DR MATTEO DELL'ACQUA (Orcid ID : 0000-0001-5703-8382)

Article type : Original Article

Molecular Diversity and Landscape Genomics of the Crop Wild Relative *Triticum urartu* Across the Fertile Crescent

Brunazzi Alice[1], Scaglione Davide[2], Talini Rebecca Fiorella[1], Mara Miculan[1], Magni Federica[2], Poland Jesse[3], Pè Mario Enrico[1], Brandolini Andrea[4], Dell'Acqua Matteo[1*]

[1] Institute of Life Sciences, Scuola Superiore Sant'Anna, Pisa, Italy
[2] Institute of Applied Genomics, Udine, Italy

[3] Wheat Genetics Resource Center, Department of Plant Pathology, Kansas State University, Manhattan, KS, USA

[4] Consiglio per la ricerca e la sperimentazione in agricoltura e l'analisi dell'economia agraria (CREA), Italy

Author for correspondence: Matteo Dell'Acqua; Tel: +39 3395496087; Email: m.dellacqua@santannapisa.it

Running Title: Diversity and adaptation in *T. urartu*

Keywords: Wheat, Adaptation, Landscape Genetics, GIS, GWAS, GBS, RAD, *Triticum urartu*, Triticum *dicoccoides*, wild emmer

Summary

Modern plant breeding can benefit from the allelic variation existing in natural populations of crop wild relatives that evolved under natural selection in varying pedoclimatic conditions. In this study, next-generation sequencing was used to generate 1.3 million genome-wide SNPs on *ex situ* collections of *Triticum urartu* L., the wild donor of the A^u sub-genome of modern wheat. A set of 75,511 high quality SNPs were retained to describe 298 *T. urartu* accessions collected throughout the Fertile Crescent. *Triticum urartu* showed a complex pattern of genetic diversity, with two main genetic groups distributed sequentially from West to East. The incorporation of geographic information of sampling points showed that genetic diversity did correlate to geographic distance (R^2 = 0.19), separating samples from Jordan and Lebanon, to samples from Syria and Southern Turkey, to samples from Eastern Turkey, Iran, and Iraq. The wild emmer genome was used to derive SNPs physical positions on the 7 chromosomes of the A^u sub-genome, allowing to describe a relatively slow linkage disequilibrium decay in the collection. Outlier loci were described on the basis of geographical distribution of the *T. urartu* accessions, identifying a hotspot of directional selection on chromosome 4A. Bioclimatic variation was derived from grid data and put in relation to allelic variation with a genome-wide association approach, identifying several marker-environment associations (MEAs). Fifty-seven MEAs were associated with altitude and temperature measures, while 358 were associated with rainfall measures. The most significant MEAs and outlier loci were used to identify genomic loci with adaptive potential, some already reported in wheat, including dormancy and frost resistance loci. We advocate the application of genomics and landscape genomics on *ex situ* collections of crop wild relatives to efficiently identify promising alleles and genetic materials to be incorporated in modern crop breeding.

Introduction

The adaptation of agriculture to climate change is among the most urgent challenges of our times. Food security in the coming years requires that the crops feeding humanity will be able to thrive in new climates (Lipper *et al.*, 2014). Since the last century, breeding efforts have been focused on the production of *elite* cultivars incorporating a combination of desirable traits, most notably high productivity. The production and extensive diffusion of these cultivars in much of the world's fields, however, may contribute to the erosion of genetic diversity (Jarvis et al., 2008) and to the consequent loss of resilience towards new abiotic (Abberton *et al.*, 2016) and biotic stresses (Saintenac *et al.*, 2013; Bebber *et al.*, 2013). Crop wild relatives (CWR), having diffused to diverse environments and adapted locally under natural selection (Vavilov and Dorofeev, 1992), harbor vast genetic diversity. Their use in breeding has long been advocated to provide favorable alleles to crop cultivars (Harlan, 1976). However, the use of CWR in breeding is hampered by limited knowledge of their genetic diversity and by the challenge of combining desirable CWR alleles to elite lines background. Nowadays, the increasing availability of genomic tools bears the promise to mine wild alleles with increased efficiency, and to use this information to produce improved crops (Brozynska *et al.*, 2016). The conservation and classification of CWR is therefore a global priority (Maxted *et al.*, 2012; Dempewolf *et al.*, 2017). Although extensive *ex situ* germplasm collections exist, much remain to be done to cover CWR taxonomic and geographic diversity (Castañeda-Álvarez *et al.*, 2016), a task made more urgent by the alteration of their spatial distribution and availability due of climate change (Jarvis *et al.*, 2008).

The molecular, geographic and phenotypic characterization of existing CWR collections may provide useful information to support their use in plant breeding. Recent approaches in statistics and genomics join genotypic and bioclimatic information to identify

genomic loci responsible to environmental adaptation (Rellstab *et al.*, 2015; Rissler, 2016). These 'landscape genomics' approaches have found application in several research fields, including evolutionary studies (Sork *et al.*, 2013), screening of diversity in non-model organisms (Dell'Acqua, Fricano, *et al.*, 2014), conservation efforts (Vincent *et al.*, 2013), and epidemiology (Schwabl *et al.*, 2017). In an agronomic perspective, landscape genomics may be either applied to model species to derive detailed information on candidate genes for environmental adaptation (Dell'Acqua, Zuccolo, *et al.*, 2014; Mattila *et al.*, 2016) or be used on crop landraces to identify adaptation alleles readily available for breeding (Pallotta *et al.*, 2014; Lasky *et al.*, 2015; Russell *et al.*, 2016). Applying landscape genomics approaches to natural populations of CWR may provide the double advantage of reducing the gap between model and crop species while benefiting of higher allelic diversity than that available in landraces (Zhou *et al.*, 2015).

Modern wheat (*Triticum aestivum* L. and *Triticum durum* Desf.) is markedly less diverse than its ancestors. A series of demographic and selective bottlenecks occurring since the initial domestication of wild emmer reduced wheat allelic diversity (Haudry *et al.*, 2007). During the second half of the 20th century, breeding focused on *elite* germplasm, further narrowing variation and increasing field uniformity (Cox *et al.*, 1986). This trend is currently slowing down, and possibly reverting, also thanks to the use of landraces and CWR to mine alleles of breeding relevance (Reif *et al.*, 2005). Wheat landraces and CWR are indeed strategic reservoirs of allelic diversity (Reynolds *et al.*, 2007; Hairat and Khurana, 2015; Mengistu *et al.*, 2016), whose breeding value may be unlocked by genomic and landscape genomics approaches leveraging environmental adaptation developed during evolutionary times. *Triticum urartu* L. (2n=2x=14; genome A^uA^u) is the donor of the A sub-genome to wild and cultivated tetraploid (2n=2x=28; genome AABB) and hexaploid wheat (2n=2x=42; genome AABBDD). Unlike its sister species *Triticum monococcum* (2n=2x=14; genome

A^mA^m), *T. urartu* was never domesticated and is still broadly distributed across the Fertile Crescent, where it contributed to originate the first wild forms subsequently domesticated (Özkan *et al.*, 2002). Having evolved under natural selection, *T. urartu* populations may have accumulated alleles providing adaptation to local conditions. Resistance genes have been already mapped in *T. urartu* (Qiu *et al.*, 2005) and other diploid A genomes (Chhuneja *et al.*, 2008). *Triticum urartu* has been also used as a model to study gliadin alleles (Zhang *et al.*, 2015), and showed a variety of glutenin alleles of promising use in wheat breeding (Cuesta *et al.*, 2015). Recent studies described the genetic diversity of natural populations of *T. urartu* also in relation to agronomic and quality traits, but were limited by the use of few dozen microsatellite markers (Wang *et al.*, 2017). The availability of a draft genome sequence for *T. urartu* (Ling *et al.*, 2013) and of the high quality genome sequence of related species such as wild emmer (2n=2x=28; genome AABB) (Avni *et al.*, 2017), discloses the possibility to extensively characterize the genetic diversity of *T. urartu*, propelling its incorporation in wheat breeding pipelines. Alleles from *T. urartu* may be then transferred to cultivated wheat either by biotechnology approaches, by amphiploids production (Ahmed *et al.*, 2014), or by direct hybridization with polyploid (Qiu *et al.*, 2005) and diploid wheat (Fricano *et al.*, 2014). Recently, *T. urartu* alleles were expressed in cultivated wheat to complement their homeologs, providing enhanced functionality (Gao *et al.*, 2017). Genome editing approaches in wheat (Zhang *et al.*, 2016) bears the promise to accelerate the use of *T. urartu* and other CWR variation in cultivated wheat .

