Exploring the Legacy of Central European Historical Winter Wheat Landraces

András Cseh (cseh.andras@atk.hu)
Centre for Agricultural Research, ELKH

Péter Poczai
University of Helsinki, Finnish Museum of Natural History

Tibor Kiss
Centre for Agricultural Research, ELKH

Krisztina Balla
Centre for Agricultural Research, ELKH

Zita Berki
Centre for Agricultural Research, ELKH

Ádám Horváth
Centre for Agricultural Research, ELKH

Csaba Kuti
Centre for Agricultural Research, ELKH

Ildikó Karsai
Centre for Agricultural Research, ELKH

Research Article

Keywords: Historical wheat landraces, heat stress tolerance and higher tillering capacity, genetic erosion

Posted Date: June 25th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-637988/v1

Version of Record: A version of this preprint was published at Scientific Reports on December 1st, 2021.
See the published version at https://doi.org/10.1038/s41598-021-03261-4.

Abstract

Historical wheat landraces are rich sources of genetic diversity offering untapped reservoirs for broadening the genetic base of modern varieties. Using a 20K SNP array, we investigated the accessible genetic diversity in a Central European bread wheat landrace collection with great drought, heat stress tolerance and higher tillering capacity. We discovered distinct differences in the number of average polymorphisms between Central and Western European collections, and identified a set of novel rare alleles present at low frequencies in the historical collection. The detected polymorphisms were unevenly distributed along the wheat genome, and polymorphic markers co-localized with genes of great agronomic importance. The efficiency of the highly diverse population for Genome-Wide Association study was confirmed and two significant marker trait associations with seed hardness were identified on the 5DS chromosome arm. The geographical distribution of the inferred Bayesian clustering revealed six genetically homogenous ancestral groups among the collection, where the Central European core bared an admixed background originating from four ancestral groups. We evaluated the effective population sizes (*N*e) of the Central European collection and assessed changes in diversity over time, which revealed a dramatic ~97% genetic erosion between 1955 and 2015.

Introduction

Wheat (Triticum aestivum L., AABBDD, 2n=6x=42) is traditionally one of the main food sources of humankind and modern cultivars provide 15% of calories consumed every day1. Despite its global impact on food security, domesticated wheat faces critical challenges generated by the changing climate. Climate change manifested in increased temperatures, drought or alteration in rainfall frequency and intensity is already affecting agriculture, posing a further barrier to efficient wheat production2.

Wheat landraces as traditional wheat varieties, may preserve a specific capacity to tolerate biotic and abiotic stresses, resulting in yield stability and an intermediate yield level under a low input agricultural system3,4. Valuable agricultural characteristics, e.g., stress tolerance and quality traits, can be readily introduced from landraces into new high-yielding wheat varieties in order to ensure food security for the rapidly growing population of the world.

Genetic variation provides the basis for crop adaptation in diverse environments. During the ‘green revolution’, wheat breeding focused on the development of high-yielding, disease-resistant wheat varieties with dwarfing genes that reduced the genetic diversity of modern elite varieties leaving a legacy of great variation behind in landraces. The unique population genetic structure of wheat landraces and the morpho-physiological traits are responsible for the better adaptation to changing climatic conditions. Population genetic studies provided extensive evidence for the greater genetic variation in bread wheat landraces and highlighted them as excellent sources of unaccustomed alleles potentially useful for modern breeding5–9. Wheat improvement programs can greatly benefit from the diverse genetic background preserved in these populations to ensure food security and sustainable, climate-smart agriculture in the future2,10.

There are clear examples of how exploiting landraces to introduce novel yield or drought characteristics into modern wheat, have resulted in global economic impacts. For instance, the introduction of dwarfing genes (Rht-B1b, Rht-D1b) from the Japanese cultivar ‘Norin 10’ during the ‘green revolution’ led to spectacular increases in yield11. ‘Norin 10’ inherited these genes from the landrace ‘Shiro Daruma’. Another example is the old Hungarian wheat cultivar ‘Bánkúti 1201’ developed in the first half of the 20th century, which contributed to the significant diversity and several unique alleles of modern Hungarian wheat cultivars, known worldwide for their special bread making quality parameters12–14.

The Central European wheat landrace collection including 199 accessions originated from six countries were collected during 1950-60 and conserved in the Gene Bank collection of the Centre for Plant Diversity (NÖDIK) Tápiószele, Hungary. This important legacy of Hungarian wheat breeding represents a vast and largely untapped source of genetic diversity. These wheat landraces are generally tall, prone to lodging and collectively considered to be highly adaptable to the agro-ecological conditions of Central and Eastern Europe. They have excellent drought and heat stress tolerance and stronger tillering ability under low nutrition input farming conditions 15,16. As a potential source of useful loci to improve wheat stress tolerance and grain quality, it is essential to characterize the phenotypic and genotypic diversity present among the Central European landrace collection. Modern marker-assisted selection (MAS) programs are based on genetic markers, but their use in wheat breeding was encumbered for a long time by the large genome size and the presence of three homoeologous genomes (ABD). The level of genetic diversity within a population can, however, be measured by single nucleotide polymorphism (SNP) genotyping arrays that offer the anticipated impetus to accelerate wheat breeding17,18. By using SNP genotyping, many lines can be cost-efficiently screened at an early stage making it possible to design more effective breeding programs. High-power Genome-Wide Association Studies (GWAS) are also based on information gained from high-throughput genotyping. GWAS enables the identification of important genes and their transfer from landraces into modern cultivars. For GWAS analysis the population structure needs to be investigated to avoid false positive associations between phenotypes and markers19.

Here we explored the genetic structure, effective population size and the available genetic diversity in the Central and Eastern European bread wheat landraces by using a 20K SNP array. We then compared these landraces to modern wheat cultivars and presented this diverse population as a promising source where agronomically important new alleles can be identified using GWAS.

Results

Novel polymorphisms among Central European landraces

To compare the variability of modern elite varieties and landraces, we genotyped the two collections using an 20K array containing 17,267 SNPs to identify polymorphisms across the 21 chromosomes of hexaploid wheat. Genotyping resulted in 15,808 high quality SNPs, while 1,459 (8.44%) SNPs were

trimmed from the final analysis during quality assessment. The selected 15,808 markers showed an average heterozygosity index (H) of 0.75, polymorphic information content (PIC) of 0.70, and had a discriminatory power (D) of 0.31 among all investigated accessions. From the selected markers, 15,565 (98.46%) were polymorphic across the two collections considered together, while there were 15,121 (95.65%) and 15,357 (97.14%) polymorphic markers in landraces and modern varieties, respectively. Most of the detected markers (14,913; 94.33%) were polymorphic in both collections. Nucleotide diversity (π), defined as the number of nucleotide differences per site between two randomly chosen sequences, was estimated to be 2.6 × 10-3 in the entire collection. Landraces showed a greater value (2.5 × 10-3) compared to modern varieties (1.03 × 10-3) in line with previous estimates20,21.

After determining that landraces contain a considerable number of polymorphisms, we were interested in investigating how many of these polymorphisms are unique to the Central European collection. We also assessed whether a geographical bias exists in this regard, resulting in some accessions being more polymorphic than others on a regional basis. Thus, the genotypic scores of Central European landraces were sequentially added to the scores of the accessions consisting of all modern varieties in order to determine the number of novel polymorphisms, following Winfield et al.22 This cumulative addition revealed that the average number of novel polymorphisms is 82 (range 10–154) (Fig.1a). The list of novel polymorphisms is available as Supplementary Table S1. The smallest number of novel polymorphisms (10) was present in CLUJ50-650 from Romania, while the addition of Banja-Luka-6 (Yugoslavia), Garljana (Bulgaria) and Martonvásári-K118 (Hungary) contributed 154 new polymorphisms to the modern varieties. The Central European wheat landraces harbored a collection of rare alleles present at low frequencies (<0.05, 78%; Supplementary Table S1 and Fig 2b). It is possible that these alleles might have had some selective advantage during breeding and improvement. A similar scenario was also reported in other landrace collections 23,24. Only a few alleles (11%) were found at higher frequencies (>0.15) among the novel polymorphisms. These alleles are likely responsible for the wide adaptation of the landraces, as shown by previous studies in wheat 6,25, maize 26 and rice 27.

Overall, landraces originating from the former Yugoslavia had the largest number of novel polymorphisms per accession, followed by Hungary and Austria (Fig. 1b). There was a distinct difference in the number of average novel polymorphisms when the Central and Western European collections were compared. When we added Central European landraces to the genotypic scores of modern varieties composed of only Western elite cultivars, on average 449 (range 10–888) new polymorphisms were detected. This suggests that these landraces possessed many novel polymorphisms compared to the Western European modern varieties. Thus, the Central European collection could be a good source to improve the genetic diversity of the Western modern elite varieties.

