High molecular weight glutenin gene diversity in Aegilops tauschii demonstrates unique origin of superior wheat quality

Emily Delorean
Kansas State University
LiangLiang Gao
Kansas State University
Jose Fausto Cervantes Lopez
International Maize and Wheat Improvement Center (CIMMYT)
The Open Wild Wheat Consortium
John Innes Centre
Brande Wulff
John Innes Centre https://orcid.org/0000-0003-4044-4346
Maria Itria Ibba
International Maize and Wheat Improvement Center (CIMMYT)
Jesse Poland (jpoland@ksu.edu)
Kansas State University https://orcid.org/0000-0002-7856-1399

Article

Keywords: plant gene diversity, breeding, wheat quality

Posted Date: February 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-230727/v1

Version of Record: A version of this preprint was published at Communications Biology on November 1st, 2021. See the published version at https://doi.org/10.1038/s42003-021-02563-7.

High molecular weight glutenin gene diversity in *Aegilops tauschii* demonstrates unique origin of superior wheat quality

Emily Delorean[1,4], LiangLiang Gao[1,4], Jose Fausto Cervantes Lopez[2], Open Wild Wheat Consortium, Brande B. H. Wulff[3], Maria Itria Ibba[2], Jesse Poland[1,*]

[1] Department of Plant Pathology, Kansas State University, Manhattan, 66506, KS, USA

[2] Global Wheat Program, International Maize and Wheat Improvement Center (CIMMYT), Apdo Postal 6-641, Mexico, D.F., Mexico

[3] The John Innes Centre, Norwich Research Park, Norwich, NR4 7UH, UK

[4] These authors contributed equally

* corresponding author: jpoland@ksu.edu

Abstract

Central to the diversity of wheat products was the origin of hexaploid bread wheat, which added the D-genome of *Aegilops tauschii* to tetraploid wheat giving rise to superior dough properties in leavened breads. The polyploidization, however, imposed a genetic bottleneck, with only limited diversity introduced in the wheat D-subgenome. To understand genetic variants for quality, we sequenced 273 accessions spanning the known diversity of *Ae. tauschii*. We discovered 45 haplotypes in *Glu-D1*, a major determinant of quality, relative to the two predominant haplotypes in wheat. The wheat allele *2+12* was found in *Ae. tauschii* Lineage 2, the donor of the wheat D-subgenome. Conversely, the superior quality wheat allele *5+10* allele originated in Lineage 3, a recently characterized lineage of *Ae. tauschii*, showing a unique origin of this important allele. These two wheat alleles were also quite similar relative to the total observed molecular diversity in *Ae. tauschii* at *Glu-D1*. *Ae. tauschii* is thus a reservoir for unique *Glu-D1* alleles and provides the genomic resource to begin utilizing new alleles for end-use quality improvement in wheat breeding programs.

Introduction

Originating in the Fertile Crescent some 10,000 years ago, hexaploid wheat (*Triticum aestivum*) is now grown and consumed around the world [1]. The global consumption of wheat as a staple crop is owed principally to the unique viscoelastic properties of wheat dough that lend it the capacity to make diverse baked products such as leavened bread, tortillas, chapati, pastries, and noodles. The uniqueness of wheat dough can also be described as the strength to resist deformation and elasticity to recover the original shape as well as the viscosity to permanently deform under persistent stress. Elasticity is important for the product to hold shape, while viscosity allows the dough to be worked and formed. The balance of the competing properties determines what baked goods a dough is suitable for, such as a dough with greater strength for leavened pan bread compared to the more extensible dough that is desired for a chapati or tortilla.

Bread wheat is an allohexaploid with the A-, B- and D-subgenomes contributed by different, but related, species. The closest relative to the wheat A-subgenome is diploid *Triticum urartu*, with other diploid A-genome species including the wild and domesticated Einkorn wheat (*Triticum monococcum*). While the exact ancestor of the B-genome is unknown and presumed extinct, it is believed that *Ae. speltoides* (S-genome) is the closest living relative. These two species were brought together to form a tetraploid wheat species with AABB genome composition, which is known as durum or pasta wheat (*Triticum durum*). The D genome from *Aegilops tauschii* was the most recent addition forming the hexaploid genome. This addition of the D-subgenome, to form hexaploid wheat, led to a much broader adaptation and superior bread making quality compared to the tetraploid and diploid ancestors [2]. However, the original hexaploid species originated from very few *Ae. tauschii* accessions and limited subsequent cross-hybridization likely caused by ploidy barriers with the diploid *Ae. tauschii* [3]. This genetic bottleneck resulted in limited genetic diversity in the wheat D-subgenome [4].

The utility of wheat and the variation of wheat products and consumption is driven by the strength and elasticity of the dough which is determined by the structure of the gluten matrix. This matrix is formed from a combination of high-molecular weight (HMW) and low-molecular weight (LMW) glutenin proteins and gliadins [5]. The backbone of the gluten matrix is developed under mixing by the covalent disulphide bonds between cysteine residues in HMW glutenins [6]. These glutenins, therefore, are some

of the most important genes giving wheat its unique dough properties. They are encoded by a relatively simple locus on the long arm of the group one chromosomes of the Triticeae. Hexaploid wheat, comprised of the A-, B- and D-genomes, thus contains three HMW glutenin loci; *Glu-A1, Glu-B1* and *Glu-D1.* Each locus harbors two HMW glutenin genes known as the x and y subunit, that are tightly linked but separated by tens to hundreds of kilobase pairs (kb) [7-9]. Each subunit consists of short, unique N and C terminal domains which flank a central highly repetitive region that accounts for 74-84% of the total protein length [10].

Allelic differences in all three gluten proteins contribute to the conformation of the gluten matrix and variable end-use quality. The D-subgenome locus, however, is the major driver of bread quality and absence of the D-genome leads to substantially different dough qualities found in tetraploid pasta wheats [11]. The two common alleles at *Glu-D1* found in bread wheat are *Glu-D1a* (SDS-PAGE allele designation *2+12)* and *Glu-D1d* (*5+10*), with the latter associated with superior breadmaking quality [12-15]. Following the domestication and breeding of wheat, there is limited variation at the *Glu-D1* locus in the D-genome with only these two alleles found in the vast majority of bread wheat throughout the world [16,17]. Of the HMW glutenin alleles on the three sub-genomes, the greatest impact on end-use quality is imparted by the *Glu-D1* locus. Thus, the addition of the wheat D-subgenome and specifically variation at *Glu-D1* has substantial impact on wheat quality globally. This is arguably the single greatest defining feature of bread wheat.

Reflecting the importance of *Glu-D1* in determining the end-use quality of wheat, focus has been given to understanding the variation present in *Ae. tauschii* for this locus. Much of the work has utilized sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) protein analysis of *Ae. tauschii* collections [18-21]. From this work, over 37 SDS-PAGE *Glu-D1* alleles have been named in *Ae. tauschii.* However, due to the limited resolution of SDS-PAGE, many of the alleles have indistinguishable SDS-PAGE mobilities from the common *Glu-D1* hexaploid alleles, *2+12* and *5+10*, or are difficult to reliably distinguish. By changing the polyacrylamide percentage or acidity in the SDS-PAGE, it was shown that the *Ae. tauschii 2+12* and *5+10* alleles were slightly different than the common wheat alleles [22]. These *Ae. tauschii* alleles are therefore given the designations *2t+12t* and *5t+10t*. In addition to the *2t+12t* and *5t+10t* alleles, a large number of SDS-PAGE alleles have been described, supporting the hypothesis that *Ae. tauschii* could be a vast resource for untapped diversity at *Glu-D1* and that this diversity could be utilized for wheat quality improvement.