In this study, we report the characterization of the most complete *ex situ* collection of wild *T. urartu* accessions currently available, spanning the entire Fertile Crescent. The use of restriction-site associated DNA (RAD) markers allowed the description of the genome-wide molecular variation among *T. urartu* natural populations. Markers were projected onto the wild emmer genome sequence to allow incorporating their positional information in linkage

and association analyses. Landscape genomics approaches making use of climatic data at sampling points provided insights into the genomic signatures of environmental adaptation, leading to the discovery of several loci whose allele frequencies are related to the spatial and climatic distribution of this species.

Results

Genotyping

The RAD sequencing of 298 *T. urartu* accessions collected for this study produced more than 2 billion reads after de-multiplexing (Table S1; Raw sequencing data can be retrieved at [European Nucleotide Archive]). After removing reads without the expected cut site downstream the barcode (0.46%), the average number of reads per sample was 5,332,961 (σ = 1,703,164). Considering the whole set of accessions, the maximum number of reads retrieved was more than 9.5 million, the minimum 4,945 (Table S1). Reads were projected on the *T. urartu* draft reference genome (Ling *et al.*, 2013), obtaining a median alignment rate of 95.8%. Aligned reads were used to call variants and yielded 1,300,216 genome-wide polymorphic sites. The list of variants was restricted to biallelic single-nucleotide polymorphisms (SNPs) in haplotypes with a maximum length of six, with minor allele frequency (MAF) above 5%, retaining 75,511 high-quality markers for downstream analyses. The high-quality marker set was distributed on 21,501 scaffolds, the most diverse containing 48 SNPs, with 3.5 SNPs per scaffold on average.

All variants called on the *T. urartu* genome were projected onto the genome of wild emmer wheat (*T. turgidum* spp. *dicoccoides*), tetraploid CWR bearing the A and B wheat sub-genomes. Among all RAD markers, 1,296,925 (99.7%) were mapped on the wild emmer genome, and 713,300 (54.9%) were uniquely aligned. Of the uniquely mapped markers, 700,949 (98.3%) were placed on the A sub-genome, leaving only 1.2% of the markers

(8,353) mapping on the B sub-genome. The remaining 0.6% of uniquely mapped markers (3,998) were placed on the *unknown* chromosome (Chr) of the wild emmer genome assembly. Among the subset of 75,511 high-quality *T. urartu* markers, 56,728 (75.1%) had a position on the A genome of wild emmer and were used for map-based analysis. Hereafter, when reporting Chr numbers we refer to the wild emmer chromsomes.

Geographic characterization of the collection

The analyses were conducted keeping track of the geographic origin of the *T. urartu* accessions via GPS coordinates of sampling points. Twenty-five samples not having GPS coordinates could be traced to approximate sampling positions using gazetteer notations, while 49 did not have any geographic information associated and could not be traced to sampling areas. A sampling order was reconstructed via multidimensional scaling (MDS) (Fig 1). The sampled area covers a broad region across the Fertile Crescent, spanning across Jordan, Lebanon, Syria, Turkey, Armenia, Iraq and Iran. The extreme sampling points West to East were more than 1,400 Km apart. Likewise, Northernmost accession were sampled more than 700 Km away from most Southern samples. The altitude of sampling points ranged from 45 to 2,419 meters above sea level, and all BioClim variables showed broad variation across the sampling points (Fig. S1). A principal component analysis (BIO-PCA) performed on climatic variation across the collection revealed a relevant structuration (Fig. 2a). When the original variables were correlated with the BIO-PC 1 to 3, the contribution of each BioClim variables to this structure became apparent (Fig. 2b). BIO-PC1 accounted for 52.2% of the original variance, and was positively correlated with temperature indexes and negatively correlated with altitude and, to a lesser extent, with rainfall indexes. Western samples and Eastern samples are both distributed across a broad range of altitudes and temperatures (Fig. 2a). BIO-PC2 (26.9% of the bioclimatic variance), orthogonal to altitude,

was faintly correlated with temperature and negatively correlated with precipitations of the wettest month (Bio 13), and the wettest (Bio 16) and coldest (Bio 19) quarter. Precipitation seasonality was also negatively related with BIO-PC2. This gradient separates the Western and the Eastern part of the sampling area (Fig. 2a). BIO-PC3, although explaining only 12.2% of the original variance, was the sole BIO-PC highly correlated with isothermality (Bio 3). Altogether, BIO-PC 1 to 3 accounted for 91.3% of the bioclimatic diversity reported by altitude and the 19 BioClim indexes.

Diversity analyses

The unrooted phylogenetic tree deriving from the set of high-quality SNPs shows three main clades (Fig 2a). Samples coming from the Eastern portion of the Fertile Crescent grouped together in a loose clade. Samples from the opposite end of the collection, mainly coming from Lebanon and Jordan, grouped in a separate clade. A number of samples with no clear geographic patterning projected out of this group in a monophyletic clade. A clustering analysis performed on molecular data confirmed these samples as a separate group (Fig. 3a). This group included the outgroups *T. monococcum* and *T. boeticum*, and accessions having an intermediate phylogenetic relationship between *T. urartu* and outgroups. Outgroups were removed from further analyses, and a PCA was used to depict the genetic relatedness of *T. urartu* accessions (Fig. 3b). Samples from the Western and Southern sampling grouped separately according to PC1, accounting for 9.17% of the molecular variation. Low PC loadings confirmed a limited population structure beyond the main geographic separation across the East-West gradient: 69 PC are needed to reach 50% of the variation originally present in the molecular dataset. Samples without a GPS location group in the vicinity of mapped samples in the bottom left corner of the PCA, suggesting a proximate, although unrecorded, geographic origin (Fig. 3b).

A Bayesian analysis of the cryptic genetic clusters supported the outcome of the phylogenetic analyses (Fig. 4). When considering the entire dataset, the most probable number of clusters (K) was two (Fig. S2), separating a West group from an East group (Fig. 4a). When exploring deeper structures existing within the two main groups, the most probable K for Western collections was either three or four (Fig. S2). In both cases, samples with contrasting cluster assignment fell in the intermediate sampling areas. Groups of samples with low genetic admixture were located at the extremes of the geographic distribution (Fig. 4b). The most probable K for the Eastern sample group were either three or six (Fig. 4c; Fig. S2). One main genetic cluster characterized intermediate samples with both K interpretations. With the lowest K, a geographic structure of genetic diversity was still visible. Interestingly, samples furthest apart shared the same genetic membership (K1). When setting K to six, at least four linearly ordered groups of samples emerged among those having GPS coordinates (Fig. 4c).

Linkage disequilibrium (LD) was studied deriving marker positions form the wild emmer genome sequence. Most of the chromosomes showed higher centromeric LD, even though regions of localized higher LD were visible in telomeric regions (Fig. S3). On all chromosomes, absolute values of LD were relatively low, but the rate of LD decay was slow. The LD decay, measured as halving distance of mean LD, was spanning from 11.2 Mb on Chr 1A to 66 Mb on Chr 2A (Fig. S4). Half the amount of initial LD was on all chromosomes close to $r^2 = 0.2$, a value generally considered null LD.

