

Genome-wide SNP diversity in Turkish bread wheat landraces: three gene pools, a climatic gradient and genotypic redundancy

Muhammad Massub Tehseen¹

muhammad.tehseen@ndsu.edu

North Dakota State University

Deniz Istipliler²

Ege University

Fatma Aykut Tonk²

Ege University

Zakaria Kehel³

International Center for Agricultural Research in the Dry Areas

Marta Silva Lopes⁴

Institute of Agrifood Research and Technology

Muhammad Salman Mubarik⁵

University of Narowal

Ezgi Kurtulus⁶

Syngenta Crop Protection

Handan Kavaz⁷

Viticulture Research Institute

Izzet Ozseven⁸

Aegean Agricultural Research Institute

Muhammad Azhar Nadeem⁹

Mersin Üniversitesi

Muzaffer Tosun²

Ege University

Kumarse Nazari¹⁰

International Center for Agricultural Research in the Dry Areas

Research Article

Keywords: Bread wheat, Landraces, Genotyping-by-sequencing, Population structure, Genetic resources conservation, Bioclimatic variables

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11295451/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Türkiye's farmer-maintained wheat landraces harbor genetic diversity largely absent from modern cultivars, but how this variation is structured across the genebank collection remains uncharacterized at a genome-wide scale. We genotyped 794 bread wheat (*Triticum aestivum* L.) landraces from all seven geographical regions of Türkiye by genotyping-by-sequencing and analyzed 2,411 quality-filtered single nucleotide polymorphisms. Bayesian clustering resolved three ancestral gene pools of 330, 204 and 260 accessions, differentiated at $F_{ST} = 0.211\text{--}0.440$ and differing markedly in breadth ($H_e = 0.164\text{--}0.343$). The narrowest and most differentiated, Cluster 2, was effectively confined to Eastern Anatolia (91.7% of its members) and absent from the Aegean and Mediterranean. Expected heterozygosity varied little among regions (0.269–0.310), but region-restricted alleles were concentrated in Eastern Anatolia at equal sampling effort (34.1 against 0.5–2.3 elsewhere). Ancestry accounted for 28.33% of molecular variance, against 11.54% for region of collection and 16.26% for province; regional differentiation was not predicted by geography. At province resolution, differentiation was associated with geographical ($r = 0.656$) and, more strongly, climatic distance ($r = 0.733$), and the climatic association held with geography controlled for. Mean Cluster 2 membership declined with provincial temperature ($r = -0.742$). Pairwise identity grouped the 794 accessions into 523 genotype groups. Conservation and core collection design for this material should be organized around inferred ancestry and region-restricted variation rather than region of collection.

1. Introduction

Bread wheat (*Triticum aestivum* L.) is among the most widely cultivated crops on earth and a foundational component of global food and nutritional security. Sustaining and increasing wheat productivity under progressively warmer, drier and more variable growing conditions has become one of the defining challenges of contemporary agriculture, and it is increasingly clear that the genetic variation required to meet that challenge is not fully represented within elite breeding germplasm (Lopes et al. 2015; Cheng et al. 2024). Landraces, locally adapted, genetically heterogeneous populations maintained by farmers under traditional cultivation and informal seed systems, are the principal repository of the variation that formal breeding has left behind, and they have consequently become a focus of renewed interest in both conservation and pre-breeding (Lopes et al. 2015; Babalola et al. 2025).

The rationale for this interest is grounded in a well-documented contraction of the cultivated wheat gene pool. Comparative marker studies spanning the twentieth century have shown that the transition from farmer-maintained landraces to systematically bred, widely adopted semi-dwarf cultivars was accompanied by measurable losses of allelic richness, with diversity bottlenecks associated both with domestication itself and with subsequent rounds of directional selection (Reif et al. 2005). This process, generally described as crop genetic erosion, is now recognized as a systemic risk to the adaptive capacity of major staple crops rather than a purely academic concern (Khoury et al. 2022). Recent work has also demonstrated that the erosion is reversible in practical terms: introgression of landrace-derived haplotypes into modern backgrounds can deliver measurable gains in yield, resilience and quality traits, confirming that historical collections retain breeding value rather than merely archival value (Cheng et al. 2024). Studies further indicate that landraces and landrace-derived evolutionary populations frequently outperform modern cultivars in yield stability under drought and low-input conditions, precisely the environments projected to expand under climate change (Babalola et al. 2025).

Türkiye occupies a position of exceptional importance in this context. The country spans a substantial part of the Fertile Crescent, and molecular evidence has repeatedly localized the domestication of the founder wheats to southeastern Anatolia: DNA fingerprinting placed einkorn domestication in the vicinity of the Karacadağ range (Heun et al. 1997), and amplified fragment length polymorphism analysis of tetraploid wheats implicated the same region in the origin of domesticated emmer and, by descent, hard wheat (Özkan et al. 2002). Broader syntheses of the genetics and geography of Near Eastern cereal domestication have consolidated this picture (Salamini et al. 2002), while more recent analyses have argued for a reticulated rather than strictly monophyletic origin, involving multiple wild populations distributed across the Fertile Crescent (Civáň et al. 2013). Superimposed on this deep history is an unusual degree of contemporary environmental heterogeneity. Türkiye's seven geographical regions encompass Mediterranean coastal zones, humid Black Sea shoreline, the semi-arid Central Anatolian plateau, and cold, high-altitude montane environments in the east, and variation among Turkish landraces has been shown to track altitude and region of collection (Karagöz and Zencirci 2005).

That heterogeneity raises a question that administrative units are poorly placed to answer. The geographical regions of Türkiye are governmental constructs, and there is no reason for their boundaries to coincide either with the seed-exchange catchments through which landrace germplasm circulates (Pautasso et al. 2013; Jensen et al. 2013) or with the agro-ecological zones along which cereal landraces are known to differentiate. Where the comparison has been made directly, environment has been the better descriptor:

Ethiopian barley collected along an altitudinal transect diverged among altitude classes while showing little divergence among administrative districts (Hadado et al. 2010), and Iranian bread wheat landraces separated primarily by climatic zone rather than by province (Alipour et al. 2017). Whether the same holds for Turkish landraces has not been tested with genome-wide markers, and the question is not merely descriptive: if environmental gradients rather than collection regions delimit the material, then a collection stratified by region will sample the wrong axis.

A number of studies have examined the molecular diversity of Turkish wheat germplasm, but the resolution and scope of these efforts have varied considerably. Microsatellite and morphological characterization established an initial baseline for Turkish bread wheat landraces (Sönmezoğlu et al. 2012), and subsequent SSR-based work extended this to ancient hulled wheats (Gurcan et al. 2017) and to national genotype panels (Türkoğlu et al. 2023). Retrotransposon-based marker systems have been applied to Turkish bread wheat germplasm and to multi-ploidy collections (Nadeem 2021; Demirel et al. 2024), and comparative analyses have contrasted Turkish material with collections from other countries (Khan et al. 2015). In parallel, in situ conserved Turkish landraces have been mined successfully for grain yield and stripe rust resistance alleles, demonstrating the applied value of the collection (Sehgal et al. 2016). Turkish accessions have also been included in wider regional SNP surveys: a genotyping-by-sequencing study of 600 landraces held at the ICARDA genebank comprised 157 Turkish accessions alongside material from Syria, Iran and neighboring countries (Tehseen et al. 2022). What is largely absent from this body of work is a high-density, genome-wide SNP characterization of a large, geographically stratified panel drawn wholly from within Türkiye. This contrasts with the situation in several neighboring and comparable countries, where genotyping-by-sequencing (Elshire et al. 2011) and SNP array platforms have already been deployed at scale, in Iranian landraces and cultivars (Alipour et al. 2017), in Afghan landrace collections (Sohail et al. 2015; Manickavelu et al. 2014; Tehseen et al. 2021) and in Ethiopian durum wheat (Alemu et al. 2020; Negisho et al. 2021). The absence of an equivalent dataset for Türkiye is a substantive gap, because the resolution of legacy marker systems is generally insufficient to distinguish shared ancestry from convergent allele frequencies, and therefore insufficient to determine whether the structure of a landrace collection is organized by geography, by ancestry, or by some combination of the two.

Resolving that question has direct practical consequences. Core collection design depends on knowing which axis of variation captures the most genetic information, since sampling stratified along the wrong axis systematically under-represents distinct material while duplicating redundant material (Brown 1989; Berkner et al. 2024). Estimates of genetic differentiation and admixture likewise determine which accessions are informative for association mapping and which crosses are most likely to generate novel allelic combinations (Wang et al. 2025). Here we report a genome-wide SNP-based analysis of 794 Turkish bread wheat landraces spanning all seven geographical regions of Türkiye. Our objectives were to (i) infer the genetic structure of the collection using complementary Bayesian, multivariate and network approaches; (ii) quantify diversity and region-restricted allelic variation with explicit correction for the unbalanced sampling; (iii) measure genotypic redundancy within the collection; (iv) partition molecular variance among regions, inferred ancestral groups and provinces of collection; and (v) test whether differentiation among provinces is associated with geographical separation, with climate, or with both.

2. Materials and Methods

2.1 Plant material

The study panel comprised 794 bread wheat (*Triticum aestivum* L.) landrace accessions held in the International Center for Agricultural Research in the Dry Areas (ICARDA) gene bank collection. Each accession was assigned to one of the seven geographical regions of Türkiye on the basis of its recorded collection locality: Aegean, Black Sea, Central Anatolia, Eastern Anatolia, Marmara, Mediterranean and Southeastern Anatolia. Sampling was markedly unbalanced across regions, reflecting the historical composition of the collection rather than a designed sampling scheme: Eastern Anatolia accounted for 560 accessions (70.5% of the panel), followed by Southeastern Anatolia (n = 70), Black Sea (n = 54), Central Anatolia (n = 44), Mediterranean (n = 30), Marmara (n = 19) and Aegean (n = 17) (Table 1). Each accession also carried a recorded province of collection, giving 44 provinces across the panel, of which 14 were represented by at least ten accessions (Table 5). Regional labels were treated as predefined population assignments throughout.

2.2 Genotyping and variant calling

Genomic DNA was extracted from young leaf tissue of each accession using a cetyltrimethylammonium bromide (CTAB) based protocol. Reduced-representation libraries were prepared and sequenced following the genotyping-by-sequencing (GBS) approach of Elshire et al. (2011). Reads were demultiplexed, aligned and processed, and variants called, with the TASSEL-GBS pipeline (Glaubitz et al. 2014) against the International Wheat Genome Sequencing Consortium reference assembly (IWGSC 2018). The sequencing was carried out at the Genetic Analysis Service for Agriculture (SAGA) of the International Maize and Wheat Improvement Center (CIMMYT) in Mexico, supported by the Seed of Discovery Project, SeeD. Variant positions were reported against the 21 bread wheat chromosomes. Genotypes were encoded as alternate-allele dosages (0, 1 or 2). The unfiltered variant call format (VCF) file contained 10,269 SNPs across 813 samples. Across the whole file 3,548,993 calls were homozygous reference and 714,318 homozygous alternate; no heterozygous call was made at any locus in any sample, so observed heterozygosity is zero throughout and the accessions are treated as inbred lines.

2.3 Marker filtering

Marker quality control was carried out in R (R Core Team 2024) using the *vcfR* package (Knaus and Grünwald 2017) for file import and the *adegenet* package (Jombart 2008) for representation of genotypes as a *genlight* object. Loci with more than 10% missing genotype calls were removed, followed by filtering of loci with a minor allele frequency (MAF) below 5%. The resulting quality-filtered marker dataset was used for all subsequent genetic diversity, population structure, differentiation, multivariate, network, and environmental association analyses.

2.4 Bayesian clustering

Population structure was inferred using STRUCTURE v2.3.4 (Pritchard et al. 2000) under the admixture model with correlated allele frequencies. Values of K from 1 to 10 were evaluated, with ten independent runs at each value of K, each comprising a burn-in of 100,000 Markov chain Monte Carlo iterations followed by 200,000 post burn-in iterations. The optimal number of genetic clusters was determined using the Δ_K statistic of Evanno et al. (2005), computed from the run-wise log-likelihoods; the mean log-likelihood $L(K)$, its first and second derivatives, and Δ_K are presented in Figure S1. Individual membership coefficient (Q) matrices from replicate runs at the selected K were aligned using CLUMPP v1.1.2 (Jakobsson and Rosenberg 2007).