Here we characterized the *Glu-D1* allelic diversity in a panel of 273 sequenced *Ae. tauschii* accessions. The panel spans the known genetic diversity of *Ae. tauschii* and is a powerful resource for association mapping and gene identification [23]. From the sequenced *Ae. tauschii* panel, we discovered hundreds of genetic variants which defined dozens of unique haplotypes. This gives the needed molecular information to track these alleles in breeding germplasm, which will in turn enable targeted assessment of the novel *Ae. tauschii* HMW glutenin alleles in hexaploid backgrounds leading to utilization of favorable alleles for wheat quality improvement.

Results and Discussion

Molecular diversity of *Glu-D1* in Ae. tauschii

Through the Open Wild Wheat Consortium, we obtained Illumina 150 bp paired-end short reads from 234 unique *Ae. tauschii* accessions each sequenced to greater than 7-fold coverage[23]. These were aligned to the *Ae. tauschii* AL8/78 reference genome and sequence variants at the annotated *Glu-D1* locus were extracted. We also included three wheat cultivars in this analysis to compare *Ae. tauschii* variants to the common *5+10* (variety 'CDC Stanley') and *2+12* (varieties 'Chinese Spring' and 'LongReach Lancer') alleles. From this panel, we identified a total of 310 variants at *Glu-D1*, which were used to generate haplotypes and evaluate molecular diversity at this locus.

From the *Ae. tauschii* germplasm collection we identified 32 and 33 haplotypes within the coding sequence for the x and y subunits of the *Glu-D1* locus, respectively (Figure 1, Supplemental File S1). When considering the complete *Glu-D1* locus with combination of the x and y subunit, a total of 45 haplotypes were identified (Table 1). The various x and y subunit haplotypes were almost exclusively associated with each other, demonstrating the close physical association and limited recombination between the two genes. We included the 2500 bp up- and downstream sequences in our analysis to see if this resulted in further differentiation of alleles as short-read sequences often do not align uniquely to the central, highly repetitive region of the HMW glutenin genes. Including the flanking regions did not result in additional haplotypes. Thus, it appears that the identified variants are sufficient for faithfully differentiating alleles at *Glu-D1*.

We then calculated genetic distances and determined a gene-level phylogeny at *Glu-D1* for all of the *Ae. tauschii* accessions (Figure 1). Haplotypes clustered into three major clades, two of which were associated predominantly with Lineage 2 and one with Lineage 1. A unique group of *Glu-D1* alleles from the newly characterized Lineage 3 accessions were found within a narrow clade with Lineage 2. Among the three major clades, we designated 16 subclades that were clearly distinguished by variants and coincided with a Euclidean distance of 4. Of the 16 subclades, eight were associated exclusively with Lineage 2, five with Lineage 1, and one with Lineage 3. The Lineage 3 accessions all fell within the Lineage 2 major clades, but occupied a unique subclade therein. Thus, the gene-level phylogeny at this locus agrees very closely with the overall previously described population structure of the *Ae. tauschii* lineages[23,24]. We also observed one clade (9) that had representative accessions from both Lineage 1 and 2. This could represent an ancestral haplotype found in both lineages which underwent incomplete lineage sorting, or a case of recent interlineage haplotype exchange. Cases of haplotypes shared across Lineages 1 and 2 were also observed for pest (*Cmc4*) and disease resistance (*Sr46*) genes[23].

Lineage 2, the recognized ancestral diploid donor of the D-subgenome of hexaploid wheat [3], had greater *Glu-D1* molecular haplotype diversity than Lineage 1. Not only were there more subclades associated with Lineage 2, there were also more haplotypes (Supplemental Tables S1 and S2). As expected, the haplotypes of wheat clustered within Lineage 2 subclades (Figure 1). Within Lineage 2, we observed *Ae. tauschii* accessions with a matching sequence haplotype to the wheat *2+12* allele consistent with the D-subgenome origin from Lineage 2. Interestingly, the wheat *5+10* allele clustered within the unique Lineage 3 sub-clade. Supporting the inheritance of the *5+10* allele from Lineage 3, Gaurav *et al*. (2021)[23] observed genome-wide contribution of Lineage 3 to wheat ancestry. These findings reveal that the Lineage 3 contribution to the wheat D-subgenome included the very valuable *Glu-D1 5+10* allele, arguably one of the most important genes defining the quality of bread wheat.

Given the large difference in quality between wheat cultivars carrying *2+12* and *5+10* alleles, we hypothesized that these two haplotypes would not be similar at a molecular level. However, we found that *2+12* and *5+10* clustered relatively closely within major-clade III, with much greater overall diversity detected across *Ae. tauschii* particularly when including the Lineage 1 accessions which had very different haplotypes. When comparing the *2+12* and *5+10* haplotypes to those found in Lineage 1, it

becomes apparent that *Ae. tauschii* carries alleles that are very unlike anything seen in bread wheat and may offer unique functional characteristics when introgressed into hexaploid backgrounds.

Geographic diversity

Given the known geographic structure and distribution of *Ae. tauschii* which is associated with various levels of population structure [24], we evaluated the *Glu-D1* diversity relative to the geographic origin of the *Ae. tauschii* accessions. Molecular haplotypes were strongly associated with geographic origin, consistent with the overall genome-wide picture [24], and genetic distances between alleles increased with the geographic distance between collection sites of the *Ae. tauschii* accessions (Figure 2). The greatest concentration of haplotype diversity was located along the shores of the Caspian Sea in Iran (Figure 2). Consistent with a hypothesis of admixture between Lineage 1 and Lineage 2 leading to shared gene-level haplotypes across the lineages, the accessions from Lineage 1 and 2 with the same *Glu-D1* haplotype (within subclade 9) were collected very near one another.

Molecular haplotypes identify novel *Glu-D1* alleles

We employed SDS-PAGE analysis, the traditional standard for differentiating HMW glutenin loci, to determine if the haplotype molecular sequence diversity would also reflect differences in protein mobility. We evaluated at total of 72 unique accessions with SDS-PAGE and differentiated 9 alleles for the x subunit and 8 alleles at the y subunit from this protein mobility assay. Analysis of the Lineage 1 and Lineage 2 variants revealed that molecular haplotypes were consistent with the proteins differentiated by SDS-PAGE (Supplemental Tables 1 and 2). For the majority of the alleles that were differentiated by SDS-PAGE, we were able to unambiguously correlate the observed SDS-PAGE alleles with the molecular variants. Although specific molecular haplotypes were associated with specific SDS-PAGE mobilities, there was little concordance between gene level variation and SDS-PAGE mobility as similar alleles at the molecular level were observed with very different SDS-PAGE mobilities. Alternatively, very different molecular haplotypes were observed with the same SDS-PAGE. This supports our hypothesis that the observed sequence variants are effectively in complete linkage disequilibrium and tagging the size variants from the central repeat region. Similarly, the SDS-PAGE diversity was lower having less differentiating power than the molecular haplotypes. As noted, the same SDS-PAGE mobilities were observed in both Lineage 1 and Lineage 2 haplotypes, but the molecular haplotypes were clearly differentiated (Figure 1). The protein mobility differences are considered to be primarily due to variation in the central repetitive region and therefore are not directly detectable with

short-read sequencing, though the variable central repeats are completely phased with diagnostic haplotype variants within the terminal coding regions. Thus, we conclude that a sequence-based resource such as this *Ae. tauschii* panel provides a superior tool for identification and tracking of unique *Glu-D1* alleles in molecular breeding.