Landscape genomics

To focus on the relation between genetic diversity and geographic diversity, 239 accessions (80% of the initial collection) with geographic information were grouped in 19 demes according to their geographical distance, and were used to compute population genetics

indexes. The genetic distance was related to geographic distance, with an R^2 of 0.19 ($p << 0.001$) (Fig. 5a). Indeed, a spatial principal component analysis (sPCA) reported that a global structure of allelic frequencies was predominant over local structures (Fig. 5b). Three main clusters of allele frequencies were linearly distributed from West to East across the sampling area (Fig. 5b). The spatial-genetic clusters identified by the sPCA separated samples from Jordan and Lebanon, samples from Northern Syria and neighboring Turkey, and samples from Eastern Turkey, Iraq and Iran (Fig. 5c).

Genomic loci under putative selection were detected joining SNPs information with the geographic distribution of accessions sorted on a Gabriel Graph (Fig. 5b). The power spectrum deriving from the MSOD method, marker-specific, was used to detect SNPs significantly deviating from the average power spectrum, defined outlier loci (Table S2). The 100 most extreme outlier loci were all significant at the $p < 0.005$ level (Fig. S5) and when projected onto the wild emmer sequence were scattered across the 7 chromosomes of the A sub-genome (Fig. S5). On Chr 1A, only two markers featured in the most significant associations, at 419 Mb and 578 Mb. On Chr 2A, two outliers were detected 5 Mb apart around 135 Mb and two additional outliers at around 498 and 528 Mb. Chr 3A featured 25 outliers clustering in five loci at approximately 60, 280, 470, 610, and 650 Mb. Chr 4A had the highest number of extreme outliers, 36 in total, spread from 163 to 608 Mb. Chr 5A reported seven outliers, three of which clustering around 350 Mb. Twelve more outliers were found in three clusters on Chr 6A, centering at approximately 30, 260, and 590 Mb. Chr 7A featured 13 outliers of which three grouped at 572 Mb and the others interspersed along the chromosome. The remaining outlier of the 100 most significant ones mapped on Chr *unknown*.

A genome-wide association (GWA) scan was performed to detect marker-environment associations (MEA) using SNPs and the first three BIO-PC derived from BioClim data. Since the association statistics showed some inflation (Fig. S6), a stringent significance threshold based on multiple test correction was used. Altogether, the GWA analysis on the three BIO-PC reported 535 MEAs reaching the suggestive threshold of FDR 10^{-6}, of which 63 surpassed the high significance threshold of FDR 10^{-8} (Table S3). The MEAs identified by the GWA scans were scattered on 407 different scaffolds of the *T. urartu* genome assembly, but when were projected on the wild emmer chromosomes they reported a limited number of clear significance peaks of multiple MEAs in mutual LD and close genomic position (Fig. 6).

BIO-PC1, the most important component of climatic variation, reported 57 significant associations (Table S3). This environmental measure is positively correlated with temperature measures and negatively correlated with altitude and rainfall measures (Fig. 2). Three MEAs were detected on Chr 1A at 40, 50, and 533 Mb. One MEA was detected on Chr 2A at 134 Mb, and seven MEAs clustered between 406 and 463 Mb. These MEAs were originally mapped on seven unordered scaffolds on the *T. urartu* genome assembly. Two MEA appeared on Chr 3A at 703 and 750 Mb. Several MEAs mapped on Chr 4A, five individually mapping at 116, 178, 361, and 415 Mb, and 415, and 17 clustering from 487 to 506 Mb. On Chr 5A, one MEA mapped at 111 Mb and five MEAs co-mapped between 639 and 641 Mb. Of these, one surpassed the high significance threshold (Fig. 6). Seven MEAs mapped in scattered positions on Chr 6A, and ten MEAs were found between 504 and 586 Mb.

BIO-PC2 reported 358 MEAs, 23 of which surpassing the most stringent threshold (Table S3; Fig. 6). This variable is mainly contributed by rainfall measures (Fig. 2). Among highly significant MEAs, five appeared on Chr 1A at 350 and 527 Mb. On Chr 2A, three

highly significant MEAs clustered at around 635 Mb, while a single MEA appeared at 678 Mb on Chr 3A. Three highly significant MEAs were located on a clear association peak on Chr 4A at around 503 Mb (Fig. 6). This peak is contributed by 32 MEAs surpassing the suggestive threshold from 472 to 515 Mb, previously mapping on 24 unordered scaffolds on *T. urartu* sequence (Table S3) and overlapping a signal from BIO-PC1. Three highly significant MEAs were found in close vicinity on Chr 5A at 80, 105, 134 Mb, and an additional one mapped at 661 Mb. On Chr 7A, highly significant MEAs individual mapped at 123, 151, 234, and 270 Mb, and three clustered from 508 to 526 Mb.

The GWA on BIO-PC3, the least significant bioclimatic component mainly accounting for temperature seasonality (Fig. 2), generally provided lower significance and higher background noise (Fig. 6). The GWA scan in this case identified 149 MEAs, mostly scattered across the genome, one surpassing the high significance threshold at 204 Mb on Chr 7A (Table S3). Notable significance peaks appeared on Chr 2A and 3A, where multiple significant MEAs mapped around 690 Mb and 150 Mb, respectively (Fig. 6).

The outliers and most significant MEAs were used to identify candidate genes considering the chromosome-specific LD halving distance as confidence interval. Outlier loci targeted 1,680 unique gene models, 91 of which identified by more than one outlier locus (Table S4). The high significance MEAs identified a comparable number of 1,418 unique genes, often targeted by multiple associations (Table S5). Among those, 642 were identified by between 2 and 10 MEAs, and 69 were identified by more than 10 MEAs each. Approximately one third of the candidate genes (530) was jointly targeted by outlier loci and MEAs.

Discussion

The high number of SNPs showed elevated genetic diversity within the *T. urartu* collection. By reducing the set of molecular markers to those deriving from reads having haplotypes shorter than six SNPs some information was lost, but the reliability of the retained SNPs increased. The rationale of removing long haplotypes is that the occurrence of multiple SNPs in *cis* in relatively short RAD reads (110 bp) may derive from the misalignment of such reads in multiple, repeated regions of the *T. urartu* genome. The slow LD decay in *T. urartu* (Fig. S4), similarly to that of wheat (Crossa *et al.*, 2007) and related species (Sela *et al.*, 2014), allows to represent most haplotype blocks without the need of an exceedingly dense genotyping. Since at the time of writing the genome sequence of *T. urartu* was a draft arranged in scaffolds with N50 only slightly above 60 kb (Ling *et al.*, 2013), we used the high quality genome assembly of wild emmer (*Triticum turgidum* spp. *dicoccoides*) (Avni *et al.*, 2017) to derive chromosomal position of SNPs. Wild emmer originated by the hybridization between *T. urartu* and the B genome ancestor, a close relative of *Aegilops speltoides*, some 500,000 years ago (Peng *et al.*, 2011). Since the close phylogenetic relationship, it may be expected high sequence homology and collinearity of the A sub-genomes in the two species. This is confirmed by the high specificity of *T. urartu* sequences to the A sub-genome of wild emmer, with only 1.2% of the SNPs univocally matching the B sub-genome. The high confidence mapping of *T. urartu* sequences on the B sub-genome may be contributed by intergenomic invasions already observed in wild emmer, where they may contribute to stabilize allopolyploidy (Nevo, 2014). As expected, several *T. urartu* markers (45.1%) could not be mapped univocally on the *T. turgidum* spp. *dicoccoides* genome. This is due to the stringent filters employed in the alignment procedure, and to the several polymorphisms present in some of the *T. urartu* scaffolds. This figure was drastically reduced (24.9%) when we focused on the subset of high-quality markers later used for the analyses, confirming the

goodness of the quality filtering. Once an improved version of the *T. urartu* genome will be available, it will be possible to assign these markers to one of the seven *T. urartu* linkage group, further improving the characterization of the genomic landscape of *T. urartu* diversity.

