nature genetics

ARTICLES

https://doi.org/10.1038/s41588-018-0266-x

Genebank genomics highlights the diversity of a global barley collection

Sara G. Milner[1,12], Matthias Jost[1,9,12], Shin Taketa[2], Elena Rey Mazón[1], Axel Himmelbach[1], Markus Oppermann[1], Stephan Weise[1], Helmut Knüpffer[1], Martín Basterrechea[1], Patrick König[1], Danuta Schüler[1], Rajiv Sharma[1,10], Raj K. Pasam[1,11], Twan Rutten[1], Ganggang Guo[3], Dongdong Xu[3], Jing Zhang[3], Gerhard Herren[4], Thomas Müller[4], Simon G. Krattinger[4,5], Beat Keller[4], Yong Jiang[1], Maria Y. González[1], Yusheng Zhao[1], Antje Habekuß[6], Sandra Färber[6], Frank Ordon[6], Matthias Lange[1], Andreas Börner[1], Andreas Graner[1], Jochen C. Reif[1], Uwe Scholz[1], Martin Mascher[1,7]* and Nils Stein[1,8]*

Genebanks hold comprehensive collections of cultivars, landraces and crop wild relatives of all major food crops, but their detailed characterization has so far been limited to sparse core sets. The analysis of genome-wide genotyping-by-sequencing data for almost all barley accessions of the German ex situ genebank provides insights into the global population structure of domesticated barley and points out redundancies and coverage gaps in one of the world's major genebanks. Our large sample size and dense marker data afford great power for genome-wide association scans. We detect known and novel loci underlying morphological traits differentiating barley genepools, find evidence for convergent selection for barbless awns in barley and rice and show that a major-effect resistance locus conferring resistance to bymovirus infection has been favored by traditional farmers. This study outlines future directions for genomics-assisted genebank management and the utilization of germplasm collections for linking natural variation to human selection during crop evolution.

Large collections of plant genetic resources for most, if not all, crop plants are maintained in germplasm collections (so-called genebanks or seedbanks) around the world. Even for a single species, the holdings of the major national genebanks can amount to tens of thousands of accessions. The main source of information on individual accessions are 'passport' descriptions detailing taxonomic status, collection sites and provenance of the material. However, these records often date back to decades before electronic data processing. Thus, some of the information originally kept on paper files may have been lost in translation both from analog to digital format and during material exchange between genebanks. Numerous studies in plant and animal species[1,2] have shown that patterns of genetic differentiation reflect geographic origins and major germplasm divisions created by agricultural practices. Assignments to genetically defined populations can thus complement written records and expert knowledge of curators in charge of maintaining and evaluating accessions. Genetic profiles for many genetically diverse genebank accessions can guide conservation decisions and supplement incomplete passport records. Moreover, they constitute a permanent resource for connecting genetic diversity and phenotypic variation by means of association mapping[3,4], supporting the use of traits locked in genebank material for use in plant breeding.

Here, we report the collection of genetic profiles for the entire barley (*Hordeum vulgare* L.) collection of the German federal ex situ genebank hosted at IPK Gatersleben. We used this comprehensive dataset to understand the composition of IPK's barley collection in the context of global barley diversity and combined it with historic and newly collected phenotypic data on morphological and agronomic characters to find genes and loci selected during crop evolution.

Results

Molecular passport data for an entire genebank collection. We analyzed genotyping-by-sequencing (GBS) data from a total of 22,626 DNA samples (Supplementary Table 1). The majority of these were derived from single plants of 21,405 accessions of the IPK barley collection[5]. We also included single-plant samples for 297 accessions of the collection of the National Crop Genebank of China at the Institute of Crop Sciences of the Chinese Academy of Agricultural Sciences, 684 barley accessions of the Swiss national genebank of Agroscope and 240 GBS samples from a previous study[2,6]. Our panel includes both domesticated barley and its conspecific wild progenitor *H. vulgare* ssp. *spontaneum* (K. Koch) Thell. (henceforth 'wild barley'). All GBS experiments were performed

[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Seeland, Germany. [2]Institute of Plant Science and Resources, Okayama University, Kurashiki, Japan. [3]Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China. [4]Department of Plant and Microbial Biology, University of Zurich, Zurich, Switzerland. [5]Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. [6]Institute for Resistance Research and Stress Tolerance, Julius Kühn Institute (Federal Research Centre for Cultivated Plants), Quedlinburg, Germany. [7]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany. [8]Center for Integrated Breeding Research, Georg-August-Universität Göttingen, Göttingen, Germany. [9]Present address: Agriculture and Food, The Commonwealth Scientific and Industrial Research Organisation, Canberra, Australia. [10]Present address: University of Dundee at the James Hutton Institute, Invergowrie, UK. [11]Present address: Department of Economic Development, Jobs, Transport and Resources, Centre for AgriBioscience, Agriculture Victoria Research, Bundoora, Victoria, Australia. [12]These authors contributed equally: Sara G. Milner, Matthias Jost. *e-mail: mascher@ipk-gatersleben.de; stein@ipk-gatersleben.de

Table 1 | Number of segregating SNPs detected by GBS in wild and domesticated barley samples

	Number of SNPs	SNPs with MAF ≥ 1%	SNPs with MAF ≥ 5%
All samples (n = 22,626)	171,263	23,908	15,683
Domesticated barleys (n = 19,778)	76,102	22,356	15,872
Wild barleys (n = 1,140)	127,408	46,392	20,511

using the same two-enzyme (PstI-MspI) protocol[7,8]. After read alignment to the reference genome sequence of barley cv. Morex[9] (Supplementary Fig. 1), we detected 171,263 bi-allelic SNPs that passed our filters for missing rate (<10%) and heterozygosity (<10%). Most of the variants (86%) had a minor allele frequency (MAF) below 1%, although there were also 15,683 variants with a MAF ≥5% (Table 1).

A principal component analysis (PCA) of wild and domesticated germplasm recapitulated the previously reported clear-cut genetic differentiation according to domestication status (Fig. 1a). As in other studies[2,10], we found discrepancies between our genetic clustering and the taxonomic status reported in the passport records, such as wild accessions purportedly originating from outside the Fertile Crescent or even from the Americas. Several regions other than the Middle East—the primary habitat of wild barley—such as Tibet, Morocco and Ethiopia have been proposed as centers of origin of the barley crop. However, the genetic and archeological evidence for these scenarios is scant[10–14]. PCA-based reassignment of domestication status guided by the high-confidence set of wild barleys of Russell et al. [2] defined bona fide sets of 1,140 wild and 19,778 domesticated barley accessions for further analysis.

A PCA on domesticated barleys showed that geography at the continental scale is the most important correlate of genetic structure. The first four principal components (PCs), which together explain 7.1% of the variance, correspond to geographic factors: PC1 separates Eastern and Western barleys, PC2 sets Ethiopian barley apart (Fig. 1b), and PC3 and PC4 correspond to further geographic subdivisions (Supplementary Fig. 2a). When only frequent variants (MAF ≥5%) were considered, the results remained qualitatively unchanged, although the proportion of variance explained by the first two PCs increased, a pattern also observed in other species[15] (Supplementary Fig. 3). Germplasm groups defined by ADMIXTURE[16] with the number of ancestral populations (k) ranging from 2 to 12 (Fig. 1c and Supplementary Fig. 4a) corresponded to discrete clusters in the PCA space (Supplementary Fig. 4b). At k = 3, a division according to Western, Eastern or Ethiopian origin was evident. As k increased, the cross-validation error decreased, and the proportion of samples assigned to populations reached a plateau at k = 7 (Supplementary Fig. 5). Finer subdivisions at higher k could be meaningfully interpreted. In addition to geographic factors, annual growth habit and morphological characters related to end-use quality (row type, grain cover) were major determinants of population divisions (Supplementary Fig. 2 and Supplementary Table 2). Samples from Southwest Asia, barley's center of origin, harbored several ancestry components also found in other parts of the world. For example, all major ancestry components of Northern European barleys were also found in Middle Eastern material. The higher genetic diversity in Middle Eastern material is also reflected by a faster decay of linkage disequilibrium compared with other germplasm groups (Supplementary Fig. 6). As expected[9], European barleys were divided into six-rowed and two-rowed types, and the majority of ancestry components present in North American barleys trace back to Europe. Notably, two-rowed spring barleys (red shading for k = 12 in Fig. 1c) were under-represented in North American germplasm. Ethiopian barleys were divided into naked and hulled types (Fig. 1c). Interestingly, samples from the Arabian Peninsula shared an ancestry component (cyan shading in Fig. 1c) otherwise restricted to Ethiopian barleys. Apparent discrepancies between ancestry assignment and recorded provenance (such as Western ancestry in Ethiopian barleys) can be explained by erroneous passport data or the use of exotic germplasm in elite breeding programs since the beginning of the twentieth century. Finally, 373 landrace accessions without recorded countries of origin were assigned to ADMIXTURE groups (k = 12, ancestry coefficient q ≥ 0.7), indicating that genetic analyses can complement traditional genebank documentation.