We also examined the connection between the glutenin protein mobility in *Ae. tauschii* compared to hexaploid wheat. *Ae. tauschii* haplotype *Dx1a+Dy1a* matched with the wheat *2+12* haplotype and exhibited the same SDS-PAGE mobility. Although we found an *Ae. tauschii* haplotype identical to the wheat *2+12* allele haplotype, the exact wheat *5+10* haplotype was not detected in this panel, although a very closely related Lineage 3 haplotype was found. Additionally, no *5+10* SDS-PAGE mobilities were observed. This was a surprising observation given that previous studies reported *Ae. tauschii* alleles with a *5+10* SDS-PAGE mobility [18]. However, Williams *et al.* (1993)[18] did not reveal the identities of the *Ae. tauschii* accessions with 5+10 SDS-PAGE mobility. Interestingly, the haplotype *Dx7a+Dy7a* in the newly characterized Lineage 3[23] was most similar to *5+10* on the molecular level, however it carried eight variant differences. This current panel, however, only has five unique accessions representing Lineage 3. It is possible therefore that exploration of additional Lineage 3 accessions would reveal a haplotype exactly matching the wheat *5+10* with the same mobility.

Cryptic haplotypes

One of the most valuable findings of this study was the high prevalence of cryptic molecular haplotypes hidden within SDS-PAGE mobilities. Within every SDS-PAGE mobility pattern there were multiple molecular haplotypes, often from very different subclades and occasionally from entirely different clades (Figure 1). The cryptic SDS-PAGE haplotypes, accordingly, were geographically disperse (Figure 3). For example, within SDS-PAGE *2+12* were four haplotypes; one which was the same as wheat *2+12* (*Dx1a + Dy1a*), another which was within the same subclade (*Dx1c+Dy1d*), and two from entirely different major-clades (*Dx9a+Dy9b* and *Dx13b+Dy13a*). Also, within subclade 9 were the SDS-PAGE mobilities *Dx2+Dy10* and *Dx2+Dy11*, and within subclade 13 were the SDS-PAGE mobilities *1t+12*, *2.1*+12.1**, and *4+10* further supporting that these haplotypes are not all similar to the wheat *2+12* haplotype at the molecular level. However, the proteins still migrate similarly on an SDS-PAGE. These results suggest that SDS-PAGE alone is insufficient when characterizing HMW glutenin diversity in wild relatives and will not be a suitable tool for tracking novel alleles in the hexaploid wheat germplasm.

While most molecular haplotypes delineated along the three *Ae. tauschii* lineages (Figure 1), a notable exception was within the predominantly Lineage 1 major-clade, subclade 9, where the same three haplotypes (*Dx9a+Dy9a*, *Dx9a+Dy9b*, and *Dx9a+Dy9c*) were observed in both Lineage 1 and Lineage 2 accessions. Interestingly, while there were three haplotypes at the y subunit, there was only a single x haplotype associated with all three of these. The x subunit mobility was the same for all three haplotypes, indicating that the x allele is in fact the same. However, the y subunit was differentiated with the mobility *Dy9b* was faster than that of *Dy9a* and *Dy10c* (Supplemental Figure S1).

R*ecombinant haplotypes identified*

The close proximity of the glutenin genes results in such tight linkage that recombination is extremely rare. To date, a recombination between the x and y subunit of any HMW-GS locus has yet to be verified. Among the 242 *Ae. tauschii* accessions studied here, we found a clear example of a historical recombination at *Glu-D1* in the accession TA1668 (Lineage 2). SDS-PAGE mobility of TA1668 matches that of TA10081 (*Dx2+Dy10.2*), and though the y haplotype of TA1688 is the same as the y haplotype of TA10081, the x subunit is very different and matching the Lineage 1 clade (Figure 4). Within this clade, the subclade 9 contains both Lineage 1 and Lineage 2 accessions, indicating that there was incomplete lineage sorting or admixture between the two lineages that lead to the introgression of a lineage *Glu-D1* haplotype into the Lineage 2 population. In the presence of both haplotypes, it appears there was a rare recombination between the Lineage 1 and Lineage 2 *Glu-D1* haplotypes, leading to the recombinant haplotype *Dx9a+Dy5a* found in TA1688.

The Lineage 3 accession TA2576 also appears to carry a recombinant haplotype (*Dx7b + Dy15b*) (Figure 4). However, our dataset did not contain the exact haplotypes involved in the recombination that led to *Dx7b + Dy15b.* The closest x subunit haplotype is *Dx7a*, the only other Lineage 3 haplotype, from major-clade III and the closest y subunit is the Lineage 2 haplotype *Dy15a* from major-clade II (Lineage 2). We therefore designated the x and y subunit haplotypes of TA2576 haplotypes within subclades 7 and 15. Geographical analysis reveals that TA2576 was collected from a region shared with other Lineage 3 accessions. However, the accessions containing *Dy15a* haplotype were not collected from a shared region with the L3 accessions. Although not conclusive, the most parsimonious explanation is therefore that *Dx7b + Dy15b* represents a recombinant haplotype between the x and y subunits from two different alleles. Within our current panel, however, we are unable to differentiate exactly which original haplotypes gave rise to this recombinant haplotype.

Conclusions

Importance of Glu-D1 diversity. The *Glu-D1* locus of wheat provides the greatest contribution to gluten strength, regardless of the allele present [11]. The allelic diversity of *Glu-D1* in wheat is limited to two predominant alleles *2+12* and *5+10*, and a few rare alleles (*3+12, 4+10*) which are associated with similar end-use quality as *2+12* [25,26] . The unique *2.2+12* SDS-PAGE allele, which is found at high frequency in Japanese wheat, was shown to be identical to the *2+12* haplotype with the exception of additional repeats in the internal repeat domain of the x subunit[25,27,28]. The x subunit protein from *5+10* has a unique cysteine residue just within the central repeat domain which is suspected to increase disulfide bonds in the forming dough. The early expression and greater transcription of this allele is also greater than that of the other *Glu-D1* alleles, in particular *2+12*[14]. It is unclear which of these characteristics, or the combination of the two, lend *5+10* the superior quality characteristics. The unique origin of *5+10* from Lineage 3, however, further supports the important contributions of this lineage to the wheat D genome consistent with the findings by Gaurav *et al*. (2021)[23].

Unique and valuable sources of diversity. Our haplotype analysis revealed that the x and y subunits are strongly associated even in diverse germplasm and that the *Glu-D1* haplotypes were clustered to specific geographic origins. Consistent with the findings of Gaurav *et al.* (2021)[23], we found evidence of two lineages (Lineage 2 and Lineage 3) contributing to the D genome of wheat with the superior *5+10* allele found associated with Lineage 3 accessions. Given the excellent end-use quality imparted by *5+10,* understanding this unique origin of the wheat allele support the further exploration and evaluation of novel *Glu-D1* alleles to further improve wheat quality. This also greatly supports the potential of novel alleles and unique haplotypes from the breadth of *Ae. tauschii* diversity.

Wheat grain quality remains one of the most important targets for breeders to develop superior wheat cultivars. Wild wheat relatives have been shown as a valuable resource for accessing novel genetic diversity to improve a range of wheat breeding targets including yield and disease resistance. For quality evaluation, however, the large quantity of grain needed for milling and baking and the confounding morphological characteristic needed for quality evaluation, such as suitable seed size for milling, make direct evaluation of various end-use quality traits intractable to phenotype directly these

wild relatives, including *Ae. tauschii*. In this work, we therefore took the first step in a reverse genetics approach in *Ae. tauschii* by identifying and characterizing variants at the important *Glu-D1* locus. This demonstrated the unique origin of the *Glu-D1* allele in wheat as well as uncovering novel allele variants and haplotypes that can now be targeted for breeding. We also established the relation of wheat alleles to those of *Ae. tauschii* and have shown that *Ae. tauschii* contains a trove of unique *Glu-D1* alleles very unlike the alleles in current wheat germplasm. With accessible germplasm resources such as synthetic hexaploids[23], the diagnostic variants will enable marked-assisted selection of novel *Ae. tauschii* introgressions into wheat, characterization of their end-use quality, and utilization in wheat improvement.