Polymorphic markers co-localize with agronomically important genes

The 20K array was constructed from SNPs that were previously mapped in the wheat genome, enabling us to assess their distribution and allele frequencies among the 21 wheat chromosomes and seven homoeologous groups 28 (Fig. 2a,b,c). The detected polymorphisms were unevenly distributed along the

wheat genome, considering the size of the chromosomes using the estimates of Šafář et al.29. The number of polymorphisms was highest on the A-genome and lowest on the D-genome compared to the Mv Ménrót (Martonvásár, Hungary) reference (Fig. 2e and Supplementary Table S2). In general, the A and B genomes were more diverse and showed more uniform distributions of polymorphisms across the genome than the D genome, in agreement with Akhunov et al.21. The average marker density was also the lowest in the D-genome compared to the others (Fig. 2f and Supplementary Table S2), in line with previous studies.30 Chromosome 5A had the highest number of polymorphic SNPs, while 4D had the least. In general, a relatively high positive correlation was observed between the number of polymorphic SNPs and chromosome size. This has similarly been reported by Alipour et al.24

A higher number of polymorphisms was concentrated on the homoeologous group 2 of landraces when the genetic origin of the accession was considered (Supplementary Table S2). In the case of group 4, 6 and 7, only chromosomes of the B and D genomes were more diverse in landraces compared to the modern elite varieties. On the contrary, the A genome from all homoeologous groups except 1 and 2 showed more polymorphisms in modern varieties. Mapping the positions of polymorphic markers along several important genes of great agronomic importance indicated that novel polymorphisms co-localize with these genes (Fig. 2 and Supplementary Table S3). These included reduced height genes (Rht1, Rht2), photoperiod response genes (Ppd-B1, Ppd-D1), as well as several genes associated with disease resistance (Lr, Sr, Yr, Pm), suggesting that they might have provided the basis of selection during the breeding programs.

Genetic and geographical structuring of wheat accessions

The collection of winter wheat accessions was divided into ten groups according to their country of origin; they were also distinguished based on their genetic origin (landraces vs modern elite varieties) in order to allow further comparison between geographical and genotypic data (Fig. 3b,c and Supplementary Table S4). Modern elite varieties from successor states of dissolved countries (Czechoslovakia, Yugoslavia) were maintained in the same geographical unit to allow further comparison. Principal component analysis (PCA), Bayesian cluster (STRUCTURE) and maximum likelihood (ML) tree analyses were conducted to determine the population structure of the wheat collection, using the set of 15,808 markers and after removing 285 SNPs with minor allele frequencies (<0.01). We observed the loose clustering of the accessions when country of origin was overlaid on the resulting patterns (Fig. 3e), which depicted close relationships and admixture of the accessions. This was also suggested by the low bootstrap values (<35%) obtained for most of the resulting groups in the ML analysis, indicating a lack of consistent signal to cluster these accessions to fine-scale inner groups. This could be attributed to recombination or high divergence within a short timescale, resulting in relatively high homoplasy compared to the number of informative sites, ultimately making the signals too weak. These patterns could be expected with our wheat collection; however, we retained two well supported groups (>90%) in the tree (Fig. 3e, indicated with arrows). Similar patterns were observed in the PCA analyses where both the first (PC1, 19.08%) and second component axis (PC2, 9.2%) explained very little

at the regional level (Supplementary Figure S1), but were informative when chronological genetic origin (landraces vs modern) of the collection was considered (Fig. 3a).

Overlaying the chronological genetic origin of the accessions on the ML tree (Fig. 3d) and the PCA showed clear differentiation between the modern varieties and landraces. The PCA summarizes the dominant components of variation in the genomic data, showing the difference between sampled regions but also including the variation within groups of accessions, thus limiting the amount of between-population variation explained by the two principal component axes 31. In this respect, the first component axis explained the genomic variation found in the modern varieties, while the second component axis explained the variation mostly found in the landraces, in accordance with the ML tree. Interestingly, the Austrian landraces grouped together with the modern varieties in the upper right corner of the PCA plot (Supplementary Figure S1). Tightly grouped modern elite varieties were principally located at the upper right corner of the plot (Fig. 3a; marked with blue), while the landraces were evenly distributed among the two axes (marked with red). This division among the accessions was highly supported by high ML bootstrap values (>90%). The modern elite variety ‘Divana’ from Croatia closely grouped with the landraces (Fig. 3a, indicated with an arrow), while a restricted number of landraces either appeared as a first branching group to the modern varieties or they were nested within this larger cluster on both the tree and the PCA plot. The latter divided the landraces with mixed regional affiliations into three groups, one located on the upper right corner, consisting of mostly Hungarian landraces. The second admixed group was clustered along the center of the second axis, while the third group – also of mixed regional affiliation – was intercalated between the two clusters.

A well-supported ML split (>90%) can also be observed among the modern elite varieties, which divided the accessions into two major groups. This split was less prominent in the PCA, but loosely corresponded with the accessions clustering at the upper right corner of the plot and grouped closer to the first axis. The Bayesian cluster (STRUCTURE) analyses indicated a peak in the mean posterior probabilities LnP (K) at K = 6, with the lowest variance among replicates (Supplementary Figure S2). The optimal number of clusters for the SNP dataset based on ΔK (Evanno et al. 2005) showed the highest peak at K = 2, with the second highest peak appearing at K = 6. This revealed that there are two genetically homogenous groups (K = 2) followed by six (K = 6) with the lowest variance among replicates in the Bayesian analysis. As our uneven sampling can bias the inferences on the number of Bayesian clusters, efforts were made to have comparable numbers of accessions from all countries evened across genetic origin groups. We carried out subsampling by removing closely related accessions based on the PCA clustering, and evaluated the Bayesian analyses based on the statistics described by Puechmaille et al.32 These statistics provided peaks for K values either at 2 (MedMedK and MedMeanK) or 6 (MaxMedK and MaxMeanK). The K = 2 clustering divided the wheat collection into two differentiated groups, with a clear pattern of subdivision among the landraces and modern elite varieties (figure not shown). This grouping further corroborated the results obtained in the preceding ML and PCA analyses.

We chose to describe K = 6, as we were interested in the overall clustering and differentiation among the entire collection at a regional level. However, we have also overlaid the genetic origin information on

the results (Fig. 4). Thus, the average individual membership proportions (Q) of each region to the inferred clusters were divided into six distinct ancestral groups. The geographical distribution of the inferred Bayesian clustering was also investigated by projecting the inferred Q scores of the ancestral grouping on a map. Accessions were assigned membership to each of these six ancestral groups (Q1-6) if they had >0.50 membership to that group.

Mean Q scores ranged from 0.001 to 0.995 for accessions across the six inferred clusters (Supplementary Table S5). The Bayesian clustering aligned with the patterns obtained in the ML and PCA analyses; it supported the distinct separation of modern elite varieties from landraces, and revealed further information about the fine-scale structure of the Central European wheat collection. Four ancestral groups (Q1, 3, 4 and 6) were predominant among the landraces, while Q2 and Q5 were mostly found among the modern elite varieties. Accessions with >90% membership to ancestral group 5 were principally modern varieties from Western Europe, while group 2 was dominated by modern varieties from Romania and Bulgaria, and found in mixed proportions among the Central European accessions. This is in line with previous results of Winfield et al.22. Moreover, our ancestral group 5 possibly coincides with their ancestral group 2, mainly found in Western European varieties. The Central European landraces were shown to have admixed origins of four ancestral groups, where group 4 was the rarest – it was found in only 3% of the collection (>0.50). However, some proportion (<0.50) of this group was almost always found among the entire wheat collection. It had the highest proportion in a few landraces originating from the territory of the former Yugoslavia (Dunav, Banja-Luka-6), Bulgaria (Sadovo-N-159, Stalinca), and it was dominantly present only in one Hungarian landrace (Pitvaros). Ancestral group 6, with accessions having >90% membership, was exclusive to Romanian and Hungarian landraces. It was highly dominant in the Hungarian landrace – 43% of the landraces were assigned to this group (>0.50), while it was found to be predominant (>0.90) in 34% of the accessions.

Recent loss of genetic diversity in modern elite varieties

We estimated the changes in genetic diversity through time and calculated the effective population sizes (Ne) of the Central European wheat collection between 1955 and 2015 (Fig. 5). The Bayesian Skyline Plot (BSP) analysis estimated the effective population size of the entire wheat collection to be 26.37 (95% HPD 36.64–18.15). A similarly small number (~30) was estimated by He et al.33 from a larger sample with exome sequencing, demonstrating the accuracy of our estimation despite our sparser sampling. The results showed that the effective population size of wheat landraces kept stable at a plateau of 51.03 (95% HPD 69.53–36.85) during the 1980s until a bottleneck appeared and reduced the population size by ~21% (40.45, 95% HPD 65.22–10.56), with the effective population size remaining constant in the 1990s. A second bottleneck occurred in the 2000s that reduced the population size by ~28% (14.05, 95% HPD 30.51–7.12), dropping down to 1.72 (95% HPD 3.75–0.78) by 2015. The population size of modern wheat dramatically eroded by ~97% between 1955 and 2015.

The estimates of the effective population size for the modern elite varieties supported the findings of the ML tree that divided the collection into two groups associated with ancestral group 2 and 5

originating from Bulgaria and Yugoslavia, as inferred from the Bayesian cluster analysis. It should be noted that our sampling was limited among the Western European landraces and modern elite varieties; this effect can be seen in the wider HPD bonds obtained from the 2000s and the tighter values appearing for landraces from 1955 until the mid-1980s. Thus, a wider representative sampling of Western European wheat might uncover a broader diversity that was not captured by our SNP array, and tighten the bonds of the HPD values. However, other studies conducted on a larger collection of wheat accessions showed that Western elite wheat has a small number of novel SNPs22, since breeding programs mainly used local landraces1. A good example of this pattern can be seen in the PCA grouping (Fig. 3a,d) where the Austrian landraces cluster together with the modern varieties.

