barley accessions from the USDA National Small Grains Collection. Crop Science 45(4): 1271-1280
- Bowman J, Blake T, Surber L, Habernicht D and Bockelman H (2001) Feed-Quality Variation in the Barley Core Collection of the USDA National Small Grains Collection. Crop Sci 41: 863-870
- Brown A (1989) Core collections: a practical approach to genetic resources management. Genome 31:818-824
- Crossa J, de Los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713-724
- Crossa J, Jarquin D, Franco J, Perez-Rodriguez P, Burgueno J, Saint-Pierre C, Vikram P, Sansaloni C, Petroli C, Akdemir D, Sneller C, Reynolds M, Tattaris M, Payne T, Guzman C,

Pena RJ, Wenzl P, Singh S (2016) Genomic Prediction of Gene Bank Wheat Landraces. G3:Genes Genom Genet 6:1819-1834

- Darrier B, Russell J, Milner SG, Hedley PE, Shaw PD, Macaulay M, Ramsay LD, Halpin C, Mascher M, Fleury DL (2019) A comparison of mainstream genotyping platforms for the evaluation and use of barley genetic resources. Frontiers in plant science 10:544
- de Carvalho MAAP, Bebeli PJ, Bettencourt E, Costa G, Dias S, Santos TMMD, Slaski JJ (2013) Cereal landraces genetic resources in worldwide GeneBanks. A review. Agronomy for Sustainable Development 33:177-203
- Dreiseitl A and Zavřelová M (2018) Identification of barley powdery mildew resistances in gene bank accessions and the use of gene diversity for verifying seed purity and authenticity. PloS one 13(12), e0208719
- Estaghvirou SBO, Ogutu JO, Piepho H-P (2014) Influence of outliers on accuracy estimation in genomic prediction in plant breeding. G3: Genes, Genomes, Genetics 4:2317-2328
- FAO (2020) Food outlook: biannual report on global food markets. http://www.fao.org/documents/card/es/c/ca9509en/ Accessed 12 Mar 2020
- Finch HJS, Samuel AM, Lane GPF (2014) Cereals. In: Finch HJS, Samuel AM, Lane GPF (eds) Lockhart & Wiseman’s Crop Husbandry Including Grassland. 9th edn. Woodhead Publishing, Cambridge, UK, pp 287-336
- Galiano-Carneiro AL, Kessel B, Presterl T, Miedaner T (2021) Intercontinental trials reveal stable QTL for Northern corn leaf blight resistance in Europe and in Brazil. Theor Appl Genet 134: 63–79
- García-Ruiz A, Cole JB, VanRaden PM, Wiggans GR, Ruiz-López FJ, Van Tassell CP (2016) Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proceedings of the National Academy of Sciences 113:E3995-E4004
- González MY, Philipp N, Schulthess AW, Weise S, Zhao Y, Börner A, Oppermann M, Graner A, Reif JC (2018a) Unlocking historical phenotypic data from an *ex situ* collection to enhance the informed utilization of genetic resources of barley (*Hordeum* sp.). Theor Appl Genet 131:2009-2019
- Gonzalez MY, Weise S, Zhao Y, Philipp N, Arend D, Börner A, Oppermann M, Graner A, Reif JC, Schulthess AW (2018b) Unbalanced historical phenotypic data from seed regeneration of a barley *ex situ* collection. Scientific Data 5:180278
- Gonzalez MY, Zhao Y, Jiang Y, Stein N, Antje Habekuss A, Reif JC, Schulthess AW (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.
- Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. Crop Science 49:1-12
- Hellewell KB, Rasmusson DC, Gallo-Meagher M (2000) Enhancing yield of semidwarf barley. Crop Science 40:352-358
- Jarquin D, Specht J, Lorenz A (2016) Prospects of genomic prediction in the USDA soybean germplasm collection: historical data creates robust models for enhancing selection of accessions. G3 (Bethesda) 6:2329-2341