Methods

Plant Material

This study included 273 *Aegilops tauschii* accessions, of which 241 were from the Wheat Genetics Resource Center (WGRC) collection at Kansas State University in Manhattan, KS, USA. Another 28 were from the National Institute for Agricultural Botany (NIAB) in Cambridge, United Kingdom. An additional 2 were from the Commonwealth Scientific and Industrial Research Organisation (CSIRO) in Canberra, Australia. The final accession, AL8/78, was obtained from the John Innes Center (JIC) in Norwich, Norfolk, England. Data regarding original collection sites for the WGRC accessions is detailed in Supplementary Data 1[24]. *Aegilops tauschii* is divided into two subspecies, spp. *tauschii* (Lineage 1) and the wheat D-genome donor spp. *strangulata* (Lineage 2). In this data set, 117 accessions were Lineage 1 and 143 Lineage 2. An additional eight accessions (five non-redundant) belonged to the newly described Lineage 3[23].

SDS-PAGE Analysis

The sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) analysis of 72 of the *Ae. tauschii* accessions, was conducted at the Wheat Chemistry and Quality Laboratory at the International Maize and Wheat Improvement Center (CIMMYT) Texcoco, Mexico according to Singh *et al.* (1991)[29] with the following modifications. Specifically, 20 mg of whole meal flour were mixed at 1,400 rpm with 0.75 ml of 50% propanol (v/v) for 30 min at 65°C in a Thermomixer Comfort (Eppendorf). The tubes were then centrifuged for 2 min at 10,000 rpm, and the supernatant containing the gliadins was discarded. The pellet was then mixed with 0.1 ml of a 1.5% (w/v) DTT solution in a Thermomixer for 30 min at 65°C, 1,400 rpm, and centrifuged for 2 min at 10,000 rpm. A 0.1 ml volume of a 1.4% (v/v) vinylpyridine solution was then added to the tube which was subsequently placed again in a Thermomixer for 15 min at 65°C, 1,400 rpm, and centrifuged for 5 min at 13,000 rpm. The supernatant was mixed with the same volume of sample buffer (2% SDS (w/v), 40% glycerol (w/v), and 0.02% (w/v) bromophenol blue, pH 6.8) and incubated in the Thermomixer for 5 min at 90°C and 1,400 rpm. Tubes were centrifuged for 5 min at 10,000 rpm, and 8 ml of the supernatant were used for the glutenin gel. Glutenins were separated in polyacrylamide gels (15% or 13% T) prepared using 1 M Tris buffer, pH of 8.5. Gels were run at 12.5 mA for ~19 h. Alleles were identified using the nomenclatures proposed by Payne and Lawrence (1983)[17] for bread wheat high molecular weight glutenins and Lagudah and Halloran (1988)[22] for previously described *Ae. tauschii* high molecular weight glutenins.

DNA Sequencing

Whole genome Illumina paired-end sequencing to 10x coverage for most accessions, and 30x coverage for select accessions, was obtained from TruSeq PCR-free libraries with 350 bp insert with Illumina paired end sequencing of 150 bp according to manufactures recommendations. Sequence datasets are detailed in Gaurav *et al.* (2021)[23].

Variant Calling and Duplicated Accessions

Paired-end reads of the *Ae. tauschii* samples were aligned to the *Ae. tauschii* AL8/78 genome assembly (Aet v4.0; NCBI BioProject PRJNA341983, accession AL8/78) and hexaploid wheat samples aligned to an *in silico* reference assembly including the hexaploid wheat A and B genomes from 'Jagger' [30] (Aet v4.0; NCBI BioProject PRJNA341983, accession AL8/78) combined with the Aetv4.0 D genome using HISAT2 version 2.1.0 with default parameters [31]. Alignments were sorted and indexed using samtools v1.9 [32]. Variants for coding regions of the x and y subunits of *Glu-D1* were called using bcftools version 1.9 [33] 'mpileup' and 'call' commands with a minimum alignment quality of 20 (--q 20) [34]. Duplicated accessions were identified as sharing greater than 99.8% variant calls.

Molecular Haplotype Analysis

Ae. tauschii and hexaploid wheat variant call format (vcf) files were merged in R and variant calls were recoded to reference (-1) and alternate (1) alleles in R and heterozygous calls were set to missing. Variants were filtered on the following criteria: a variant must be present in either hexaploid wheat or *Ae. tauschii,* must have a quality score greater than 30 and be present in greater than 50% of samples. Given that we expected novel alleles present in single accessions, no minimum minor allele frequency was set. Samples sharing the same variants were considered to share the same molecular haplotype.

Genetic distances were calculated as the Euclidean distance on the A matrix of the variants in R. The A matrix was calculated with 'A.mat()' from the rrBLUP package [35] and Euclidean distances with 'dist()'. Hierarchical clustering of the genetic distances were found using hclust() and converted to a dendrogram object before plotting with the dendextend package [36].

Molecular haplotypes were designated by the subclade number of the x and y subunits together and then by the letter corresponding to the individual gene level haplotype within. For example, molecular haplotype *x1a + y1b* represents the a^{th} x haplotype and b^{th} y haplotype within the subclade 1. It should be noted that letter designations across subclades have no correspondence. The *ath* x haplotype of subclade 1 is different than that of subclade 2.

Supplemental Materials

Supplemental Table 1: *Glu-D1* gene positions in Aet v4 assembly

Supplemental Table 2: *Glu-D1x* Molecular haplotypes and associated SDS-PAGE alleles

Supplemental Table 3: *Glu-D1y* Molecular haplotypes and associated SDS-PAGE alleles

Supplemental Figure 1: SDS-PAGE images

Supplemental Data 1: Haplotypes and accession info (excel of TAs, SDS-PAGE allele, molecular allele, passport information)

Supplemental Data 2: vcf of haplotypes

Code available in github


Glu-D1
x subunit
y subunit
major clade
subclade
-2 kb
-1 kb
end
start
+1 kb
+2 kb
57 kb
3
1
2
1
3
4
5
7
6
8
9
10
11
12
13
14
15

Figure 1: Molecular haplotypes of *Glu-D1 Aegilops tauschii* accessions. Variant positions within the x and y subunit coding sequences and their 2.5 kb flanking sequences are marked with purple bars on schematic of the genes. The molecular haplotypes with position as column and accession as row are shown to the right of the dendrogram, reference allele is in gray and alternate allele is in purple. Dendrogram of combined x and y subunit haplotypes is shown to the left. The corresponding lineage of each accession is colored in blue (Lineage 1), red (Lineage 2) and green (Lineage 3). Haplotypes with major clades and subclades are designated by numbers. The wheat alleles *2+12* and *5+10* are shown in purple and recombinant haplotypes are shown in grey with asterisk.