The SNP markers revealed a complex pattern of genetic diversity across the Fertile Crescent. Accessions are highly differentiated and poorly structured, as indicated by the long edges and deep relations in the phylogenetic tree (Fig. 3a). The outgroup samples may be contributed by *ex situ* erroneous taxonomic assignment or by hybridization events. *Triticum urartu* can cross with *T. monococcum* producing fertile progeny (Baum and Bailey, 2013; Fricano *et al.*, 2014; Nasernakhaei *et al.*, 2015) and the diversity of these samples may indeed report such occurrence in our collection. Previous studies considering smaller collections and using less advanced molecular markers already reported high variability among *T. urartu* natural accessions (Castagna *et al.*, 1997; Mizumoto *et al.*, 2002; Wang *et al.*, 2017), however this is the first time that a genomic approach is used to characterize a collection representative of the whole geographic distribution of *T. urartu*. The extensive, genome-wide molecular characterization of *ex situ* collections of CWR supports their potential employment in crop breeding (Henry, 2014; Brozynska *et al.*, 2016). When merged with the geographic characterization of the accessions' sampling points, this information may improve the efficiency in selecting CWR germplasm to be prioritized in breeding schemes (Jones *et al.*, 2013). In our *T. urartu* collection, the cryptic genetic clusters (Fig. 4) overlapped the grouping emerged from the phylogenetic analysis (Fig. 3), and clearly separated Western from Eastern accessions. Five accessions (1.6%) showed unexpected genetic clustering based upon their sampling locations (Fig. 4). A parsimonious interpretation makes us speculate that such outliers could be due to human error rather than to complex evolutionary forces into play. Indeed, while we carefully checked both seed lots and wet lab practices, we cannot completely rule out a contamination either at the genebank or at the DNA level. This limited

number of outliers, however, may also represent *T. urartu* lineages migrated to the collection area from elsewhere. Indeed, the collection here analyzed does not derive from a continuous transect, but rather represents a sparse sampling of natural accessions, and it cannot provide a full representation of the geographic structuration of *T. urartu* diversity. New sampling campaigns are required to fill gaps in the current *ex situ* collections of *T. urartu*, even though the precarious political situation in the Fertile Crescent hampers similar efforts at the time of writing.

The pattern of genetic diversity is consistent with an isolation by distance model (Wright, 1943), in which the geographically distant populations tend to be more genetically different than the nearby ones (Fig. 5). The sampling area is twice as long from the East to the West than from the North to the South, hence the geographic separation across longitudes may be more evident. However, allelic frequencies are also separated on a latitudinal gradient, as reported by the sPCA analysis based on individual GPS coordinates (Fig. 5). These partitions may be contributed by geographic segregation as well as by climatic specificities of sampling locations. *Triticum urartu* is an autogamous species, with infrequent cross-pollinations. The autogamy of this species is reflected in the slow LD decay that was observed when anchoring markers on the wild emmer genome sequence (Fig S5). Our collection features haplotypes spanning tens of Mb, suggesting relatively rare recombination events. Localized regions of high LD (Fig. S3) may be due to genomic regions with suppressed recombination similarly to what observed in modern wheat (Darrier *et al.*, 2017), and are likely contributed by some degree of approximation introduced by the cross-mapping of markers using the wild emmer sequence. Once a high quality genome sequence of *T. urartu* will be available, these SNPs may better characterize LD features of this collection. The recombination landscape of *T. urartu* is very relevant in relation to its possible use in

wheat breeding, and could be further studied with the production of *ad hoc* segregant populations.

The dispersal strategy of *T. urartu* is focused on efficient seed germination in the vicinity of the mother plant rather than on long distance hauling of seeds (Elbaum *et al.*, 2007). The similarity found across distant populations (Fig. 4, Fig. 5c), however, suggests a leveling role of gene flow. Selection overlaps the geographic separation in counteracting the homogenizing effect of gene flow, reducing allelic diversity at loci that improve fitness in specific climatic conditions (Garant *et al.*, 2007). The sampling scheme underlying our *T. urartu* collection is not optimal for outlier detection approaches, as demes of individuals are not uniform in spatial distribution and membership (Lotterhos and Whitlock, 2015). In our case, the choice of the maximum distance to group individuals in demes depended on the uneven coverage of the sampling in the region. A denser sampling would allow to compute diversity indexes with higher confidence. For this reason, when studying outlier loci, we decided to employ the MSOD outlier detection method (Wagner *et al.*, 2017). This approach has advantage of explicitly dealing with spatial relations among individuals in a graph form, thus relying on individual-based information.

In this work, we aimed at describing outstanding allele frequencies in relation to the spatial distribution of the *ex situ* collection. Using individual sampling positions rather than on geographic or genetic clustering of individuals allows us to put in relation outlier loci with bioclimatic variation at the accession level. Both the MSOD and GWA analyses reported several significant markers at the lower significance threshold (Table S2 and Table S3). In both analyses, we decided to focus on a subset of highly significant markers, so to reduce Type I errors while discussing loci potentially involved in environmental adaptation. It is likely that many more loci are indeed under selection, and that the stringency used did not allow us to report them. Still, the highly significant outlier loci and MEAs reported several

notable genomic regions and candidate genes, often overlapping (Table S4 and Table S5). Although there is no univocal correspondence between the cM position of the many wheat QTL reported in literature with the Mb positions derived in this study, it is possible to speculate about the occurrence of shared molecular mechanisms on the basis of the approximate chromosomal positions of signals.

A notable concentration of outlier markers meeting the significance criteria are visible on Chr 4A (Fig. S5), spanning for several Mb in the central part of the chromosome. This position is compatible to that of major peaks observed in the association analysis with BIO-PC1 and BIO-PC2 on Chr 4A at around 500 Mb (Fig. 6). Several studies reported the presence of a major seed dormancy locus, *Phs1*, on the long arm of Chr 4A of modern wheat (Torada *et al.*, 2008; Torada *et al.*, 2016). Previous studies on a wild emmer x durum wheat population identified on the homeologous Chr 4B a major QTL influencing grain size and spikelet germination uniformity (Nave *et al.*, 2016). This QTL maps in a pericentromeric position compatible to our peak and was likely fixed during wheat domestication. It is likely that in the *T. urartu* collection the allele at this gene is associated to an environmental gradient related with altitude and temperature (BIO-PC1), as well as with rainfall regimes (BIO-PC2) (Fig. 2). The role of seed dormancy in adaptation is well known (Vidigal *et al.*, 2016), and may contribute in dampening environmental variability on fitness and dispersal (Venable and Brown, 1988). Linkage drag deriving form selection at this locus may have originated the several outlier signals across the pericentromeric region of Chr 4A, whose allele frequencies are influenced by the direct selection exerted on the locus. Several other outstanding candidates from MSOD and GWA analyses may be put in relation with previous literature on wheat. The highly significant MEAs identified on Chr 5A by the BIO-PC1 scan may correspond to the vernalization and frost resistance QTL identified in the distal portion of Chr 5AL in polyploid (Galiba *et al.*, 1995; Zhu *et al.*, 2014) and diploid (Vágújfalvi *et al.*,

2003) wheat. Indeed, BIO-PC1 accounts for most of the temperature variance across the sampling points (Fig. 2). This locus is not reported by outlier loci analysis; it is not always the case that geographic segregation of alleles overlaps the distribution of environmental measures. The distal signal emerging on Chr 1A in outlier loci analysis (Fig. S5) is close to highly significant MEAs reported by BIO-PC2 (Fig. 6), a measure mostly accounting for seasonality of rainfall (Fig. 2 and Table S8). Previous literature reported in this position a thermo-sensitive locus for earliness in diploid wheat (Bullrich *et al.*, 2002), that may be put in relation with climatic conditions at sampling points. The highly significant MEA detected by BIO-PC3 on Chr 7A (Fig. 6 and Table S5) may be related to several agronomic QTL for phenology and productivity (Gahlaut *et al.*, 2017) as well as for meta-QTL for drought and heat stress (Acuna *et al.*, 2015).