Association mapping of seed hardness

Seed hardness strongly correlates with the bread making quality parameters. Harder seed usually means better end use quality. The normal distribution of the seed hardness values in the wheat collection were presented on a histogram (Suplementary Figure S3). We performed GWAS to identify the genetic background of this trait in Central European wheat landrace collection. We identified two significant (-log10(P-value)=15.28) marker trait association (MTA) on the 5DS chromosome arm (BS00000020_51 and TG0028 SNP markers) (Fig. 6). We used the online Ensembl Plants database (https://plants.ensembl.org) for BLAST analysis on IWGSC 1.0 wheat sequence to identify the linked gene. In agreement with previous studies we found that they represent two alleles of the Puroin-doline-b gene (Pinb-D1) (TraesCS5D02G004300) located on the 5DS chromosome arm34,35. The minor alleles were carried by 27.5% of the lines with an average of 41.70 seed hardness volume and the lines with the major alleles showed means of 61.90 (Suplementary Figure S4). We selected landraces with high seed hardness and good bread making quality parameters to improve the modern wheat cultivars by traditional crossing and MAS.

Discussion

Meeting the nutritional demand of the growing population with limited land use, decreasing water resources under the threat of climate change is the defining challenge of humanity in the 21st century36. Agriculture must simultaneously intensify, become more sustainable, and achieve greater resilience towards pests and climate change7. Environmental change is one of the worldwide difficulties confronting humankind today, as temperatures keep rising, setting off a large group of extraordinary climate fluctuations such as heat waves, dry seasons, and flooding37. Environmental changes are already affecting crop production levels and altering food security, which coupled together with the loss of genetic diversity in most crop species caused a domestication bottleneck38. Such changes are currently decreasing the worldwide yield of wheat, posing a major threat to global production that is estimated to decrease by ~6% for each 1°C of temperature elevation39–41. Compensating such loss in production cannot be substituted by taking more land under cultivation considering further serious effects on biodiversity loss and ecosystem services. Breeders will need to include as much genetic variation as they can to fulfill future needs in such scenarios. Genetic erosion caused by domestication

can be tackled by unlocking the genetic potential from historical landrace collections and wild relatives of wheat22,42. Many landraces kept in seed banks are not adequately characterized to attract breeders interest in their effective utilization. In most cases, the patterns of genetic diversity within and among such collections are unclear. However, comparison of the levels of nucleotide diversity in the modern elite varieties and landraces may provide valuable information for inferring the demographic history of wheat, the patterns of past breeding efforts, and signatures of selection events. Our genetic investigation of the 199 Central European landraces contrasted earlier studies reporting substantial duplication of germplasm accessions obtained from gene banks43. The landraces included in the present study were genetically distinct and their comparison with modern cultivars revealed a genetic signature that was defined by regions under selection. Mapping the positions of polymorphic markers along several important genes of great agronomic importance indicated that novel polymorphisms co-localize with these chromosome regions (Fig. 2). This indicates that during the modern breeding process these genome regions probably undergo intensive selection pressure.

Demographic events such as expansions or reductions can have long-term effects on the effective size of a population44. Thus, modeling the history of such demographic events in wheat can help to identify population differentiation by inferring past population-specific demographic changes. It has been shown that domestication and further agronomic improvement has resulted in population size reductions, typically referred to as genetic bottlenecks, in wheat33,44,45. Such bottlenecks help to explain the value of germplasm exchange and the use of landraces in modern breeding programs. In our comparative assessment based on the genetic origin and regional level using the Central European wheat landrace collection we revealed a split that lacks geographical division. However, the close clustering of the Bulgarian (e.g. ‘Ahtopol’ and ‘Goz’) and Yugoslavian (e.g ‘Beljska’, ‘Panonija’) landraces with the modern elite varieties stood out in our analyses. Landraces appearing in the most basal position on the ML tree (Fig. 3) originated from Bulgaria and Yugoslavia, implicating that all modern elite varieties might share ancestry with landraces originating from these regions. This reflects well the historical data showing that following World War II Italian wheat varieties were imported by the Yugoslav government with an aim to make the country self-sufficient in wheat production. The imported Italian lines were used by local plant breeders resulting in the import of mutant Rht8 alleles from Italy to South and Central Europe46. Another ML split also divided the landraces into two distinguished groups based on their basal groupings having varieties from either Yugoslavia or Bulgaria in their first branching ancestry (Fig 3e, two-way arrow). Hence, Yugoslavian and Bulgarian landraces might have had a greater impact on the genetic structure of the Central European landraces and modern wheat cultivars.

The geographical projection of the Q scores also revealed a change in the genetic composition of wheat accessions over time. According to this the Central European landraces were admixed from four ancestral groups, while the modern Central and Eastern European wheat cultivars were composed of only two ancestral groups. There was a distinct East-West division among the collection, suggesting that Western elite varieties might have been selected from accessions originating from ancestral group 5 (Fig. 4). To further study this division, we evaluated the effective population size (Ne), which is an important population genetic parameter measuring the genetic diversity that can be maintained among the

circumscribed collection. In this respect it defines the size of an ideal collection that would present the same amount of genetic drift as the collection of induvial under study; together with the mutation rate, it determines the number of alleles expected in collection. We detected a dramatic bottleneck in our investigated timeline (1955–2015), which might relate to the selection of a small number of founder lines for modern elite wheat breeding programs that relied on only a very limited number of parental lines in variety development (Fig. 5). This could be explained by the methods of plant breeders – they want to create new high-yielding varieties, and tend to make crosses among elite lines that have the highest likelihood of developing a new variety47,48.

The unprecedented availability of large-scale genomic resources, genome-wide association studies (GWAS) are now a valuable alternative to QTL mapping defining with high precision the genetic architecture of the quantitative traits49. The application of GWAS in elite germplasm is generally limited to the identification of smaller-effect marker trait associations (MTAs), as major effect MTAs might have already become fixed within the modern wheat cultivars50. This limitation can be overcome by using a very diverse association panel composed of landraces and modern wheat cultivars. In the present study, we performed an efficient mixed-model genome wide association (EMMA) analysis by using a set of 266 very diverse winter wheat accessions, each genotyped for 15,808 high quality SNP markers. This variance component approach can correct for a wide range of sample structures by explicitly accounting for pairwise relatedness between individuals51. The effectiveness of this method is clear from our results that corroborate with earlier data showing Pinb-D134,35,52, a gene located at the Ha locus on chromosome 5DS, as one of the major genes underlying grain hardness.

Our investigation reveals an unexplored diversity within the Central European wheat landrace collection and points out the regions under a considerable selection pressure during modern breeding. This could provide a fertile ground to develop wheat varieties in the future based on germplasm conserving allelic diversity currently missing in breeding programs.

Conclusions

Progress in genomics has resulted in new concepts and techniques that have the potential to improve the precision and efficiency of plant breeding. Reference genome assemblies, in conjunction with germplasm high-throughput SNP genotyping or sequencing, can help identify breeding targets that might help secure future food supplies. Significant advancements in plant genome sequencing explain how the availability of such resources, along with gene editing tools is transforming trait identification and modification operations. New techniques to breeding, such as genomic selection and speed breeding, may be able to overcome some of the constraints of traditional breeding. When genetics and genomics are integrated into breeding such as genotyping by sequencing, SNP genotyping, genomic selection, gene editing, rapid generation turnover, and haplotype-based breeding, the pace of genetic gains in breeding programs is projected to accelerate. Our work focusing on the genetic diversity of Central European wheat provided valuable information for understanding the relationships between landraces and modern elite varieties. We facilitated their characterization and determined their population structure and ancestral origins. Our

results could enrich breeding strategies for future crop improvement through helping breeders to develop new varieties by reducing pre-breeding activities. Our data can be used, for example, to explore selective sweeps for any specific gene or chromosome region, analyze footprints defining divergence of landraces from distinct ecologies, or identify germplasm groups conserving allelic diversity missing in current breeding programs. The genomic data and analysis tools made public with this paper can assist wheat researchers to discover and use functional diversity that may be essential for meeting these challenges.

Methods

Plant material

The plant material consisted of 199 Middle-East European winter wheat landraces originating from 6 countries and 67 modern winter wheat cultivars originating from 10 countries. Part of the landrace collection originated from dissolved countries such as Yugoslavia (present day Bosnia and Herzegovina, Croatia, Kosovo, North Macedonia, Montenegro, Serbia and Slovenia) and Czechoslovakia (Czechia and Slovakia). The landraces were collected during 1950-60 and conserved in the Gene Bank collection of the Centre for Plant Diversity (NÖDIK) Tápiószele, Hungary. The modern wheat cultivars used in the present study were registered between 1970 and 2015. The description of the accessions is detailed in the Supplementary Table S4. The plant material used in our study is not regulated by CITES and IUCN, and sampling followed national and international guidelines for germplasm management outlined by FAO. Before using the landraces for any analysis, the lines were homogenized three times via using the single seed descent (SSD) methodology. Field trials were conducted in the season of 2018/2019 Martonvásár (Hungary) and the plot sizes were 1.5 m2. The plots were sown with a length of 2.5 m with five rows spaced 0.12 m apart. All standard agronomic management practices were performed in all plots and yearly nitrogen input in active ingredient of fertilizer was 120 kg/ha. The grains were harvested and the seeds were randomly used from each plot to grain hardness measurements.

SNP Genotyping

DNA was extracted from 6 weeks old seedlings using the DNeasy® Plant Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. SNP genotyping was performed by TraitGenetics GmbH (http://www.traitgenetics.com/en/), using a 20k Illumina SNP chip, which represents a subset of markers from the 90K SNP array28. SNPs with greater than 10% missing values and 5% heterozygosity were removed from the subsequent analysis, which left a set of 15,808 high quality SNP markers.