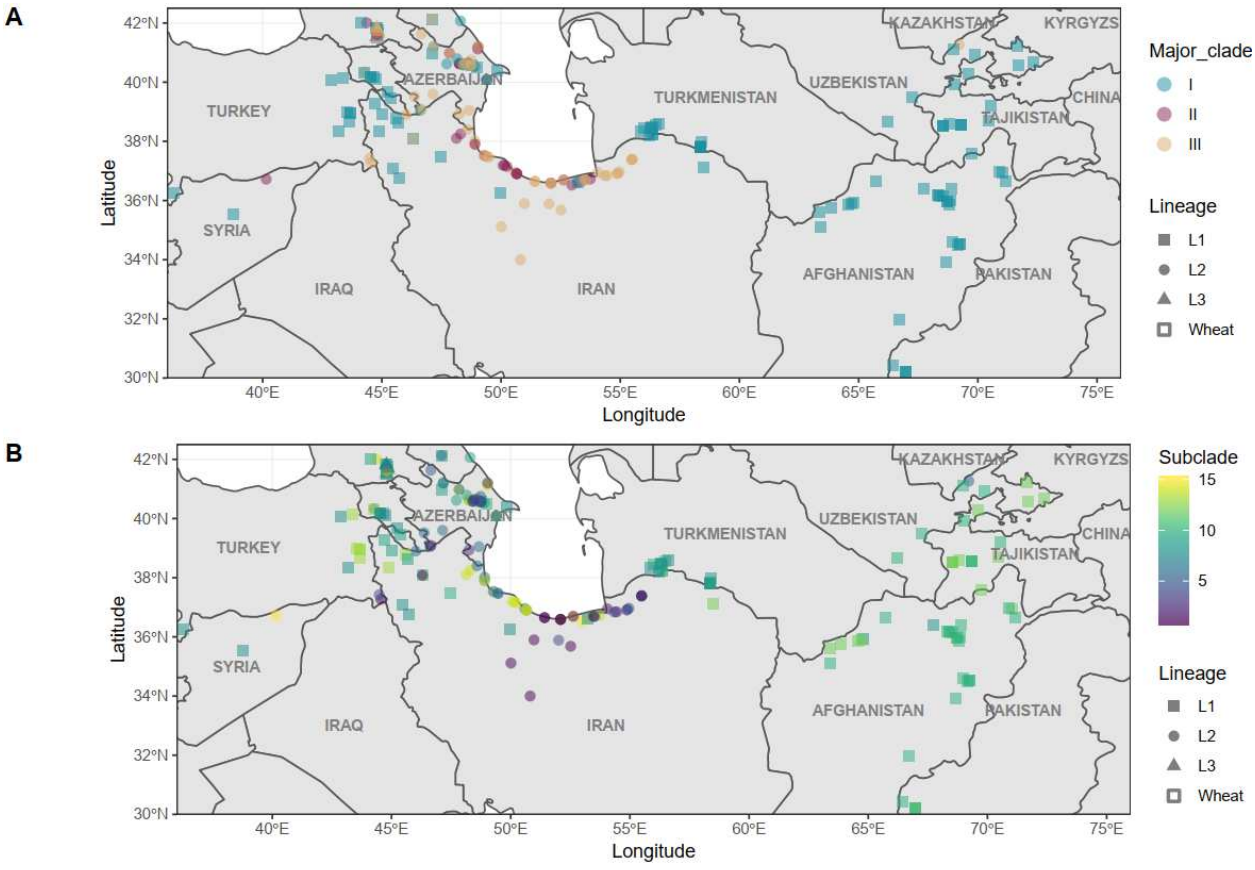

Figure 2: Geographic distribution of *Glu-D1* haplotypes. Molecular haplotypes for *Glu-D1* shown at the collection site for the given *Ae. tauschii* accession. (**a**) Distribution of accessions according to *Glu-D1* major clades with Clade I in blue, Clade II in red and Clade III in orange. (**b**) Distribution of accessions according to *Glu-D1* haplotype subclades. Haplotypes are shown on a scale from purple to yellow according to dendrogram order (Figure 1). Lineages are shown as circles for Lineage 1, triangles for Lineage 2 and squares for Lineage 3.

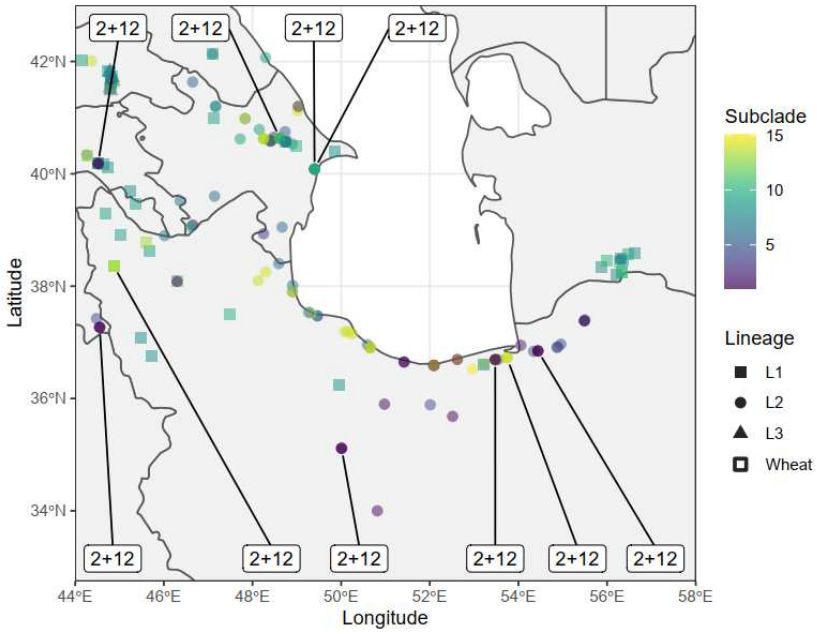

Figure 3: Cryptic haplotypes within SDS-PAGE alleles. The molecular haplotypes for *Ae. tauschii* accessions at the sites where the accessions were collected. Corresponding SDS-PAGE allele is noted for *2+12* mobility alleles.

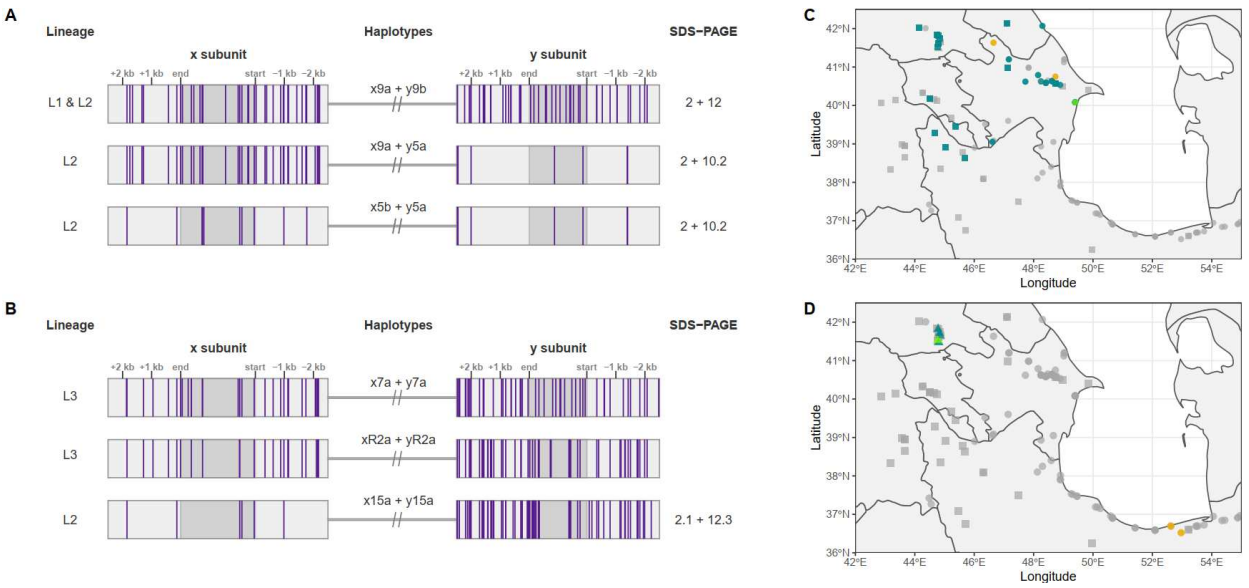

Figure 4: *Glu-D1* recombinants. Molecular haplotype representation of recombinant *Glu-D1* haplotypes for accessions (**a**) TA1668 (Lineage 2) and (**b**) TA2576 (Lineage 3). Vertical purple bars represent the alternate allele variants as called against the AL8 7/8 genome assembly. SDS-PAGE allele for the given haplotypes are show to right of each gene model. Haplotype *Dx9a* is present in both Lineage 1 and Lineage 2 accessions. The closest potential x and y subunit haplotypes involved in the recombinant haplotype *xR2+yR2a* of TA2576 are *x7a* (Lineage 3) and *y15a* (Lineage 2). SDS-PAGE protein mobilities for *x7a + y7a* and *xR2+yR2a* were not analyzed. Geographical distribution of recombinant Glu-D1 haplotypes for (**c**) TA1668 (Lineage 2) and (**d**) TA2576 (Lineage 3). Collection site of recombinant accessions is marked in lime green, whereas turquoise and orange designate the collection sites of accessions carrying x subunit and y subunit haplotypes, respectively. Accessions with unrelated haplotypes are in light gray. Lineages are shown in squares (Lineage 1), circles (Lineage 2) or triangles (Lineage 3).