Implications for conservation management. To assess how well the IPK collection (as represented by single plants per accession) captures global barley diversity, we compared it to two independent collections, the International Barley Core Collection (BCC) and 79 diverse wild barleys from the panel of Russell et al.[2]. The BCC was compiled by an international panel of expert curators and encompasses material from ex situ collections around the world[17]. As IPK maintains a copy of the BCC, single-plant samples of 1,107 BCC accessions were included in all our analyses (Supplementary Table 1). We projected the domesticated barley accessions of the IPK genebank onto the eigenvectors defined by a PCA on the BCC samples. The diversity space spanned by the BCC was well covered by the IPK samples (Fig. 2a). By contrast, we detected a pronounced under-representation of some regions of the world in IPK's wild barley collection. While 383 of 1,140 bona fide wild barleys of IPK's collection originated from Israel, other regions of the Fertile Crescent were under-represented. For example, the IPK genebank does not host a single wild barley accession from Turkey, and coverage for the Central Asian eastern range of wild barley is also sparse (Fig. 2b).

Genetic profiles can be used to determine the similarity between samples to find pairs of genetically nearly identical samples, that is, potential duplicates. Identifying and handling redundancies within and among germplasm collections has long been recognized as one of the key challenges of genebank management[18]. Differences in maintenance practices as well as incomplete documentation of material exchange between genebanks complicate the identification of duplicates solely based on passport records[18]. Based on pairwise identity-by-state (IBS) comparisons (Supplementary Fig. 7), we clustered domesticated samples from the IPK collection with very few differences at our GBS SNP loci in 2,229 groups of closely related accessions containing between 2 and 112 members and comprising 8,804 samples in total. Thus, the proportion of potential duplicates in IPK's barley collection (33%) exceeds previous estimates obtained from the perusal of passport records of IPK and other genebanks[19]. A likely reason is that duplicates were not tracked when merging the former national genebanks of East and West Germany in the early 2000s. Our IBS analysis did not take into account intra-accession diversity, which has been reported in ex situ genebank accessions of barley[20]. To get a glimpse into the genetic diversity within accessions, we selected at random 32 domesticated accessions for which we genotyped 10 individual plants (Supplementary Table 1). We observed varying degrees of intra-accession diversity: 11 accessions had fewer than 20 homozygous differences between any of their individuals, while the maximum divergence between sample pairs from 5 accessions was in the range of inter-accession diversity (Supplementary Fig. 8). Thus, our set of highly similar single-plant samples will be a starting point for further phenotypic and genotypic analysis of the corresponding accessions to inform genebank management decisions, always keeping in mind that duplicates can serve as safety backups[19].

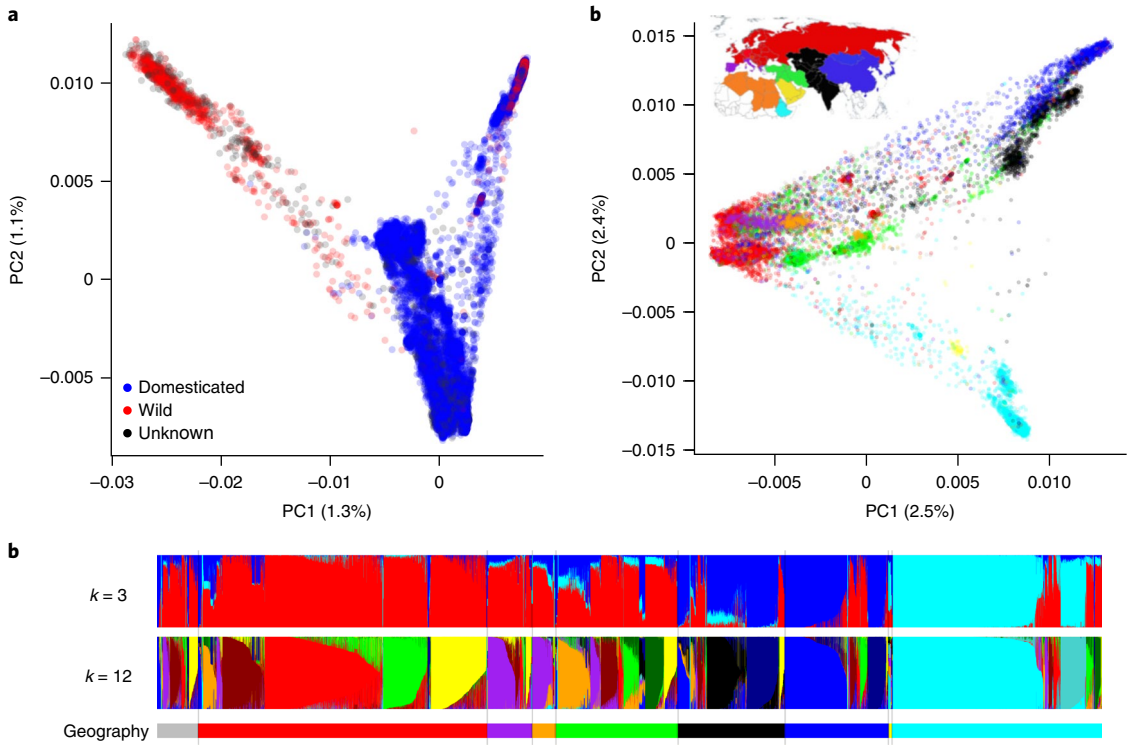

Fig. 1 | Genetic structure of barley ex situ accessions. a, PCA of 22,621 samples of wild and domesticated barley based on 171,263 SNP markers. Samples are colored according to domestication status in the genebank passport data. Five samples were excluded, likely due to incorrect species assignments. **b**, PCA of 19,778 domesticated barleys based on 76,102 markers. Samples are colored according to geographic origin. The color code is defined in the inset map, which was created with the R package mapdata. The proportion of variance explained by the PCs is indicated in the axis labels in **a** and **b**. **c**, ADMIXTURE ancestry coefficients ($k=3$, 12) for 17,640 samples of known provenance. The colored blocks below the bar plots correspond to the regional groupings in **b**. The gray block corresponds to North American samples.

In addition to identifying potential duplicates, our IBS analysis also allowed us to compare genetic similarity between geographically defined germplasm groups. Distribution of IBS values were usually multimodal (Fig. 2c), in line with the co-occurrence of divergent genepools (e.g. six- versus two-rowed and hulled versus naked types) in major barley-growing regions. IBS distributions between regional groups differed also in mean. Notably, Ethiopian barley samples were more closely related to each other than barleys within other groups even after discarding potential duplicates. Our panel includes 5,201 Ethiopian accessions from the IPK genebank (24.3% of its collection). Possible reasons for the over-representation of Ethiopian accessions are past collaborations between the German and Ethiopian genebanks and a preference for collection trips in a well-known center of crop diversity[21].

In summary, our analyses advocate for a reallocation of genebank management resources from the maintenance of duplicated or highly similar material towards a targeted augmentation of the collection with accessions of crop wild relatives[22], either by creation of new collections or exchange with other genebanks.

Genome-wide association studies with genebank material. One reason for maintaining large collections of plant genetic resources is to provide breeders with the raw material for crop improvement. Our high-density marker data for a large number of genotypes in combination with a commensurate quantity of phenotype observations lends it well to genome-wide association studies (GWASs) to define genetic loci where natural sequence diversity translates into variation of agronomic characters.

To conduct association scans with our GBS data, we first performed imputation of missing genotype calls to increase marker density. We defined a sparse genotype matrix with up to 95% missing data (that is, at least ~1,000 present calls) and filled missing values with an algorithm designed for inbreeding crops[23] with high accuracy ($R^2=0.97$). We defined a core set of 1,000 domesticated samples covering the diversity space of our total collection (Supplementary Tables 3 and 4 and Supplementary Fig. 9). Below, we report association scans on our core set using imputed GBS data (Figs. 3–5) as well as stringently filtered GBS data without imputation (Supplementary Fig. 10a–c). We assessed the feasibility of GWASs with highly heritable morphological traits that are routinely scored during genebank propagation and are, as components of infraspecific taxonomy[24], even part of passport records.