2.5 Multivariate and network analyses

Discriminant analysis of principal components (DAPC; Jombart et al. 2010) was applied using the seven regional groupings as prior population assignments. DAPC transforms the data by principal component analysis and applies linear discriminant analysis to the retained components; because it assumes neither Hardy-Weinberg nor linkage equilibrium suits a predominantly self-pollinating species. Missing genotypes were replaced by the locus mean before decomposition. The number of principal components retained was chosen by α -score optimization over a grid from 5 to 300 components, each candidate being scored as the difference between the proportion of accessions correctly reassigned to their own region and the proportion reassigned under randomized group labels, averaged over 30 randomizations; the optimum was 100 components, representing 79.0% of the total variance, and the six resulting discriminant functions were retained.

An unrooted neighbor-joining dendrogram (Saitou and Nei 1987) was constructed from a genetic distance defined for each pair of accessions as the proportion of jointly genotyped loci carrying different calls, with missing calls excluded pairwise. Branch support was obtained from 100 bootstrap replicates in which loci were resampled with replacement, the distance matrix recomputed and the tree rebuilt, and support for each internal branch of the reference tree was recorded as the proportion of replicate trees recovering the same bipartition. A minimum spanning network was derived from the same distance matrix as the minimum spanning tree over all 794 accessions, so that every edge joins the pair of accessions of least genetic distance consistent with a spanning structure. Concordance between network topology and region of collection was quantified as the proportion of network edges joining accessions of the same region, compared with the mean proportion obtained under 500 random permutations of the regional labels, and as the attribute assortative coefficient.

2.6 Diversity, rarefaction and redundancy

The mean number of observed alleles per locus (N_a), the mean number of effective alleles (N_e), Shannon's information index (I), expected heterozygosity (H_e) and unbiased expected heterozygosity (uH_e ; Nei 1978) were computed for each predefined regional population and for each inferred genetic cluster from the observed allele frequencies at the 2,411 filtered SNPs. Polymorphism information content (PIC; Botstein et al. 1980) was computed per locus as $PIC = 2pq(1 - pq)$, where p and q denote the frequencies of the two alleles, and averaged across loci within each group.

Because regional sample sizes differed by more than an order of magnitude, two rarefaction analyses supplemented the sample-size-corrected index uH_e . First, H_e was recomputed on 1,000 random draws of 17 accessions per region, the smallest regional sample, recording how often each region ranked highest. Second, region-restricted alleles were counted on the same draws: an allele was scored as restricted when observed in a region's draw and in none of the other six. Both analyses used 1,000 replicates; rarefaction of allele counts follows the approach of Kalinowski (2004). Alleles were also counted without rarefaction across the complete collection; an allele being scored as exclusive to a region when it was observed there and in no accession from any other region. Both the reference and the alternate allele were considered at every locus.

Genotypic redundancy within the collection was assessed from pairwise identity by state, calculated for every pair of accessions as the proportion of jointly genotyped loci carrying the same homozygous call. Accessions linked at an identity above 0.99 were grouped by single linkage, and the number of resulting groups was taken as the number of genetically distinguishable genotypes in the collection. Sensitivity to this threshold was examined between 0.95 and exact identity, and the minimum within-group identity was recorded for every group to confirm that grouping did not chain dissimilar accessions together (Table S6).

To establish whether the principal findings depend on repeated genotypes, all diversity, differentiation and variance-partitioning analyses were repeated on a deduplicated panel in which one accession was retained at random from each redundancy group within each region, so that a genotype occurring in more than one region was retained once per region. Cluster membership was held fixed at the assignment obtained from the complete panel. The environmental association analyses were repeated on an equivalently deduplicated panel defined within provinces.

2.7 Variance partitioning and differentiation

Analysis of molecular variance (AMOVA; Excoffier et al. 1992) was used to partition total molecular variance into among-group and within-group components under three alternative groupings of the same 794 accessions: the seven geographical regions, the three STRUCTURE-derived genetic clusters, and the 44 provinces of collection. All three analyses used the same squared Euclidean distance matrix computed from the filtered 2,411-SNP dosage data, with missing genotypes replaced by the locus mean, so that the partitions are directly comparable. Variance components, their percentage contributions and the associated Φ_{ST} statistic are reported, and significance was assessed by 999 permutations of accessions among groups.

Pairwise genetic differentiation was estimated using the F_{ST} estimator of Weir and Cockerham (1984), with missing calls retained as missing. Pairwise F_{ST} matrices were computed for the seven geographical regions, for the three genetic clusters and for the 14 provinces represented by at least ten accessions, and Nei's standard genetic distance was computed among the same groups from allele frequencies at the same 2,411 SNPs.

2.8 Geographical and climatic association

Geographical distances were computed with the Haversine formula between group reference coordinates: representative regional centroids for the seven regions (Table 1), and official province centroids for the province-level analysis (Table 5). Pairwise centroid distances ranged from 208 to 1,128 km among regions and from 70 to 1,133 km among provinces.

Climate was described by the 19 standard bioclimatic variables (BIO1–BIO19; O'Donnell and Ignizio 2012), derived from monthly temperature and precipitation normals for 1991–2020 published by the Turkish State Meteorological Service (n.d.), supplemented for Kahramanmaraş by the corresponding station record from the World Meteorological Organization archive hosted by NOAA's National Centers for Environmental Information (WMO Member Nations 2023). Values were compiled per province and used as province-level climatic proxies. The 19 variables were standardized to zero mean and unit variance across the provinces entering each analysis, and climatic distance between a pair of provinces was taken as the Euclidean distance in this standardized 19-dimensional space.

Association between genetic differentiation and each of geographical and climatic distance was assessed by Mantel tests (Mantel 1967) with 9,999 permutations of the group labels, and the independent contribution of each was assessed by partial Mantel tests in which the third matrix was held constant. Genetic distances were expressed throughout as linearized F_{ST} , calculated as $F_{ST}/(1 - F_{ST})$, the quantity expected to increase approximately linearly with geographical distance under a one-dimensional stepping-stone model (Rousset 1997). Partial Mantel tests do not account for spatial autocorrelation in the predictor matrices and can return inflated significance where such autocorrelation is present (Guillot and Rousset 2013); their results are accordingly reported as associations and are not interpreted causally. The province analysis was repeated on the deduplicated panel, on provinces outside Eastern Anatolia, and on Eastern Anatolian provinces alone (Table 6).

Two further screens examined individual bioclimatic variables. Pairwise genetic differentiation among the 14 provinces was tested against the absolute difference in each of BIO1 to BIO19 by Mantel test, and mean Cluster 2 membership per province was tested against each variable by Pearson correlation, both with 9,999 permutations. All 38 tests were corrected together by the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995), and both nominal and adjusted probabilities are reported (Table S5).

2.9 Software

STRUCTURE (v.2.3.4) and CLUMPP were run as standalone applications. Marker filtering and genotype handling were carried out in R with the packages cited above. Diversity estimation, rarefaction, redundancy grouping, AMOVA, F-statistics, the multivariate and network analyses, the Mantel and partial Mantel tests and all figures were implemented in Python 3.11 using *NumPy*, *SciPy*, *scikit-learn*, *NetworkX*, *pandas* and *Matplotlib*. Analysis scripts and intermediate outputs are available as described under Data Availability.

3. Results

3.1 Marker set and genetic structure

Quality filtering reduced the raw GBS variant set from 10,269 SNPs to 3,230 after removal of loci with more than 10% missing genotype calls, and to 2,411 after the additional removal of loci with minor allele frequency below 5%. The retained markers were distributed across all 21 bread wheat chromosomes (Figure S2). Every call in the dataset was homozygous; no heterozygote was recorded at any locus in any accession, so observed heterozygosity was zero in every group examined and is not reported further. Among the calls made in the unfiltered file, 3,548,993 were homozygous for the reference allele and 714,318 for the alternate allele.

The mean log-likelihood of the STRUCTURE runs increased steeply from $K = 1$ to approximately $K = 5$ and then plateaued, a profile that does not by itself identify a single best K , but the magnitude of the second-order rate of change peaked at $K = 3$ and the Δ_K statistic showed a single sharp maximum at that value, exceeding 340 against values below approximately 30 elsewhere (Figure S1). $K = 3$ was therefore adopted for all subsequent cluster-based analyses.

At $K = 3$ the 794 accessions were assigned by maximum membership coefficient to Cluster 1 ($n = 330$, 41.6%), Cluster 2 ($n = 204$, 25.7%) and Cluster 3 ($n = 260$, 32.7%). Assignments were generally well supported: 767 accessions (96.6%) had a maximum membership coefficient of at least 0.50, 645 (81.2%) at least 0.60 and 424 (53.4%) at least 0.80, with a panel mean maximum coefficient of 0.811. The clusters differed in internal coherence, mean membership being 0.918 within Cluster 2 against 0.823 for Cluster 3 and 0.735 for Cluster 1.

The three clusters were strongly differentiated and differed markedly in breadth (Table 2). Pairwise F_{ST} was 0.289 between Clusters 1 and 2, 0.211 between Clusters 1 and 3 and 0.440 between Clusters 2 and 3, with Nei's standard genetic distances of 0.157, 0.109 and 0.206 following the same ordering. Cluster 1 was the most diverse ($H_e = 0.343$, $PIC = 0.275$), Cluster 3 intermediate (0.212 and 0.173) and Cluster 2 by far the narrowest (0.164 and 0.137). Expected heterozygosity in Cluster 2 was less than half that of Cluster 1.

Cluster membership was distributed very asymmetrically across regions of collection (Table 2, Figure 1). Cluster 2 was largely confined to Eastern Anatolia: 187 of its 204 members (91.7%) originated there, against the 70.5% expected were membership independent of region. It was recorded in no Aegean and no Mediterranean accession, and in only two Marmara, three Southeastern Anatolian, five Central Anatolian and seven Black Sea accessions. Eastern Anatolia was also the only region in which all three clusters occurred at comparable frequency, with 197, 187 and 176 accessions in Clusters 1, 2 and 3 and mean membership coefficients of

0.325, 0.349 and 0.326. The six remaining regions were composed almost entirely of Clusters 1 and 3 but differed in which predominated: Cluster 1 was the majority component in Marmara (17 of 19 accessions), the Mediterranean (23 of 30), Southeastern Anatolia (42 of 70), the Black Sea (35 of 54) and the Aegean (11 of 17), while Central Anatolia was the one region dominated by Cluster 3, which accounted for 34 of its 44 accessions.

3.2 Multivariate and network structure

Discriminant analysis of principal components using the seven regions as prior groups retained 100 principal components and six discriminant functions, the first two accounting for 39.0% and 21.4% of the discriminant variance. On these two axes (Figure 2A) the first separated Eastern Anatolia, whose centroid lay at -0.93, from the other six centroids, which fell between 0.06 (Central Anatolia) and 4.16 (Marmara); the second separated Black Sea and Marmara, at 2.69 and 3.20, from Southeastern Anatolia, Mediterranean and Aegean, at -2.39, -1.50 and -1.07. The separation was one of density rather than of boundary, since individual Eastern Anatolia accessions spanned the first axis from -2.57 to 4.20 and were distributed throughout the space occupied by the other six regions, which themselves overlapped extensively. Reassignment to the region of collection succeeded for 87.0% of the panel overall but for 95.5% of Eastern Anatolia accessions against 57.4% to 75.7% elsewhere, and the membership bar plot (Figure 2B) showed the same asymmetry.

The unrooted neighbor-joining dendrogram resolved structure at the level of terminal and intermediate clades but not at the base of the tree (Figure 3A). Of its 791 internal branches, 325 (41.1%) were recovered in at least 70 of 100 bootstrap replicates and 405 (51.2%) in at least 50, whereas no bipartition dividing the collection into two comparably sized halves attracted appreciable support. Forty-three clades combined at least ten accessions with at least 70% support; 22 of these were at least 90% Eastern Anatolian and 16 contained accessions of a single region only, the largest being four Eastern Anatolian clades of 69 to 76 accessions. Accessions from the six smaller regions did not form region-specific deep lineages but occupied adjacent sectors of the tree, with Southeastern Anatolia, Mediterranean and Aegean accessions concentrated in one sector and Black Sea and Marmara accessions in another.

The minimum spanning network reproduced this arrangement (Figure 3B). Its 793 edges spanned genetic distances of 0.000 to 0.325 with a median of 0.057, the 180 edges of zero length joining accessions of identical genotype. Eastern Anatolia occupied the interior of the network together with several dense peripheral clusters, the largest single node connecting 63 accessions, while Black Sea and Marmara accessions formed one distinct peripheral lobe and Mediterranean, Aegean and Southeastern Anatolia accessions another. Edges joined accessions of the same region in 70.1% of cases against a permuted expectation of 51.6%, giving a regional assortativity of 0.418: neighboring accessions in the network shared a region more often than chance, but nearly a third of the shortest genetic connections in the collection crossed regional boundaries.