Analysis of genetic diversity

To determine which landraces are particularly polymorphic compared to the modern varieties, we calculated the genotypic scores for the accessions and the number of additional polymorphisms using the SNP call function of Geneious Prime (Biomatters Ltd, Auckland, New Zealand) with default settings. Polymorphic sites among landraces, as well as their chromosomal positions and allele frequencies were determined through a comparison with all modern varieties, and separately with a small subset

consisting of only modern varieties originating from Western Europe or the USA. The average number of nucleotide differences per site (nucleotide diversity; π, Jukes and Cantor 1969)53 was calculated with DnaSP v654 using default settings. This was done for each chromosome based on size, as described by Šafář et al. (2010)29. The number of SNPs and marker density of all wheat chromosomes based on 'MV Ménrót' (RMF209) reference were calculated with Geneious Prime. 'MV Ménrót' is one of the leading cultivars in Hungary and it is used as a general standard by the National Food Chain Safety Office (Hungary) during the plant variety registration process. The position of several genes of great agronomic importance55,56 were plotted alongside polymorphic sites to identify co-localizing regions. Circular ideograms from the calculated values were created with shinyCircos57. General descriptive statistics of the SNPs such as the heterozygosity index (H), polymorphic information content (PIC) and discriminatory power (D) was calculated with iMEC58.

Population structure analysis

Principal component analysis (PCA) was used to summarize patterns of variations among the wheat accession. The analysis was carried out with Tassel v5 according to Bradbury et al. (2007)59 excluding SNPs with minor allele frequency <0.01. Relationships among wheat varieties were also inferred using a maximum likelihood (ML) approach implanted in IQ-TREE v1.6.1260. The general time reversible (GTR+F) nucleotide evolutionary model with direct base frequency counts was chosen as best-fitting for the dataset inferred with the -TESTMERGEONLY and -AICc options in the built-in ModelFinder61. The analyses were performed using the ultrafast bootstrap approximation (UFBoot2)62 with 1,000 replicates to provide relatively unbiased bootstrap estimates under mild model misspecifications reducing computing time and achieving unbiased branch supports. Unrooted trees were visualized with FigTree v1.4.463.

For the analysis of population structure, a model-based Bayesian cluster analysis was performed using STRUCTURE v2.3.464. Despite being orders of magnitude slower, STRUCTURE was chosen as an analysis platform since it performs well with polyploid data and provides an unbiased inference when differentiation is weak65. We used the admixture model with an inner alpha incorporating prior pedigree-based data assuming for the relative mixed ancestry of the accessions from multiple sources. The best fitting number of assumed clusters (K) ranging from 1 to 10 was evaluated performing 10 independent Markov chain Monte Carlo (MCMC) runs of 1 × 106 iterations following a burn-in period of 1 ×105 steps for each value of K. The best K was chosen based on the estimated membership coefficients (Q) for every individual in each cluster. The best fitting number of assumed clusters (K) was estimated using lnP(D|K)64 and ΔK rate change in the log probability of data between consecutive K values66, as implemented in StructureSelector67. The results of STRUCTURE are greatly affected by sample size, since the program accounts for the most salient variation. Thus, we also used the metrics MedMeaK, MaxMeaK, MedMedK and MaxMedK32 derived from the posterior probability for each K across multiple MCMC replicates. Given that unbalanced sample sizes among landraces and modern varieties could hamper the recovery of genetic clusters, we subsampled our dataset in order to maintain sample sizes as

even as possible using the same model assumptions and parameters. We then used CLUMPAK v1.1.268 to find the best alignment of the results across the range of K values as implemented in StructureSelector.

Bayesian skyline prediction

The historical demographic changes of wheat varieties were inferred from the estimate of effective population size (Ne) over time in order to explore temporal fluctuations. This method is a non-parametric approach developed for the inference of past population size changes from genetic data built on a piecewise-linear model of Markov Chain Monte Carlo (MCMC) probability distribution sampling to reconstruct demographic history. The SNP data was loaded to BEAUTi v.2.3.3. to set parameters and specific model criteria. The years when each wheat variety was bred were extracted from archival records and were set as tip calibration points. The initially chosen best fitting GTR nucleotide substitution model was used for Ne estimations, additionally allowing for rate heterogeneity among sites by setting the Gamma Category Count to 4 and estimating the Gamma distribution shape parameter while leaving the substitution rate fixed, which allowed the clock rate to estimate the number of substitutions per site per year. Coalescent Bayesian Skyline was selected as a tree-prior, which divided the time between modern and old landraces into segments estimating Ne for each branching time in the tree. The number of parameter dimensions was set to four, allowing Ne to change three times between the root and present day. Ten Markov chains were run for 1 × 107 generations using BEAST v2.3.269 and were sampled every 1,000 steps, with the first 1 × 106 samples discarded as burn-in. Runs were analyzed using Tracer v1.6, and convergence was verified by assessing effective sample sizes (ESS) of all parameters. Independent runs were combined in LogCombiner v2.3.2.

Phenotyping of Seed Hardness

The grains were selected randomly from each plot to be used in the study of grain hardness. Hardness values were measured by the Single Kernel Characterization System (SKCS 4100, Perten Instruments, Springfield, IL, USA) according to the manufacturer instructions. Martin et al. (1993) described the SKCS 410070 and its operating principles for measuring seed hardness and the hardness index or value is calculated from measurements of the force required to crush each kernel. The Shapiro-Wilk-Test was used for testing complete samples for normality71.

Association Mapping

For our Genome-Wide Association Study (GWAS) analyses, we used 11.559 polymorphic SNPs with minor allele frequency (MAF) ≥ 10%, which had physical positions based on IWGSC RefSeq v1.072. GWAS was carried out by the Efficient Mixed-Model Association (EMMA) using easyGWAS cloud-based platform51,73. The correction for population stratification and cryptic relatedness was performed by employing first two principal components as random effects74. Appropriateness of model used for association analysis in present study was checked by drawing QQ plot between expected and observed -log10(P) values. Significance threshold was set by application of Bonferroni correction75 the -log10(P-

value) threshold rose to 5.40. The QQ plot and Circular-Manhattan plot were drawn by CMplot software (https://github.com/YinLiLin/CMplot).

Declarations

Data Availability

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Ethical standards

On behalf of all co-authors, the corresponding author states that the work described is original, previously unpublished research. All the authors listed have approved the manuscript.

Acknowledgements

This research was funded by research grants from the Hungarian Scientific Research Fund (NKFIH-K-129221) and the AGENT project (H2020-SFS-2019-2) from the Research and Innovation Action of the European Union (Grant agreement ID: 862613). A.Cs. acknowledges funding from the European Union's Marie Skłodowska-Curie Fellowship Grant (H2020-MSCA-IF-2016-752453-LANDRACES). P.P. thanks for the support of the LUOMUS Trigger fund (797810002), OECD CRP Fellowship (TAD/CRP JA00100677) and iASK Research Grant (2/2018-F-4). This work was made in memory of Laszlo Eros (RIP).

Author contributions statement

IK, AC and PP conceived the designed the experiment.

TK, BK, ZB, AH and CK conducted laboratory experiments, data organization and maintaining the accessions.

AC and IK provided resources and obtained funding for the work.

PP and AC carried out all data analyses and prepared the figures. PP and AC wrote the original manuscript and all other authors reviewed and approved the final manuscript.

Statement of plant material collection

The plant material used in our study is not regulated by CITES and IUCN. Our sampling followed national and international guidelines for germplasm collection outlined by FAO. Plant material was obtained under the permission of a Standard Material Transfer Agreement (SMTA) of the International Treaty on Plant Genetic Resources for Food and Agriculture.

References

1. Balfourier, F. *et al.* Worldwide phylogeography and history of wheat genetic diversity. *Sci. Adv.* (2019). doi:10.1126/sciadv.aav0536

2. Jatayev, S. *et al.* Green revolution “stumbles” in a dry environment: Dwarf wheat with Rht genes fails to produce higher grain yield than taller plants under drought . *Plant. Cell Environ.* (2020). doi:10.1111/pce.13819

3. Zeven, A. C. Landraces: A review of definitions and classifications. *Euphytica* (1998). doi:10.1023/A:1018683119237

4. Lopes, M. S. *et al.* Exploiting genetic diversity from landraces in wheat breeding for adaptation to climate change. *J. Exp. Bot.* (2015). doi:10.1093/jxb/erv122

5. Horvath, A. *et al.* Analysis of diversity and linkage disequilibrium along chromosome 3B of bread wheat (Triticum aestivum L.). *Theor. Appl. Genet.* **119**, 1523–1537 (2009).

6. Vikram, P. *et al.* Unlocking the genetic diversity of Creole wheats. *Sci. Rep.* **6**, 1–13 (2016).

7. Plekhanova, E. *et al.* Genomic and phenotypic analysis of Vavilov’s historic landraces reveals the impact of environment and genomic islands of agronomic traits. *Sci. Rep.* **7**, 4816 (2017).

8. Winfield, M. O. *et al.* High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotechnol. J.* **14**, 1195–1206 (2016).

9. Rosyara, U. *et al.* Genetic Contribution of Synthetic Hexaploid Wheat to CIMMYT’s Spring Bread Wheat Breeding Germplasm. *Sci. Rep.* **9**, 1–11 (2019).

10. Feuillet, C., Langridge, P. & Waugh, R. Cereal breeding takes a walk on the wild side. *Trends in Genetics* (2008). doi:10.1016/j.tig.2007.11.001

11. Hedden, P. The genes of the Green Revolution. *Trends Genet.* **19**, 5–9 (2003).

12. Juhász, A. *et al.* Identification, cloning and characterisation of a HMW-glutenin gene from an old Hungarian wheat variety, Bánkúti 1201. *Euphytica* **119**, 75–79 (2001).