Association scans for morphological characters. One of the iconic traits of barley genetics is the fertility of lateral florets, commonly referred to as row type. Each node of the inflorescence stem (rachis) of so-called two-rowed barleys bears a spikelet triplet composed of a fertile central spikelet that sets seeds and two lateral spikelets with infertile florets. The suppression of lateral spikelet development is abolished, and grain number per spike tripled, in six-rowed types, which are now prevalent in most barley-growing regions of the world. However, two-rowed barleys still predominate in the Middle East (Supplementary Table 2) and are often favored by the malting industry for the higher uniformity of their grains. We mapped row type in our 1,000-sample core set. Seedlings whose DNA was used for GBS were grown to maturity, and seed set in lateral spikelets was recorded. Genome-wide association scans (Fig. 3a and Supplementary Fig. 10a) revealed peaks close to the major row-type genes *SIX-ROWED SPIKE1* (*VRS1*; ref. [25]) and *INTERMEDIUM-C* (*INT-C*, a modifier of lateral spikelet fertility; ref. [26]). While loss-of-function alleles of *VRS1* arose in six-rowed barleys after domestication, the *INT-C* allele predominant in six-rowed types (*Int-c.a*)

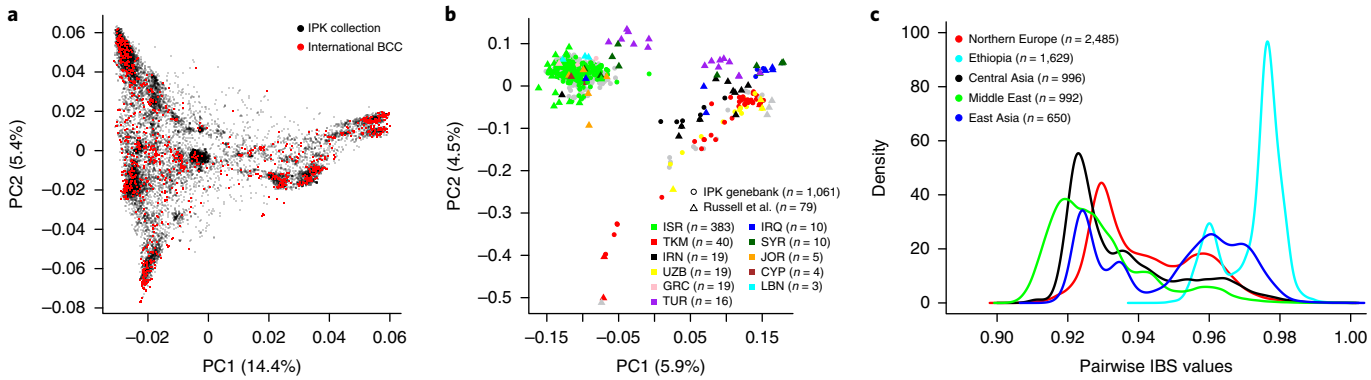

Fig. 2 | Genebank genomics. a, Domesticated barley samples from the IPK collection (n = 17,612; gray) were projected on the PCs defined by 1,102 samples from the BCC (red) using 17,944 markers with a MAF ≥5% in the BCC samples. Five BCC accessions were omitted because their domestication status did not agree with PCA results (Fig. 1). **b**, A total of 1,140 wild barleys from the IPK collection were projected on the PCs defined by 79 geographically diverse wild barleys from the panel of Russell et al.[2] using markers with a MAF ≥5% in the latter set. Samples are colored according to geographic origin as indicated in the legend. Country codes are used according to ISO 3166-1 α-3. **c**, Distribution of pairwise IBS values in regional germplasm groups.

was likely selected from standing variation in the wild progenitor[27]. In addition to these two loci long known to barley geneticists[28], we detected a third strong peak in the proximal region of chromosome 1H, coincident with a region of strong genetic differentiation between two-rowed and six-rowed barleys (Fig. 3b). We speculate that the 1H locus harbors allelic variation in a gene that was selected in six-rowed types to increase the fertility of lateral florets or the size of lateral grains, which evolved only recently from ancestral rudiments.

In wild barley and hulled types, the awn-bearing lemma is firmly attached to the grain, while it can be easily separated in hull-less (or 'naked') types. Maltsters have traditionally preferred covered grains because hulls protect the grain and act as a filtration medium. If barley is processed for direct human consumption, separating the tough, fibrous husks from the edible grain is desirable. Naked barleys carry a loss-of-function allele in the *NUD* gene, an ethylene response factor required for the formation of a lipid layer between caryopsis and lemma[29]. We determined lemma adherence to the grain as a binary character in the core set and performed an association scan (Fig. 3c and Supplementary Fig. 10c). The most highly associated marker was 453 kilobases away from the *NUD* gene. The GWAS peak coincided with a region of high genetic differentiation (F_{ST}) between naked and hulled types (Supplementary Fig. 10d). Both highly associated markers and F_{ST} signals were also found on other chromosomes, which might be explained by founder effects subsequent to the monophyletic origin[29] of naked barleys or divergent selection pressures between naked and hulled types.

Convergent selection for smooth awns in barley and rice. Encouraged by the collocation of our GWAS peaks for row type and hull adherence with the underlying genes, we attempted an association scan for a third, more subtle trait that had evolved under domestication and whose molecular basis had not been elucidated. Among the most recognizable features of the barley plant are its long, bristling awns—a nuisance for farmers during manual harvesting and for animals chewing the barley grain. Awnless varieties of barley exist but are not widely grown. While cereal awns are considered beneficial for seed dispersal in the wild[30], their persistence in domesticated barley is commonly attributed to their photosynthetic activity, whose loss is accompanied by significant yield penalties[31]. To mitigate the discomfort caused by awns, barbless varieties with smooth awns lacking silicified trichomes (Fig. 4a) are grown in some regions of the world, although they have not risen to worldwide prominence. To map loci underlying awn smoothness, we phenotyped our core set of 1,000 accessions by visual and haptic assessment of awn roughness. When considering awn roughness as a binary phenotype (rough versus smooth), an association scan detected a strong peak on the long arm of chromosome 5H (Fig. 4b and Supplementary Fig. 10b), which was collocated with both the mapping interval delineated in a biparental population (Fig. 4c) segregating for awn roughness and the *raw1* locus of traditional barley genetics[32]. We searched the vicinity of the peak for plausible candidate genes and came across a gene (HORVU5Hr1G086520) annotated as a cytokinin riboside 5′-monophosphate phosphoribohydrolase, which was homologous to the *LONG AND BARBED AWN1* (*LABA1*) gene of rice. Hua et al.[33] showed that *LABA1* is involved in cytokinin biosynthesis and that its functional alleles increase cytokinin content in epidermal cells of awn primordia in rice. Knockout mutants of *LABA1* were favored in domesticated rice to abolish barb formation and awn elongation. The sequence of its barley homolog differed between the parents of our mapping population by a non-synonymous variant (c.1186G>A) (Fig. 4d). By screening a population of chemically induced mutants in a rough-awned background[34], we found a loss-of-function allele (disrupted splice junction) reducing the size of trichomes mainly at the base of the awn (Supplementary Table 5 and Fig. 4e–g).

In addition to this *ROUGH AWN1* locus on chromosome 5H, we found another GWAS peak on 7H, which was not detected in the biparental population (Fig. 4b,c). The awns of the smooth parent Morex are not devoid of barbs (Fig. 4a), and a distinction between completely smooth and semi-smooth types with hairs only at the tip of the awn is made by breeders, who have observed a partially quantitative inheritance of awn roughness[35]. We speculate that the 7H locus may correspond to a second mutation decreasing barb formation in carriers of loss-of-function alleles at *raw1*. Further work will delimit the 7H locus through genetic mapping in crosses between completely smooth and semi-smooth types selected from our core set.

Association scans with legacy data. In contrast to the morphological characters we have focused on so far, most traits related to plant performance in present-day intensive agriculture, such as yield, plant height, flowering time and disease resistance, are quantitatively inherited and interact with environmental factors. Collecting phenotypic data for an entire genebank collection would be an immense undertaking. The outcome would be uncertain, as the majority of genebank holdings are landraces unadapted to current agricultural practices, complicating the assessment of agronomic parameters[36]. Evaluation data collected by genebank managers and researchers in past decades offer a ready opportunity to assess the