Acknowledgements

ED was supported though the Monsanto's Beachell-Borlaug International Scholars Program. This material is based upon work supported by the National Science Foundation under Award No. 1822162 "Phase II IUCRC at Kansas State University Center for Wheat Genetic Resources WGRC" and Award No. 1339389 "GPF-PG: Genome Structure and Diversity of Wheat and Its Wild Relatives". BW was supported by the UK Biotechnology and Biological Sciences Research Council Designing Future Wheat Institute Strategic Programme BB/P016855/1. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Conflict of Interest: The authors declare no conflict of interest.

References

1 Salamini, F., Özkan, H., Brandolini, A., Schäfer-Pregl, R. & Martin, W. Genetics and geography of wild cereal domestication in the near east. *Nature Reviews Genetics* **3**, 429-441, doi:10.1038/nrg817 (2002).

2 Dubcovsky, J. & Dvorak, J. Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science (New York, N.Y.)* **316**, 1862-1866, doi:10.1126/science.1143986 (2007).

3 Wang, J. *et al.* Aegilops tauschii single nucleotide polymorphisms shed light on the origins of wheat D-genome genetic diversity and pinpoint the geographic origin of hexaploid wheat. *The New Phytologist* **198**, 925-937, doi:10.1111/nph.12164 (2013).

4 Zhou, Y. *et al.* Triticum population sequencing provides insights into wheat adaptation. *Nat Genetics* **52**, 1412-1422, doi:10.1038/s41588-020-00722-w (2020).

5 Shewry, P. What Is Gluten—Why Is It Special? *Frontiers in Nutrition* **6**, doi:10.3389/fnut.2019.00101 (2019).

6 Lutz, E., Wieser, H. & Koehler, P. Identification of disulfide bonds in wheat gluten proteins by means of mass spectrometry/electron transfer dissociation. *Journal of Agricultural and Food Chemistry* **60**, 3708-3716, doi:10.1021/jf204973u (2012).

7 Anderson, O., Rausch, C., Moullet, O. & Lagudah, E. The wheat D-genome HMW-glutenin locus: BAC sequencing, gene distribution, and retrotransposon clusters. *Functional & Integrative Genomics* **3**, 56-68, doi:10.1007/s10142-002-0069-z (2003).

8 Kong, X. Y., Gu, Y. Q., You, F. M., Dubcovsky, J. & Anderson, O. D. Dynamics of the evolution of orthologous and paralogous portions of a complex locus region in two genomes of allopolyploid wheat. *Plant Molecular Biology* **54**, 55-69, doi:10.1023/B:PLAN.0000028768.21587.dc (2004).

9 Gu, Y. Q. *et al.* Types and rates of sequence evolution at the high-molecular-weight glutenin locus in hexaploid wheat and its ancestral genomes. *Genetics* **174**, 1493-1504, doi:10.1534/genetics.106.060756 (2006).

10 Shewry, P. R., Halford, N. G., Belton, P. S. & Tatham, A. S. The structure and properties of gluten: an elastic protein from wheat grain. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **357**, 133-142, doi:10.1098/rstb.2001.1024 (2002).

11 Wang, Z. *et al.* New insight into the function of wheat glutenin proteins as investigated with two series of genetic mutants. *Scientific Reports* **7**, 3428, doi:10.1038/s41598-017-03393-6 (2017).

12 Rooke, L. *et al.* Overexpression of a Gluten Protein in Transgenic Wheat Results in Greatly Increased Dough Strength. *Journal of Cereal Science* **30**, 115-120, doi:https://doi.org/10.1006/jcrs.1999.0265 (1999).

13 Lafiandra, D., D'Ovidio, R., Porceddu, E., Margiotta, B. & Colaprico, G. New Data Supporting High Mr Glutenin Subunit 5 as the Determinant of Quality Differences among the Pairs 5 + 10 vs. 2 + 12. *Journal of Cereal Science* **18**, 197-205, doi:https://doi.org/10.1006/jcrs.1993.1046 (1993).

14 Don, C., Lookhart, G., Naeem, H., MacRitchie, F. & Hamer, R. J. Heat stress and genotype affect the glutenin particles of the glutenin macropolymer-gel fraction. *Journal of Cereal Science* **42**, 69-80, doi:https://doi.org/10.1016/j.jcs.2005.01.005 (2005).

15 Zhang, P.-P., Ma, H.-X., Yao, J.-B. & He, Z.-H. Effect of Allelic Variation and Expression Quantity at Glu-1 Loci on Size Distribution of Glutenin Polymer in Common Wheat. *Acta Agronomica Sinica* **35**, 1606-1612, doi:https://doi.org/10.1016/S1875-2780(08)60104-2 (2009).

16 Payne, P. I., Holt, L. M. & Law, C. N. Structural and genetical studies on the high-molecular-weight subunits of wheat glutenin : Part 1: Allelic variation in subunits amongst varieties of wheat (*Triticum aestivum*). *Theoretical and Applied Genetics* **60**, 229-236, doi:10.1007/bf02342544 (1981).

17 Payne, P. I. & Lawrence, G. J. Catalogue of alleles for the complex gene loci, Glu-A1, Glu-B1, and Glu-D1 which code for high-molecular-weight subunits of glutenin in hexaploid wheat. *Cereal Research Communications* **11**, 29-35 (1983).

18 William, M. D. H. M., Peña, R. J. & Mujeeb-Kazi, A. Seed protein and isozyme variations in Triticum tauschii (*Aegilops squarrosa*). *Theoretical and Applied Genetics* **87**, 257-263, doi:10.1007/BF00223774 (1993).

19 Xu, S., Khan, K., Klindworth, D. & Nygard, G. Evaluation and characterization of high-molecular weight 1D glutenin subunits from *Aegilops tauschii* in synthetic hexaploid wheats. *Journal of Cereal Science* **52**, doi:10.1016/j.jcs.2010.05.004 (2010).

20 Mackie, A. M., Lagudah, E. S., Sharp, P. J. & Lafiandra, D. Molecular and biochemical characterisation of HMW glutenin subunits from *T. tauschii* and the D genome of hexaploid wheat. *Journal of Cereal Science* **23**, 213-225, doi:https://doi.org/10.1006/jcrs.1996.0022 (1996).

21 Gianibelli, M. C., Gupta, R. B., Lafiandra, D., Margiotta, B. & MacRitchie, F. Polymorphism of high Mr glutenin subunits in *Triticum tauschii*: Characterisation by Chromatography and

Electrophoretic Methods. *Journal of Cereal Science* **33**, 39-52, doi:https://doi.org/10.1006/jcrs.2000.0328 (2001).

22 Lagudah, E. S. & Halloran, G. M. Phylogenetic relationships of *Triticum tauschii* the D genome donor to hexaploid wheat. *Theoretical and Applied Genetics* **75**, 592-598, doi:10.1007/BF00289125 (1988).

23 Gaurav, K. *et al.* Evolution of the bread wheat D-subgenome and enriching it with diversity from *Aegilops tauschii*. *bioRxiv*, 2021.2001.2031.428788, doi:10.1101/2021.01.31.428788 (2021).

24 Singh, N. *et al.* Genomic analysis confirms population structure and identifies inter-Lineage Hybrids in Aegilops tauschii. *Frontiers in Plant Science* **10**, doi:10.3389/fpls.2019.00009 (2019).

