

Strategy for exploiting exotic germplasm using genetic, morphological, and environmental diversity: the *Aegilops tauschii* Coss. example

H. Jones · N. Gosman · R. Horsnell · G. A. Rose · L. A. Everest · A. R. Bentley · S. Tha · C. Uauy · A. Kowalski · D. Novoselovic · R. Simek · B. Kobiljski · A. Kondic-Spika · L. Brbaklic · O. Mitrofanova · Y. Chesnokov · D. Bonnett · A. Greenland

Received: 28 June 2012 / Accepted: 21 March 2013 / Published online: 5 April 2013
© Springer-Verlag Berlin Heidelberg 2013

Abstract Hexaploid bread wheat evolved from a rare hybridisation, which resulted in a loss of genetic diversity in the wheat D-genome with respect to the ancestral donor, *Aegilops tauschii*. Novel genetic variation can be introduced into modern wheat by recreating the above hybridisation; however, the information associated with the *Ae. tauschii* accessions in germplasm collections is limited, making rational selection of accessions into a re-synthesis programme difficult. We describe methodologies to identify novel diversity from *Ae. tauschii* accessions that combines Bayesian analysis of genotypic data, sub-species diversity and geographic information that summarises variation in climate and habitat at the collection point for each accession. Comparisons were made between diversity discovered amongst a panel of *Ae. tauschii* accessions, bread wheat varieties and lines from the CIMMYT

synthetic hexaploid wheat programme. The selection of *Ae. tauschii* accessions based on differing approaches had significant effect on diversity within each set. Our results suggest that a strategy that combines several criteria will be most effective in maximising the sampled variation across multiple parameters. The analysis of multiple layers of variation in ex situ *Ae. tauschii* collections allows for an informed and rational approach to the inclusion of wild relatives into crop breeding programmes.

Introduction

The wild relatives of crop species have contributed novel diversity to breeding in a range of crops including barley, cassava, potato, rice, tomato, and wheat. It has been estimated that the contribution from crop wild relatives to genetic improvement is 1 % per year, with a value to global agriculture of \$1 billion (Heywood, 2011). It has been estimated that 29 crop species have benefitted from the transfer of useful traits from crop wild relatives (Maxted

Communicated by J. Dubcovsky.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-013-2093-x) contains supplementary material, which is available to authorized users.

H. Jones (✉) · N. Gosman (✉) · R. Horsnell · G. A. Rose · L. A. Everest · A. R. Bentley · C. Uauy · A. Greenland
NIAB, Huntingdon Road, Cambridge CB1 0LE, UK
e-mail: huw.jones@niab.com

N. Gosman
e-mail: nick.gosman@niab.com

S. Tha · C. Uauy · A. Kowalski
John Innes Centre, Norwich Research Park, Colney,
Norwich NR4 7UH, UK

D. Novoselovic · R. Simek
Agricultural Institute of Osijek, Južno Predgrade 17,
Osijek, Croatia

B. Kobiljski · A. Kondic-Spika · L. Brbaklic
Institute of Field and Vegetable Crops, M/Gorkog 30, Novi Sad,
Serbia

O. Mitrofanova · Y. Chesnokov
Vavilov Institute of Plant Industry, 42-44, B.Morskaya Street,
190000 St. Petersburg, Russia

D. Bonnett
CIMMYT, Km. 45, Carretera Mexico-Veracruz, El Batan,
CP 56130 Texcoco, Edo. de México, Mexico

and Kell 2009). A functional form of the grain protein content gene Gpc-B1 identified in *Triticum turgidum* ssp. *dicoccoides* is found to offer improved nutrition and changes in processing quality when introgressed into bread wheat and durum wheat (Brevis et al. 2010). Strategies based on extensive phenotyping of germplasm collections to identify wild relatives that might enhance productivity or adaptation of agricultural crops has been compared to searching for a ‘needle in a haystack’ and population-based genetics has been suggested as an alternative approach (Prada 2009). For example, the advent of next generation sequencing facilitates allele mining where novel SNP loci or haplotypes can be used to select accessions for targeted phenotypic investigations (Kilian and Graner 2012).