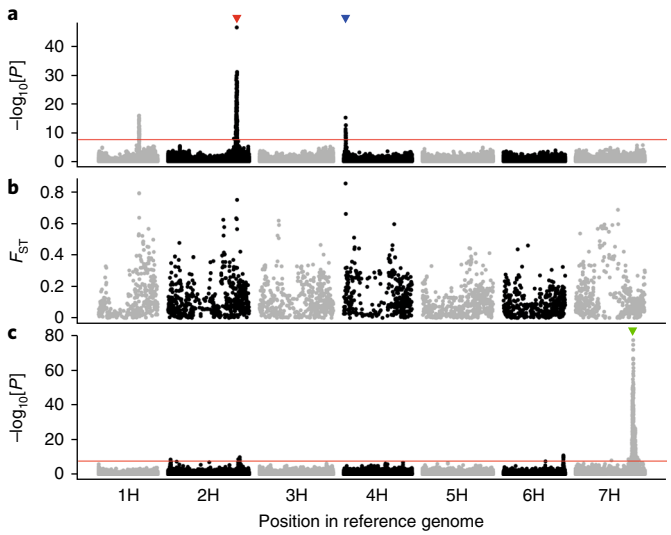

Fig. 3 | Genome-wide association scans for morphological characters. **a**, Significance of marker–trait associations for lateral spikelet fertility ('row type'). The red and blue arrowheads mark the positions of the cloned row-type genes *VRS1* (ref. [25]) and *INT-C* (ref. [26]), respectively. **b**, Genetic differentiation (F_{ST}) between six-rowed and two-rowed types in 1-megabase bins. **c**, Significance of marker–trait association for hull adherence. The green arrowhead marks the position of the *NUD* gene[29]. The red lines in **a** and **c** indicate the significance threshold after correction for multiple testing using the Bonferroni method. Imputed variant matrices were used. Genome-wide association scans were done using a mixed-linear model approach with a sample set of 1,000 biologically independent individuals.

power of association mapping of agronomic traits without additional investments and can thus guide the design of future efforts for linking phenotype and genotype in plant genetic resources. In the present study, we took advantage of digitized records of 69 propagation cycles of the IPK collection dating back to as early as 1946 and extracted flowering dates for 9,903 spring barley accessions. After data curation and outlier removal, we used a mixed-linear model approach[37] to account for the effects of environmental factors (different years of propagation) and variance inhomogeneity. A genome-wide association scan detected three peaks close to the known flowering-time genes *PHOTOPERIOD-H1* (*PPD-H1*) on chromosome 2H, *VRN-H1* on 5H and *VRN-H3* on 7H (Fig. 5a). As a component of the photoperiod pathway, *PPD-H1* is a major regulator of flowering time in both wheat and barley[38] and has pleiotropic effects on plant architecture[39]. *VRN-H1* and *VRN-H3* are homologous to *APETALA1* and *FLOWERING LOCUS T* of *Arabidopsis thaliana*, respectively. Sequence variation in *VRN* genes underlies the differences in vernalization requirement between spring- and autumn-sown cereals[40,41]. *VRN-H1* has been implicated in frost tolerance in spring barley[42]. Our results suggest that allelic diversity at *VRN-H1* and *VRN-H3* correlates with flowering-time variation in spring-sown barleys. An important caveat is that landraces from diverse agricultural environments were grown at a single field site in central Germany, an environment to which they were not adapted.

The identification of resistance genes from plant genetic resources by association genetics has received considerable attention in recent years[43,44]. Although several commonly deployed resistance genes or alleles trace back to traditional landraces preserved in ex situ germplasm collections[45], historical field records of genebanks are of limited use for assaying the response to pathogen stress. Prevalent pathogen strains change over time, as do pesticide regimes and disease scoring schemes. To evaluate the suitability of our genetic profiles to map resistance gene loci, we used evaluation data for resistance to bymoviruses of the barley yellow mosaic virus (BaYMV) complex collected in field trials (BaYMV/barley mild mosaic virus (BaMMV)) at three locations in central Germany and by artificial inoculation in growth chambers (BaMMV) for a large subset (1,894 winter barley accessions) of the IPK collection across a 31 year time period (1985–2016, with intermissions; Supplementary Table 6). BaYMV and BaMMV are transmitted by the soil-borne protist *Polymyxa graminis*. Since the vector is infectious to a soil depth up to 70 cm and no pesticides effective against the viruses are known to date, only resistant cultivars can be grown on virus-infested fields. Best linear unbiased predictors were calculated for BaYMV and BaMMV resistance and used for association scans. A strong peak on the long arm of chromosome 3H was detected for BaYMV resistance (Fig. 5b). The most highly associated marker was located close to the well-characterized *rym4/5* locus encoding the eukaryotic translation initiation factor 4 (*Hv-eIF4E*; ref. [46]). The rather large distance (1.6 megabases) between the peak marker and *Hv-eIF4E* is likely a consequence of extensive linkage disequilibrium in this region due to recent breeding practices[47]. Alleles of *Hv-eIF4E* are effective against isolates of both BaYMV and BaMMV[48]. In line with this, the *rym4/5* locus was also highly associated with resistance to BaMMV (Fig. 5c). In addition, a second peak on the long arm of chromosome 4H was found. The most highly associated marker fell within a broadly defined mapping interval for BaMMV resistance in the Taihoku A × Plaisant population (ref. [49]; S.F. and F.O., unpublished results), in which segregants carrying the *rym13* allele of the Taiwanese cultivar Taihoku A are resistant to BaMMV strains that have overcome the commonly deployed *Hv-eIF4E* allele. Our association scan in a natural diversity panel suggests that the resistance of Taihoku A to BaMMV is not an isolated event, but that alleles of *rym13* have played an important role in managing BaMMV infection in the field and may have been selected by farmers in East Asia, where many resistant accessions originate[48].

Discussion

Our analysis has established the feasibility of genome-wide high-density genotyping of an entire genebank collection. Molecular passport data provide invaluable complementary information to traditional passport records of genebank managers and will be crucial in the transformation of genebanks from living archives into biodigital resource centers. Genome-wide genetic marker data enable the identification of candidate duplicates, highlight collection gaps and enable the informed selection of core sets for deeper study. In combination with phenotypic data for many accessions, GBS data are a permanent resource for investigating the genes underlying crop evolution and selection for agronomic traits. The methodology employed in the current study—reduced representation sequencing, reference-based variant calling, imputation and genome-wide association—has been used in many crop species and scales well to large sample sizes. Thus, similar efforts are feasible, if not already under way, for other crop species[50,51]. If molecular passports are compiled for other barley collections, it will be possible to assess redundancy and complementarity at the international level to allow informed decisions on material exchange and planning of new collections. As sequencing methodologies evolve, whole-genome sequencing of entire collections[51] may also become affordable in large-genome species, enabling in silico allele mining for the vast majority of genes and possibly also zooming in on GWAS peaks to candidate gene resolution[52] without complementary resources such as biparental populations. Future research should focus on concepts for discovering beneficial genetic variation contributing to complex traits such as quantitative disease resistance and yield components, and on their practical implementation[53,54]. Genebank genomics will also be an attractive avenue to elucidate the molecular basis of crop evolution in species whose domestication history has not been studied as well as that of the major cereal crops.