3.3 Diversity and allelic uniqueness

Diversity indices for the seven regional populations are given in Table 1. Expected heterozygosity spanned a narrow range, from 0.269 in Central Anatolia to 0.310 in the Mediterranean, with Eastern Anatolia at 0.307; polymorphism information content ranged from 0.219 to 0.248, with the Mediterranean and Eastern Anatolia tied at the upper limit. The effective number of alleles ranged from 1.445 in Central Anatolia to 1.529 in the Mediterranean and Shannon's information index from 0.416 to 0.469. The mean number of observed alleles per locus rose monotonically with sample size, from 1.820 in the Aegean to 2.000 in Eastern Anatolia and is reported for completeness rather than as a comparative measure.

Because regional sample sizes differed by a factor of 33, H_e was recomputed on 1,000 balanced draws of 17 accessions per region (Table 3). Under rarefaction the regional means converged further, to 0.259-0.301, and their 95% intervals overlapped extensively. The Mediterranean region had the highest mean rarefied H_e (0.301) and Eastern Anatolia the second highest (0.288), but Eastern Anatolia ranked first in only 19.7% of draws and the Mediterranean in 72.0%. No region held the highest average diversity reproducibly, and the ordering in Table 1 is therefore not interpreted as a ranking.

Region-restricted alleles behaved very differently. On the same balanced draws, Eastern Anatolia carried an average of 34.1 alleles found in no other region (95% interval 16-57), while every other region carried between 0.5 and 2.3 (Table 3). Eastern Anatolia therefore held far more region-restricted alleles than any other region at equal sampling effort, and counted without rarefaction across the complete collection, 12 alleles were exclusive to Eastern Anatolia and none was exclusive to any other region. Both counts

are conditional on the filtered marker set. The rarefaction equalizes sampling effort at the counting stage, but the minor allele frequency filter was applied earlier and across the whole panel, so an allele confined to one of the smaller regions cannot reach the 5% threshold and is removed before any count is made. Repeating the analysis on the 3,230 markers retained before that filter reduced the contrast to between 3.6 and 9.7 fold and recovered a small number of exclusive alleles in other regions, but left the ranking unchanged.

Pairwise F_{ST} among the seven regions ranged from 0.025 to 0.196 with a mean of 0.108 and a median of 0.102 across the 21 region pairs (Figure 4A; the corresponding cluster matrix is Figure 4B, discussed above; Table S3). The largest values involved Marmara and the two interior regions, Central Anatolia-Marmara (0.196) and Eastern Anatolia-Marmara (0.190), followed by Eastern Anatolia-Mediterranean (0.165), Central Anatolia-Mediterranean (0.146) and Aegean-Eastern Anatolia (0.145). The smallest was between the Aegean and Mediterranean regions (0.025); other low values were Marmara-Mediterranean (0.055), Central Anatolia-Eastern Anatolia (0.061), Black Sea-Marmara (0.063) and Aegean-Southeastern Anatolia (0.066). Seven of the 21 pairs exceeded 0.13 and five fell below 0.07, so most of the spread was contributed by a minority of pairs. Nei's standard genetic distances among regions followed the same ordering, from 0.033 for Aegean-Mediterranean to 0.122 for Eastern Anatolia-Marmara.

The contrast between the two groupings is the most consistent feature of the diversity data. Expected heterozygosity spanned 0.040 across the seven geographical regions but 0.178 across the three genetic clusters, and polymorphism information content spanned 0.029 and 0.139 respectively. On the frequency-based indices, N_e , I , H_e , uH_e and PIC, Cluster 2 fell below every regional value and Cluster 1 exceeded every regional value. Grouping the same 794 accessions by inferred ancestry therefore separated them into units of very different genetic breadth, whereas grouping them by region of collection produced seven units of closely similar breadth.

3.4 Genotypic redundancy

Pairwise identity by state grouped the 794 accessions into 523 genotype groups at a threshold of 0.99 SNP identity. The 77 groups containing multiple accessions comprised 348 accessions (43.8%), including 271 additional accessions (34.1%) beyond one representative per group. The largest group contained 70 accessions. Redundancy was present in every region but varied more than fivefold in extent, from 10.5% of accessions in Marmara and 24.1% in the Black Sea to 46.6% in Eastern Anatolia and 55.7% in Southeastern Anatolia (Table 1). It varied more still among the genetic clusters, from 27.9% in Cluster 2 to 38.5% in Cluster 1 and 63.1% in Cluster 3 (Table 2).

The count of distinct genotypes was not sensitive to the choices made in the grouping (Table S6). Raising the identity threshold from 0.99 to 0.995 and 0.999 changed the count from 523 to 534 and 550, and requiring exact identity gave 614; lowering it to 0.98 gave 504. Replacing single linkage with complete linkage at the same threshold changed the count from 523 to 525, only two of the 77 duplicate groups contained any pair falling below the threshold, and the lowest within-group identity observed anywhere was 0.985, confirming that the grouping did not chain dissimilar accessions together.

Repeating the principal analyses on a deduplicated panel of 552 accessions left every headline result intact (Table S2). Cluster 2 remained an Eastern Anatolian gene pool (90.5% of its members, and still absent from the Aegean and Mediterranean), the regional spread of H_e remained narrow, the Mediterranean remained the most diverse region, and the ordering of cluster diversity and of cluster F_{ST} was unchanged. Molecular variance among the seven regions was essentially unaltered (11.91% against 11.54%) and among the three clusters remained more than twice as large (25.39% against 28.33%).

3.5 Spatial and environmental structure

Analysis of molecular variance partitioned the same 794 accessions under three groupings, all significant at $p = 0.001$ with 999 permutations (Table 4). Region of collection accounted for 11.54% of molecular variance ($\Phi_{ST} = 0.115$), the three genetic clusters for 28.33% ($\Phi_{ST} = 0.283$) and province of collection for 16.26% ($\Phi_{ST} = 0.163$). Ancestry therefore explained roughly two and a half times as much variance as region and around three quarters more than province, while province in turn explained about 40% more than region. Under every grouping the majority of molecular variance resided within groups rather than among them.

Geographical distance did not predict differentiation at the regional scale. A Mantel test of linearized F_{ST} against Haversine distance between the seven regional centroids returned $r = 0.259$ with a one-sided p of 0.118 from 9,999 permutations (Figure 5D). Several of the most differentiated pairs were geographically proximate: Central Anatolia and Marmara, separated by 368 km, were the most differentiated of the 21 pairs ($F_{ST} = 0.196$), and Central Anatolia and the Mediterranean, separated by 283 km, the fourth most (0.146). Conversely Central Anatolia and Eastern Anatolia were among the least differentiated pairs (0.061) despite a centroid separation of approximately 690 km, and the Aegean and Southeastern Anatolia were separated by 970 km yet differentiated at only 0.066.

At the finer resolution of the province, the picture changed. Fourteen provinces were represented by at least ten accessions, together comprising 682 accessions (Table 5); pairwise F_{ST} among them ranged from 0.011 to 0.407 with a mean of 0.189. Across the 91 province pairs, differentiation was associated with geographical distance (Mantel $r = 0.656$, $p = 0.0001$; Figure 5A) and more strongly with climatic distance ($r = 0.733$, $p = 0.0001$; Figure 5B). The two predictors were themselves strongly collinear ($r = 0.661$, $p = 0.0001$), as expected for a country in which climate varies systematically from the coasts to the eastern highlands.

Partial Mantel tests separated the two predictors as far as the data permit. With geographical distance held constant the climatic association remained substantial ($r = 0.528$, $p = 0.0001$), whereas with climatic distance held constant the geographical association fell to $r = 0.336$ ($p = 0.0204$). The same asymmetry recurred in three further panels (Table 6). Deduplicating genotypes within provinces reduced the panel to 490 accessions in 12 provinces, Adiyaman and Muş falling below ten accessions; the climatic partial association remained clear ($r = 0.515$, $p = 0.0006$) and the geographical partial association rose to $r = 0.487$ ($p = 0.0023$). Among the six provinces outside Eastern Anatolia the climatic partial association was of similar magnitude but fell short of conventional significance ($r = 0.416$, $p = 0.075$) while the geographical partial association was negligible ($r = 0.060$, $p = 0.397$). Within the eight Eastern Anatolian provinces alone, where geographical distance had no detectable effect ($r = 0.137$, $p = 0.288$), the climatic association was weak but still detectable ($r = 0.333$, $p = 0.038$; partial $r = 0.322$, $p = 0.047$). The geographical partial association therefore moved between $r = 0.060$ with $p = 0.397$ and $r = 0.487$ with $p = 0.002$ depending on which accessions and provinces were retained and is reported here as a modest positive association whose magnitude is not stable.

Screening the 19 bioclimatic variables individually placed the association predominantly on winter cold and dry-season moisture (Table S5). Out of the 38 tests corrected together, 16 survived at $q < 0.05$. Differentiation among provinces was most strongly associated with temperature annual range ($r = 0.592$, $q = 0.011$), mean temperature of the coldest quarter ($r = 0.558$, $q = 0.002$), precipitation of the driest quarter ($r = 0.558$, $q = 0.014$) and of the driest month ($r = 0.556$, $q = 0.014$), with minimum temperature of the coldest month ($r = 0.552$, $q = 0.002$), temperature seasonality ($r = 0.546$, $q = 0.011$) and precipitation of the warmest quarter ($r = 0.538$, $q = 0.016$). Isothermality and the wet-season precipitation variables showed no association.

The same gradient was visible in ancestry itself. Mean Cluster 2 membership per province declined as provincial annual mean temperature rose ($r = -0.742$, $p = 0.0010$, $q = 0.010$; Figure 5C), and correlations of similar magnitude were obtained against mean temperature of the driest quarter ($r = -0.771$, $q = 0.013$), of the warmest quarter ($r = -0.721$, $q = 0.013$) and minimum temperature of the coldest month ($r = -0.698$, $q = 0.013$). Mean Cluster 2 membership reached 0.698 in Kars, at an annual mean of 5.9 °C, and stood at 0.338 in Erzurum, the coldest province in the set at 5.4 °C, and 0.361 in Ağrı (7.0 °C), against 0.050 in Hatay (19.1 °C) and 0.018 in Adiyaman (18.0 °C), the two warmest. Two provinces departed from the trend: Muş, at 10.6 °C, held almost no Cluster 2 membership (0.014), while Hakkâri, at a similar 10.8 °C, held 0.321. The gradient was not simply a contrast between Eastern Anatolia and the rest of the country: within the eight Eastern Anatolian provinces alone, which span annual mean temperatures of 5.4 to 12.0 °C, mean Cluster 2 membership still declined with temperature ($r = -0.700$).

4. Discussion

4.1 Three gene pools of unequal breadth

The collection resolves into three genetic groups that are differentiated from one another more strongly than the seven administrative regions are, and that differ markedly in internal breadth. Differentiation of the magnitude recovered here sits within the range reported for inferred subpopulations of bread wheat elsewhere: SSR analysis of Turkish bread wheat genotypes yielded F_{ST} values of 0.244 to 0.330 among subpopulations (Türkoğlu et al. 2023), GBS analysis of Iranian landraces gave 0.23 among structure-derived groups (Alipour et al. 2017), and an ICARDA landrace panel spanning eight countries gave 0.21 to 0.23 between its two inferred subpopulations (Tehseen et al. 2022). The values obtained here are therefore ordinary for this species and this kind of material and

should not be read as evidence of deep divergence; under every grouping examined the majority of molecular variance remained within groups rather than among them.

What distinguishes the three groups is not their differentiation but their asymmetry. One is broad and pervasively admixed and forms the majority component in five of the seven regions; second is intermediate and concentrated on the central plateau; the third is narrow, sharply defined and regionally confined. The combination of strong differentiation with low internal diversity in that third group is the signature of a gene pool propagated within a limited set of populations rather than circulated through the collection at large. The multivariate and network analyses recovered the same asymmetry independently of the clustering model. Discriminant analysis separated Eastern Anatolia from the remaining regions along the leading axis while leaving the six smaller regions overlapping, and the minimum spanning network joined accessions of the same region more often than chance without resolving regions into separate components. Neither analysis assumes Hardy-Weinberg or linkage equilibrium, which matters for a predominantly self-pollinating species (Jombart et al. 2010), and their agreement with the Bayesian result argues that the structure is a property of the data rather than of the admixture model.