An alternative is to follow the approach of Vavilov, viewing plants as having a “tendency to geographic evolution, and divided into a greater or smaller number of eco-geographic or agro ecological groups” (Vavilov 1957). In wild barley, population sub-structure appears to be related to variation in rainfall (Hübner et al. 2009). Similarly, in barley landraces, associations between climatic factors and population structure were detected (Jones et al. 2011). Such eco-geographic data has long been considered as a tool to discover novel adaptive variation; however, population sub-structure must be taken into account. Its omission may explain the limited success in identifying salt tolerant barley or in detecting frost tolerance in potato wild relatives in studies by Peeters et al. (1990) and Hijmans et al. (2003). The value of eco-geographic data has also been demonstrated in experiments aimed at assessing its predictive power for a set of agronomic traits (days to heading, days to maturity, plant height, harvest index, and volumetric weight) in Nordic barley landraces (Endresen 2010). The statistical techniques applied in that study have been extended and applied to prediction of responses to biotic stresses (stem rust in wheat and net blotch) in a wider set of landraces where a calibration calculated using a ‘training set’ successfully predicted susceptibility in a second ‘test set’ of accessions (Endresen et al. 2011). These genomic and eco-geographic approaches have been integrated into the Focussed identification of germplasm strategy where eco-geographic data is used as a pre-screening tool to select a panel of ‘focussed’ accessions that are taken forward for allele mining. The power of this strategy is illustrated by the isolation of novel resistances to powdery mildew, identified in a panel of wheat landraces by Bhullar et al. (2009). In that study, eco-geographic data were used to reduce a virtual panel of 16,000 accessions to 1,300 that were taken forward for phenotypic evaluation and these in turn were reduced to 111 candidates for allele mining.

Wheat is an allohexaploid, combining within a single genome those of three diploid grass species, the A-genome of *Triticum urartu*, the B-genome from a species within the

Sitopsis section and related to *Aegilops speltoides* and the D-genome of *Aegilops tauschii* (Dvorak et al. 1992; Feldman et al. 1995). In addition, it has been shown that the wheat D-genome is more closely related to *Ae. tauschii* ssp. *strangulata* than to ssp. *tauschii* (Dvorak et al. 1998a). These three species, therefore, constitute wheat’s primary gene pool (Jiang et al. 1994). Of the three, the D-genome is considered to be the least genetically diverse (Dubcovsky and Dvorak 2007), thereby restricting the potential for breeder selection and improvement. To increase D-genome diversity in bread wheat, the International Center for Maize and Wheat Improvement (CIMMYT) has generated in the past 25 years more than 1,200 synthetic hexaploid wheat (SHW) lines by crossing elite tetraploid durum with *Ae. tauschii* accessions (Trethowan and Mujeeb-Kazi 2008). However, it is unclear how much additional D-genome diversity has been reintroduced into modern varieties by the CIMMYT-SHW programme compared to what is available within the *Ae. tauschii* diversity range as a whole.

Aegilops tauschii, distributed across Eurasia from the Caucasus region between the Black and Caspian Sea in the west through central Asia to China in the east, is widely conserved ex situ in germplasm collections (Maxted et al. 2008). It has previously been shown that diversity at 18 microsatellite loci is sufficient to divide *Ae. tauschii* into eastern and western clusters with Iranian accessions partitioning between the two groups (Pestsova et al. 2000). Similarly, chloroplast sequence diversity, assessed at five loci, divides *Ae. tauschii* into two groups, broadly congruent with the sub-species *tauschii* and *strangulata* designations (Dudnikov 2011). The geographic organisation of haplotype diversity suggests that *Ae. tauschii* originated within the Caucasus and migrated eastward. Lineages associated with ssp. *tauschii* are found across central Asia, while lineages associated with ssp. *strangulata* are confined to that part of the range located to the west of the Kopet Dag Mountains (located between modern day Turkmenistan and Iran). Within ssp. *strangulata*, Dudnikov (2011) notes a division in the geographic organisation of lineages between those that include the region around the Caspian Sea within their range and those that do not. A similar pattern is described by nuclear diversity calculated using AFLP markers (Mizuno et al. 2010). In both studies, the greatest diversity was observed within the Caucasus with both lineages present (Mizuno et al. 2010; Dudnikov 2011). Commensurate with its extensive eco-geographic range, *Ae. tauschii* carries important sources of biotic (Kishii et al. 2007; Mujeeb-Kazi et al. 2007) and abiotic stress tolerance (Reynolds et al. 2007; Reynolds and Trethowan 2007) that are of potential value for wheat breeding.