13. Juhász, A. *et al.* Bánkúti 1201—an old Hungarian wheat variety with special storage protein composition. *Theor. Appl. Genet.* **107**, 697–704 (2003).

14. Rakszegi, M. *et al.* Starch Properties in Different Lines of an old Hungarian Wheat Variety, Bánkúti 1201. *Starch - Stärke* **55**, 397–402 (2003).

15. Lelley, J. & Rajhathy, T. *Wheat and its breeding.* (Akademiai Kiado, 1955).

16. Lelley, J. *The variety issue and Hungarian wheat*. (Mezogadasagi Kiado, 1967).

17. Collard, B. C. Y. & Mackill, D. J. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos. Trans. R. Soc. B Biol. Sci.* **363**, 557–572 (2008).

18. Ladejobi, O. *et al.* Reference Genome Anchoring of High-Density Markers for Association Mapping and Genomic Prediction in European Winter Wheat. *Front. Plant Sci.* **10**, 1–13 (2019).

19. Zhao, K. *et al.* An Arabidopsis Example of Association Mapping in Structured Samples. *PLoS Genet.* **3**, e4 (2007).

20. Haudry, A. *et al.* Grinding up Wheat: A Massive Loss of Nucleotide Diversity Since Domestication. *Mol. Biol. Evol.* **24**, 1506–1517 (2007).

21. Akhunov, E. D. *et al.* Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. *BMC Genomics* **11**, 702 (2010).

22. Winfield, M. O. *et al.* High-density genotyping of the A.E. Watkins Collection of hexaploid landraces identifies a large molecular diversity compared to elite bread wheat. *Plant Biotechnol. J.* **16**, 165–175 (2018).

23. Kabbaj, H. *et al.* Genetic Diversity within a Global Panel of Durum Wheat (Triticum durum) Landraces and Modern Germplasm Reveals the History of Alleles Exchange. *Front. Plant Sci.* **8**, (2017).

24. Alipour, H. *et al.* Genotyping-by-Sequencing (GBS) Revealed Molecular Genetic Diversity of Iranian Wheat Landraces and Cultivars. *Front. Plant Sci.* **8**, (2017).

25. Rufo, R., Alvaro, F., Royo, C. & Soriano, J. M. From landraces to improved cultivars: Assessment of genetic diversity and population structure of Mediterranean wheat using SNP markers. *PLoS One* **14**, e0219867 (2019).

26. Romero Navarro, J. A. *et al.* A study of allelic diversity underlying flowering-time adaptation in maize landraces. *Nat. Genet.* **49**, 476–480 (2017).

27. Leung, H. *et al.* Allele mining and enhanced genetic recombination for rice breeding. *Rice* **8**, 34 (2015).

28. Wang, S. *et al.* Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnol. J.* **12**, 787–796 (2014).

29. Šafář, J. *et al.* Development of Chromosome-Specific BAC Resources for Genomics of Bread Wheat. *Cytogenet. Genome Res.* **129**, 211–223 (2010).

30. Edae, E. A., Byrne, P. F., Haley, S. D., Lopes, M. S. & Reynolds, M. P. Genome-wide association mapping of yield and yield components of spring wheat under contrasting moisture regimes. *Theor. Appl.*

Genet. **127**, 791–807 (2014).

31. Barth, J. M. I., Damerau, M., Matschiner, M., Jentoft, S. & Hanel, R. Genomic Differentiation and Demographic Histories of Atlantic and Indo-Pacific Yellowfin Tuna (Thunnus albacares) Populations. *Genome Biol. Evol.* **9**, 1084–1098 (2017).

32. Puechmaille, S. J. The program <scp>structure</scp> does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. *Mol. Ecol. Resour.* **16**, 608–627 (2016).

33. He, F. *et al.* Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. *Nat. Genet.* **51**, 896–904 (2019).

34. Pasha, I., Anjum, F. M. & Morris, C. F. Grain Hardness: A Major Determinant of Wheat Quality. *Food Sci. Technol. Int.* **16**, 511–522 (2010).

35. Muqaddasi, Q. H. *et al.* Prospects of GWAS and predictive breeding for European winter wheat's grain protein content, grain starch content, and grain hardness. *Sci. Rep.* **10**, 12541 (2020).

36. Urruty, N., Tailliez-Lefebvre, D. & Huyghe, C. Stability, robustness, vulnerability and resilience of agricultural systems. A review. *Agron. Sustain. Dev.* **36**, 15 (2016).

37. Feulner, G. Global Challenges: Climate Change. *Glob. Challenges* **1**, 5–6 (2017).

38. Olsen, K. M. & Gross, B. L. Detecting multiple origins of domesticated crops. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 13701–13702 (2008).

39. Asseng, S. *et al.* Rising temperatures reduce global wheat production. *Nat. Clim. Chang.* **5**, 143–147 (2015).

40. Kumar, D. & Kalita, P. Reducing Postharvest Losses during Storage of Grain Crops to Strengthen Food Security in Developing Countries. *Foods* **6**, 8 (2017).

41. Vikram, P. *et al.* Unlocking the genetic diversity of Creole wheats. *Sci. Rep.* **6**, 23092 (2016).

42. King, J. *et al.* A step change in the transfer of interspecific variation into wheat from Amblyopyrum muticum. *Plant Biotechnol. J.* **15**, 217–226 (2017).

43. Singh, N. *et al.* Efficient curation of genebanks using next generation sequencing reveals substantial duplication of germplasm accessions. *Sci. Rep.* **9**, 650 (2019).

44. Thuillet, A. C., Bataillon, T., Poirier, S., Santoni, S. & David, J. L. Estimation of long-term effective population sizes through the history of durum wheat using microsatellite data. *Genetics* **169**, 1589–1599 (2005).

45. Joukhadar, R., Daetwyler, H. D., Bansal, U. K., Gendall, A. R. & Hayden, M. J. Genetic Diversity, Population Structure and Ancestral Origin of Australian Wheat. *Front. Plant Sci.* **8**, (2017).

46. Borojevic, K. & Borojevic, K. The transfer and history of 'reduced height genes' (Rht) in wheat from Japan to Europe. *J. Hered.* **96**, 455–459 (2005).

47. Baenziger, P. S. & Depauw, R. M. Wheat Breeding: Procedures and Strategies. in *Wheat Science and Trade* 273–308 (Wiley-Blackwell, 2009). doi:10.1002/9780813818832.ch13

48. Mascher, M. *et al.* Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. *Nat. Genet.* **51**, (2019).

49. Saïdou, A. A., Thuillet, A. C., Couderc, M., Mariac, C. & Vigouroux, Y. Association studies including genotype by environment interactions: Prospects and limits. *BMC Genet.* **15**, (2014).

50. Salvi, S. & Tuberosa, R. The crop QTLome comes of age. *Curr. Opin. Biotechnol.* **32**, 179–185 (2015).

51. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).

52. Morris, C. F. Puroindolines : the molecular basis of wheat grain hardness . Plant Mol Biol Puroindolines : the molecular genetic basis of wheat grain hardness. *Plant Mol. Biol.* **48**, 633–647 (2015).

53. JUKES, T. H. & CANTOR, C. R. Evolution of Protein Molecules. in *Mammalian Protein Metabolism* 21–132 (Elsevier, 1969). doi:10.1016/B978-1-4832-3211-9.50009-7

54. Rozas, J. *et al.* DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol. Biol. Evol.* **34**, 3299–3302 (2017).

55. Liu, S. *et al.* Molecular markers linked to important genes in hard winter wheat. *Crop Sci.* **54**, 1304–1321 (2014).

56. Juliana, P. *et al.* Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics. *Nat. Genet.* **51**, 1530–1539 (2019).

57. Yu, Y., Ouyang, Y. & Yao, W. ShinyCircos: An R/Shiny application for interactive creation of Circos plot. *Bioinformatics* **34**, 1229–1231 (2018).

58. Amiryousefi, A., Hyvönen, J. & Poczai, P. iMEC: Online Marker Efficiency Calculator. *Appl. Plant Sci.* **6**, e01159 (2018).

59. Bradbury, P. J. *et al.* TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).

60. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

61. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

62. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

63. Rambaut, A. FigTree, a graphical viewer of phylogenetic trees for producing publication-ready figures. (2020). Available at: http://tree.bio.ed.ac.uk/software/figtree/.

64. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–59 (2000).

65. Stift, M., Kolář, F. & Meirmans, P. G. Structure is more robust than other clustering methods in simulated mixed-ploidy populations. *Heredity (Edinb).* **123**, 429–441 (2019).

66. EVANNO, G., REGNAUT, S. & GOUDET, J. Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol. Ecol.* **14**, 2611–2620 (2005).

67. Li, Y.-L. & Liu, J.-X. <scp>StructureSelector</scp>: A web-based software to select and visualize the optimal number of clusters using multiple methods. *Mol. Ecol. Resour.* **18**, 176–177 (2018).

68. Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A. & Mayrose, I. <scp>Clumpak</scp>: a program for identifying clustering modes and packaging population structure inferences across *K. Mol. Ecol. Resour.* **15**, 1179–1191 (2015).

69. Bouckaert, R. *et al.* BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014).

70. C. R. Martin, R. Rousser & D. L. Brabec. Development of a Single-kernel Wheat Characterization System. *Trans. ASAE* **36**, 1399–1404 (1993).