Comparable decoupling between inferred genetic groups and administrative geography recurs across landrace collections analyzed at similar marker density. Bayesian clustering of Iranian wheat landraces did not align with province of origin (Alipour et al. 2017); the same was reported for Afghan landraces characterized with DArT and SNP platforms (Sohail et al. 2015; Manickavelu et al. 2014) and for a GBS-genotyped Afghan panel (Tehseen et al. 2021). Ethiopian durum wheat landraces analyzed with high-density SNPs showed their strongest structural signal separating landraces from improved material rather than separating regions of origin (Alemu et al. 2020; Negisho et al. 2021), and the regional survey that included Turkish accessions among ICARDA gene bank collection recovered structure cutting across national and provincial boundaries (Tehseen et al. 2022). Across these studies the decoupling is consistent enough to suggest a common feature of farmer-maintained cereal germplasm rather than a peculiarity of this collection.

What the three groups correspond to biologically cannot be settled from genotype data alone. A broad admixed component overlying older, more localized lineages is the configuration expected where freely exchanged material coexists with germplasm that has not circulated, and seed-exchange networks are known to move crop germplasm along social rather than strictly geographical pathways (Pautasso et al. 2013; Jensen et al. 2013). The alternative, that the groups reflect when and where collection expeditions sampled rather than differences in the germplasm, cannot be excluded without expedition records, which were not available. Membership was also taken from a single aligned STRUCTURE run rather than from coefficients averaged across runs, which avoids label-switching artefacts (Jakobsson and Rosenberg 2007) but does not quantify run-to-run variability in assignment.

Two properties of the marker data qualify the absolute values throughout. All statistics were computed under a diploid model, whereas bread wheat is an allohexaploid carrying three closely related sub-genomes, and reads originating from homeologous loci can collapse onto a single reference position, producing apparent variation that is in fact fixed difference between sub-genomes. This is a recognized difficulty in reduced-representation genotyping of polyploid cereals, which the two-enzyme GBS protocol was developed partly to mitigate (Poland et al. 2012) and which sub-genome-aware SNP platforms were designed specifically to resolve in hexaploid wheat (Winfield et al. 2016). Because homeologous interference acts through the same loci in every accession it affects all groups alike, so relative comparisons are considerably more robust to it than absolute values, and the diversity estimates reported here should not be compared directly with values from sub-genome-resolved datasets. The complete absence of heterozygous calls is in turn consistent with a predominantly self-pollinating species maintained as inbred lines (de Vries 1971; Waines and Hegde 2003), and means the accessions are best treated as fixed genotypes rather than segregating populations. Whether the underlying landrace populations were genuinely uniform or were represented in this collection by single-plant derivatives, cannot be determined from genotype data alone, a distinction that matters for anyone seeking to recover within-accession variation from the original field material.

4.2 A restricted Eastern Anatolian pool and uneven allelic uniqueness

The clearest geographical signal in the collection is the confinement of the narrow third gene pool to Eastern Anatolia and its complete absence from the Aegean and Mediterranean regions. Absence from a region is a stronger observation than a difference in frequency, because it does not depend on how the remaining membership is distributed. Eastern Anatolia was simultaneously the only region carrying all three components at comparable frequency, so it contains both the material shared across the country and a component found essentially nowhere else in these accessions.

Two non-exclusive processes would produce this. The first is historical. Eastern and southeastern Anatolia lie within the geographical envelope in which the founder wheats were domesticated: einkorn domestication was localized to the Karacadağ region by DNA fingerprinting (Heun et al. 1997), and AFLP analysis implicated southeastern Turkey in the origin of domesticated emmer and hard wheat (Özkan et al. 2002). Later syntheses placed these events within a broader Fertile Crescent framework (Salamini et al. 2002), and subsequent work argued for a reticulated origin drawing on several wild populations across the region (Civáň et al. 2013). Germplasm descended from continuous cultivation in that zone would be expected to retain lineages never carried westward. The second is ecological. Eastern Anatolia is topographically the most complex of the seven regions, and Turkish wheat landraces vary systematically with altitude and collection site (Karagöz and Zencirci 2005). Cultivation fragmented across steep environmental gradients limits the homogenising effect of seed exchange and allows locally differentiated populations to persist, a pattern documented directly in Ethiopian barley collected along an altitudinal transect (Hadado et al. 2010). The present data cannot weigh the historical against the ecological contribution, and no attempt is made to do so. Whichever process dominates, the practical consequence is identical: this gene pool is available almost exclusively from Eastern Anatolian collection sites and is not represented in the western regions from which germplasm is most easily obtained.

Average diversity and allelic uniqueness gave sharply different answers about that region, and the divergence is instructive. Expected heterozygosity varied little among regions, and rarefaction showed that even this narrow range does not constitute a stable ranking. Reporting a sample-size-corrected index such as uH_e adjusts the expectation of a statistic but does not establish that an ordering is reproducible, and resampling to a common size shows that in this collection it is not. We therefore make no claim that any region is the most diverse. This matters because the sampling design is severely unbalanced, and unbalanced designs are precisely the circumstance in which uncorrected allelic and diversity comparisons mislead (Kalinowski 2004).

Region-restricted alleles behaved very differently, separating Eastern Anatolia from every other region at equal sampling effort and identifying it as the region holding by far the most alleles found nowhere else in the complete collection. The two statistics measure different things. Average heterozygosity is dominated by common variants, which are shared throughout a collection of this kind, whereas allelic counts capture rare variants confined to particular populations. Conservation genetics has long held that allelic richness rather than heterozygosity should guide the choice of populations to conserve, because alleles present in a single population are those a sampling scheme can actually lose (Petit et al. 1998; Brown 1989), and allele-specificity measures have been developed explicitly for germplasm bank management on the same reasoning (Reyes-Valdés et al. 2018). The contrast recovered here is a direct illustration: the region a heterozygosity-based scheme would treat as unremarkable is the one holding nearly all the irreplaceable variation. Absolute diversity values are broadly comparable with those reported for other landrace collections analyzed with SNP markers, though such comparisons are constrained by differences in platform, filtering and ascertainment (Alipour et al. 2017; Sohail et al. 2015; Tehseen et al. 2021, 2022); the consistent finding across these studies is that average heterozygosity discriminates poorly among sub-populations of a landrace collection.

The allelic result and the clustering result are not independent, but they are different summaries of the same data and they agree: the region holding an ancestry component absent elsewhere is also the region holding alleles absent elsewhere.

4.3 Environment tracks differentiation more closely than administrative geography

Region of collection accounted for a modest share of molecular variance, and pairwise differentiation among regions did not decline with proximity: two interior regions separated by nearly 700 km were among the least differentiated pairs, while a pair barely half that distance apart was the most differentiated. A conservative reading is warranted. The regional test rests on 21 comparisons derived from seven reference points, a small number from which to detect a moderate correlation, so its outcome is not evidence that distance is irrelevant. Administrative regions are also not evolutionary units: they are defined for statistical and governmental purposes, and there is no reason to expect them to coincide with the seed-exchange catchments through which landrace germplasm moves (Pautasso et al. 2013) or with the agro-ecological zones along which cereal landraces differentiate (Hadado et al. 2010; Alipour et al. 2017). Seed moves through channels that geography does not describe: a barley study that traced exchange networks directly found gene flow between communities the farmers themselves did not report, alongside weak inter-community differentiation (Jensen et al. 2013). Because bread wheat is predominantly self-pollinating, with limited anther extrusion and brief stigma receptivity (de Vries 1971; Waines and Hegde 2003), such movement occurs through seed lots rather than pollen, which is consistent with distinct ancestral components persisting within populations even as allele frequencies are homogenized among them.

Repeating the analysis at province resolution changed the picture. Province explained appreciably more molecular variance than region, and among the 14 provinces with adequate representation, differentiation rose clearly with both geographical and climatic separation. The two results are complementary rather than contradictory, and it would overstate the evidence to describe the regional null as an established artefact of aggregation: moving from regions to provinces changes the number of groups, their identity and the spatial scale at once, and these data cannot attribute the difference to any one of those changes. What can be said is that the axis along which this collection is differentiated is finer than the administrative region, and that a stratification built on regions will not capture it.

Of the two predictors, climate was the more closely associated, and the association survived holding geography constant while the reverse was not consistently true. The interpretation of that asymmetry requires care on three counts. First, the predictors are strongly collinear, as they must be in a country where climate varies systematically from the coasts to the eastern highlands, so neither can be isolated cleanly. Second, partial Mantel tests do not account for spatial autocorrelation in the matrices being compared and can return inflated significance where it is present, a limitation demonstrated explicitly for this class of test (Guillot and Rousset 2013); the partial coefficients reported here are therefore descriptive summaries of a conditional association, not estimates of an independent causal contribution. Third, the geographical partial association was not stable: it was nominally significant on the full and deduplicated panels but negligible among the provinces outside Eastern Anatolia and within Eastern Anatolia alone. The difference between a result that crosses a conventional threshold and one that does not is not itself a difference in finding (Gelman and Stern 2006), and we therefore report a modest positive geographical association of uncertain significance rather than treating either outcome as decisive. Geography plainly contributes; what these data will not support is a claim that it contributes independently of climate.

The individual variables implicated are coherent with what limits wheat cultivation on the Anatolian plateau. Differentiation was associated most strongly with dry-season moisture and with winter cold and its seasonal amplitude, while isothermality and wet-season precipitation showed nothing. Both classes of variable correspond to recognized constraints: winter survival governs whether autumn-sown material persists at altitude in eastern Türkiye, and terminal moisture availability governs grain filling across the interior. That the same variables emerged from a screen with no agronomic input is reassuring, but it is a consistency argument and not evidence of selection. Nothing in these data identifies a locus, a phenotype or a selective mechanism, and a marker set of this density in a collection with this degree of ancestry structure could not do so.

The ancestry gradient is the clearest expression of the association. Mean membership in the narrow gene pool declined steadily with provincial temperature, and the gradient held within Eastern Anatolia alone, where the provinces span nearly seven degrees of annual mean temperature. That internal consistency makes a purely compositional explanation (the gene pool being eastern, and the east being cold) insufficient on its own. It does not make the association causal. Two provinces departed sharply from the trend, and the climate values used here are official province-labelled normals rather than measurements at collection sites, so a province is characterized by a single climatic description that its constituent localities need not share. A within-province analysis using accession-level coordinates and interpolated climate surfaces would test the same hypothesis at the resolution at which selection would actually act and remains the natural next step.

Read together with the structural results, the environmental analysis supports a modest and specific claim: the genetic organization of this collection aligns better with an environmental gradient, principally cold and dry-season moisture, than with the administrative units by which the accessions are catalogued. Whether that alignment reflects local adaptation, the historical geography of cultivation, or the spatial pattern of the collection missions themselves is not resolved here, and the balance among those explanations should not be inferred from the strength of the correlations.

4.4 Implications for conservation, core collection design and pre-breeding

Roughly a third of the accessions duplicate material already present elsewhere in the collection, and a small number of duplicate groups are very large. Redundancy on this scale is neither unusual nor peculiar to this collection: marker-based screening has revealed comparable levels across major genebanks, including an estimate that around a third of one large barley collection is likely duplicated (Anglin et al. 2025), and equivalent analyses in other crops have used the same identity-threshold approach to guide curation (Carvajal-Yepes et al. 2024). Genotyping platforms are now routinely assessed for exactly this purpose in wheat genebank genomics (Chu et al. 2020). The operational consequence is straightforward: regeneration, storage, phenotyping and distribution costs scale with the number of accessions, while the genetic content of the collection scales with the number of distinct genotypes

(Anglin et al. 2025; Chu et al. 2020). Redundancy arises through routes these data cannot separate, and which are documented across genebank collections: the same landrace collected more than once under different vernacular names, accessions split and re-registered during regeneration, and material acquired from several sources descending from a common origin (Anglin et al. 2025; Carvajal-Yepes et al. 2024). Identifying duplicates does not by itself justify discarding accessions; the conservative response, and the one adopted in genebank practice, is to flag redundancy groups, retain one representative in the active collection for routine distribution and evaluation, and hold the remainder in base storage (Anglin et al. 2025).

Two qualifications apply. An identity threshold of 0.99 across 2,411 SNPs identifies accessions with highly similar marker profiles, although differences may remain both at the assayed loci and elsewhere in the genome; denser genotyping would resolve some groups further, and morphological or phenological differences may exist that this marker set cannot detect. The count of distinct genotypes is therefore a lower bound on genetic distinctness and an upper bound on redundancy. Single-linkage grouping can in principle also chain accessions together through intermediates, although in this collection that affected only two groups, whose most distant members remained more than 98% identical.