In this work, we explore *Ae. tauschii* diversity by considering genetic, phenotypic, geographic and environmental

data, using this information to select accessions from a wider collection of germplasm. By using these data in combination it is possible to make rational selections of exotic germplasm for inclusion in a plant breeding programmes.

Materials and methods

Plant materials

A set of 429 *Ae. tauschii* accessions from diverse collection sites in 14 countries was assembled from seven international germplasm collections. Available passport data were collated and cross-referenced to identify duplicated entries and, where possible, to enhance the passport information for that accession. A single accession was taken forward for genotyping if duplicate accessions were identified. Where the information was available, the accessions' sub-species, collection site (description, latitude and longitude), and germplasm collection references were collated (Table S1). Where the full botanical descriptions were unavailable, plants were grown and a morphological examination was made to determine each accession's sub-species (see below). For those accessions with a collection site description only, geographic coordinates were abstracted from National Geospatial-Intelligence Agency (USA) online databases available at <https://www.1.nga.mil> (Jones et al. 2008). A panel of 269 *Ae. tauschii* accessions was taken forward for genotyping and, of these, 254 accessions were taken through for detailed analysis, eliminating accessions with a high rate of missing genotype data and, where possible, eliminating duplicate entries.

Forty-eight SHW lines originating from independent *Ae. tauschii* accessions (CIMMYT WX prefix) were selected through diversity analysis to be representative of the D-genome diversity from a larger collection of 430 SHW lines supplied by CIMMYT. The D-genome donors for the CIMMYT-SHW programme were selected by screening the available *Ae. tauschii* accessions for disease resistance. The information supplied with the CIMMYT-SHW lines was cross-referenced against germplasm collection databases and, where the *Ae. tauschii* parent could be securely identified, collection site data were collated (Table S2).

A diverse collection of 166 bread wheat varieties from northern, central, and eastern Europe was included, encompassing commercial varieties released between 1970 and 2010 from breeding programmes in the UK, France, Germany, Serbia, and Croatia (Table S3). This germplasm was also selected from a larger collection on the basis of D-genome SSR marker diversity. Varieties from UK, France, and Germany were selected from the association genetics panel assembled for the TriticeaeGenome project (<http://www.triticeaegenome.eu/>). Other bread wheat lines

were sourced from the Agricultural Institute of Osijek and the Institute of Field and Vegetable Crops of Novi Sad.

Sub-species classification of *Ae. tauschii* accessions

Sub-species classification was provided by germplasm collections for 136 of the *Ae. tauschii* accessions taken for genotyping. For the remaining 118, classification into ssp. *strangulata* and ssp. *tauschii* was carried out with reference to work previously published in Knaggs et al. (2000). To provide material for study, four naturally vernalised plants of each accession were grown under long days (16 h photoperiod) in a heated (temperature range 17–25 °C) glasshouse between February and May 2010. Height and spike length were measured and spike appearance was visually assessed. Direct observation of the following characters was used for differentiation: (1) spikelet width and length (excluding scurs) where ssp. *strangulata* is characterised by a broader spikelet in comparison to ssp. *tauschii*, (2) spiculation on the inner-lateral nerves where spicules were fewer and weaker in ssp. *strangulata*, (3) seed shape where ssp. *tauschii* seeds are elongated while those of ssp. *strangulata* are oval or round, (4) appearance of glumes where those of ssp. *strangulata* glumes tend to curl out from the shoulder. Where visual assessment gave ambiguous results, image analysis of spikelet shape was carried out using a bespoke image analysis system (NIAB DUS Imaging system). Psi Soft image analysis software (v2.2.1.1) was used to reduce the data to width, length and area measurements for statistical analysis, calibrated with a training set of including both ssp. *strangulata* and ssp. *tauschii*. Analysis of means differentiated ssp. *strangulata* from *tauschii* as the former had significantly broader spikelets ($p < 0.05$).