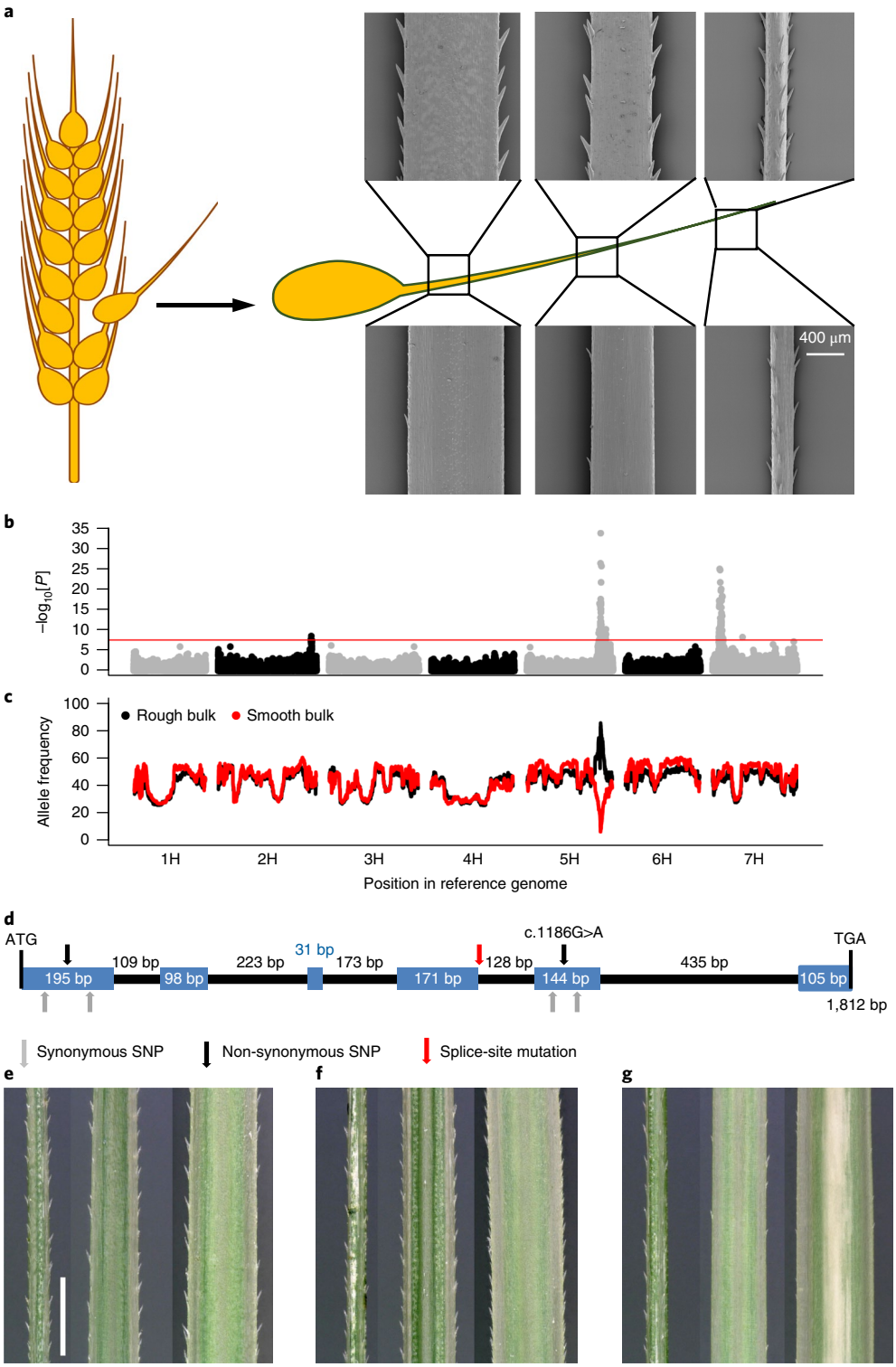

Fig. 4 | Isolation of the *ROUGH AWN1*. **a**, Awn roughness in barley. Scanning electron micrographs from the base, middle and tip of the awn in the rough-awned cultivar Barke (top row) and the smooth-awned cultivar Morex (bottom row). Images were obtained for representative individuals of cultivars Morex and Barke. **b**, Significance of marker–trait associations for awn roughness. The red line indicates the significance threshold after correction for multiple testing using the Bonferroni method. Genome-wide association scans were done using a mixed-linear model approach with a sample set of 1,000 biologically independent individuals. **c**, Mapping-by-sequencing with two phenotypically defined bulks, each comprising 180 rough and smooth awns recombinants, respectively, of the Morex × Barke recombinant inbred line (F_8) population. **d**, Exon–intron structure of the *ROUGH AWN1* gene with positions of a non-synonymous SNP between Barke and Morex (c.1186G>A) and mutations in TILLING plants. **e**–**g**, Segregation of awn roughness in offspring of a TILLING mutant heterozygous for the splice-site mutation in the fourth exon: homozygous wild type (**e**), heterozygous (**f**), homozygous mutant (**g**). The photographs show images of representative individuals from a population of 104 F_2 plants derived from a selfed M_3 plant. There were 24 plants homozygous for the wild-type allele, 55 heterozygotes and 25 homozygous for the mutant allele. Phenotypes and genotypes at the causal SNP were scored in all plants. Scale bar for **e**, **f**, **g**, 1 mm.

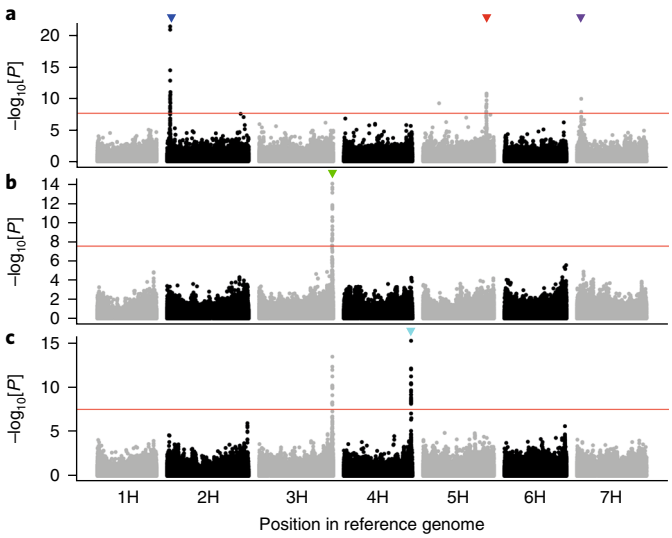

Fig. 5 | Association mapping of agronomic characters. a, Significance of marker–trait associations for flowering time in spring-sown barleys. Blue, red and purple arrowheads mark the locations of the major flowering-time genes *PPD-H1* (ref. [38]), *VRN-H3* (ref. [41]) and *VRN-H1* (ref. [40]), respectively. **b**,**c**, Significance of marker–trait associations for resistance to BaYMV (**b**) and BaMMV (**c**) in winter barley accessions of the IPK genebank. The arrowheads mark the position of *Hv-eIF4E* (*rym4/5*; green) and the location of the *rym13* locus (cyan). Imputed variant matrices were used. The red lines indicate the significance threshold after correction for multiple testing using the Bonferroni method. Genome-wide association scans were done using a mixed-linear model approach with phenotypic data from historic field trials for 8,825 (**a**), 1,852 (**b**) and 1,894 (**c**) accessions.

URLs

NovoSort, http://www.novocraft.com/products/novosort/; Picard, https://broadinstitute.github.io/picard/; script for filtering VCF files, https://bitbucket.org/ipk_dg_public/vcf_filtering; BRIDGE portal, http://bridge.ipk-gatersleben.de.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-018-0266-x.

Received: 2 August 2018; Accepted: 26 September 2018;
Published online: 12 November 2018