25 Wan, Y. *et al.* Comparative analysis of the D genome-encoded high-molecular weight subunits of glutenin. *Theoretical and Applied Genetics* **111**, 1183-1190, doi:10.1007/s00122-005-0051-y (2005).

26 Dong, Z. *et al.* Haplotype Variation of Glu-D1 Locus and the Origin of Glu-D1d Allele Conferring Superior End-Use Qualities in Common Wheat. *PLoS ONE* **8**, e74859, doi:10.1371/journal.pone.0074859 (2013).

27 Payne, P. I., Holt, L. M. & Lawrence, G. J. Detection of a novel high molecular weight subunit of glutenin in some Japanese hexaploid wheats. *Journal of Cereal Science* **1**, 3-8, doi:https://doi.org/10.1016/S0733-5210(83)80003-4 (1983).

28 Shimizu, K. K. *et al.* De Novo Genome Assembly of the Japanese Wheat Cultivar Norin 61 Highlights Functional Variation in Flowering Time and Fusarium Resistance Genes in East Asian Genotypes. *Plant and Cell Physiology*, doi:10.1093/pcp/pcaa152 (2020).

29 Singh, N. K., Shepherd, K. W. & Cornish, G. B. A simplified SDS-PAGE procedure for separating LMW subunits of glutenin. *Journal of Cereal Science* **14**, 203-208, doi:https://doi.org/10.1016/S0733-5210(09)80039-8 (1991).

30 Walkowiak, S. *et al.* Multiple wheat genomes reveal global variation in modern breeding. *Nature* **588**, 277-283, doi:10.1038/s41586-020-2961-x (2020).

31 Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**, 907-915, doi:10.1038/s41587-019-0201-4 (2019).

32 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

33 Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)* **27**, 2987-2993, doi:10.1093/bioinformatics/btr509 (2011).

34 Gao, L. (Zenodo, 2020).

35 Endelman, J. B. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome* **4**, doi:https://doi.org/10.3835/plantgenome2011.08.0024 (2011).

36 Galili, T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31**, 3718-3720, doi:10.1093/bioinformatics/btv428 (2015).

Figures

Figure 1

Molecular haplotypes of Glu-D1 Aegilops tauschii accessions. Variant positions within the x and y subunit coding sequences and their 2.5 kb flanking sequences are marked with purple bars on schematic of the genes. The molecular haplotypes with position as column and accession as row are shown to the

right of the dendrogram, reference allele is in gray and alternate allele is in purple. Dendrogram of combined x and y subunit haplotypes is shown to the left. The corresponding lineage of each accession is colored in blue (Lineage 1), red (Lineage 2) and green (Lineage 3). Haplotypes with major clades and subclades are designated by numbers. The wheat alleles purple and recombinant haplotypes are shown in grey with asterisk.

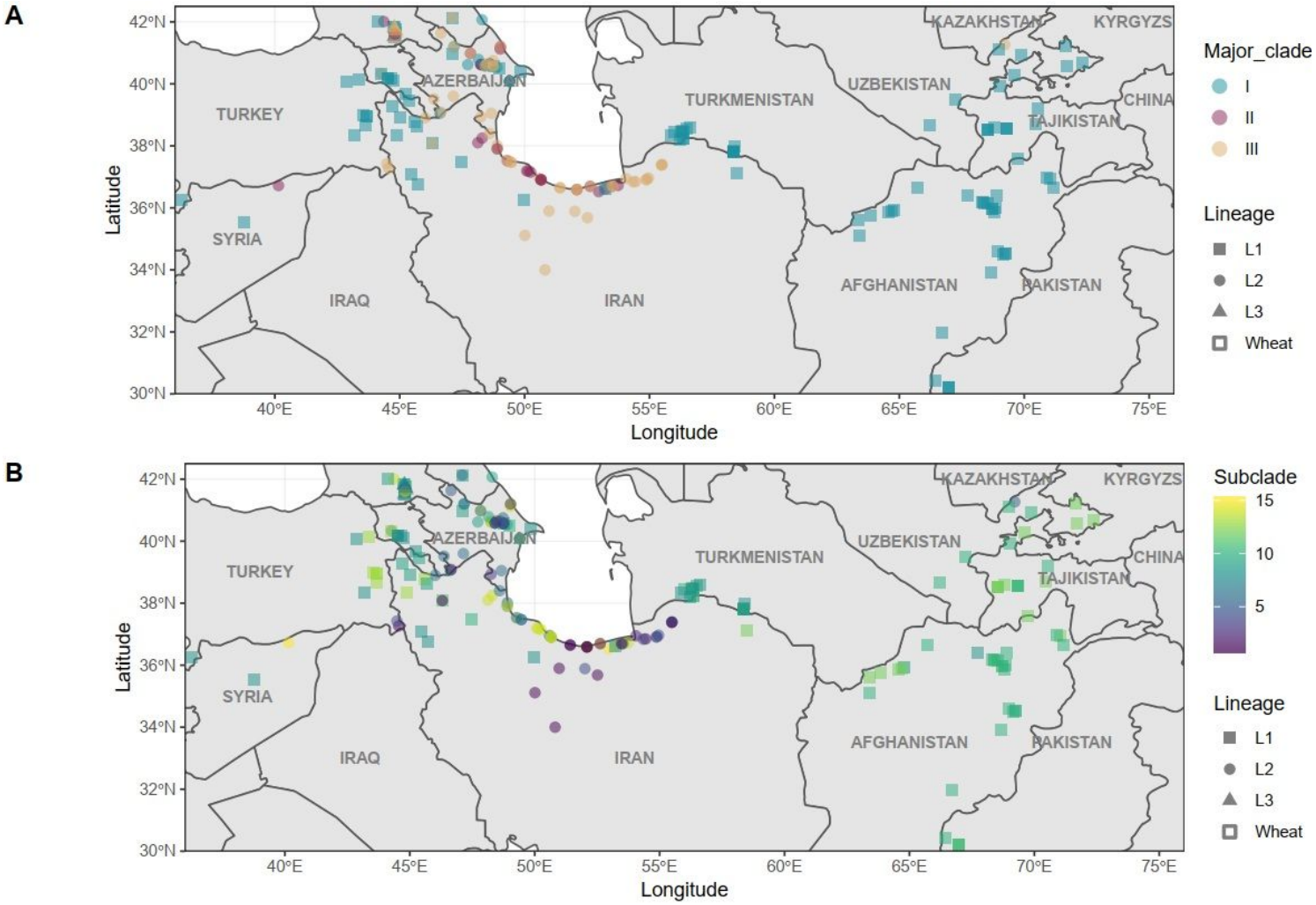

Figure 2

Geographic distribution of Glu-D1 haplotypes. Molecular haplotypes for Glu-D1 shown at the collection site for the given Ae. tauschii accession. (a) Distribution of accessions according to Glu-D1 major clades with Clade I in blue, Clade II in red and Clade III in orange. (b) Distribution of accessions according to Glu-D1 haplotype subclades. Haplotypes are shown on a scale from purple to yellow according to dendrogram order (Figure 1). Lineages are shown as circles for Lineage 1, triangles for Lineage 2 and squares for Lineage 3.