Genotyping

Fifteen D-genome-specific microsatellite (SSR) markers were chosen for the analysis. These loci provided coverage of all the bread wheat D-genome chromosomes with the exception of 7D (Table S4). Based on previous studies, selection of SSR primer pairs was made on the basis of high polymorphism information content (PIC) values in both SHW and bread wheat. Sequences and PCR conditions for microsatellite markers are published in Eujayl et al. (2002), Gupta et al. (2002), Röder et al. (1998), Song et al. (2002, 2005). A subset of 232 *Ae. tauschii* accessions were also genotyped at 62 D-genome specific SNP loci using KASPar assays. These markers were developed for use by plant breeders by the BBSRC funded UK wheat SNP consortium (Allen et al. 2011, 2013).

The plant material was assayed for diversity at the *Photoperiod-1* (*Ppd-1*) locus on chromosome 2D using

previously described primers and amplification conditions (Beales et al. 2007), modified for capillary electrophoresis by the addition of a FAM fluorophore to the 5' primer. This locus was selected for analysis both as an exemplar and because variation at photoperiod loci has a significant influence on environmental adaptation, and, hence, yield in UK conditions.

DNA isolation, PCR, and fragment analysis

Genomic DNA was extracted from fresh leaf material and prepared for PCR according to Fulton et al. (1995). PCR was carried out with fluorescently labelled primer pairs and amplified fragments were analysed using an ABI Prism™ 3730 sequencer (Applied Biosystems). Allele calling was carried out with GeneMapper® version 4 0.0 (Applied Biosystems).

Proxies derived from geography

ArcGIS 10.0 (ESRI 2011) was used to analyse the geographical distributions of accessions and the biome (IUCN 1974) at each accession's point of origin was abstracted using a spatial query. Climate data for each accession's point of origin was abstracted from a 10 min resolution (approximately 10–15 km at these latitudes) climate data set Climate Research Unit (CRU) CL 2.0 (downloaded from <http://www.cru.uea.ac.uk/>) (New et al. 2002), including monthly mean data for precipitation, wet day frequency, temperature, diurnal temperature range, relative humidity, sunshine, frost frequency and wind-speed. These climate data were augmented with derived values for potential evapotranspiration, moisture deficit, and minimum temperature.

Modelling alternative selection strategies

We modelled alternative selection strategies by selecting sets of 20 accessions using differing criteria. The decision to select 20 accessions represented a pragmatic choice based on resources available to characterise the populations generated post re-synthesis. Our selection criteria included selections intended to maximise genetic, environmental or geographic diversity, either as de novo selections or selections to complement the extant CIMMYT-SHW lines. Selections were also made for their similarity to the climates of wheat growing areas of interest to our breeding programmes and selections made to represent the biomes found within the natural range of *Ae. tauschii*. We calculated the genetic and environmental variations for each selection and tabulated the distributions for sub-species, Ppd-D1 diversity, membership of distinct sub-populations, and biomes descriptions for each accession selected.

Within a subset of our accessions, we also compared selections made using genetic diversity using two alternative marker systems. We evaluated each approach to selection by comparing these results with results for the same metrics calculated for sets of 20 accessions, drawn by random sampling, in 1,000 replications (Table 1).

To select a set that maximises geographic diversity we found the central point of the geographic distribution and then calculated the geographic (great circle) distance of all accessions from the central point. The first accession selected was that found to be furthest from the central point. The second accession selected was that found to be furthest from the first accession, the third accession selected was that found to be furthest from the first and second accession and so on until 20 accessions had been selected (Table S10). Geographic (great circle) distances were used in this instance as they best represent inter-accession distances on the curved surface of the earth.