References

1. Novembre, J. et al. Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
2. Russell, J. et al. Exome sequencing of geographically diverse barley landraces and wild relatives gives insights into environmental adaptation. *Nat. Genet.* **48**, 1024–1030 (2016).
3. Brachi, B., Morris, G. P. & Borevitz, J. O. Genome-wide association studies in plants: the missing heritability is in the field. *Genome. Biol.* **12**, 232 (2011).
4. Lopes, M. S. et al. Exploiting genetic diversity from landraces in wheat breeding for adaptation to climate change. *J. Exp. Bot.* **66**, 3477–3486 (2015).
5. Oppermann, M., Weise, S., Dittmann, C. & Knüpffer, H. GBIS: the information system of the German Genebank. *Database* **2015**, bav021 (2015).
6. Pourkheirandish, M. et al. Evolution of the grain dispersal system in barley. *Cell* **162**, 527–539 (2015).
7. Poland, J. A., Brown, P. J., Sorrells, M. E. & Jannink, J.-L. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE* **7**, e32253 (2012).
8. Wendler, N. et al. Unlocking the secondary gene-pool of barley with next-generation sequencing. *Plant Biotechnol. J.* **12**, 1122–1131 (2014).
9. Mascher, M. et al. A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**, 427–433 (2017).
10. Jakob, S. S. et al. Evolutionary history of wild barley (*Hordeum vulgare* subsp. *spontaneum*) analyzed using multilocus sequence data and paleodistribution modeling. *Genome Biol. Evol.* **6**, 685–702 (2014).
11. Chen, F. H. et al. Agriculture facilitated permanent human occupation of the Tibetan Plateau after 3600 BP. *Science* **347**, 248–250 (2015).
12. Badr, A. et al. On the origin and domestication history of barley (*Hordeum vulgare*). *Mol. Biol. Evol.* **17**, 499–510 (2000).
13. Blattner, F. R. & Méndez, A. G. B. RAPD data do not support a second centre of barley domestication in Morocco. *Genet. Resour. Crop Evol.* **48**, 13–19 (2001).
14. Pourkheirandish, M. et al. Elucidation of the origin of 'agriocrithon' based on domestication genes questions the hypothesis that Tibet is one of the centers of barley domestication. *Plant J.* **94**, 525–534 (2018).
15. Galinsky, K. J. et al. Fast principal-component analysis reveals convergent evolution of *ADH1B* in Europe and East Asia. *Am. J. Hum. Genet.* **98**, 456–472 (2016).
16. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
17. Knüpffer, H. & van Hintum, T. J. L. in *Core Collections of Plant Genetic Resources* (eds Hodgkin, T., Brown, A. H. D., van Hintum, T. J. L. & Morales, E. A. V.) 171–178 (John Wiley and Sons, Chichester, UK, 1995).
18. van Hintum, T. J. L. & Knüpffer, H. Duplication within and between germplasm collections. I. *Genet. Resour. Crop Evol.* **42**, 127–133 (1995).
19. van Hintum, T. J. L. & Visser, D. L. Duplication within and between germplasm collections. II. *Genet. Resour. Crop Evol.* **42**, 135–145 (1995).
20. Parzies, H., Spoor, W. & Ennos, R. Genetic diversity of barley landrace accessions (*Hordeum vulgare* ssp. *vulgare*) conserved for different lengths of time in ex situ gene banks. *Heredity* **84**, 476 (2000).
21. Harlan, J. R. Ethiopia: a center of diversity. *Econ. Bot.* **23**, 309–314 (1969).
22. Castañeda-Álvarez, N. P. et al. Global conservation priorities for crop wild relatives. *Nat. Plants* **2**, 16022 (2016).
23. Swarts, K. et al. Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. *Plant Genome* **7**, https://doi.org/10.3835/plantgenome2014.05.0023 (2014).
24. Mansfeld, R. Das morphologische System der Saatgerste, *Hordeum vulgare* L. sl. *Der Züchter* **20**, 8–24 (1950).
25. Komatsuda, T. et al. Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene. *Proc. Natl Acad. Sci. USA* **104**, 1424–1429 (2007).
26. Ramsay, L. et al. *INTERMEDIUM-C*, a modifier of lateral spikelet fertility in barley, is an ortholog of the maize domestication gene *TEOSINTE BRANCHED 1*. *Nat. Genet.* **43**, 169–172 (2011).
27. Youssef, H. M. et al. Natural diversity of inflorescence architecture traces cryptic domestication genes in barley (*Hordeum vulgare* L.). *Genet. Resour. Crop Evol.* **64**, 843–853 (2017).
28. Lundqvist, U. Hexastichon and intermedium mutants in barley. *Hereditas* **92**, 229–236 (1980).
29. Taketa, S. et al. Barley grain with adhering hulls is controlled by an ERF family transcription factor gene regulating a lipid biosynthesis pathway. *Proc. Natl Acad. Sci. USA* **105**, 4062–4067 (2008).
30. Elbaum, R., Zaltzman, L., Burgert, I. & Fratzl, P. The role of wheat awns in the seed dispersal unit. *Science* **316**, 884–886 (2007).
31. Liller, C. B. et al. Fine mapping of a major QTL for awn length in barley using a multiparent mapping population. *Theor. Appl. Genet.* **130**, 269–281 (2017).
32. Franckowiak, J. D. BGS 312; Smooth awn 1. *Barley Genet. Newsl.* **26**, 261 (1997).
33. Hua, L. et al. *LABA1*, a domestication gene associated with long, barbed awns in wild rice. *Plant Cell* **27**, 1875–1888 (2015).
34. Gottwald, S., Bauer, P., Komatsuda, T., Lundqvist, U. & Stein, N. TILLING in the two-rowed barley cultivar 'Barke' reveals preferred sites of functional diversity in the gene *HvHox1*. *BMC Res. Notes* **2**, 258 (2009).
35. Åberg, E. & Wiebe, G. A. *Classification of Barley Varieties Grown in the United States and Canada in 1945* (US Department of Agriculture, Washington, DC, USA, 1946).
36. Longin, C. F. H. & Reif, J. C. Redesigning the exploitation of wheat genetic resources. *Trends Plant Sci.* **19**, 631–636 (2014).
37. González, M. Y. et al. Unlocking historical phenotypic data from an ex situ collection to enhance the informed utilization of genetic resources of barley (*Hordeum* sp.). *Theoret. Appl. Genet.* **131**, 2009–2019 (2018).
38. Turner, A., Beales, J., Faure, S., Dunford, R. P. & Laurie, D. A. The pseudo-response regulator Ppd-H1 provides adaptation to photoperiod in barley. *Science* **310**, 1031–1034 (2005).
39. Digel, B. et al. Photoperiod1 (Ppd-H1) controls leaf size. *Plant Physiol.* **172**, 405–415 (2016).
40. Fu, D. et al. Large deletions within the first intron in *VRN-1* are associated with spring growth habit in barley and wheat. *Mol. Genet. Genom.* **273**, 54–65 (2005).

41. Yan, L. et al. The wheat and barley vernalization gene *VRN3* is an orthologue of *FT*. *Proc. Natl Acad. Sci. USA* **103**, 19581–19586 (2006).
42. Tondelli, A. et al. Allelic variation at *Fr-H1*/*Vrn-H1* and *Fr-H2* loci is the main determinant of frost tolerance in spring barley. *Environ. Exp. Bot.* **106**, 148–155 (2014).
43. Richards, J. K., Friesen, T. L. & Brueggeman, R. S. Association mapping utilizing diverse barley lines reveals net form net blotch seedling resistance/susceptibility loci. *Theoret. Appl. Genet.* **130**, 915–927 (2017).
44. Arora, S. et al. Resistance gene discovery and cloning by sequence capture and association genetics. Preprint at bioRxiv https://doi.org/10.1101/248146 (2018).
45. Jørgensen, I. H. Discovery, characterization and exploitation of Mlo powdery mildew resistance in barley. *Euphytica* **63**, 141–152 (1992).
46. Stein, N. et al. The eukaryotic translation initiation factor 4E confers multiallelic recessive *Bymovirus* resistance in *Hordeum vulgare* (L.). *Plant J.* **42**, 912–922 (2005).
47. Stracke, S. et al. Effects of introgression and recombination on haplotype structure and linkage disequilibrium surrounding a locus encoding *Bymovirus* resistance in barley. *Genetics* **175**, 805–817 (2007).
48. Friedt, W. & Foroughi-Wehr, B. Genetics of resistance to barley yellow mosaic virus. In: *Barley Genetics V* (Yasuda, S. & Konishi, T., eds.) 659–664 (Sanjo Press, Okayama, Japan, 1987).
49. Humbroich, K. et al. Mapping of resistance against *Barley mild mosaic virus*-Teik (BaMMV)—an *rym5* resistance breaking strain of BaMMV—in the Taiwanese barley (*Hordeum vulgare*) cultivar 'Taihoku A'. *Plant Breeding* **129**, 346–348 (2010).
50. Romay, M. C. et al. Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biol.* **14**, R55 (2013).
51. Wang, W. et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43–49 (2018).
52. Yano, K. et al. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nat. Genet.* **48**, 927–934 (2016).
53. Jiang, Y., Schmidt, R. H., Zhao, Y. & Reif, J. C. A quantitative genetic framework highlights the role of epistatic effects for grain-yield heterosis in bread wheat. *Nat. Genet.* **49**, 1741–1746 (2017).
54. Navarro, J. A. R. et al. A study of allelic diversity underlying flowering-time adaptation in maize landraces. *Nat. Genet.* **49**, 476 (2017).

Acknowledgements

We thank G. Matzig, J. Pohl, M. Ziems, C. Fricke, M. Kretschmann, S. König, I. Walde, G. Schütze, A. Fiebig, J. Bauernfeind, T. Münch and D. Grau for technical assistance and G. Proeseler for initiating the long-term virus testing. We are grateful to H. de Beukelaer for Corehunter support. We thank B. Schierscher-Viret from the Swiss national genebank for providing seeds and K. Lipfert for artwork. This work was supported by a grant from the Leibniz Association to N.S., U.S., H.K., A.B., A.G. and J.C.R. (Pakt für Forschung und Innovation: SAW-2015-IPK-1 'BRIDGE'); by the German Ministry of Education and Research (BMBF; grant 031A536 'de.NBI' to U.S.); by the Young Elite Scientists Sponsorship Program (2015QNRC001) from the China Association for Science and Technology (CAST); by a grant from the China Scholarship Council to G.G.; by funding from the China Agriculture Research System (CARS-05) and the Agricultural Science and Technology Innovation Program to J.Z.; and by the Swiss Federal Office for Agriculture in the framework of the National Plan of Action for the conservation and sustainable utilization of plant genetic resources (NAP-PGREL). S.G.M. acknowledges support from the German Academic Exchange service (DAAD) through a Leibniz-DAAD fellowship. Y.J. and M.Y.G. were supported by BMBF grants 031B0184A and 031B0190A, respectively. S.F. was supported by BMBF grants to F.O. and A.Habekuß (ViReCrop, FKZ: 0315708B; COBRA, FKZ: 031A323B).