- Jiang Y, Weise S, Graner A and Reif JC (2021) Using Genome-Wide Predictions to Assess the Phenotypic Variation of a Barley (*Hordeum* sp.) Gene bank collection for important agronomic traits and passport information. Frontiers in plant science 11: 2180
- Kanyuka K, Ward E, Adams MJ (2003) Polymyxa graminis and the cereal viruses it transmits: a research challenge. Molecular Plant Pathology 4:393-406
- Keilwagen J, Kilian B, Özkan H, Babben S, Perovic D, Mayer KFX, Walther A, Poskar CH, Ordon F, Eversole K, Börner A, Ganal M, Knüpffer H, Graner A, Friedel S (2014) Separating the wheat from the chaff – a strategy to utilize plant genetic resources from *ex situ* genebanks. Scientific Reports 4:5231
- Kilian B, Graner A (2012) NGS technologies for analyzing germplasm diversity in genebanks. Briefings in Functional Genomics 11:38-50
- König P, Beier S, Basterrechea M, Schüler D, Arend D, Mascher M, Stein N, Scholz U, Lange M (2020) BRIDGE – A visual analytics web tool for barley genebank genomics. Frontiers in Plant Science 11
- Krajewski P, Chen D, Ćwiek H, van Dijk AD, Fiorani F, Kersey P, Klukas C, Lange M, Markiewicz A, Nap JP (2015) Towards recommendations for metadata and data handling in plant phenotyping. Journal of Experimental Botany 66:5417-5427
- Kreide S, Oppermann M, Weise S (2019) Advancement of taxonomic searches in the European search catalogue for plant genetic resources. Plant Genetic Resources: Characterization and Utilization 17:559-561
- Kunze W (2010) Technology brewing and malting, 4th updated edn. VLB, Berlin, p 609
- Lafiandra D, Riccardi G, Shewry PR (2014) Improving cereal grain carbohydrates for diet and health. J Cereal Sci 59(3):312-326
- Li CD, Cakir M, Lance R. (2010) Genetic Improvement of Malting Quality through Conventional Breeding and Marker-assisted Selection. In: Zhang G, Li C (eds) Genetics and Improvement of Barley Malt Quality. Advanced Topics in Science and Technology in China. Springer, Berlin, Heidelberg, pp 260-292
- Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, Bayer M, Ramsay L, Liu H, Haberer G, Zhang X-Q, Zhang Q, Barrero RA, Li L, Taudien S, Groth M, Felder M, Hastie A, Šimková H, Staňková H, Vrána J, Chan S, Muñoz-Amatriaín M, Ounit R, Wanamaker S, Bolser D, Colmsee C, Schmutzer T, Aliyeva-Schnorr L, Grasso S, Tanskanen J, Chailyan A, Sampath D, Heavens D, Clissold L, Cao S, Chapman B, Dai F, Han Y, Li H, Li X, Lin C, McCooke JK, Tan C, Wang P, Wang S, Yin S, Zhou G, Poland JA, Bellgard MI, Borisjuk L, Houben A, Doležel J, Ayling S, Lonardi S, Kersey P, Langridge P, Muehlbauer GJ, Clark MD, Caccamo M, Schulman AH, Mayer KFX, Platzer M, Close TJ, Scholz U, Hansson M, Zhang G, Braumann I, Spannagl M, Li C, Waugh R, Stein N (2017) A chromosome conformation capture ordered sequence of the barley genome. Nature 544:427
- Mascher M, Schreiber M, Scholz U, Graner A, Reif JC, Stein N (2019) Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. Nature Genetics 51:1076-1081
- Maurer A, Draba V, Jiang Y, Schnaithmann F, Sharma R, Schumann E, Kilian B, Reif JC and Pillen K (2015) Modelling the genetic architecture of flowering time control in barley through nested association mapping. BMC Genomics 16:290