To simplify the selections intended to maximise genetic or environmental diversity, the complexity of genetic distances (SSR) and environmental dissimilarities were reduced to two dimensions by principle coordinate ordination (PCO). Geometric distances within this two dimensional representation of diversity were calculated from the starting position and selections made using the same method applied to the geographic data. For de novo selections, the starting point was the origin (0, 0). When making selections to complement the CIMMYT-SHW lines, we used the position of those CIMMYT-SHW lines within diversity space as a starting point from which to identify those individuals furthest from previously sampled regions and, as sampling continued, furthest from each other (Fig. 1) (Tables S11 and S12). Euclidean distances were used in this instance as they best represent inter-accession distances once genetic distances (SSR) and environmental dissimilarities had been reduced to two dimensions by PCO.

In order to compare the effect of using different types of genetic data, de novo selections were made within diversity space calculated using SSR data or KASPar data.

When making selections for their similarity to the climates of wheat growing areas of interest to our breeding programmes, we abstracted climate data for Cambridge (UK) and Tours (France), representing north west Europe or Novi Sad (Serbia) and Osijek (Croatia) representing the Danube basin. In this case we selected 20 accessions for their proximity within climate space to Cambridge and Tours or Novi Sad and Osijek (Table S13).

Finally, the accessions were grouped according to their biome descriptions, sub-species, Ppd-D1 diversity and STRUCTURE sub-population membership, and sets drawn at random (Table S14).

Table 1 Effects on genetic, environmental, or geographic diversity in the sets of *Ae. tauschii* accessions selected by each approach compared to random selections

	Genetic variation			Climate variation Total variance	Sub-species	Ppd-D1	STRUCTURE sub-populations	Biome Description
	H	Heterozygosity	PIC					
CIMMYT SHW	ns	ns	ns	+	*	**	**	**
Selection within the geo-located accessions	De novo selection to							
Maximise genetic diversity	++	++	++	ns	*	**	**	ns
Maximise geographic diversity	–	–	–	++	ns	ns	ns	ns
Maximise climate diversity	ns	ns	ns	++	ns	ns	ns	ns
Composite of diversity selections	ns	ns	ns	++	ns	ns	ns	*
Selections to complement CIMMYT-SHW								
Maximise genetic diversity	ns	ns	ns	ns	ns	ns	ns	ns
Maximise geographic diversity	ns	ns	ns	ns	ns	ns	ns	ns
Maximise climate diversity	–	–	–	++	ns	ns	ns	*
Composite of diversity selections	ns	ns	ns	++	ns	ns	ns	**
Targeted selection by climate								
Cambridge/Tours	++	++	++	–	**	**	**	**
Osijek/Novi Sad	+	+	+	–	ns	ns	ns	**
Targeted selection by								
Desert and Xeric shrubland	ns	ns	ns	–	ns	ns	ns	**
Montane grasslands	ns	ns	ns	++	ns	ns	ns	**
Temperate broadleaf forests	++	++	++	ns	**	**	**	**
Temperate grasslands	ns	ns	ns	–	ns	ns	ns	**
Targeted selection by Ppd diversity								
Ppd allele 415	ns	ns	ns	ns	**	**	**	**
Ppd allele 430	ns	ns	ns	ns	**	**	**	*
Ppd allele 454	–	–	–	ns	ns	ns	ns	ns
Targeted selection by subspecies diversity								
ssp <i>stragulata</i>	++	++	++	ns	**	**	**	*
ssp <i>tauschii</i>	ns	ns	ns	ns	*	ns	ns	*
Targeted selection by sub population membership								
Sub-population II	–	–	–	ns	*	ns	**	ns
Sub-population III	++	++	++	ns	**	**	**	**