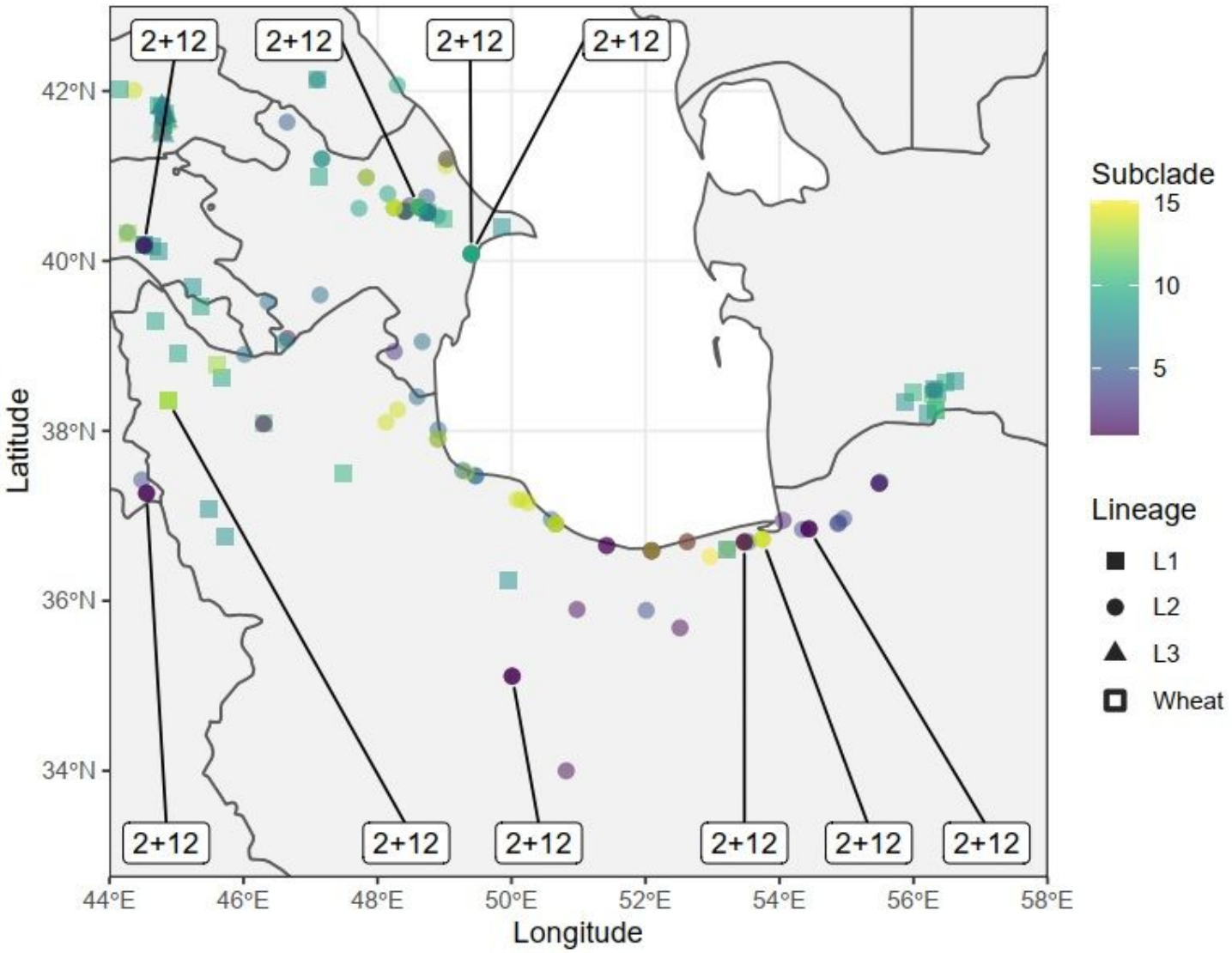

Figure 3

Cryptic haplotypes within SDS-PAGE alleles. The molecular haplotypes for Ae. tauschii accessions at the sites where the accessions were collected. Corresponding SDS-PAGE allele is noted for 2+12 mobility alleles.

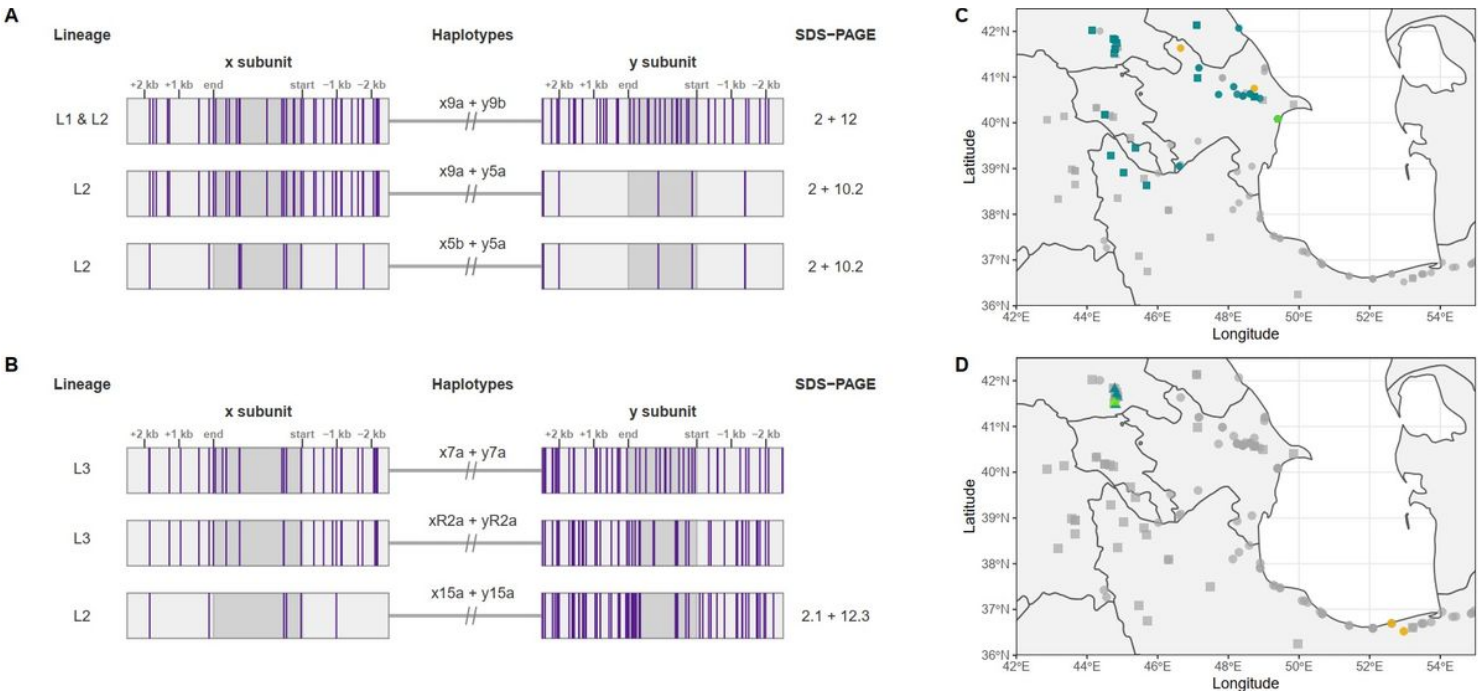

Figure 4

Glu-D1 recombinants. Molecular haplotype representation of recombinant Glu-D1 haplotypes for accessions (a) TA1668 (Lineage 2) and (b) TA2576 (Lineage 3). Vertical purple bars represent the alternate allele variants as called against the AL8 7/8 genome assembly. SDS-PAGE allele for the given haplotypes are show to right of each gene model. Haplotype Dx9a is present in both Lineage 1 and Lineage 2 accessions. The closest potential x and y subunit haplotypes involved in the recombinant haplotype xR2+yR2a of TA2576 are x7a (Lineage 3) and y15a (Lineage 2). SDS-PAGE protein mobilities for x7a + y7a and xR2+yR2a were not analyzed. Geographical distribution of recombinant Glu-D1 haplotypes for (c) TA1668 (Lineage 2) and (d) TA2576 (Lineage 3). Collection site of recombinant accessions is marked in lime green, whereas turquoise and orange designate the collection sites of accessions carrying x subunit and y subunit haplotypes, respectively. Accessions with unrelated haplotypes are in light gray. Lineages are shown in squares (Lineage 1), circles (Lineage 2) or triangles (Lineage 3).

Supplementary Files

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

- SupplementalMaterials.pdf
- SupplementalData1popmap.csv
- SupplementalData2GluD1.txt