71. Shapiro, S. S. & Wilk, M. B. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* **52**, 591 (1965).

72. Appels, R. *et al.* Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science (80-.).* **361**, eaar7191 (2018).

73. Grimm, D. G. *et al.* easyGWAS: A cloud-based platform for comparing the results of genome-wide association studies. *Plant Cell* **29**, 5–19 (2017).

74. Hyun, M. K. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).

75. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).

Figures

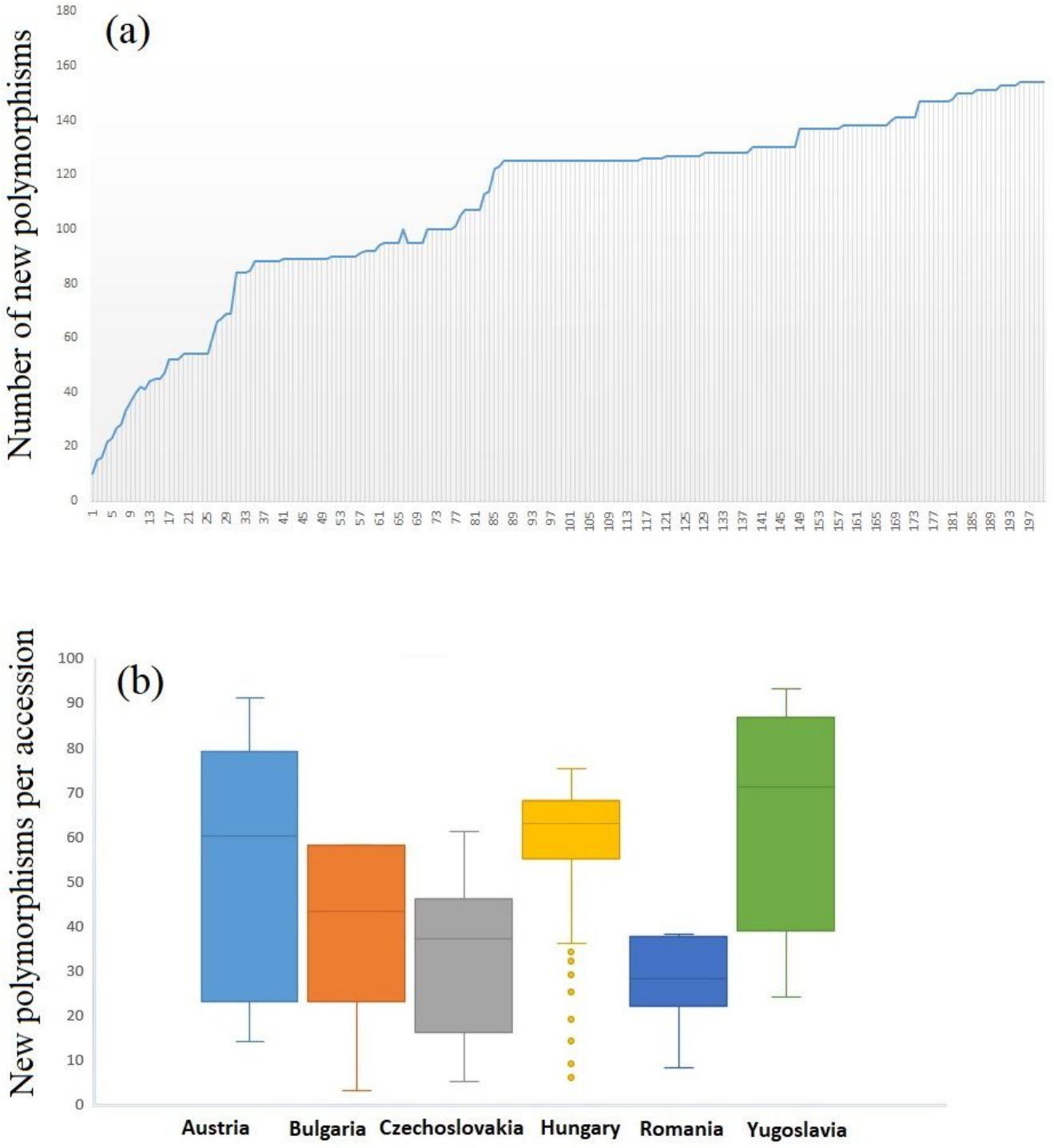

Figure 1

(a) Plot depicting the gradual incorporation of novel polymorphisms when more Central European accessions are added to the scores of modern lines. (b) Box and whisker plots depicting the amount of extra polymorphic markers introduced as each accession from each country was added to the modern varieties one by one. Outliers among Hungarian accessions are marked with yellow dots.

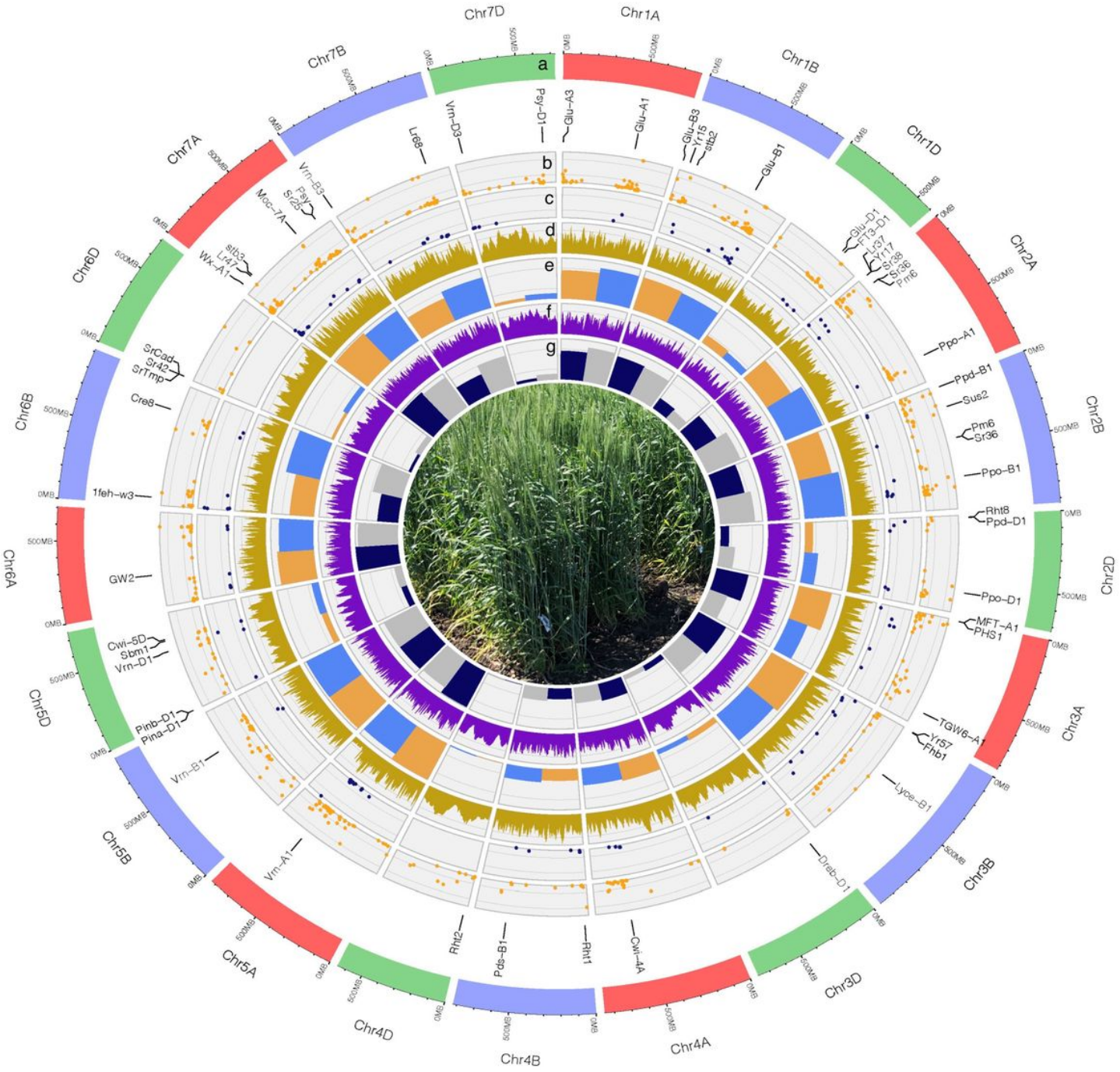

Figure 2

Circos diagram showing the physical map of twenty-one wheat chromosomes. (a) The physical scale (Mbp) of the A (red), B (blue) and D (green) genome chromosomes of wheat. Chromosomes are displayed by size proportioned bars in the outer circle. The position of genes encoding important agronomic traits are marked under the bars (see text for further details). (b) the chromosome position and frequency of polymorphic SNPs (yellow dots) found in Central European landraces compared to modern Western varieties. (c) the chromosome position and frequency of polymorphic SNPs (dark blue dots) found in landraces compared to modern Central and Western European varieties. (d) Nucleotide diversity (π) of

Central European landraces compared to Western varieties in 3Mb window intervals (purple). (e) Bar chart showing the total number of SNPs compared to the 'MV Ménrót' reference in modern varieties (orange) and landraces (light blue). (f) Nucleotide diversity (π) of Central European landraces in 3Mb window intervals (light brown). (g) Marker density of polymorphic SNPs based on the 'MV Ménrót' reference in modern varieties (dark lilac) and landraces (dark blue). Tracks are marked alphabetically (a-g) from top to bottom.