Redundancy also bears on how the diversity comparisons should be read. A region in which nearly half the accessions duplicate one another contributes fewer independent genotypes than its accession count implies, which explains why the region contributing most of the panel showed no diversity advantage: much of its apparent size is repetition rather than variation. The same reasoning does not weaken the allelic result, because a duplicated accession contributes the same alleles as its twin and cannot inflate a count of alleles unique to a region. Repeating the analyses on a deduplicated panel confirmed this directly, and the one appreciable change was in the intermediate pool, whose diversity rose and whose differentiation from the broad pool fell once duplicates were removed, indicating that part of its apparent narrowness was repetition rather than genuine constraint.

A core collection aims to represent the variation of a larger collection in a subset small enough to characterize and distribute intensively (Brown 1989), and its value depends on whether the stratification used to select it corresponds to the real axes of variation. Four of the present results bear on that choice. Stratifying by region of collection would be inefficient, because region explained little of the molecular variance and regional diversity was effectively flat, so a regionally balanced sample would distribute effort in proportion to administrative boundaries that carry little genetic information; marker-informed sampling has been shown to outperform passport-data stratification in wheat genebank collections (Berkner et al. 2024; Wang et al. 2025). Sampling should nevertheless be weighted towards Eastern Anatolia, but for allelic rather than diversity reasons: the region is not demonstrably the most diverse, yet it holds the great majority of region-restricted alleles and is effectively the sole source of the narrow third gene pool, which a core selected on average heterozygosity would not prioritize at all (Petit et al. 1998; Reyes-Valdés et al. 2018). Redundancy should be removed before, not after, the core is drawn, since a core assembled without a duplicate check would spend roughly a third of its slots on repeats, and redundancy is concentrated in precisely the regions a diversity-weighted scheme would sample most heavily. And where passport geography is used at all, province is the more informative unit, and coverage of the climatic gradient is a better secondary criterion than coverage of the administrative map.

These considerations suggest a concrete sequence: collapse the accessions to their distinct genotypes; allocate slots across the three gene pools rather than the seven regions, weighting the narrow pool above its numerical share; within each pool prioritize accessions carrying region-restricted alleles over those maximizing average heterozygosity; and use province and climatic coverage as secondary checks on the resulting sample rather than as the primary stratification. Such a scheme uses genotype data already in hand and requires no additional characterization, though its performance would need to be benchmarked against random and regionally stratified samples of equal size before adoption. Low regional diversity is not evidence of advancing genetic erosion. The present analysis has no temporal sampling and no information on the current *in situ* status of landrace cultivation in any region; establishing erosion requires comparison across time, not across space (Khoury et al. 2022).

For pre-breeding, differentiation among the three gene pools implies that representatives of different pools carry appreciably divergent allele frequencies, so inter-pool crosses sample complementary variation rather than reshuffling the same alleles. At the same time more than half the panel was appreciably admixed, so the pools are not reproductively separated, and such crosses remain ordinary intraspecific crosses. The narrow Eastern Anatolian pool is the most attractive donor on these grounds: it is the most divergent from the other two, the least redundant internally, and the group least likely to be represented in material already circulating in breeding programs. Its association with the cold end of the climatic gradient makes it a reasonable place to begin looking for winter-hardiness and drought-related variation, but that is a hypothesis for phenotypic evaluation rather than a result of this study. The applied case for exploiting this germplasm is established independently of the present data: Turkish landraces conserved *in situ*

have yielded useful alleles for grain yield and stripe rust resistance (Sehgal et al. 2016), landrace-derived variation delivers measurable gains when introgressed into modern backgrounds (Cheng et al. 2024), and landrace panels have been advanced specifically as sources of adaptation to projected drought and heat regimes (Lopes et al. 2015; Babalola et al. 2025). What this analysis contributes is a sampling framework: it identifies which accessions are genetically distinct, which are duplicates, and which carry a gene pool scarce elsewhere in the collection.

5. Conclusions

Genotyping-by-sequencing of 794 Turkish bread wheat landraces at 2,411 quality-filtered SNPs resolved three ancestral gene pools of markedly unequal breadth, differentiated at $F_{ST} = 0.211-0.440$. The narrowest and most strongly differentiated, Cluster 2, was effectively restricted to Eastern Anatolia, which supplied 91.7% of its members and from which the Aegean and Mediterranean regions held none. Expected heterozygosity varied little among the seven geographical regions and no region ranked highest reproducibly under rarefaction, but region-restricted alleles were concentrated in Eastern Anatolia at equal sampling effort. Inferred ancestry explained 28.33% of molecular variance against 11.54% for region of collection and 16.26% for province, and geographical distance did not predict differentiation among regional centroids. At province resolution, differentiation was associated with both geographical and climatic distance, but only the climatic association was consistently maintained when the other was held constant, and mean membership of the narrow gene pool declined with provincial temperature. Grouping accessions at a threshold of 0.99 SNP identity yielded 523 genotype groups from the 794 accessions. Taken together these results indicate that conservation and core collection design for this material should be organized around inferred ancestry and region-restricted variation rather than average diversity or region of collection, that Eastern Anatolian material is the priority for both conservation and pre-breeding, and that prioritizing representatives of groups with highly similar SNP profiles could reduce evaluation costs while retaining the original accessions for conservation. Whether the alignment between genetic structure and climate reflects local adaptation remains to be tested with accession-level coordinates and phenotypic evaluation.

Declarations

Funding

This research was funded by Delivering Genetic Gain in Wheat project supported by the Bill and Melinda Gates Foundation (BMGF) and the UK Department for International Development (DFID), grant number OPP1133199.

Competing Interests

The authors declare no competing interests.

Author Contributions

M.M.T. and K.N. conceived and designed the study. M.M.T., M.S.M. and D.I. performed the formal analyses. M.M.T. wrote the original draft of the manuscript. Z.K., M.d.S.L., I.O. and F.A.T. provided resources. E.K., H.K. and M.A.N. contributed to the investigation and data curation. M.T. and K.N. provided supervision. K.N. acquired funding and coordinated the study. All authors reviewed and edited the manuscript, and read and approved the final version.

Data Availability

The analysis scripts, the filtered genotype matrix, the derived distance matrices and all intermediate output files are available from the corresponding author on reasonable request. Bioclimatic values were compiled from published Turkish State Meteorological Service normals for 1991-2020; all 19 variables are given per province in Table S7, and four of them alongside the sampling and ancestry summaries in Table 5.

Compliance with Ethical Standards

This study involved genetic characterization of accessions held in ICARDA genebank collection; no new plant material was collected from the wild or from farmers' fields for this work.

References

1. Alemu A, Feyissa T, Letta T, Abeyo B (2020) Genetic diversity and population structure analysis based on the high density SNP markers in Ethiopian durum wheat (*Triticum turgidum* ssp. durum). *BMC Genetics* 21:18. <https://doi.org/10.1186/s12863-020-0825-x>
2. Alipour H, Bihamta MR, Mohammadi V, Peyghambari SA, Bai G, Zhang G (2017) Genotyping-by-sequencing (GBS) revealed molecular genetic diversity of Iranian wheat landraces and cultivars. *Frontiers in Plant Science* 8:1293. <https://doi.org/10.3389/fpls.2017.01293>
3. Anglin NL, Wenzl P, Azevedo V, Lusty C, Ellis D, Gao D (2025) Genotyping genebank collections: Strategic approaches and considerations for optimal collection management. *Plants* 14(2):252. <https://doi.org/10.3390/plants14020252>
4. Babalola KO, Monacelli N, Gozzi M, Ceccarelli S, Folloni S, Galaverna G (2025) Genetic diversity and climate change adaptation in wheat: A systematic review of landraces, composite cross populations, and evolutionary populations. *Frontiers in Sustainable Food Systems* 9:1504922. <https://doi.org/10.3389/fsufs.2025.1504922>
5. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57(1):289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
6. Berkner MO, Jiang Y, Reif JC, Schulthess AW (2024) Trait-customized sampling of core collections from a winter wheat genebank collection supports association studies. *Frontiers in Plant Science* 15:1451749. <https://doi.org/10.3389/fpls.2024.1451749>
7. Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* 32(3):314–331
8. Brown AHD (1989) Core collections: A practical approach to genetic resources management. *Genome* 31(2):818–824. <https://doi.org/10.1139/g89-144>
9. Carvajal-Yepes M, Ospina JA, Aranzales E, Velez-Tobon M, Correa Abondano M, Manrique-Carpintero NC, Wenzl P (2024) Identifying genetically redundant accessions in the world's largest cassava collection. *Frontiers in Plant Science* 14:1338377. <https://doi.org/10.3389/fpls.2023.1338377>
10. Cheng S, Feng C, Wingen LU, Cheng H, Riche AB, Jiang M, Leverington-Waite M, Huang Z, Collier S, Orford S, ... Griffiths S (2024) Harnessing landrace diversity empowers wheat breeding. *Nature* 632(8026):823–831. <https://doi.org/10.1038/s41586-024-07682-9>
11. Chu J, Zhao Y, Beier S, Schulthess AW, Stein N, Philipp N, Röder MS, Reif JC (2020) Suitability of single-nucleotide polymorphism arrays versus genotyping-by-sequencing for genebank genomics in wheat. *Frontiers in Plant Science* 11:42. <https://doi.org/10.3389/fpls.2020.00042>
12. Civián P, Ivaničová Z, Brown TA (2013) Reticulated origin of domesticated emmer wheat supports a dynamic model for the emergence of agriculture in the Fertile Crescent. *PLOS ONE* 8(11):e81955. <https://doi.org/10.1371/journal.pone.0081955>
13. de Vries APh (1971) Flowering biology of wheat, particularly in view of hybrid seed production: A review. *Euphytica* 20(2):152–170. <https://doi.org/10.1007/BF00056076>
14. Demirel F, Yıldırım B, Eren B, Demirel S, Türkoğlu A, Haliloğlu K, Nowosad K, Bujak H, Bocianowski J (2024) Revealing genetic diversity and population structure in Türkiye's wheat germplasm using iPBS-retrotransposon markers. *Agronomy* 14(2):300. <https://doi.org/10.3390/agronomy14020300>
15. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6(5):e19379. <https://doi.org/10.1371/journal.pone.0019379>
16. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology* 14(8):2611–2620. <https://doi.org/10.1111/j.1365-294X.2005.02553.x>
17. Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* 131(2):479–491. <https://doi.org/10.1093/genetics/131.2.479>