- Meints B, Hayes PM (2020) Breeding naked barley for food, feed, and malt. In: Goldman I (ed) Plant Breeding Reviews. Wiley, Hoboken, pp 95-119
- Milner SG, Jost M, Taketa S, Mazón ER, Himmelbach A, Oppermann M, Weise S, Knüpffer H, Basterrechea M, König P, Schüler D, Sharma R, Pasam RK, Rutten T, Guo G, Xu D, Zhang J, Herren G, Müller T, Krattinger SG, Keller B, Jiang Y, González MY, Zhao Y, Habekuß A, Färber S, Ordon F, Lange M, Börner A, Graner A, Reif JC, Scholz U, Mascher M, Stein N (2019) Genebank genomics highlights the diversity of a global barley collection. Nature Genetics 51:319-326
- Mühleisen J, Maurer HP, Stiewe G, Bury P and Reif JC (2013) Hybrid breeding in barley. Crop Science 53: 819-824
- Muñoz-Amatriaín M, Cuesta-Marcos A, Endelman JB, Comadran J, Bonman JM, Bockelman HE, Chao S, Russell J, Waugh R, Hayes PM (2014) The USDA barley core collection: genetic diversity, population structure, and potential for genome-wide association studies. PLoS One 9:e94688
- Neuhauser S, Bulman S, Kirchmair M (2010) Plasmodiophorids: The challenge to understand soil-borne, obligate biotrophs with a multiphasic life cycle. In: Gherbawy Y, Voigt K (eds) Molecular Identification of Fungi. Springer, Berlin, pp 51-78
- Neuweiler JE, Maurer HP, WürschumT (2020) Long-term trends and genetic architecture of seed characteristics, grain yield and correlated agronomic traits in triticale (×Triticosecale Wittmack). Plant Breed 139: 717– 729
- Nevo E (2013) Evolution of wild barley and barley improvement. In: Zhang GL, Liu X (eds) Advance in Barley Sciences, Proceedings of 11th International Barley Genetics Symposium. Springer, Dordrecht pp 1-23
- Nielsen NH, Jahoor A, Jensen JD, Orabi J, Cericola F, Edriss V, Jensen J (2016) Genomic prediction of seed quality traits using advanced barley breeding lines. PLoS One 11:e0164494
- Ordon F, Schiemann A, Pellio B, Dauck V, Bauer E, Streng S, Friedt W, Graner A (1999) Application of molecular markers in breeding for resistance to the Barley yellow mosaic virus complex/Einsatzmöglichkeiten molekularer Marker in der Resistenzzüchtung gegen den Gelbmosaikviruskomplex der Gerste. Zeitschrift für Pflanzenkrankheiten und Pflanzenschutz/Journal of Plant Diseases and Protection 256-264
- Patterson H, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. Biometrika 58(3):545–554
- Perovic D, Kopahnke D, Habekuss A, Ordon F, Serfling A (2019) Marker-based harnessing of genetic diversity to improve resistance of barley to fungal and viral diseases. In: Miedaner T and Korzun V (eds) Applications of genetic and genomic research in cereals. Woodhead Publishing, Series in Food Science, Technology and Nutrition, pp 137-164
- Philipp N, Liu G, Zhao Y, He S, Spiller M, Stiewe G, Pillen K, Reif JC, Li Z (2016) Genomic prediction of barley hybrid performance. The plant genome 9:1-8
- Philipp N, Weise S, Oppermann M, Börner A, Graner A, Keilwagen J, Kilian B, Zhao Y, Reif JC, Schulthess AW (2018) Leveraging the use of historical data gathered during seed regeneration of an *ex situ* genebank collection of wheat. Frontiers in Plant Science 9
- Philipp N, Weise S, Oppermann M, Börner A, Keilwagen J, Kilian B, Arend D, Zhao Y, Graner A, Reif JC, Schulthess AW (2019) Historical phenotypic data from seven decades of seed regeneration in a wheat *ex situ* collection. Scientific Data 6:137