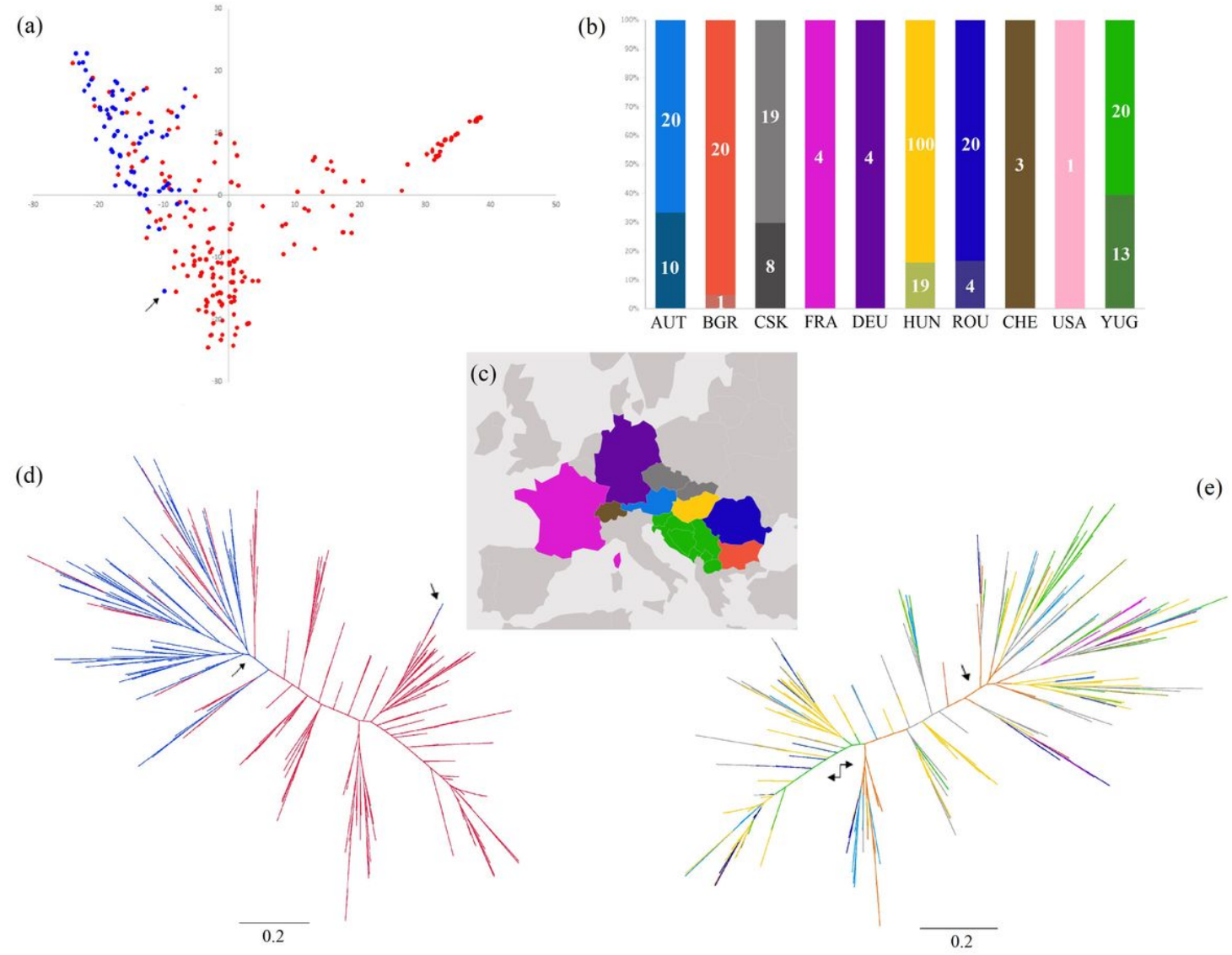

Figure 3

Genetic and geographical structuring of wheat accessions. (a) PCA plot showing the relationship between the accessions belonging to the historical collection (red dots) and modern varieties (blue dots). The modern elite variety 'Divana' from Croatia closely grouped with the landraces (black arrow). (b) Using representative coloring the number of accessions is shown as bars for each region. Darker shades and inserted numbers indicate modern varieties. A total of 199 accessions were included in our study (see Supplementary Table S4). (c) The ten regions are marked on the map from which the wheat accessions

were collected using the same coloring. (d) Unrooted maximum likelihood (ML) tree generated with IQ-Tree, with overlayed genetic origin of the accessions; branches representing modern varieties (blue) and landraces (red). The well-supported (bootstrap >90%) division (blue split) of modern varieties is marked with a black arrow. The position of ‘Divana’ is indicated with the second black arrow pointing to the blue branch nested within the red group of landraces confirming its close affinity. (e) Unrooted ML tree with overlayed representative coloring (see b and c) corresponding to the country of origin of the accessions. Arrows indicate well-supported groupings (bootstrap >90%) in both ML trees (d and e) while the rest of the nodes received weak signal (<35%). A general time reversible nucleotide evolutionary model with direct base frequency counts was used to infer topologies. The trees are draw to scale with branch lengths measured in the number of substitutions per site. The scale bar represents 0.2 substitution per site. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

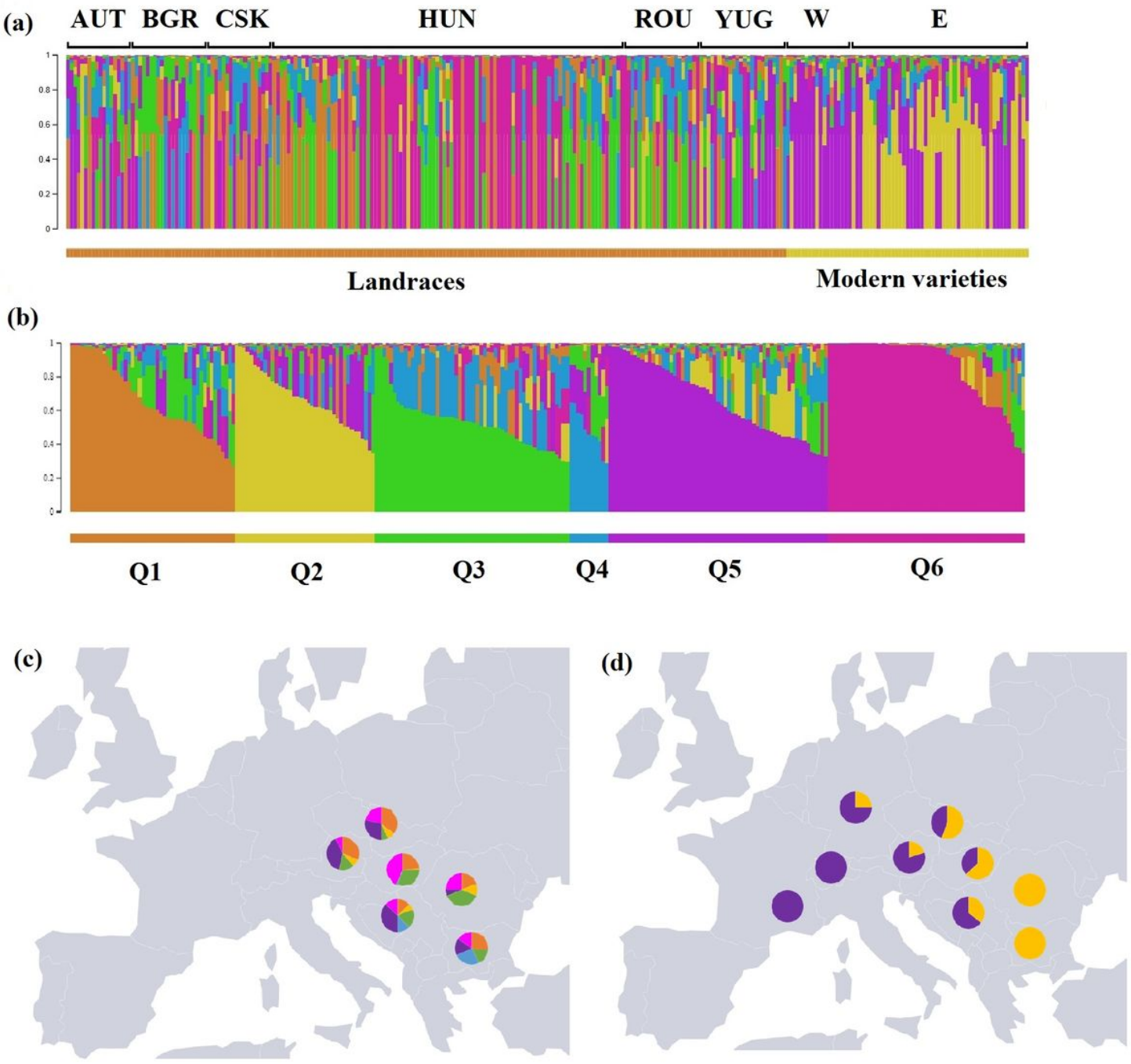

Figure 4

Patterns of admixture and population structure in the Central European historical wheat collection. (a) Plotted are STRUCTURE models with indicated K=6 optimal clustering. Each accession is represented by an individual vertical line divided into K colored segments with heights according to genotype memberships in the clusters. Landraces are grouped according to their regional origins, with international three letter country codes. Modern varieties are divided into Western (W) and Central/Eastern European grouping (E). (b) Accessions and K clustering vertical lines were given membership coefficients for each of the six clusters (Q1-6) if they had 50% or more participation in that group. (c) Pie charts depicting the average individual membership proportions (Q) in each of the six inferred ancestral groups identified by