18. Gelman A, Stern H (2006) The difference between “significant” and “not significant” is not itself statistically significant. *The American Statistician* 60(4):328–331. <https://doi.org/10.1198/000313006X152649>
19. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES (2014) TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS ONE* 9(2):e90346. <https://doi.org/10.1371/journal.pone.0090346>
20. Guillot G, Rousset F (2013) Dismantling the Mantel tests. *Methods in Ecology and Evolution* 4(4):336–344. <https://doi.org/10.1111/2041-210x.12018>
21. Gurcan K, Demirel F, Tekin M, Demirel S, Akar T (2017) Molecular and agro-morphological characterization of ancient wheat landraces of Turkey. *BMC Plant Biology* 17:171. <https://doi.org/10.1186/s12870-017-1133-0>
22. Hadado TT, Rau D, Bitocchi E, Papa R (2010) Adaptation and diversity along an altitudinal gradient in Ethiopian barley (*Hordeum vulgare* L.) landraces revealed by molecular analysis. *BMC Plant Biology* 10:121. <https://doi.org/10.1186/1471-2229-10-121>
23. Heun M, Schäfer-Pregl R, Klawan D, Castagna R, Accerbi M, Borghi B, Salamini F (1997) Site of einkorn wheat domestication identified by DNA fingerprinting. *Science* 278(5341):1312–1314. <https://doi.org/10.1126/science.278.5341.1312>
24. International Wheat Genome Sequencing Consortium (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361(6403):eaar7191. <https://doi.org/10.1126/science.aar7191>
25. Jakobsson M, Rosenberg NA (2007) CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23(14):1801–1806. <https://doi.org/10.1093/bioinformatics/btm233>
26. Jensen HR, Belqadi L, De Santis P, Sadiki M, Jarvis DI, Schoen DJ (2013) A case study of seed exchange networks and gene flow for barley (*Hordeum vulgare* subsp. *vulgare*) in Morocco. *Genetic Resources and Crop Evolution* 60(3):1119–1138. <https://doi.org/10.1007/s10722-012-9909-4>
27. Jombart T (2008) adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics* 24(11):1403–1405. <https://doi.org/10.1093/bioinformatics/btn129>
28. Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genetics* 11:94. <https://doi.org/10.1186/1471-2156-11-94>
29. Kalinowski ST (2004) Counting alleles with rarefaction: Private alleles and hierarchical sampling designs. *Conservation Genetics* 5(4):539–543. <https://doi.org/10.1023/B:COGE.0000041021.91777.1a>
30. Karagöz A, Zencirci N (2005) Variation in wheat (*Triticum* spp.) landraces from different altitudes of three regions of Turkey. *Genetic Resources and Crop Evolution* 52(6):775–785. <https://doi.org/10.1007/s10722-004-3556-3>
31. Khan MK, Pandey A, Thomas G, Akkaya MS, Kayış SA, Özsensoy Y, Hamurcu M, Gezgin S, Topal A, Hakki EE (2015) Genetic diversity and population structure of wheat in India and Turkey. *AoB PLANTS* 7:plv083. <https://doi.org/10.1093/aobpla/plv083>
32. Khoury CK, Brush S, Costich DE, Curry HA, de Haan S, Engels JMM, Guarino L, Hoban S, Mercer KL, Miller AJ, Nabhan GP, Perales HR, Richards C, Riggins C, Thormann I (2022) Crop genetic erosion: Understanding and responding to loss of crop diversity. *New Phytologist* 233(1):84–118. <https://doi.org/10.1111/nph.17733>
33. Knaus BJ, Grünwald NJ (2017) vcfr: A package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources* 17(1):44–53. <https://doi.org/10.1111/1755-0998.12549>
34. Lopes MS, El-Basyoni I, Baenziger PS, Singh S, Royo C, Ozbek K, Aktas H, Ozer E, Ozdemir F, Manickavelu A, Ban T, Vikram P (2015) Exploiting genetic diversity from landraces in wheat breeding for adaptation to climate change. *Journal of Experimental Botany* 66(12):3477–3486. <https://doi.org/10.1093/jxb/erv122>
35. Manickavelu A, Jighly A, Ban T (2014) Molecular evaluation of orphan Afghan common wheat (*Triticum aestivum* L.) landraces collected by Dr. Kihara using single nucleotide polymorphic markers. *BMC Plant Biology* 14:320. <https://doi.org/10.1186/s12870-014-0320-5>
36. Mantel N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research* 27(2, Part 1):209–220
37. Nadeem MA (2021) Deciphering the genetic diversity and population structure of Turkish bread wheat germplasm using iPBS-retrotransposons markers. *Molecular Biology Reports* 48(10):6739–6748. <https://doi.org/10.1007/s11033-021-06670-w>
38. Negisho K, Shibru S, Pillen K, Ordon F, Wehner G (2021) Genetic diversity of Ethiopian durum wheat landraces. *PLOS ONE* 16(2):e0247016. <https://doi.org/10.1371/journal.pone.0247016>

39. Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89(3):583–590. <https://doi.org/10.1093/genetics/89.3.583>
40. O'Donnell MS, Ignizio DA (2012) Bioclimatic predictors for supporting ecological applications in the conterminous United States. U.S. Geological Survey Data Series 691. <https://doi.org/10.3133/ds691>
41. Özkan H, Brandolini A, Schäfer-Pregl R, Salamini F (2002) AFLP analysis of a collection of tetraploid wheats indicates the origin of emmer and hard wheat domestication in southeast Turkey. *Molecular Biology and Evolution* 19(10):1797–1801. <https://doi.org/10.1093/oxfordjournals.molbev.a004002>
42. Pautasso M, Aistara G, Barnaud A, Caillon S, Clouvel P, Coomes OT, Delêtre M, Demeulenaere E, De Santis P, Döring T, Eloy L, Emperaire L, Garine E, Goldringer I, Jarvis D, Joly HI, Leclerc C, Louafi S, et al (2013) Seed exchange networks for agrobiodiversity conservation. A review. *Agronomy for Sustainable Development* 33(1):151–175. <https://doi.org/10.1007/s13593-012-0089-6>
43. Petit RJ, El Mousadik A, Pons O (1998) Identifying populations for conservation on the basis of genetic markers. *Conservation Biology* 12(4):844–855. <https://doi.org/10.1046/j.1523-1739.1998.96489.x>
44. Poland JA, Brown PJ, Sorrells ME, Jannink J-L (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLOS ONE* 7(2):e32253. <https://doi.org/10.1371/journal.pone.0032253>
45. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155(2):945–959. <https://doi.org/10.1093/genetics/155.2.945>
46. R Core Team (2024) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/> [DATA NEEDED: confirm the R version and release year actually used.]
47. Reif JC, Zhang P, Dreisigacker S, Warburton ML, van Ginkel M, Hoisington D, Bohn M, Melchinger AE (2005) Wheat genetic diversity trends during domestication and breeding. *Theoretical and Applied Genetics* 110(5):859–864. <https://doi.org/10.1007/s00122-004-1881-8>
48. Reyes-Valdés MH, Burgueño J, Singh S, Martínez O, Sansaloni CP (2018) An informational view of accession rarity and allele specificity in germplasm banks for management and conservation. *PLOS ONE* 13(2):e0193346. <https://doi.org/10.1371/journal.pone.0193346>
49. Rousset F (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* 145(4):1219–1228. <https://doi.org/10.1093/genetics/145.4.1219>
50. Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4(4):406–425. <https://doi.org/10.1093/oxfordjournals.molbev.a040454>
51. Salamini F, Özkan H, Brandolini A, Schäfer-Pregl R, Martin W (2002) Genetics and geography of wild cereal domestication in the Near East. *Nature Reviews Genetics* 3(6):429–441. <https://doi.org/10.1038/nrg817>
52. Sehgal D, Dreisigacker S, Belen S, Küçüközdemir Ü, Mert Z, Özer E, Morgounov A (2016) Mining centuries old in situ conserved Turkish wheat landraces for grain yield and stripe rust resistance genes. *Frontiers in Genetics* 7:201. <https://doi.org/10.3389/fgene.2016.00201>
53. Sohail Q, Manickavelu A, Ban T (2015) Genetic diversity analysis of Afghan wheat landraces (*Triticum aestivum*) using DArT markers. *Genetic Resources and Crop Evolution* 62(8):1147–1157. <https://doi.org/10.1007/s10722-015-0219-5>
54. Sönmezoğlu ÖA, Bozmaz B, Yıldırım A, Kandemir N, Aydın N (2012) Genetic characterization of Turkish bread wheat landraces based on microsatellite markers and morphological characters. *Turkish Journal of Biology* 36(5):589–597. <https://doi.org/10.3906/biy-1202-43>
55. Tehseen MM, Istiqlil D, Kehel Z, Sansaloni CP, da Silva Lopes M, Kurtulus E, Muazzam S, Nazari K (2021) Genetic diversity and population structure analysis of *Triticum aestivum* L. landrace panel from Afghanistan. *Genes* 12(3):340. <https://doi.org/10.3390/genes12030340>
56. Tehseen MM, Tonk FA, Tosun M, Istiqlil D, Amri A, Sansaloni CP, Kurtulus E, Mubarik MS, Nazari K (2022) Exploring the genetic diversity and population structure of wheat landrace population conserved at ICARDA genebank. *Frontiers in Genetics* 13:900572. <https://doi.org/10.3389/fgene.2022.900572>
57. Turkish State Meteorological Service (n.d.) Official climate statistics: provincial climate normals for 1991–2020. Meteoroloji Genel Müdürlüğü, Ankara. <https://www.mgm.gov.tr/veridegerlendirme/il-ve-ilceler-istatistik.aspx?k=H>. Accessed 2 September 2026

58. Türkoğlu A, Haliloğlu K, Mohammadi SA, Öztürk A, Bolouri P, Özkan G, Bocianowski J, Pour-Aboughadareh A, Jamshidi B (2023) Genetic diversity and population structure in Türkiye bread wheat genotypes revealed by simple sequence repeats (SSR) markers. *Genes* 14(6):1182. <https://doi.org/10.3390/genes14061182>
59. Waines JG, Hegde SG (2003) Intraspecific gene flow in bread wheat as affected by reproductive biology and pollination ecology of wheat flowers. *Crop Science* 43(2):451–463. <https://doi.org/10.2135/cropsci2003.0451>
60. Wang X, Lee S, Choi Y-M, Yi J, Desta KT, Shin M-J, Lee G-A, Jeon Y-A, Yoo E (2025) Evaluation of genetic diversity and population structure of wheat core collection using SNP and KASP markers for major agronomic traits. *BMC Plant Biology* 25:904. <https://doi.org/10.1186/s12870-025-06926-8>
61. Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38(6):1358–1370. <https://doi.org/10.1111/j.1558-5646.1984.tb05657.x>
62. Winfield MO, Allen AM, BurrIDGE AJ, Barker GLA, Benbow HR, Wilkinson PA, Coghill J, Waterfall C, Davassi A, Scopes G, Pirani A, Webster T, Brew F, Bloor C, King J, West C, Griffiths S, King I, Bentley AR, Edwards KJ (2016) High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotechnology Journal* 14(5):1195–1206. <https://doi.org/10.1111/pbi.12485>
63. WMO Member Nations (2023) WMO climatological standard normals for 1991–2020 (NCEI Accession 0253808). Version 1.1, Türkiye: Kahramanmaraş station 17255. NOAA National Centers for Environmental Information. <https://doi.org/10.25921/800j-vn07>. Accessed 2 September 2026

Tables

Table 1. Sampling, genetic diversity and genotypic redundancy of the seven geographical regions, from 2,411 filtered SNPs. N_a , mean number of alleles per locus; N_e , effective number of alleles; I , Shannon's information index; H_e , expected heterozygosity; uH_e , unbiased expected heterozygosity; PIC, polymorphism information content. An accession is counted as redundant when it shares an identity of at least 0.99 with at least one other accession in the collection.

Region	n	% of panel	Lat. (°N)	Long. (°E)	N_a	N_e	I	H_e	uH_e	PIC	In a duplicate group (n)	% redundant
Aegean	17	2.1	38.50	28.00	1.820	1.466	0.416	0.276	0.285	0.222	7	41.2
Black Sea	54	6.8	41.20	35.00	1.961	1.483	0.447	0.291	0.294	0.236	13	24.1
Central Anatolia	44	5.5	39.00	33.00	1.932	1.445	0.416	0.269	0.273	0.219	18	40.9
Eastern Anatolia	560	70.5	39.50	41.00	2.000	1.513	0.469	0.307	0.307	0.248	261	46.6
Marmara	19	2.4	40.20	29.00	1.821	1.484	0.426	0.284	0.292	0.227	2	10.5
Mediterranean	30	3.8	37.00	31.00	1.920	1.529	0.466	0.310	0.315	0.248	8	26.7
Southeastern Anatolia	70	8.8	37.50	39.00	1.973	1.472	0.441	0.287	0.289	0.233	39	55.7
Total	794	100.0	-	-							348	43.8

Centroid coordinates are representative regional reference points used for the geographical distance calculations. Percentages of the panel sum to 99.9 owing to rounding.

Table 2. Composition, diversity, redundancy and differentiation of the three genetic clusters inferred by STRUCTURE at $K = 3$. Columns AEG to SEA give the number of accessions of each cluster collected in each region: AEG, Aegean; BLS, Black Sea; CAN, Central Anatolia; EAN, Eastern Anatolia; MAR, Marmara; MED, Mediterranean; SEA, Southeastern Anatolia. Pairwise F_{ST} follows Weir and Cockerham (1984).

Cluster	n	AEG	BLS	CAN	EAN	MAR	MED	SEA	H _e	PIC	In a duplicate group (n)	% redundant
Cluster 1	330	11	35	5	197	17	23	42	0.343	0.275	127	38.5
Cluster 2	204	0	7	5	187	2	0	3	0.164	0.137	57	27.9
Cluster 3	260	6	12	34	176	0	7	25	0.212	0.173	164	63.1
Total	794	17	54	44	560	19	30	70				

Cluster pair	n (pair)	F _{ST}	Nei's standard genetic distance
Cluster 1 vs. Cluster 2	330 / 204	0.289	0.157
Cluster 1 vs. Cluster 3	330 / 260	0.211	0.109
Cluster 2 vs. Cluster 3	204 / 260	0.440	0.206

Table 3. Rarefied and census comparison of allelic variation among the seven geographical regions, from 2,411 filtered SNPs. Rarefied values are means over 1,000 random draws of 17 accessions per region, the smallest regional sample. An allele is region-restricted when it appears in a region's draw and in no other region's draw, and exclusive when it appears in a region and in no accession of any other region across the complete collection.