At present, the wild emmer gene models detected in the vicinity of association signals may relate to a multitude of molecular functions, including transporters and transcription factors (Table S4 and Table S5). The predicted availability of a high quality reference sequence for *T. urartu* will increase the discrimination power of these analyses, and, together with the production of high quality sequences for durum and bread wheat, it will contribute to the genomic revolution in wheat breeding. The further characterization of the significant signals falls beyond the scopes of this study. The methods here employed aim at a synthetic description of CWR diversity and adaptation potential. The loci hereby reported may be validated by targeted re-sampling and even by phenotypic characterizations of the collection. We can anticipate that we are developing a multiparental population following a nested association mapping (NAM) crossing scheme (McMullen *et al.*, 2009), putting together the diversity of a selected subset of the accessions into an interlinked segregating population. Once the NAM population will be completed, it will represent a useful resource to push forward discoveries made on *T. urartu.*

In this study, we have shown that *T. urartu* is highly diverse, and that the study of its natural populations might provide important information on genomic loci involved in environmental adaptation. Leveraging these modern genomic approaches, *T. urartu* could play again a key role in producing better wheats, even some 500,000 years after originally hybridizing with the B genome of modern wheat.

Experimental procedures

Plant materials and DNA extraction

The plant materials used in this study (Table S6) were 298 accessions of *Triticum urartu* L. in the U.S. Department of Agriculture (USDA) National Plant Germplasm System, USA, and the Consiglio per la Ricerca e la Sperimentazione in Agricoltura e l'Analisi dell'Economia Agraria (CREA), Italy. The collection assembled for this study represents the entirety of *T. urartu* accessions available from *ex situ* germplasm banks at the time of the experiment. One accessions of *Triticum monococcum* L. (var. MONLIS) and two accessions of *Triticum boeoticum* L. (ID 1094 and ID 948 from CREA genebank) were included as outgroups. Five seeds per accession were germinated in individual petri dishes, and green tissues were pooled and used to extract genomic DNA with a GeneElute Plant Genomic DNA Miniprep Kit (Sigma-Aldrich, St Louis, MO) following manufacturer instructions. DNA was checked for quality and quantity using agarose gels and spectrophotometry.

Genotyping

Genomic DNA was shipped to IGA Technology Services (Udine, Italy) to perform genotyping using a custom double-digestion RAD sequencing (Baird *et al.*, 2008) protocol. To define the enzymes to be used in the genomic DNA digestion and size selection, restriction simulations were carried on the reference genome, as available from

EnsemblPlants (GCA_000347455.1.26 build), using custom scripts. The combination of *Sph*I and *Bst*YI enzymes together with a size selection of fragments in the range of 230-330 bp predicted to generate loci in the order of 100,000. For each sample 250 ng of genomic DNA were digested in a 30 µL reaction with 2U of each *Sph*I and *Bst*YI enzymes (New England Biolabs) in SmartCut buffer for 1 h at 37 °C, followed by 1 h at °60 C and heat inactivation at 65 °C for 15 min. One and a half volumes of AmpureXP beads (Agencourt) were added to the reaction mix and put on a magnetic rack. Beads pellet was washed twice with 70% ethanol and DNA was re-suspended in 20 µL of Tris-HCl 10 mM (pH 8.5). For each sample 10 µL of restriction product were mixed with 2 and 5 pmoles of adapters P1 and P2 respectively (Table S; variable length inline barcodes on both sides) and 200 U of T4 DNA ligase (New England Biolabs) in a final reaction volume of 30 µL and incubated for 1 h at 23 °C and 1 h at 20 °C. Purification was done as described above. Samples were pooled in 24-plex by means of P1 inline barcodes and concentrated using a SpeedVac centrifuge. 400 ng of ligated DNA were loaded on 1X low-melting agarose gel. For each pool a gel band in the range of 300-400 bp was cut and purified in a QIAquick column (QIAGEN). Recovered DNA was amplified in the following PCR reaction: 3 min at 95 °C, 8 cycles at 95 °C (30 sec) – 60 °C (30 sec) 72 °C (45 sec), 2 min at 72 °C using custom primers (Table S7) to incorporate flowcell hybridization sequences with inter-pool barcodes (Illumina i7 index). After purification, libraries were validated on Agilent Bioanalyzer 2100. Sequencing was performed on an Illumina HiSeq2500 platform with 125 bp paired reads.

Bioinformatics analysis and SNP calling

A first de-multiplexing step - to divide pools by means of Illumina i7 index - was carried using the CASAVA software 1.8.2 (Illumina). Each pool (pair of "fastq" files) was then de-multiplexed at sample level by means of inline barcodes using the stacks package v1.08

(Catchen *et al.*, 2011) with a maximum of 1 bp mismatch per inline barcode. After removal of the leading barcode sequences, all reads were trimmed to the first 110 bp by removing 3'-ends. Alignments to the *T. urartu* draft genome (Ling *et al.*, 2013), assembly GCA_000347455.1.26, were performed using Bowtie2 (Langmead *et al.*, 2009) and filtered for a minimum mapping quality of 10. Aligned reads were processed with the stacks pipeline, including the following steps: *pstacks* (min. 2 reads, bounded SNP model with upper_bound=0.10 and alpha=0.05)*, cstacks, ssstacks and rx_stacks*. In the latter, a minimum average likelihood threshold of -15 was imposed to filter low quality sites. The *populations* module was used to generate the genotype matrix, requiring a minimum of two reads supporting the genotype with a minimum individual likelihood of -15. Polymorphic sites were retained only when calling requirements were met for at least 75% of the samples. A working set of SNPs was obtained after further filtering to reduce the number of molecular markers but increase their reliability. Haplotypes longer than 6 SNPs were discarded as possibly contributed by mis-alignment of sequencing reads. Markers with a MAF lower that 5% were also removed from the dataset. A reduced set of markers to be used for genome-wide surveys of diversity was obtained by random sampling of 20,000 SNPs from the working set. Data management and filtering were performed in R 3.3.2 (R Core Team, 2013).

Physical map of markers

Molecular markers developed on the *T. urartu* panel were ordered using the recently published reference sequence of the wild emmer (*T. turgidum* spp. *dicoccoides*) genome (Avni *et al.*, 2017), to which *T. urartu* contributed the A sub-genome. The nucleotide sequence 100bp upstream and 100bp downstream of each marker was derived from the *T. urartu* genome sequence (Ling *et al.*, 2013) by means of a Python script available upon request. Sequences were transformed into single-end reads in fastq format using the

fasta_to_fastq.pl Perl script publicly available (https://github.com/ekg/fasta-to-fastq). These synthetic reads were then mapped on the wild emmer genome. Since reads were longer than 70 bp, they were mapped by means of the *bwa mem* aligner with default parameters (Li, 2013). Reads with low mapping quality (MAPQ<10) secondary alignments, and/or reads with multiple hits on the *T. turgidum* ssp. *dicoccoides* genome were filtered out using *samtools* (Li *et al.*, 2009) applying the following command line. *samtools view -q 10 -F 4 -F 256 -F 2048*. Only markers mapping on the A sub-genome were retained.

Geographic characterization

The available passport data associated to the accessions were used to retrieve geographic information of sampling points (Table S6). The GPS coordinates of samples only having gazetteer information were manually derived at the highest precision possible from Google Maps (Google Maps, 2017). Sampling points with native or derived GPS coordinates were analyzed with a geographic information system (GIS). Bioclimatic (BioClim; Table S8) variables for the sampling area were obtained from Worldclim 30 arc-seconds data (Hijmans *et al.*, 2004) projected in QGIS 2.4 (QGIS Development Team, 2017). Altitude and 19 BioClim variables were assigned to each accession based on sampling coordinates. In order to reduce redundancy in the dataset, we reduced the environmental variables with a principal component analysis (BIO-PCA). The most significant BIO-PCs were retained for further analysis. Accessions having spatial information were organized in demes (also referred as populations) by grouping all samples collected within a 25 Km radius using R/raster (Hijmans and van Etten, 2012). Geographic position of demes was obtained from the average of the sampling coordinates of the samples they contained. Demes containing only one sample were not considered in diversity and landscape genomics analyses.