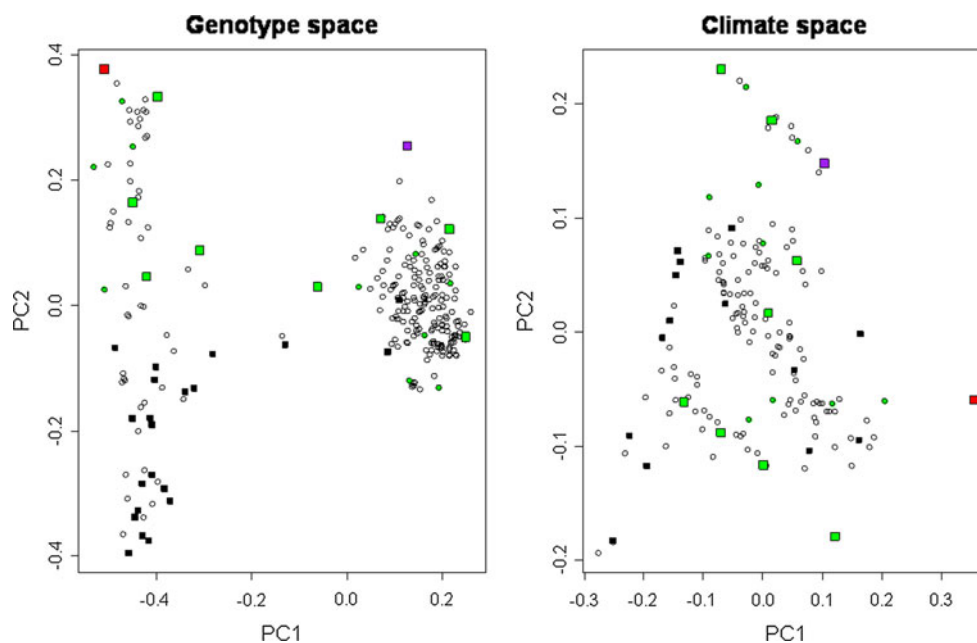
Selections used three strategies: selecting accessions to maximise geographic, genetic or climate diversity either de novo or to complement the CIMMYT SHW; selection targeting climates most similar to that in the region of interest to a breeding programme or selection according the biome description at the accessions' point of origin. For genetic and climate variation diversity is reported as similar (no significant difference: ns), lower (–) or higher (+) by comparison to randomly selected accessions. For category data variation results are classed as similar (no significant difference: ns) or skewed (*) by comparison to randomly selected accessions. Significant results ($p < 0.05$: *, – or +), highly significant results ($p < 0.01$: **, – or ++)

Data analysis

Statistical analyses were carried out in Excel or within the R-statistics package. Population structure was determined using STRUCTURE software (Pritchard et al. 2000) with a burn-in of 200,000 followed by 1,500,000 Markov Chain Monte Carlo iterations. The haploid setting and admixture model for ancestry between individuals were chosen, a degree of admixture being a reasonable expectation for populations that may have had opportunities for cross-pollination. Solutions for the number of sub-populations

identified were assessed using CorrSieve (Campana et al. 2010) and by calculating Akaike's information criterion (Sakamoto et al. 1986) for logistic regression of the *Ppd-D1* allele score or the sub-species designations against the STRUCTURE Q-matrices. Genetic distances (Rogers Distance) and summary statistics for the genotypic data were calculated in PowerMarker (Liu and Muse 2005) and calculated in R using the 'summary.genotype' function from the 'genetics' package. Environmental distances were calculated in R using the 'daisy' function from the 'cluster' package, using Gower's distance. Principle component

Fig. 1 Identification of novel diversity among *Ae. tauschii* accessions by successively selecting the ‘furthest’ accession within genetic diversity space. Each previously exploited D-genome donor for the CIMMYT-SHW lines is shown as a *black square*. The first selection from *Ae. tauschii* accessions is shown as a *red square*, the second in *mauve*, the following eight as *green squares*, the second ten as *green circles*



analysis was carried out in R using the ‘pco’ function from the ‘labdsv’ package. Calculations for geometric distances were carried out in R using the ‘sp’ package.