- Piepho H-P, Möhring J (2006) Selection in Cultivar Trials—Is It Ignorable? Crop Science 46:192-201
- Plumb RT, Lennon EA, Gutteridge RA (1986) The effects of infection by barley yellow mosaic virus on the yield and components of yield of barley. Plant Pathol 35:314-318
- Saisho D, Tanno K-i, Chono M, Honda I, Kitano H, Takeda K (2004) Spontaneous Brassinolide-insensitive Barley Mutants 'uzu' Adapted to East Asia. Breeding Science 54:409-416
- Sallam AH, Endelman JB, Jannink J-L, Smith KP (2015) Assessing genomic selection prediction accuracy in a dynamic barley breeding population. The Plant Genome 8(1)
- Silvar C, Casas AM, Kopahnke D, Habekuß A, Schweizer G, Gracia MP, Lasa JM, Ciudad FJ, Molina-Cano J, Igartua E and Ordon F (2010) Screening the spanish barley core collection for disease resistance. Plant Breeding 129: 45-52
- Šurlan-Momirović G, Flat K, Silvar C Branković G, Kopahnke D, Knežević D, Schliephake E, Ordon F and Perović D (2016) Exploring the Serbian genbank barley (*Hordeum vulgare* L. subsp. vulgare) collection for powdery mildew resistance. Genet Resour Crop Evol 63: 275–287
- Taketa S, Kikuchi S, Awayama T, Yamamoto S, Ichii M, Kawasaki S (2004) Monophyletic origin of naked barley inferred from molecular analyses of a marker closely linked to the naked caryopsis gene (nud). Theoretical and Applied Genetics 108:1236-1242
- Thorwarth P, Ahlemeyer J, Bochard A-M, Krumnacker K, Blümel H, Laubach E, Knöchel N, Cselényi L, Ordon F, Schmid KJ (2017) Genomic prediction ability for yield-related traits in German winter barley elite material. Theoretical and Applied Genetics 130:1669-1683
- Thorwarth P, Yousef EAA, Schmid KJ (2018) Genomic prediction and association mapping of curd-related traits in gene bank accessions of cauliflower. G3:Genes Genomes Genetics 8:707-718
- Tilman D, Balzer C, Hill J, Befort BL (2011) Global food demand and the sustainable intensification of agriculture. Proceedings of the national academy of sciences 108:20260-20264
- Trini J, Maurer HP, Weissmann S, Würschum T (2020) Hybrid breeding for biomass yield in winter triticale: II. Combining ability and hybrid prediction. Plant Breed 139: 906– 915.
- Verstegen H, Köneke O, Korzun V, von Broock R (2014) The world importance of barley and challenges to further improvements. In: Kumlehn J, Stein N (eds) Biotechnological approaches to barley improvement. Springer, Berlin, pp 3-19
- Voss-Fels KP, Cooper M, Hayes BJ (2019) Accelerating crop genetic gains with genomic selection. Theoretical and Applied Genetics 132:669-686
- Watson A, Ghosh S, Williams MJ, Cuddy WS, Simmonds J, Rey M-D, Asyraf Md Hatta M, Hinchliffe A, Steed A, Reynolds D, Adamski NM, Breakspear A, Korolev A, Rayner T, Dixon LE, Riaz A, Martin W, Ryan M, Edwards D, Batley J, Raman H, Carter J, Rogers C, Domoney C, Moore G, Harwood W, Nicholson P, Dieters MJ, DeLacy IH, Zhou J, Uauy C, Boden SA, Park RF, Wulff BBH, Hickey LT (2018) Speed breeding is a powerful tool to accelerate crop research and breeding. Nature Plants 4:23-29
- Waugh R. Thomas B, Flavell A, Ramsay L, Comadran J, Russell J (2014) Genome-Wide Association Scans (GWAS). In: Kumlehn J, Stein N (eds) Biotechnological Approaches to Barley Improvement. Springer, Berlin, pp 345-365

- Werner K, Friedt W, Ordon F (2005) Strategies for Pyramiding Resistance Genes Against the Barley Yellow Mosaic Virus Complex (BaMMV, BaYMV, BaYMV-2). Molecular Breeding 16:45-55
- Wiegmann M, Maurer A, Pham A, March TJ, Al-Abdallat A, Thomas WTB, Bull HJ, Shahid M, Eglinton J, Baum M, Flavell AJ, Tester M, Pillen K (2019) Barley yield formation under abiotic stress depends on the interplay between flowering time genes and environmental cues. Scientific Reports 9:6397
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Neuhauser alez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, ’t Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3:160018
- Wolc A, Zhao HH, Arango J, Settar P, Fulton JE, O’sullivan NP, Preisinger R, Stricker C, Habier D, Fernando RL (2015) Response and inbreeding from a genomic selection experiment in layer chickens. Genetics Selection Evolution 47:59
- Yu X, Li X, Guo T, Zhu C, Wu Y, Mitchell SE, Roozeboom KL, Wang D, Wang ML, Pederson GA, Tesso TT, Schnable PS, Bernardo R, Yu J (2016) Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nat Plants 2:16150
- Zhao Y, Mette MF, Gowda M, Longin CFH, Reif JC (2014) Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. Heredity 112:638-645

10. List of general abbreviations

Abbreviation	Explanation
AGENT	Activated GEnebank NeTwork
BaMMV	Barley Mild Mosaic Virus
BaYMV	Barley Yellow Mosaic Virus
BCC	Barley Core Collection
BLUE	Best Linear Unbiased Estimate
BRIDGE	Data warehouse and exploratory data analysis tool for genebank genomics of barley
CIMMYT	Maize and Wheat Improvement Center, Mexico
CV	Coefficient of Variation
DH	Doubled Haploid
EURISCO	European Search Catalogue for Plant Genetic Resources
FAIR	Findable, Accessible, Interoperable, Reusable
FAO	Food and Agriculture Organisation
FT	Flowering Time
GEBVs	Genomic estimate breeding values
GBIS	Genebank Information System
GBS	Genotyping by Sequencing
GBLUP	Genome-Wide Best Linear Unbiased Prediction
IPK	Leibniz Institute of Plant Genetics and Crop Plant Research
ICARDA	International Center for Agriculture Research in the Dry Areas
JKI	Julius Kühn-Institut
ISA-Tab	Investigation-Study-Assay: data and metadata format
MAS	Marker-Assisted Selection
MIAPPE	Minimum Information About a Plant Phenotyping Experiment
NGS	Next Generation Technologies
PGP	Genomics and Phenomics Research Data Repository
PGR	Plant Genetic Resources
PH	Plant Height
PDAs	Personal Digital Assistants
REML	Restricted Maximum Likelihood
SNP	Single Nucleotide Polymorphism
TGW	Thousand Grain Weight
USDA	United States Department of Agriculture
W-BLUP	Weighted Genome-wide Best Linear Unbiased Prediction