Region	Rarefied H _e	95% interval	Region-restricted alleles	95% interval	Alleles exclusive to the region	% of draws ranked highest for H _e
Aegean	0.276	0.276-0.276	0.5	0-3	0	0.0
Black Sea	0.278	0.251-0.299	1.0	0-4	0	3.1
Central Anatolia	0.259	0.230-0.284	2.3	0-8	0	0.1
Eastern Anatolia	0.288	0.256-0.316	34.1	16-57	12	19.7
Marmara	0.282	0.276-0.286	0.8	0-3	0	0.1
Mediterranean	0.301	0.286-0.313	1.2	0-5	0	72.0
Southeastern Anatolia	0.273	0.234-0.305	1.8	0-9	0	5.0

Table 4. Analysis of molecular variance partitioning the total variation among 794 Turkish wheat landraces under three alternative groupings of the same accessions. All three partitions use the same squared Euclidean distance matrix computed from 2,411 filtered SNPs, and significance was assessed by 999 permutations of accessions among groups.

Grouping	Source of variation	df	Variance component (σ^2)	% of total	Φ_{ST}	p
Seven geographical regions (k = 7)	Among groups	6	181.87	11.54	0.115	0.001
	Within groups	787	1,394.83	88.46		
Three STRUCTURE clusters (k = 3)	Among groups	2	465.72	28.33	0.283	0.001
	Within groups	791	1,178.11	71.67		
Province of collection (k = 44)	Among groups	43	244.79	16.26	0.163	0.001
	Within groups	750	1,260.42	83.74		

Table 5. The 14 provinces represented by at least 10 accessions (682 accessions in total), with the province centroid used for geographical distances and four of the nineteen bioclimatic variables used in the climatic analysis: BIO1, annual mean temperature; BIO6, minimum temperature of the coldest month; BIO12, annual precipitation; BIO17, precipitation of the driest quarter. Climate values are official Turkish State Meteorological Service normals for 1991-2020, and for Kahramanmaraş the corresponding WMO/NOAA record. Q_1 to Q_3 are mean STRUCTURE membership coefficients and H_e is expected heterozygosity within the province.

Province	Region	n	Lat. (°N)	Long. (°E)	BIO1 (°C)	BIO6 (°C)	BIO12 (mm)	BIO17 (mm)	H_e	Q_1	Q_2	Q_3
Sinop	Black Sea	17	41.627	34.878	15.0	4.4	728	109.1	0.242	0.550	0.117	0.334
Zonguldak	Black Sea	10	41.269	31.825	14.2	3.6	1239	199.6	0.177	0.599	0.085	0.316
Sivas	Central Anatolia	31	39.561	37.285	10.1	-6.2	455	37.4	0.257	0.165	0.202	0.632
Ağrı	Eastern Anatolia	36	39.553	43.256	7.0	-14.8	513	57.2	0.291	0.257	0.361	0.381
Bitlis	Eastern Anatolia	89	38.527	42.378	8.9	-8.5	1047	31.8	0.299	0.635	0.156	0.209
Erzincan	Eastern Anatolia	24	39.663	39.291	12.0	-5.7	380	38.2	0.242	0.171	0.175	0.654
Erzurum	Eastern Anatolia	86	40.060	41.547	5.4	-15.9	396	56.9	0.265	0.125	0.338	0.537
Hakkâri	Eastern Anatolia	50	37.458	44.068	10.8	-7.5	777	25.6	0.290	0.393	0.321	0.285
Kars	Eastern Anatolia	152	40.452	43.062	5.9	-14.8	530	70.2	0.235	0.102	0.698	0.200
Muş	Eastern Anatolia	12	39.001	41.845	10.6	-9.9	782	32.0	0.189	0.134	0.014	0.852
Van	Eastern Anatolia	103	38.460	43.588	10.2	-6.0	410	30.1	0.293	0.565	0.122	0.313
Hatay	Mediterranean	16	36.427	36.253	19.1	5.0	1124	33.8	0.271	0.682	0.050	0.268
Adıyaman	Southeastern Anatolia	14	37.795	38.288	18.0	1.8	729	13.6	0.215	0.496	0.018	0.485
Diyarbakır	Southeastern Anatolia	42	38.096	40.341	15.9	-2.0	498	9.8	0.292	0.530	0.036	0.434

Table 6. Mantel and partial Mantel tests of pairwise genetic differentiation against geographical and climatic distance among provinces, with 9,999 permutations. Climatic distance is the Euclidean distance over the 19 bioclimatic variables standardised across the provinces analysed; geographical distance is the Haversine distance between province centroids. The upper block describes each panel and the lower block gives the tests. The final line of each block gives the corresponding test at the coarser regional scale, between the seven regional centroids.

Analysis	Provinces	Pairs	Accessions	F _{ST} range
All provinces with n≥10	14	91	682	0.011-0.407
Deduplicated genotypes, n≥10	12	66	490	0.004-0.382
Excluding Eastern Anatolia	6	15	130	0.056-0.372
Within Eastern Anatolia	8	28	552	0.011-0.289
Seven geographical regions (coarse-scale comparison)	7	21	794	0.025-0.196

Analysis	Test	r	p
All provinces with n≥10	Genetic ~ geographic	0.656	0.0001
All provinces with n≥10	Genetic ~ climatic	0.733	0.0001
All provinces with n≥10	Climatic ~ geographic	0.661	0.0001
All provinces with n≥10	Genetic ~ climatic, geographic held constant	0.528	0.0001
All provinces with n≥10	Genetic ~ geographic, climatic held constant	0.336	0.0204
Deduplicated genotypes, n≥10	Genetic ~ geographic	0.725	0.0003
Deduplicated genotypes, n≥10	Genetic ~ climatic	0.736	0.0001
Deduplicated genotypes, n≥10	Climatic ~ geographic	0.640	0.0009
Deduplicated genotypes, n≥10	Genetic ~ climatic, geographic held constant	0.515	0.0006
Deduplicated genotypes, n≥10	Genetic ~ geographic, climatic held constant	0.487	0.0023
Excluding Eastern Anatolia	Genetic ~ geographic	0.453	0.0653
Excluding Eastern Anatolia	Genetic ~ climatic	0.583	0.0131
Excluding Eastern Anatolia	Climatic ~ geographic	0.719	0.0047
Excluding Eastern Anatolia	Genetic ~ climatic, geographic held constant	0.416	0.0747
Excluding Eastern Anatolia	Genetic ~ geographic, climatic held constant	0.060	0.3966
Within Eastern Anatolia	Genetic ~ geographic	0.137	0.2882
Within Eastern Anatolia	Genetic ~ climatic	0.333	0.0375
Within Eastern Anatolia	Climatic ~ geographic	0.125	0.2140
Within Eastern Anatolia	Genetic ~ climatic, geographic held constant	0.322	0.0472
Within Eastern Anatolia	Genetic ~ geographic, climatic held constant	0.101	0.3317
Seven geographical regions (coarse-scale comparison)	Genetic ~ geographic	0.259	0.1176

Probabilities are one-sided for the directional Mantel and partial Mantel tests, from 9,999 permutations. F_{ST} ranges are given for the untransformed estimator; all correlations were computed on linearised F_{ST}.

Figures



Figure 1

STRUCTURE membership coefficients at K = 3 for 794 Turkish wheat landraces, from CLUMPP-aligned coefficients of a single representative run. Each vertical bar is one accession, partitioned into the three ancestry components. Accessions are grouped by region of collection, abbreviated AEG (Aegean), BLS (Black Sea), CAN (Central Anatolia), EAN (Eastern Anatolia), MAR (Marmara), MED (Mediterranean) and SEA (Southeastern Anatolia), and ordered within each region by assigned cluster and then by membership in that cluster.

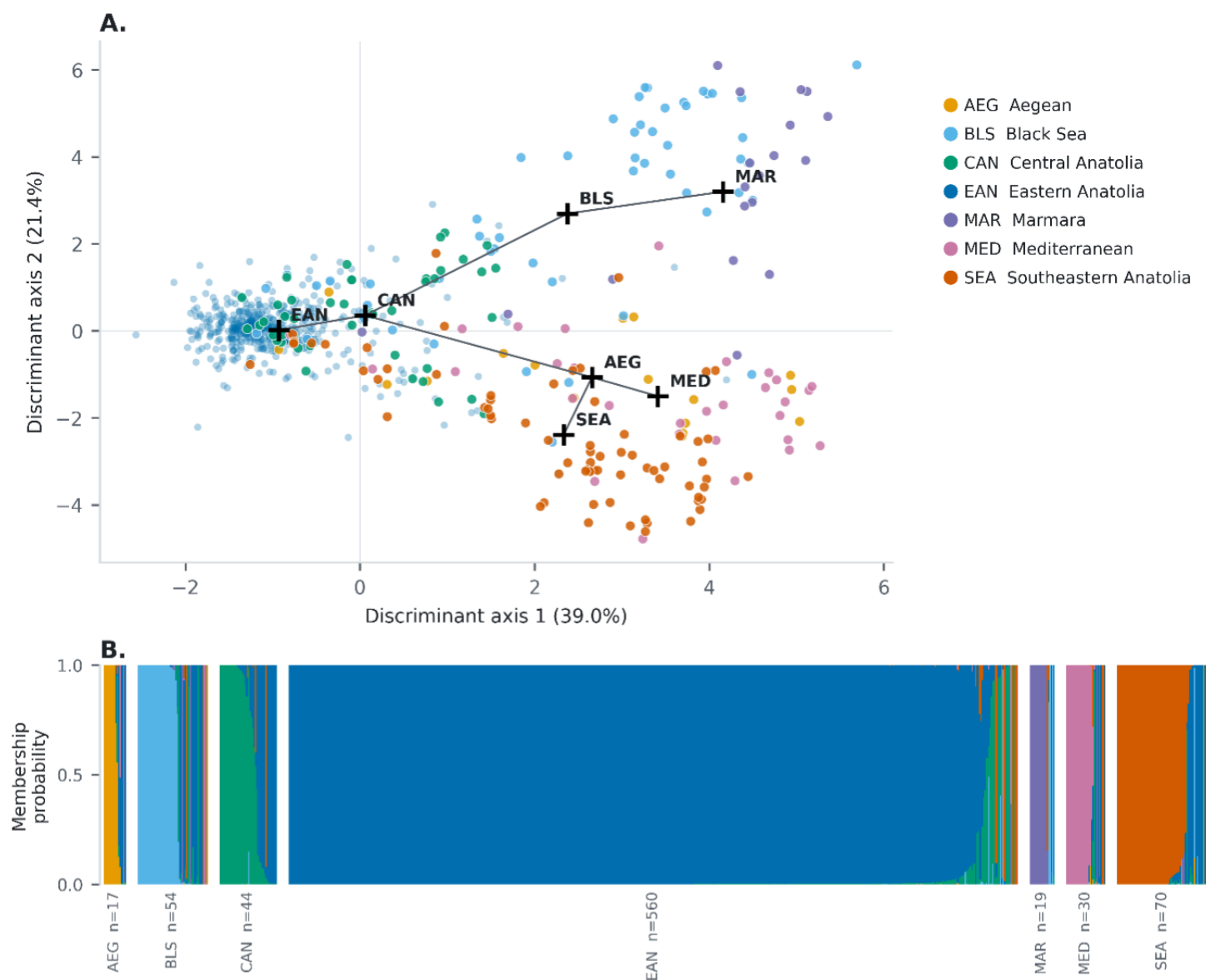


Figure 2

Discriminant analysis of principal components of 794 Turkish wheat landraces, using the seven geographical regions as prior groups and retaining 100 principal components. (A) Scatter of individuals on the first two discriminant axes, which account for 39.0% and 21.4% of the discriminant variance; points are colored by region, crosses denote group centroids, and connecting segments show the minimum spanning tree among centroids. (B) Membership probability bar plot; each bar is one accession; bars are grouped by region of collection and colors denote the region to which the accession was assigned. Region abbreviations are as in Figure 1.