Results

By cross-referencing passport data available from the germplasm collections, duplication could be seen among entries from differing sources. Among the full collection of 429 *Ae. tauschii* accessions, 26 pairs were shared between germplasm held at the Vavilov Institute and IPK at Gaterslaben, seven pairs were shared between the Vavilov Institute and the USDA ARS GRIN collection, five pairs were shared between the IPK and the USDA ARS GRIN, one entry shared between ICARDA, IPK and the Vavilov Institute.

Distribution of sub-species amongst the accessions

Sub-species designations were available from passport data for 136 *Ae. tauschii* accessions from the accession set and, by cross referencing, 24 D-genome donors of the CIMMYT-SHW lines. The phenotypes of a further 107 *Ae. tauschii* accessions were examined and 17 were classified as *ssp. strangulata* based on seed shape, glume shape and spiculation on the inner lateral nerves, where 90 were classified as *ssp. tauschii*. Taking the passport data and the results of phenotypic classification together, the *Ae. tauschii* accessions comprised 199 *Ae. tauschii ssp. tauschii* and 44 *Ae. tauschii ssp. strangulata* while among the D-genome donors of the CIMMYT-SHW lines we were

able to identify 15 *ssp. tauschii* and nine *ssp. strangulata*. (Table S5). The composition of sub-species between these two groups differs significantly (χ^2 test, $p < 0.05$) with *ssp. tauschii* being less represented in the SHW lines.

Genetic diversity amongst the accessions

We genotyped the collection of plant material with a suite of 15 microsatellites, each displaying variability among the *Ae. tauschii* accessions and the bread wheat varieties. The number of alleles observed per locus ranged between 5–33 and 2–13 in each group, respectively (Table 2). Across all loci, the total number of alleles was highest for *Ae. tauschii* (252), lower for SHW (154), and lowest for bread wheat (79). Average gene diversity index (H) calculated across all markers for each germplasm pool was 0.56, 0.80 and 0.72 for bread wheat, synthetic wheat and *Ae. tauschii*, respectively, and 0.68 and 0.83 for *ssp. tauschii* and *ssp. strangulata*, respectively. The D-genome diversity among the D-genome donors of the CIMMYT-SHW lines and *Ae. tauschii* accessions was similar. The average D-genome gene diversity of bread wheat germplasm was similar in accessions originating in western or eastern Europe, although three loci that were variable among the eastern European set were monomorphic among the western European varieties. The greater diversity observed among the combined D-genome donors of the CIMMYT-SHW lines and *Ae. tauschii* accessions compared to bread wheat varieties was consistent with a reduction of D-genome diversity during the hybridisation event that gave rise to bread wheat. Analysis of the distribution of SSR diversity within germplasm groups identified 93 alleles that were

Table 2 Summary of the D-genome diversity revealed by genotyping within the study material

	Number	PIC	Gene diversity	Heterozygosity
SSR diversity				
D-genome donor CIMMYT-SHW lines	48	0.78 (0.03)	0.80 (0.02)	0.34 (0.10)
<i>Ae. tauschii</i> accessions	255	0.70 (0.05)	0.72 (0.05)	0.23 (0.08)
<i>Ae. tauschii</i> summary	303	0.75 (0.04)	0.76 (0.04)	0.24 (0.08)
<i>Ae. tauschii</i> ssp. <i>tauschii</i>	212	0.66 (0.06)	0.68 (0.06)	0.24 (0.09)
<i>Ae. tauschii</i> ssp. <i>strangulata</i>	53	0.81 (0.02)	0.83 (0.02)	0.20 (0.06)
TriticeaeGenome association panel	89	0.40 (0.07)	0.45 (0.07)	0.28 (0.11)
Eastern European diversity study	77	0.41 (0.05)	0.45 (0.05)	0.25 (0.10)
Bread wheat summary	166	0.51(0.05)	0.56 (0.05)	0.27 (0.09)
All accessions	469	0.80 (0.02)	0.82 (0.02)	0.25 (0.07)
SNP diversity				
<i>Ae. tauschii</i> accessions	232	0.19 (0.01)	0.22 (0.02)	0.01 (0.00)
<i>Ae. tauschii</i> ssp. <i>tauschii</i>	182	0.12 (0.01)	0.13 (0.02)	0.00 (0.00)
<i>Ae. tauschii</i> ssp. <i>strangulata</i>	40	0.30 (0.01)	0.38 (0.02)	0.03 (0.00)