Diversity analyses

Diversity analyses were conducted on the reduced set of SNPs, considering samples regardless of their geographic origin. A consistent color-coding representing samples position across the Fertile Crescent, used in all graphical outputs, was derived from MDS of the latitude and longitude values of sampling points. A neighbor joining (NJ) phylogeny including outgroups was produced with R/adegenet (Jombart and Ahmed, 2011). Outlier samples were detected with the *find.clusters()* function in R/adegenet, and were removed from further analyses. A PCA was performed to check the structure existing within the *T. urartu* dataset, and to survey the existence of spatial segregation of genetic groups. The software Structure 2.3.4 (Pritchard *et al.*, 2000) was used with the reduced marker set to assign individuals to cryptic genetic clusters following a Bayesian procedure detecting the number of clusters best describing the data. Structure was run with standard settings (length of burn-in 10,000, and number of MCMC reps 100,000) and admixture model. The number of clusters tested was from K=1 to K=20, with 10 replications each. The method from Evanno (Evanno *et al.*, 2005) implemented in Structure Harverster (Earl and vonHoldt, 2012) was used to identify the most probable number of clusters. Once the most probable clusters were identified by the global analysis, Structure was run again with the same setting within each of the clusters.

Linkage disequilibrium (LD) was calculated among all markers having a position on the A sub-genome of wild emmer according to the parameter reported above. The R package LDheatmap (Shin *et al.*, 2006) was used to calculate pairwise r^2, a measure accounting for allele frequency at loci. A custom R script available upon request was used to join pairwise LD measures with physical distances of markers within each chromosome. LD decay was studied interpolating the Hill and Weir equation to LD measures as a function of genetic distance (Marroni *et al.*, 2011; Mengistu *et al.*, 2016), and LD halving distance for each

chromosome was recorded. For each chromosome, pairwise LD measures were averaged for markers falling within LD halving distance and plotted in a rolling window of size 100 markers to display LD evolution along chromosomes. A custom R script, available upon request, was used to conduct the analysis.

Landscape genomics

Landscape genomics analyses focused on samples having spatial information. R/adegenet was used to compute Nei's distance (Nei, 1972) among demes. A linear regression was used to study the relation between molecular and geographic distance among demes. A sPCA (Jombart *et al.*, 2008), implemented in R/adegenet, was used to characterize the pattern of allelic variation in relation to spatial data and to survey global and local structures of genotypic diversity. Maps of genetic clines were obtained interpolating the three principal sPC across the sampling area.

Putative outlier loci, *i.e.* genomic loci subjected to directional selection, were discovered sorting all georeferenced samples on a Gabriel graph and using the Moran spectral outlier detection (MSOD) method (Wagner *et al.*, 2017). The MSOD is aimed at detecting SNPs loci responding to directional selection on a geographic base, whilst accounting for the spatial structure of allelic distribution reported by the graph. The subset of the high quality SNPs with a position on the A sub-genome of wild emmer was used. Custom R script were used to produce plots and numerical outputs, and only the most extreme 100 markers are discussed.

In order to test association between markers and environmental variation, molecular diversity data was input in the R package Genome Association and Prediction associated Tools (GAPIT) (Lipka *et al.*, 2012). In this analysis, the subset of the high quality SNPs mapping on wild emmer A sub-genome was used. The GWA scan was run with a mixed

linear model on the first three BIO-PC derived from BioClim variables as phenoytpes. The GWA was run using 1 to 10 principal components derived from molecular data in fixed part of the model. A kinship matrix calculated with the VanRaden method was fitted in the random part of the model. R/GAPIT with the SUPER method (Wang *et al.*, 2014). Quantile-quantile plots were visually evaluated to determine the goodness of fit of the model with varying PCs, and to choose the best run to be discussed in the main text. Multiple test correction was performed with the R package q-value (Dabney *et al.*, n.d.) according to Storey's method (Storey, 2002). Arbitrary thresholds at 1×10^{-6} (suggestive) and 1×10^{-8} (high significance) were chosen to discuss the most relevant associations to minimize type I errors.

The wild emmer wheat annotation WEWseq_PGSB_v1 (Avni *et al.*, 2017) was used to derive genes models for outlier loci and MEAs surpassing the high significance threshold. The chromosome-specific LD halving distance was used as window size upstream and downstream each significant marker, and gene models were searched in that window with a custom R script available upon request.

Acknowledgements

This work is funded by the Doctoral Programme in Agrobiodiversity at the Scuola Superiore Sant'Anna of Pisa, Italy. We acknowledge the U.S. Department of Agriculture (USDA) National Plant Germplasm System for providing part of the collection. The Authors declare no conflict of interest.

Short Supporting Information Legends

Fig. S1. Distribution of BioClim variables across the sampling area.

Fig. S2. Evanno outcome of Structure analysis.

Fig. S3. LD evolution across chromosomes.

Fig. S4. Chromosome-specific LD decay.

Fig. S5. Outliers distribution across chromosomes.

Fig. S6. Quantile-quantile plots for association tests.

Table S1. Count of total and retained sequencing reads.

Table S2. Numerical outcome of the MSOD method for outlier detection.

Table S3. Numerical outcome of the GWA scan on BIO-PC 1 to 3.

Table S4. Candidate genes deriving from the MSOD method for outlier detection

Table S5. Candidate genes deriving from the GWA scan on BIO-PC 1 to 3.

Table S6. Accessions used in this study and relative information.

Table S7. Adapters and primers used in the generation of the molecular data.

Table S8. Meaning of BioClim variables.

References

Abberton, M., Batley, J., Bentley, A., et al. (2016) Global agricultural intensification during climate change: a role for genomics. *Plant Biotechnol. J.*, **14**, 1095–1098.

Acquaah, G. (2012) *Principles of Plant Genetics and Breeding*, John Wiley & Sons.

Ahmed, S., Bux, H., Rasheed, A., Kazi, A.G., Rauf, A., Mahmood, T. and Mujeeb-Kazi, A. (2014) Stripe rust resistance in Triticum durum-T. monococcum and T. durum-T. urartu amphiploids. *Australas. Plant Pathol. AAP*, **43**, 109.

Avni, R., Nave, M., Barad, O., et al. (2017) Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science*, **357**, 93–97.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A. and Johnson, E.A. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One*, **3**, e3376.

Baum, B.R. and Bailey, L.G. (2013) Genetic diversity in the Red wild einkorn: T. urartu Gandilyan (Poaceae: Triticeae). *Genet. Resour. Crop Evol.*, **60**, 77–87.

Bebber, D.P., Ramotowski, M.A.T. and Gurr, S.J. (2013) Crop pests and pathogens move polewards in a warming world. *Nat. Clim. Change*, **3**, 985–988.

Brozynska, M., Furtado, A. and Henry, R.J. (2016) Genomics of crop wild relatives: expanding the gene pool for crop improvement. *Plant Biotechnol. J.*, **14**, 1070–1085.

Castagna, R., Gnocchi, S., Perenzin, M. and Heun, M. (1997) Genetic variability of the wild diploid wheat Triticum urartu revealed by RFLP and RAPD markers. *Theor. Appl. Genet.*, **94**, 424–430.

Castañeda-Álvarez, N.P., Khoury, C.K., Achicanoy, H.A., et al. (2016) Global conservation priorities for crop wild relatives. *Nat. Plants*, **2**, nplants201622.

Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W. and Postlethwait, J.H. (2011) Stacks: building and genotyping loci de novo from short-read sequences. *G3 Genes Genomes Genet.*, **1**, 171–182.

Chhuneja, P., Kaur, S., Garg, T., et al. (2008) Mapping of adult plant stripe rust resistance genes in diploid A genome wheat species and their transfer to bread wheat. *Theor. Appl. Genet.*, **116**, 313–324.