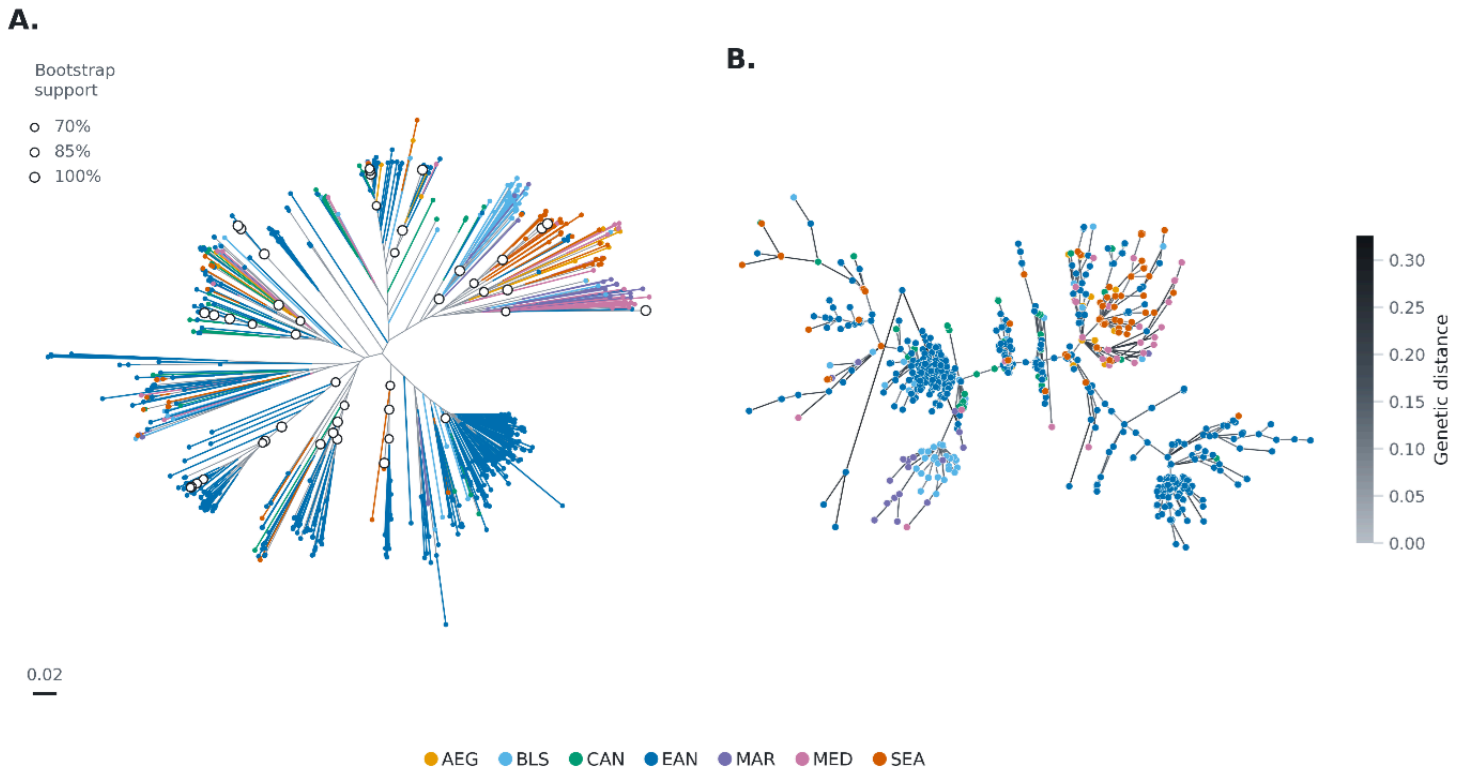


Figure 3

Genetic relationships among 794 Turkish wheat landraces based on a genetic distance defined as the proportion of jointly genotyped loci differing between two accessions. (A) Unrooted neighbor-joining dendrogram; tips and terminal branches are colored by geographical region and open circles mark internal branches recovered in at least 70 of 100 bootstrap replicates and subtending at least ten accessions, circle size being proportional to support. The scale bar denotes 0.02 substitutions per locus. (B) Minimum spanning network constructed from the same distance matrix; node color indicates geographical region, and edge shading indicates genetic distance (0.000-0.325). Region abbreviations are as in Figure 1.

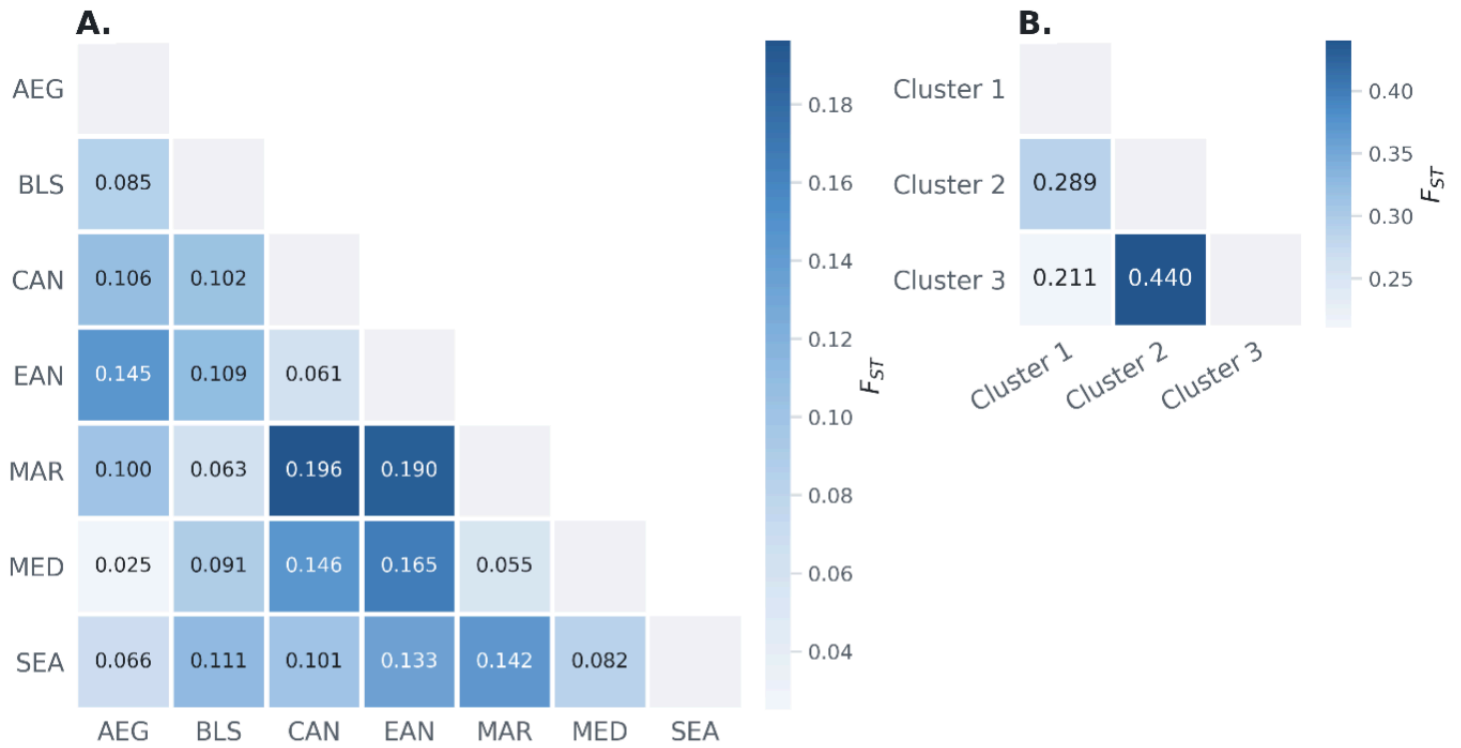


Figure 4

Heatmaps of pairwise F_{ST} (Weir and Cockerham 1984). (A) Among the seven geographical regions, abbreviated as in Figure 1. (B) Among the three STRUcTURE-inferred genetic clusters at $K = 3$. Cell shading is proportional to F_{ST} on a single-hue scale and every value is annotated within its cell; the two panels use independent color scales.

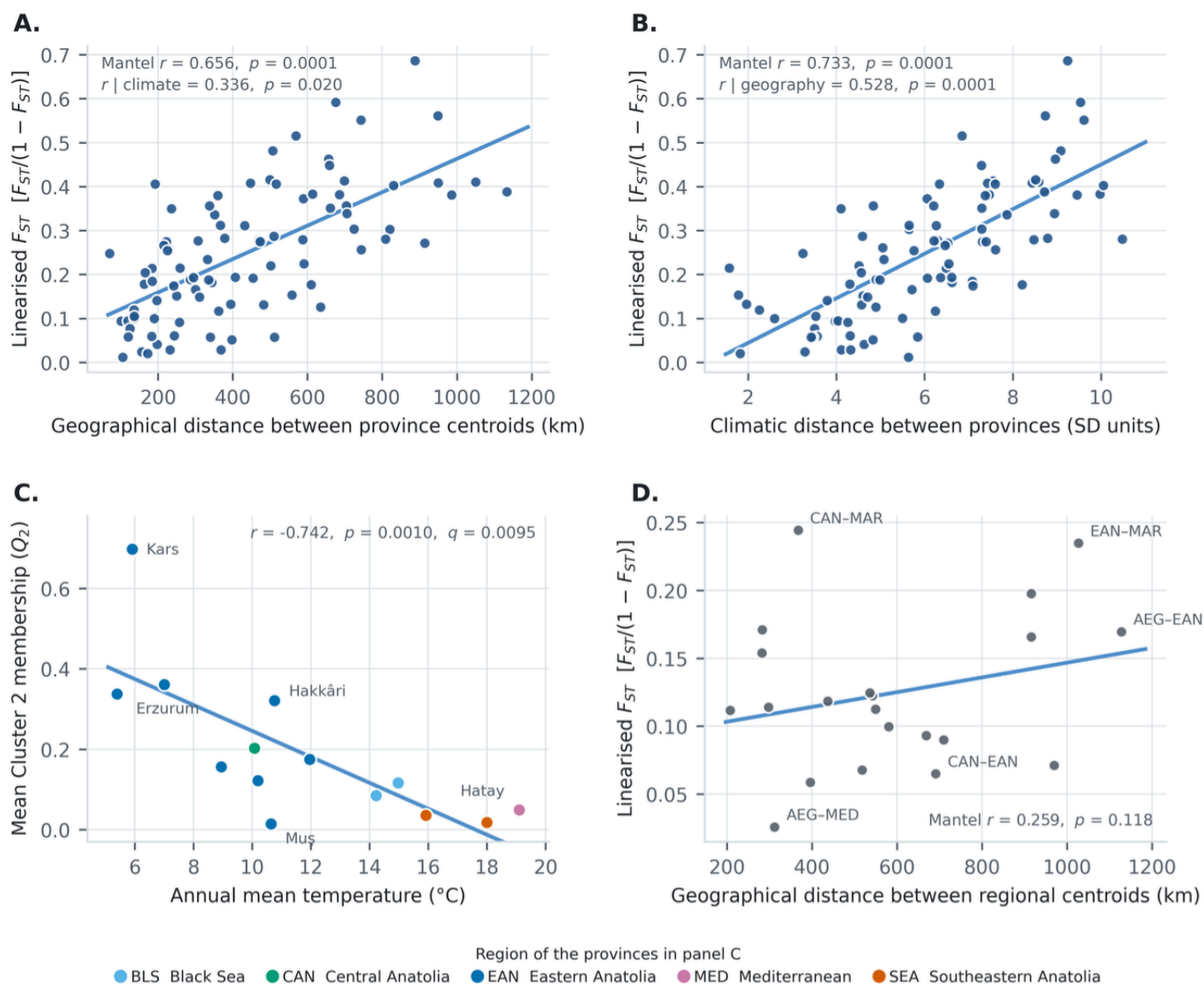


Figure 5

Genetic differentiation against geography and climate. (A) Linearized F_{ST} between the 14 provinces represented by at least ten accessions, against Haversine distance between province centroids, for the 91 province pairs. (B) The same differentiation against climatic distance, the Euclidean distance over the 19 bioclimatic variables standardized across these provinces. (C) Mean Cluster 2 membership per province against provincial annual mean temperature; points are colored by the region containing the province and the five provinces discussed in the text are labelled. (D) The corresponding test at the coarser regional scale, linearized F_{ST} between the seven regions against distance between regional centroids, for the 21 region pairs; the five pairs discussed in the text are labelled. Fitted lines are ordinary least squares and are shown to indicate direction; the statistics quoted are Mantel, partial Mantel or Pearson coefficients with permutation probabilities as stated.

Supplementary Files

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

- [TableS3regionalFstNeiD.csv](#)
- [TableS2deduplicationsensitivity.csv](#)
- [TableS7provincebioclimaticvariables.csv](#)

- [TableS6redundancythresholdsensitivity.csv](#)
- [TableS5bioclimaticscreen.csv](#)
- [TableS4provincepairwise.csv](#)
- [FigureS1evannoDeltaK.tiff](#)
- [TableS1accessionleveloutput.csv](#)
- [FigureS2SNPdistribution.tiff](#)
- [SupplementaryMaterial.docx](#)