11. Acknowledgements

I would like to thank my first supervisor Prof. Dr. Jochen Reif for his support and kindness. I have been feeling fortunate for being a member of QG group. I want to gratefully thank Dr. Albert Schulthess, Dr. Norman Philipp, Dr. Yusheng Zhao and Dr. Yong Jiang because of their valuable contributions to my research. Thank you to all co-authors for their contribution to the peer-reviewed research articles. I want to thank my colleagues at the quantitative genetics group because of the friendship and support. Thank you, Fang Liu, Moritz Lell, Maximilian Rembe, Jie Zhang, Johannes Schneider, Valentin Hinterberger, Jianting Chu, Lars Otto, and Beate Fraust. I would like to thank Britt Leps because of the support for all needed for my stay in Germany during the PhD.

Thank you to my old friends Lisney, Leidy, Yohana, Carlos, Miguel, Jose, Eloina, Daniel, Luis Guillermo, Ivan, Juan, Ivona, Alberth, Antonio, Fernando and Jessica. I have enjoyed many years of your friendship and it was very important during my PhD studies. Thanks to my family. I appreciate the persistence, hard work, ethics, and courage of my family. Thanks to Jose, Elber and Marisol because despite the distance we were connected.

12. *Curriculum vitae*

Personal information

Name:	Maria Yuli Gonzalez
Date of birth:	20th May 1983
Nationality:	Colombian

Work Experience

Ph.D. Student in Agricultural Sciences **Mar 2017 - Present**

The Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Department of breeding research, Quantitative genetics group.

- Analysis and interpretation of phenotypic and marker data of barley germplasm.
- Assessment of breeding tools such as genome-wide prediction and marker assisted selection on barley germplasm.

Research Coordinator **May 2016 - Feb 2017**

Colombian Rubber Research Center (CENICAUCHO).

- Project management in crop physiology and technology transfer in three locations. Working in collaboration with research, educational and commercial organizations, and growers.
- Collaborator in the organization of an International Conference.

Master Researcher **Oct 2013 - Aug 2015**

Colombian Agricultural Research Center (AGROSAVIA formerly CORPOICA).

- Management of field trials of commercial materials of oil palm stablished in four locations nationwide, including evaluations of yield and plant growth, and coordination of agronomic practices. Training and supervising technicians and field workers. Data analysis and experimental design. Budget management.
- Submitting proposals for collection and evaluation of rubber tree germplasm of diverse origins as the basis for the Colombian breeding program.
- Cooperation and collaboration in an interdisciplinary environment and across research stations.
- Networking with crop growers and members of government institutions.

Assistant Researcher **Nov 2006 - Oct 2013**

Colombian Oil Palm Research Center (CENIPALMA), Department of Biology and Plant Breeding.

- Management of field trials of commercial materials stablished in two locations, including evaluations of yield and plant growth, and coordination of agronomic practices with agronomists of commercial plantations.
- Designing and implementing strategies for evaluation and management of germplasm of African and American oil palm origins.
- Parental selection based on phenotypic and molecular marker data. Production and distribution of seeds for research proposes. Implementing a methodology for seed conservation and germination.
- Improving data collection by implementing trials scoring with PDAs and supervision. Experimental design and data analysis using Statistix, SAS, PowerCore, PopGene, and R packages. Training and supervising technicians and field workers.
- Cooperation and collaboration in experiments of water management, drought resistance, pollination, cloning and techniques of replanting oil palm. Interdisciplinary work with colleagues from areas such as plant pathology, biostatistics, physiology, biotechnology, soil and geomatics.
- Writing reports and publication of the results in peer-reviewed journals.

Education

Ph.D. Student in Agricultural Sciences **Mar 2017 - Present**

The Martin Luther University of Halle-Wittenberg, Germany. Research topic: “Populating a biodigital resource center for barley (*Hordeum sp.*) using historical records and genomic prediction”.