STRUCTURE analysis for each region from which landraces were gathered. The landraces were dominated by four ancestral groups (Q1,3,4, and 6) supporting the distinct separation of landraces from modern elite varieties. (d) Pie charts depicting the percentage memberships of modern wheat varieties. Ancestral group 5 dominates accessions from Western Europe, while group 2 and 5 are mixed in Central Europe representing a junction between Easter Europe where group 2 has the upper hand. The chart shows how the previously dominant four ancestral groups (Q1,3,4, and 6) in landraces were replaced by group Q2 and 5 over time in modern varieties. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

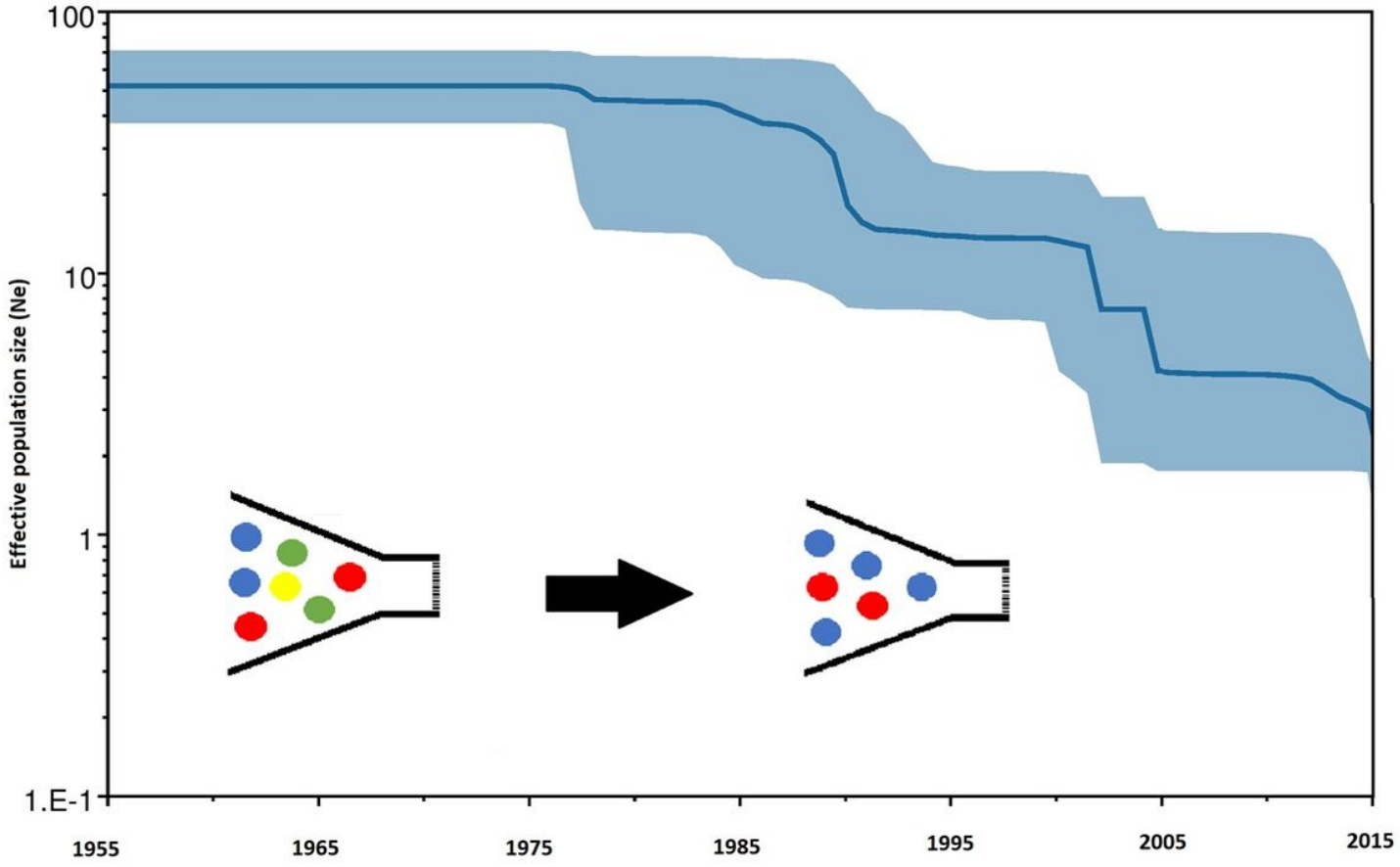

Figure 5

Bayesian Skyline plot (BSP) of the wheat collection depicting population size fluctuation over time. The vertical axis represents a time scale of calendar years between 1955–2015, while the horizontal axis represents changes in the inferred value of the effective population size over time (Ne). The median estimate is shown as a black line, while the 95% highest posterior density intervals are shown in blue. The genetic erosion of the accessions due to domestication bottleneck is predominant in the figure.

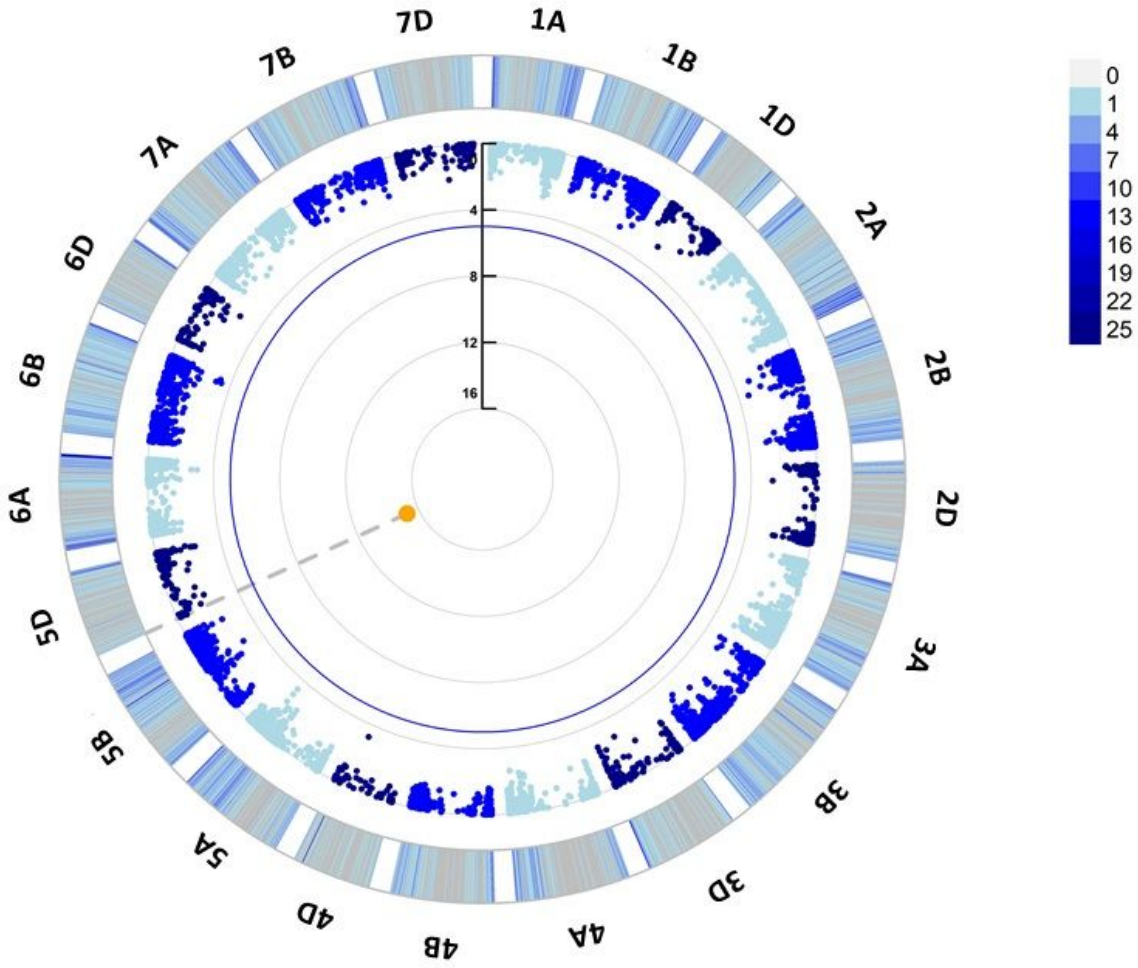

Figure 6

Circular-Manhattan plot of genome-wide association study (GWAS) for the grain hardness. The significance level of marker-trait associations (−log10(P-values) is represented by the vertical scale bar. Significance threshold (blue line) was set by application of Bonferroni correction (-log10(P)=5.40). The yellow dot represents the two significant markers (BS00000020_51 and TG0028). Individual chromosomes are represented on the outer circle and separated from each other by white borders and the difference tones of blue are shown in separated windows with different SNP density.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigureS1PCAplotcountries.pdf
- SupplementaryFigureS2DeltaK.jpg
- SupplementaryFigureS3.pdf

- SupplementaryFigureS4.pdf
- SupplementaryTableS1NovelPolymorphisms.xlsx
- SupplementaryTableS2Chromosomediversity.xlsx
- SupplementaryTableS3Listofimportantgenes.xlsx
- SupplementaryTableS4Originofwheataccessions.xlsx
- SupplementaryTableS5AncestralGroupsandAdmixedAccessions.xlsx