Author contributions

N.S., M.M., U.S., J.C.R. and A.G. designed research. S.G.M., M.J., M.Y.G., Y.J., Y.Z. and M.M. analyzed data. A.B. supervised germplasm retrieval. G.H. optimized DNA extraction methods. A.Himmelbach performed GBS experiments. S.W., M.O. and H.K. managed, digitized and validated passport and phenotypic data. D.S., M.L. and U.S. managed sequence and phenotypic data. G.H., T.M., S.G.K. and B.K. contributed Swiss genebank accessions. G.G., D.X. and J.Z. contributed Chinese genebank accessions. M.J. and E.R.M. led phenotyping efforts. T.R. carried out microscopy. R.S. and R.K.P. mapped awn roughness. S.T. contributed expert knowledge on awn roughness. M.B., P.K., M.L. and U.S. implemented the online portal. A.Habekuß, S.F. and F.O. contributed data on virus resistance. S.G.M., M.J., N.S. and M.M. wrote the paper. All authors have read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0266-x.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.M. or N.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Plant cultivation and DNA isolation. Individual plants were grown in a greenhouse[55] for 2 weeks using Quickpot QP 96 T propagation trays (Hermann Meyer). Two leaf segments (each approximately 3 cm in length and 0.5 to 0.8 cm wide) were placed in a 96-well polypropylene cluster tube (1.2 ml, 8 tube stripe format; Corning) containing one 4 mm glass bead (ROTH). A second glass bead was placed on top, and the tube was closed with 8-cap tube strips (Heinemann Labortechnik). Samples were frozen in liquid nitrogen and stored at −80 °C. The racks with the frozen samples were inserted into the aluminum 96-adapter set (precooled in liquid nitrogen) of the ball-mill (Retsch, Model MM 400) and ground for 30 s at maximum speed (frequency, 30 Hz). In order to obtain even homogenization, the adapter-rack sandwich was disassembled, and the rack containing the samples was turned by 180°. The samples were ground for another 30 s at maximum speed and stored at −80 °C. Prior to DNA extraction the plate was kept at room temperature for 10 min. Pipetting was performed in the 96-well format using the Platemaster P220 (Gilson) and tips recommended by the manufacturer. Preheated (65 °C) GTC buffer (600 µl; 1 M guanidine thiocyanate, 2 M NaCl, 30 mM sodium acetate pH 6.0) was added to the frozen powder. The plate was shaken to ensure complete suspension of the plant material. Following a pulse-spin to remove liquid from the lid, the extracts were incubated at 65 °C for 30 min. Cell debris was removed by centrifugation (10 °C, 3,450*g*, 30 min), and the clear supernatant was transferred into a 96-well EconoSpin plate (Epoch Life Science). The DNA was bound to the silica membrane of the plate by using a NucleoVac 96 (Macherey-Nagel) manifold (vacuum, 600 mbar) and washed twice with 900 µl wash buffer (50 mM NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA, 70% ethanol). Vacuum was applied to remove the liquid completely from the wells. The plate was mounted on top of a standard 96-well microtiter plate (flat bottom), and the residual liquid was removed by centrifugation (2,550 *g*, 3 min). Finally, the plate was placed on top of a fresh U96 MicroWell Plate (Th. Geyer), and the DNA was eluted by the addition of 100 µl TE light buffer (10 mM Tris-HCl pH 8.0, 0.1 mM EDTA). After 5 min incubation the DNA was collected by centrifugation (2,550*g*, 10 min). Plates were capped using closure mats (Th. Geyer) and stored at −20 °C.

GBS library construction and sequencing. Genomic DNA was digested with PstI and MspI (New England Biolabs) and processed for GBS library construction essentially as described previously[56]. Typically, 180 individually barcoded samples were pooled per lane in an equimolar manner and sequenced on the Illumina HiSeq 2500, 1 × 107 cycles, single read, using a custom sequencing primer[56] according to the manufacturer's instructions (Illumina).

GBS read alignment and variant calling. After adapter trimming with cutadapt[57], reads were aligned to the reference genome sequence of barley cultivar Morex[58] with BWA-MEM version 0.7.12a (ref. [59]). After conversion to binary alignment map (BAM) format with SAMtools[60], alignment records were sorted by reference position and indexed with NovoSort (see URLs). To reduce the number of alignment files, BAM files for samples sequenced on the same flowcell were merged with Picard (see URLs). Variant calling was performed with SAMtools/BCFtools version 1.3 (ref. [61]) using the parameter '-DV' for SAMtools mpileup. Variant calling was parallelized across genomic windows using GNU Parallel[62]. Genotypes at bi-allelic sites with a minimum QUAL (i.e., mapping quality) score of 40 were called based on read depth ratios calculated from the DP (total read depth) and DV (depth of the alternative allele) fields using an AWK script. For the analysis of population structure and genetic similarity of samples, homozygous and heterozygous calls had to be supported by two and four reads, respectively. SNP sites were retained if they had less than 10% missing data and less than 10% heterozygous calls and the number of heterozygous calls did not the exceed the number of homozygous calls for either allele. The filtered SNP matrix was exported as a VCF file and converted into GDS format with seqArray[63]. Digital object identifiers for SNP matrices were registered with e!DAL[64] in the Plant Genome and Phenomics Research Data Repository[65].

Analysis of population structure and genetic similarity. PCA was done in the R statistical environment[66] with the snpgdsPCA() function of SNPRelate[66] implementing the FastPCA algorithm of Galinski et al.[67]. IBS was calculated using the snpgdsIBSNum function() of SNPRelate. Linkage disequilibrium (r^2) was calculated with snpgdsLDmat(). Model-based estimation of ancestry coefficient was done with ADMIXTURE[68]. F_{ST} values were computed using the method of Bhatia et al.[69]. Clustering of nearly identical samples used functions of the igraph R package[70] for graph operations.

Imputation of missing values. Imputation of missing genotype calls was done with FILLIN[23] using default parameters. Bi-allelic SNP sites with more than 95% missing data, less than 10 homozygous genotype calls for the minor allele and more than 1% heterozygous calls were discarded, as were sites not assigned to chromosomes. Only domesticated samples with less than 0.3% heterozygous calls and at least 10% present calls were considered. Domestication status was determined by setting a threshold on PC1 in a PCA across a set of all (wild and domesticated) samples. This procedure resulted in a genotype matrix with 856,437 SNPs and 20,458 samples. Imputation accuracy was assessed by masking known genotypes as implemented in FILLIN. Missing rates and MAFs were determined after imputation, and markers with less than 90% present calls or a MAF below 1% were discarded, yielding a matrix with 306,049 SNPs and 20,458 samples. The remaining missing calls (average, 2.9% per marker) were imputed with allele frequencies for use in genome-wide association scans.

Selection of a core set. Core set selection was done with CoreHunter3 (ref. [71]) using the average entry-to-nearest-entry criterion. The entries of the distance matrix were Euclidian distances in the space spanned by the first 16 eigenvectors determined by PCA of the SNP matrix. The sample universe contained only one representative from clusters of highly similar samples. We first selected 960 samples that were assigned to one of three populations defined by an ADMIXTURE run with $k = 3$ (corresponding to Eastern, Western and Ethiopian barleys) and then added 40 samples that were not assigned to any of these groups[72].

Phenotypic observations. Plants used for DNA extraction and genotyping were grown to full maturity. Morphological traits were observed on the harvest spike prior to threshing. Row type was scored according to the descriptors of Mansfeld[73]. Awn roughness was assessed visually and by sliding one's finger along the central part of the awn in the direction from top to bottom.

Flowering dates were retrieved from digitized records of genebank propagation cycles[5]. Seed regenerations were performed between 1946 and 2015 in Gatersleben (51° 49′ N 11° 16′ E). During this process, date of flowering for spring-sown barley was determined as the number of days after sowing when 50% of the plants had reached flowering. The phenotypic data were highly unbalanced, with 57 to 4,783 data points per year. Accessions were evaluated in unreplicated field trials under a randomized experimental design. A total of 43,814 phenotypic records were available for 9,903 spring barley accessions grown across 69 years.

Evaluation for BaYMV-1 (referred to in rest of this section as BaYMV)/BaMMV resistance was carried out on naturally virus-contaminated fields at different locations (Morgenrot, Saxony-Anhalt and Sunstedt, Lower Saxony) and on an artificially laid-out field at Aschersleben (Saxony-Anhalt) from 1985 to 2001 and continued in2015 and 2016. About 40 seeds per accession were sown in 2-rowed plots, 1 m long, in the middle of September. In February and March of the following year the expression of mosaic symptoms was estimated (score 1 (resistant) to score 9 (susceptible)), and the presence of the different viruses was analyzed by double antibody sandwich enzyme-linked immunosorbent assay (DAS–ELISA) according to Clark and Adams[74] using BaYMV/BaMMV-specific polyclonal antibodies.

Furthermore, the reaction to BaMMV was also tested by mechanical inoculation with the BaMMV-ASL1 isolate in a growth chamber at 12 °C and 16 h photoperiod. At the three-leaf stage, the plants were inoculated twice at an interval of 5–7 d using sap of infected leaves of the cv. 'Maris Otter' homogenized on ice in K_2HPO_4 buffer (1:10; 0.1 M, pH 9.1) after adding silicon carbide (carborundum, mesh 400, 0.5 g per 25 ml sap). Five weeks after the first inoculation, the number of plants with mosaic symptoms was scored and DAS–ELISA was carried out to estimate the infection rate.