M.Sc. In Agricultural Sciences, Genetics and Plant Breeding **Feb 2009 - Apr 2012**

National University of Colombia, Bogotá, Colombia.

B.S. In Agronomy **Sep 2001 - Sept 2007**

National University of Colombia, Bogotá, Colombia.

Honors

- Academic excellence. 1st Colombian Congress of Biochemistry and Molecular Biology: “Genetic Diversity of Natural Accessions of Oil Palm *Elaeis Oleifera* (H.B.K) Cortés. (Poster).
- Honorific mention (best oral presentation for students) at the Colombian Association of Plant Breeding and Crop Production conference: Morpho-agronomic characterization and evaluation of oil palm *Elaeis guineensis* Jacq from Angola (Conference).

- Scholarship “Young researchers and innovators - Virginia Gutierrez de Pineda”. The Colombian National Institute of Science and Technology (Colciencias).
- Sixth place on the national test of all Agronomy Faculties (ECAES).
- Undergraduate scholarship program "Prestamo beca". Agronomy College, National University of Colombia**.**

Publications

- **González, M.Y.,** Zhao, Y. *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. *Theoretical and Applied Genetics*.
- Milner, S., Jost, M., Taketa, S., Mazón, E., Himmelbach, A., Oppermann, M., Weise, S., Knüpffer, H., Basterrechea, M., König, P., Schüler, D., Sharma, R., Pasam, R., Rutten, T., Guo, G., Xu, D., Zhang, J., Herren, G., Müller, T., Krattinger, S., Keller, B., Jiang, Y., **González, M.Y.** *et al***.** (2019). Genebank genomics highlights the diversity of a global barley collection. *Nature Genetics* 51, 319-326.
- **González, M.Y.,** Weise S., Zhao, Y. *et al*. (2018). Unbalanced historical phenotypic data from seed regeneration of a barley *ex situ* collection. *Scientific Data* 5 (180278).
- **González, M.Y.,** Philipp, N., Schulthess, A.W. *et al*. (2018). Unlocking historical phenotypic data from an *ex situ* collection to enhance the informed utilization of genetic resources of barley (*Hordeum* sp.). *Theoretical and Applied Genetics* 131(9), 2009-2019.
- Arias, D., **González, M.,** Prada, F. *et al*. (2015). Genetic and phenotypic diversity of natural American oil palm (*Elaeis oleifera* (HBK) Cortés) accessions. *Tree Genetics & Genomes* 11:122.
- Arias, D., **González, M.,** & Romero, H. (2015). Genetic diversity and establishment of a core collection of oil palm (*Elaeis guineensis* Jacq) based on molecular data. *Plant Genetic Resources* 13(3), 256-265.
- **González, M.,** Sánchez, Y., Flórez, V., & Chaves, B. (2013). Biomass distribution efficiency of rose cv. Charlotte grown in soil and substrates. *Agronomía Colombiana*, 31(3), 304-313.
- Arias, D., **González, M.** *et al*. (2012). Morpho-agronomic and molecular characterization of oil palm *Elaeis guineensis* Jacq. material from Angola. *Tree Genetics & Genomes* 9:1283.
- **González, M.,** & **Romero**, H. (2010). Evaluating different sources of potassium to reduce the doubling in oil palm leaf. *Revista Palmas 31(3),* 17-25.
- Ruiz, R., **González, M.,** & Romero, H. (2009). Effect of replanting systems in the production of oil palm in the North Zone of Colombia. *Revista Palmas 30(4),* 42-52.

Maria Yuli González

13. Eidesstattliche Erklärung / Declaration under Oath

Ich erkläre, an Eides statt, dass ich die Arbeit selbstständig und ohne fremde Hilfe verfasst, keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. / I declare under penalty of perjury that this thesis is my own work entirely and has been written without any help from other people. I used only the sources mentioned and included all the citations correctly both in word or content.

Datum / Date

Unterschrift / *Signature*

14. Erklärung über bestehende Vorstrafen und anhängige Ermittlungsverfahren/ Declaration concerning Criminal Record and Pending Investigations

Hiermit erkläre ich, dass ich weder vorbestraft bin noch dass gegen mich Ermittlungsverfahren anhängig sind. / *I hereby declare that I have no criminal record and that no preliminary investigations are pending against me.*

______________________________ ______________________________

Datum / Date Unterschrift des Antragstellers / *Signature of the applicant*