Cox, T.S., Murphy, J.P. and Rodgers, D.M. (1986) Changes in genetic diversity in the red winter wheat regions of the United States. *Proc. Natl. Acad. Sci. U. S. A.*, **83**, 5583–5586.

Crossa, J., Burgueño, J., Dreisigacker, S., et al. (2007) Association Analysis of Historical Bread Wheat Germplasm Using Additive Genetic Covariance of Relatives and Population Structure. *Genetics*, **177**, 1889–1913.

Cuesta, S., Guzmán, C. and Alvarez, J.B. (2015) Molecular characterization of novel LMW-i glutenin subunit genes from Triticum urartu Thum. ex Gandil. *Theor. Appl. Genet.*, **128**, 2155–2165.

Dabney, A., Storey, J.D. and Warnes, G.R. *qvalue: Q-value estimation for false discovery rate control*, Available at: http://www.bioconductor.org/packages/release/bioc/html/qvalue.html [Accessed May 9, 2014].

Dell'Acqua, M., Zuccolo, A., Tuna, M., Gianfranceschi, L. and Pè, M.E. (2014) Targeting environmental adaptation in the monocot model Brachypodium distachyon: a multi-faceted approach. *BMC Genomics*, **15**, 801.

Dempewolf, H., Baute, G., Anderson, J., Kilian, B., Smith, C. and Guarino, L. (2017) Past and Future Use of Wild Relatives in Crop Breeding. *Crop Sci.*, **57**, 1070–1082.

Earl, D.A. and vonHoldt, B.M. (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.*, **4**, 359–361.

Elbaum, R., Zaltzman, L., Burgert, I. and Fratzl, P. (2007) The Role of Wheat Awns in the Seed Dispersal Unit. *Science*, **316**, 884–886.

Evanno, G., Regnaut, S. and Goudet, J. (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol. Ecol.*, **14**, 2611–2620.

Fricano, A., Brandolini, A., Rossini, L., et al. (2014) Crossability of *Triticum urartu* and *Triticum monococcum* Wheats, Homoeologous Recombination, and Description of a Panel of Interspecific Introgression Lines. *G3 Genes Genomes Genet.*, **4**, 1931–1941.

Gao, F., Chen, B., Jiao, J., Jia, L. and Liu, C. (2017) Two Novel Vesicle-Inducing Proteins in Plastids 1 Genes Cloned and Characterized in Triticum urartu. *PLOS ONE*, **12**, e0170439.

Garant, D., Forde, S.E. and Hendry, A.P. (2007) The multifarious effects of dispersal and gene flow on contemporary adaptation. *Funct. Ecol.*, **21**, 434–443.

Hairat, S. and Khurana, P. (2015) Evaluation of Aegilops tauschii and Aegilops speltoides for acquired thermotolerance: Implications in wheat breeding programmes. *Plant Physiol. Biochem. PPB*, **95**, 65–74.

Harlan, J.R. (1976) Genetic Resources in Wild Relatives of Crops. *Crop Sci.*, **16**, 329–333.

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

Henry, R.J. (2014) Sequencing of wild crop relatives to support the conservation and utilization of plant genetic resources. *Plant Genet. Resour.*, **12**, S9–S11.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. and Jarvis, A. (2004) *The WorldClim interpolated global terrestrial climate surfaces. Version 1.3*,.

Hijmans, R.J. and Etten, J. van (2012) raster: Geographic analysis and modeling with raster data. R package version 2.0–12. *R Found Stat Comput Vienna HttpCRAN R-Proj. Orgpackage Raster*.

Jarvis, A., Lane, A. and Hijmans, R.J. (2008) The effect of climate change on crop wild relatives. *Agric. Ecosyst. Environ.*, **126**, 13–23.

Jombart, T. and Ahmed, I. (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, **27**, 3070–3071.

Jombart, T., Devillard, S., Dufour, A.-B. and Pontier, D. (2008) Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity*, **101**, 92–103.

Jones, H., Gosman, N., Horsnell, R., et al. (2013) Strategy for exploiting exotic germplasm using genetic, morphological, and environmental diversity: the Aegilops tauschii Coss. example. *Theor. Appl. Genet.*, **126**, 1793.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Lasky, J.R., Upadhyaya, H.D., Ramu, P., et al. (2015) Genome-environment associations in sorghum landraces predict adaptive traits. *Sci. Adv.*, **1**, e1400218.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Available at: https://arxiv.org/abs/1303.3997 [Accessed December 12, 2017].

Li, H., Handsaker, B., Wysoker, A., et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.*, **25**, 2078–2079.

Ling, H.-Q., Zhao, S., Liu, D., et al. (2013) Draft genome of the wheat A-genome progenitor Triticum urartu. *Nature*, **496**, 87–90.

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S. and Zhang, Z. (2012) GAPIT: genome association and prediction integrated tool. *Bioinformatics*, **28**, 2397–2399.

Lipper, L., Thornton, P., Campbell, B.M., et al. (2014) Climate-smart agriculture for food security. *Nat. Clim. Change*, **4**, 1068–1072.

Lotterhos, K.E. and Whitlock, M.C. (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Mol. Ecol.*, **24**, 1031–1046.

Marroni, F., Pinosio, S., Zaina, G., Fogolari, F., Felice, N., Cattonaro, F. and Morgante, M. (2011) Nucleotide diversity and linkage disequilibrium in Populus nigra cinnamyl alcohol dehydrogenase (CAD4) gene. *Tree Genet. Genomes*, **7**, 1011–1023.

Mattila, T.M., Aalto, E.A., Toivainen, T., Niittyvuopio, A., Piltonen, S., Kuittinen, H. and Savolainen, O. (2016) Selection for population-specific adaptation shaped patterns of variation in the photoperiod pathway genes in Arabidopsis lyrata during post-glacial colonization. *Mol. Ecol.*, **25**, 581–597.

Maxted, N., Kell, S., Ford-Lloyd, B., Dulloo, E. and Toledo, Á. (2012) Toward the Systematic Conservation of Global Crop Wild Relative Diversity. *Crop Sci.*, **52**, 774–785.

Mengistu, D.K., Kidane, Y.G., Catellani, M., Frascaroli, E., Fadda, C., Pè, M.E. and Dell'Acqua, M. (2016) High-density molecular characterization and association mapping in Ethiopian durum wheat landraces reveals high diversity and potential for wheat breeding. *Plant Biotechnol. J.* Available at: http://onlinelibrary.wiley.com/doi/10.1111/pbi.12538/pdf [Accessed June 14, 2017].

Mizumoto, K., Hirosawa, S., Nakamura, C. and Takumi, S. (2002) Nuclear and chloroplast genome genetic diversity in the wild einkorn wheat, Triticum urartu, revealed by AFLP and SSLP analyses. *Hereditas*, **137**, 208–214.

Nasernakhaei, F., Rahiminejad, M.R., Saeidi, H. and Tavassoli, M. (2015) Genetic structure and diversity of Triticum monococcum ssp. aegilopoides and T. urartu in Iran. *Plant Genet. Resour.*, **13**, 1–8.

Nei, M. (1972) Genetic distance between populations. *Am. Nat.*, **106**, 283–292.

Nevo, E. (2014) Evolution of wild emmer wheat and crop improvement. *J. Syst. Evol.*, **52**, 673–696.

Özkan, H., Brandolini, A., Schäfer-Pregl, R. and Salamini, F. (2002) AFLP Analysis of a Collection of Tetraploid Wheats Indicates the Origin of Emmer and Hard Wheat Domestication in Southeast Turkey. *Mol. Biol. Evol.*, **19**, 1797–1801.

Pallotta, M., Schnurbusch, T., Hayes, J., Hay, A., Baumann, U., Paull, J., Langridge, P. and Sutton, T. (2014) Molecular basis of adaptation to high soil boron in wheat landraces and elite cultivars. *Nature*, **514**, 88–91.

Peng, J.H., Sun, D. and Nevo, E. (2011) Domestication evolution, genetics and genomics in wheat. *Mol. Breed.*, **28**, 281.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

QGIS Development Team (2017) *QGIS Geographic Information System. Open Source Geospatial Foundation*, Available at: http://qgis.osgeo.org.