Genome-wide association scans. Genome-wide association scans for morphological traits (row type, adherence of grain hulls, awn roughness; unreplicated observations on single plants) and bymovirus resistance were performed with GAPIT[75]. A mixed-linear model incorporating the kinship matrix was used. Best linear unbiased predictors for resistance to virus strains (BaYMV-1 and BaMMV) were calculated with lme4 (refs [68,76]). The basic model in the GWAS for flowering time in spring barleys was a standard Q + K linear mixed model according for population structure (Q) and kinship (K)[77]. Since the phenotypic data were highly unbalanced, the heterogeneous residual variance had to be taken into account[78]. The model is described as follows:

$$y = 1_n\mu + X\beta + ma + g + e$$

where y is the vector of best linear unbiased estimators of the accessions resulting from the phenotypic data analysis; 1_n is an n-dimensional vector of ones, with n being the number of accessions; μ is a common intercept term; β is the effect of subpopulations defined by row type with design matrix X; a is the effect of the marker being tested, with m being the corresponding marker profiles coded as 0, 1, 2; g is the vector of genotypic effects; and e is the vector of residuals. In the model we assumed μ and β as fixed effects, $g \sim N(0, G\sigma_g^2)$ and $e \sim N(0, E\sigma_e^2)$. The covariance matrix G is the realized genomic relationship matrix[79]. The residual covariance matrix E is a diagonal matrix with $E^{-1} = \mathrm{diag}(V^{-1})$, where V is the actual covariance matrix of the best linear unbiased estimators obtained in the phenotypic data analysis[80]. After filtering for MAF (>0.01 in the panel of accessions with phenotypic data, 297,550 SNPs were considered for GWAS. To determine the genome-wide threshold for significant marker–trait association, the Bonferroni correction[81] was used. The genome-wide thresholds were $P < 0.05$ after corrections[81].

Positional cloning of *ROUGH AWN1 (HvRAW1)*. Exome sequencing of mutant and wild-type bulks was performed as described previously[82,83]. Read mapping

(Illumina HiSeq 2000, 2 × 100 base pairs (bp)) was performed as described above for GBS data. Allele frequencies at variant sites were extracted from the DP and DV fields of the VCF file.

A TILLING population of 7,979 pre-existing ethyl methanesulfonate-treated plants of cv. Barke[34] was screened for independent mutant alleles of *HvRaw1*. Two primer pairs (Supplementary Table 7) were used to amplify the five exons of the gene HORVU5Hr1G086520.6 by a standard PCR with a final heteroduplex step as described previously[34]. PCR products were digested with a DNF-480-3000 Double-Stranded DNA Cleavage Kit and analyzed using a DNF-910-1000T Mutation Discovery 910 Gel Kit on the AdvanCETM FS96 system according to the manufacturer's guidelines (Advanced Analytical). Identified SNPs were confirmed by Sanger sequencing.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability. Scripts for filtering VCF files are available in a Bitbucket repository: https://bitbucket.org/ipk_dg_public/vcf_filtering.

Data availability

Sequence data collected in this study have been deposited at the European Nucleotide Archive (accession numbers PRJEB23967, PRJEB24563, PRJEB24627, PRJEB26634, PRJEB26652 and PRJEB27184; Supplementary Table 1). SNP matrices and phenotypic data have been deposited at https://doi.org/10.5447/IPK/2018/9. Passport data for all accessions are reported in Supplementary Table 1. Phenotypic data used for GWAS are reported in Supplementary Table 4 (morphological characters), Supplementary Table 6 (virus resistance), and at https://doi.org/10.5447/IPK/2018/10 (flowering time). Passport, phenotypic and sequence data can be browsed in the BRIDGE web portal (http://bridge.ipk-gatersleben.de).

References

55. Zimmermann, G., Bäumlein, H., Mock, H.-P., Himmelbach, A. & Schweizer, P. The multigene family encoding germin-like proteins of barley. Regulation and function in basal host resistance. *Plant Physiol.* **142**, 181–192 (2006).
56. Wendler, N. et al. Unlocking the secondary gene-pool of barley with next-generation sequencing. *Plant Biotechnol. J.* **12**, 1122–1131 (2014).
57. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
58. Mascher, M. et al. A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**, 427–433 (2017).
59. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
60. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
61. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
62. Tange, O. GNU Parallel—the command-line power tool. *;login:* **36**, 42–47 (2011).
63. Zheng, X. et al. SeqArray—a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics* **33**, 2251–2257 (2017).
64. Arend, D. et al. e!DAL—a framework to store, share and publish research data. *BMC Bioinformatics* **15**, 214 (2014).
65. Arend, D. et al. PGP repository: a plant phenomics and genomics data publication infrastructure. *Database* **2016**, baw033 (2016).
66. R Development Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, 2017).
67. Galinsky, K. J. et al. Fast principal-component analysis reveals convergent evolution of *ADH1B* in Europe and East Asia. *Am. J. Hum. Genet.* **98**, 456–472 (2016).
68. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
69. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST: the impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013).
70. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* **1695**, 1–9 (2006).
71. De Beukelaer, H., Davenport, G. F. & Fack, V. Core Hunter 3: flexible core subset selection. *BMC Bioinformatics* **19**, 203 (2018).
72. Dray, S. & Dufour, A.-B. The ade4 package: implementing the duality diagram for ecologists. *J. Stat. Softw.* **22**, 1–20 (2007).
73. Mansfeld, R. Das morphologische System der Saatgerste, *Hordeum vulgare L.* sl. *Der Züchter* **20**, 8–24 (1950).
74. Clark, M. F. & Adams, A. Characteristics of the microplate method of enzyme-linked immunosorbent assay for the detection of plant viruses. *J. Gen. Virol.* **34**, 475–483 (1977).
75. Lipka, A. E. et al. GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28**, 2397–2399 (2012).
76. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **7**, https://doi.org/10.18637/jss.v067.i01 (2015).
77. Yu, J. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).
78. Stich, B. et al. Comparison of mixed-model approaches for association mapping. *Genetics* **178**, 1745–1754 (2008).
79. VanRaden, P. Genomic measures of relationship and inbreeding. *Interbull Bull.* **37**, 33–36 (2007).
80. Smith, A., Cullis, B. & Gilmour, A. Applications: the analysis of crop variety evaluation data in Australia. *Aust. N. Z. J. Stat.* **43**, 129–145 (2001).
81. Dunn, O. J. Multiple comparisons among means. *J. Am. Stat. Assoc.* **56**, 52–64 (1961).
82. Mascher, M. et al. Barley whole exome capture: a tool for genomic research in the genus *Hordeum* and beyond. *Plant J.* **76**, 494–505 (2013).
83. Mascher, M. et al. Mapping-by-sequencing accelerates forward genetics in barley. *Genome Biol.* **15**, R78 (2014).

nature research

Corresponding author(s): Nils Stein, Mascher

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	BWA-MEM 0.7.12, Samtools 1.5, Picard 1.128, Novosort 3.02.12, R 3.4.2, GAPIT, seqArray 1.16.0, SNPRelate 1.12.0, Corehunter3, custom code: https://bitbucket.org/ipk_dg_public/vcf_filtering/src

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequence data collected in this study have been deposited at the European Nucleotide Archive (accession numbers: PRJEB23967, PRJEB24563, PRJEB24627, PRJEB26634, PREJ26652 and PRJEB27184; Supplementary Table 1). SNP matrices and phenotypic data have been deposited under the digital object identifier (DOI)

https://doi.ipk-gatersleben.de/DOI/f4847b76-05e7-4276-a8ea-bc8e429fc7c6/f8072c00-b003-47f6-8e67-596223007de8/2/1847940088. Passport data for all accession is reported in Supplementary Table 1. Phenotypic data used for GWAS scans is reported in Supplementary Table 4 (morphological characters), Supplementary Table 6 (virus resistance), and under DOI https://doi.ipk-gatersleben.de/DOI/825aff42-d46f-42bd-8a81-2eb938d7704f/3b3777b7-fc04-43bd-a535-71b514083b1d/2/1847940088 (flowering time). Passport, phenotypic and sequence data can be browsed in an online portal (http://bridge.ipk-gatersleben.de).

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Our aim was to collect GBS data for all accession of the barley collection of German Federal ex situ genebank.
Data exclusions	Samples were excluded from analysis if passport data (e.g. species assignment, domestication status) and genetic clustering showed obvious disagreements. Genetic clustering was done by principal component analysis and outliers were identified. Outlier removal by PCA is a previously described method (implemented e.g. in EIGENSOFT)
Replication	Historic phenotypic data can be considered as replicated field trials. Statistical models for analyzing these data are described in the manuscript.
Randomization	No randomization was performed.
Blinding	Phenotypic data were collected without knowledge of passport records or genetic data.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☐	☒ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Unique biological materials

Policy information about availability of materials

Obtaining unique materials	Seeds of genebank accessions can be ordered from the German Federal ex situ genebank under the terms of the Standard Material Transfer Agreement.