Qiu, Y.C., Zhou, R.H., Kong, X.Y., Zhang, S.S. and Jia, J.Z. (2005) Microsatellite mapping of a Triticum urartu Tum. derived powdery mildew resistance gene transferred to common wheat (Triticum aestivum L.). *Theor. Appl. Genet.*, **111**, 1524–1531.

R Core Team (2013) R: A Language and Environment for Statistical Computing.

Reif, J.C., Zhang, P., Dreisigacker, S., Warburton, M.L., Ginkel, M. van, Hoisington, D., Bohn, M. and Melchinger, A.E. (2005) Wheat genetic diversity trends during domestication and breeding. *Theor. Appl. Genet.*, **110**, 859–864.

Rellstab, C., Gugerli, F., Eckert, A.J., Hancock, A.M. and Holderegger, R. (2015) A practical guide to environmental association analysis in landscape genomics. *Mol. Ecol.*, **24**, 4348–4370.

Reynolds, M., Dreccer, F. and Trethowan, R. (2007) Drought-adaptive traits derived from wheat wild relatives and landraces. *J. Exp. Bot.*, **58**, 177–186.

Rissler, L.J. (2016) Union of phylogeography and landscape genetics. *Proc. Natl. Acad. Sci.*, **113**, 8079–8086.

Russell, J., Mascher, M., Dawson, I.K., et al. (2016) Exome sequencing of geographically diverse barley landraces and wild relatives gives insights into environmental adaptation. *Nat. Genet.*, **48**, 1024–1030.

Saintenac, C., Zhang, W., Salcedo, A., Rouse, M.N., Trick, H.N., Akhunov, E. and Dubcovsky, J. (2013) Identification of Wheat Gene *Sr35* That Confers Resistance to Ug99 Stem Rust Race Group. *Science*, **341**, 783–786.

Schwabl, P., Llewellyn, M.S., Landguth, E.L., Andersson, B., Kitron, U., Costales, J.A., Ocaña, S. and Grijalva, M.J. (2017) Prediction and prevention of parasitic diseases using a landscape genomics framework. *Trends Parasitol.*, **33**, 264–275.

Sela, H., Ezrati, S., Ben-Yehuda, P., Manisterski, J., Akhunov, E., Dvorak, J., Breiman, A. and Korol, A. (2014) Linkage disequilibrium and association analysis of stripe rust resistance in wild emmer wheat (Triticum turgidum ssp. dicoccoides) population in Israel. *TAG Theor. Appl. Genet. Theor. Angew. Genet.*, **127**, 2453–2463.

Shin, J.-H., Blay, S., McNeney, B. and Graham, J. (2006) LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J. Stat. Softw.*, **16**, 1–10.

Sork, V.L., Aitken, S.N., Dyer, R.J., Eckert, A.J., Legendre, P. and Neale, D.B. (2013) Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. *Tree Genet. Genomes*, **9**, 901–911.

Storey, J.D. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **64**, 479–498.

Vavilov, N.I. and Dorofeev, V.F. (1992) *Origin and geography of cultivated plants*, Cambridge University Press.

Vincent, B., Dionne, M., Kent, M.P., Lien, S. and Bernatchez, L. (2013) Landscape Genomics in Atlantic Salmon (salmo Salar): Searching for Gene–Environment Interactions Driving Local Adaptation. *Evolution*, **67**, 3469–3487.

Wagner, H.H., Chávez-Pesqueira, M. and Forester, B.R. (2017) Spatial detection of outlier loci with Moran eigenvector maps. *Mol. Ecol. Resour.*, n/a-n/a.

Wang, Q., Tian, F., Pan, Y., Buckler, E.S. and Zhang, Z. (2014) A SUPER Powerful Method for Genome Wide Association Study. *PLOS ONE*, **9**, e107684.

Wang, X., Luo, G., Yang, W., Li, Y., Sun, J., Zhan, K., Liu, D. and Zhang, A. (2017) Genetic diversity, population structure and marker-trait associations for agronomic and grain traits in wild diploid wheat Triticum urartu. *BMC Plant Biol.*, **17**, 112.

Wright, S. (1943) Isolation by distance. *Genetics*, **28**, 114.

Zhang, Y., Liang, Z., Zong, Y., Wang, Y., Liu, J., Chen, K., Qiu, J.-L. and Gao, C. (2016) Efficient and transgene-free genome editing in wheat through transient expression of CRISPR/Cas9 DNA or RNA. *Nat. Commun.*, **7**, 12617.

Zhang, Y., Luo, G., Liu, D., Wang, D., Yang, W., Sun, J., Zhang, A. and Zhan, K. (2015) Genome-, Transcriptome- and Proteome-Wide Analyses of the Gliadin Gene Families in Triticum urartu. *PLOS ONE*, **10**, e0131559.

Zhou, Z., Jiang, Y., Wang, Z., et al. (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.*, **33**, 408–414.

Figure Legends

Figure 1. Geographic distribution of the *T. urartu* collection. The map of the sampling area is reported in shades of gray representing altitude according to the bar on the side (m.a.s.l.). Longitude and Latitude values in WGS84 degrees are reported on the x-axis and y-axis, respectively. Sampled accessions are represented by circles colored according to their position on the Fertile Crescent. Accessions without GPS coordinates are not shown.

Figure 2. Bioclimatic variation at sampling points. (a) Principal component analysis (PCA) of altitude and the 19 BioClim variables. Sampling points are represented by circles colored according to Fig. 1. (b) Correlation between altitude (alt_17) and the original BioClim variables (bio_1 to bio_19) and the derived PC axes 1 to 3. The direction and intensity of correlations is shown by circle color (according to legend on the right) and size, respectively.

Figure 3. Molecular diversity of the world collection of *T. urartu*. (a) Phylogenetic tree deriving from SNP data. Samples bottom left, marked with an asterisk, are recognized as outgroup according to a clustering analysis. Samples are colored with the same color code used in Figure 1, unmapped samples are reported in gray. Some samples are overlapping on the phylogeny. (b) Principal component analysis of the molecular diversity within the collection, excluding outgroups. Samples are colored with the same color code used in Figure 1, unmapped samples are reported in gray.

Figure 4. Structure analysis of the world collection of *T. urartu*. (a) Bar plot representing accession ancestries according to the most probable model. Each individual is represented by a vertical bar with colors proportionally to their ancestry to one of the K genetic clusters according to legend to the right. Accessions are ordered by their position on the transect, reported on the x-axis with colors according to Fig. 1. (b) Cryptic genetic structure in the Western portion of the collection (K2 in panel a). The two most probable K arrangements are shown. (c) Cryptic genetic structure in the Eastern portion of the collection (K1 in panel a), depicted as in panel b.

Figure 5. Relation among genetic and geographic features of the world collection of *T. urartu*. (a) Linear regression of geographic distance (Km, y-axis) over genetic distance (Nei's distance, x-axis) shows that distant demes are more diverse than demes in close proximity. (b) Eigenvalues resulting from a spatial PCA (sPCA). The genetic diversity is better explained by global structures (Global, red color) than by local structures (Local, blue color). The Gabriel Graph summarizing the spatial relation among samples is shown as insert. Nodes represent accessions and are colored according to the combination of sPC 1 to 3 values. (c) Spatial representation of sPC 1 to 3 interpolated across the sampling area (from yellow to blue shades, decreasing sPC values). Accessions are represented by gray dots.

Figure 6. Outcome of the GWA scan on climatic variation. Each dot represents a SNP marker tested against BIO-PC1 to BIO-PC3. Markers are ordered according to their physical position on the A sub-genome of wild emmer, with alternating colors for chromosomes 1A to 7A. Chromosome unknown is reported at the end of the graphs. The y axis represents the negative logarithm of the FDR value of the tests; the suggestive (10-6) and high significance (10-8) thresholds are depicted in blue and red color, respectively.