Averages across all markers are given for PIC, Gene Diversity and Heterozygosity, with standard errors shown in brackets

unique to the *Ae. tauschii* collection and not found in either the bread or synthetic wheat germplasm sampled in this study. Only seven alleles were unique to the CIMMYT-SHW and not found in the *Ae. tauschii* sample. There were, however, 12 alleles that were unique to bread wheat.

We genotyped a subset of the *Ae. tauschii* collection with a suite of 62 D-genome specific SNP and found 43 loci to be polymorphic. Average gene diversity index (H) calculated across all markers for each germplasm pool was 0.22 for *Ae. tauschii*, and 0.13 and 0.38 for ssp. *tauschii* and ssp. *strangulata*, respectively.

Population structure and genetic diversity

To determine population structure, we examined solutions giving a number of sub-populations between one and nine using the STRUCTURE software to dissect microsatellite diversity. A plot of log likelihood $\ln(\text{PD})$ against the number of sub-populations (k) suggested the value of k was between three and nine, while the plot of $\delta(\ln(\text{PD}))$ indicated optimum solutions at three and at six or seven sub-populations. Comparison of Akaike's information content for the regression of STRUCTURE Q-matrices against Ppd-D1 genotypes and sub-species both suggest three or seven sub-populations (Figure S1). For every STRUCTURE solution, the bread wheat varieties were grouped in a single, exclusive sub-population (Table S6). The D-genome donors of the CIMMYT-SHW lines and *Ae. tauschii* accessions partition between the remaining groups at $k = 3$ and $k = 7$ and there was a hierarchical relationship between II at $k = 3$ and B and C at $k = 7$ and between III at $k = 3$ and D-G at $k = 7$. Among the accessions in sub-population II, ssp. *tauschii* predominates while both sub-species are seen in similar proportions in sub populations III. The allele-frequency divergence among sub-

populations (Kullback–Leibler distance), computed within STRUCTURE suggests that sub-population III (at $k = 3$) is most closely related to bread wheat (sub-population I) and sub-populations D and E (at $k = 7$) have the closest relationship to bread wheat (Supplementary Figure S15, Supplementary Table S15).

The gross structure revealed by the microsatellite data, partitioning *Ae. tauschii* accessions into two sub-populations (II and III), was confirmed when we used STRUCTURE software to dissect SNP diversity. The more subtle population structures revealed by the SSR data at $k = 7$ were not confirmed by further dissection of the SNP diversity. Correlations between genetic distances calculated using the SSR data and KASPar data were high and positive (0.84). Inspection of PCO plots of genetic diversity, calculated using the SSR data, and KASPar data suggests there is reduction in diversity within sub-population II when compared with sub-population III when KASPar data are used in place of SSR data (Supplementary Figure S16). This observation is confirmed when genetic diversity (H) and polymorphism information content (PIC) values are calculated (SSR data: sub-population II $H = 0.585$, $\text{PIC} = 0.565$, sub-population III $H = 0.795$, $\text{PIC} = 0.763$; KASPar data: sub-population II $H = 0.087$, $\text{PIC} = 0.075$, sub-population III $H = 0.365$, $\text{PIC} = 0.288$). The reduction in diversity among one sub-population relative to another may be due to ascertainment bias in the D-genome specific SNPs. The KASPar SNPs were discovered in a panel of elite bread wheat varieties and these varieties are more closely related to sub-population III than sub-population II; in this case, an apparent reduction in diversity among sub-population II is unsurprising (Morgante et al. 2010).

Frequency based distance measures separated the bread wheat varieties from north western Europe from those

originating in south eastern Europe. The two groups formed a distinct clade among the *Ae. tauschii* accessions suggesting they share a common D-genome ancestry (Fig. S2).

Three haplotypes at the *Ppd-D1* locus were observed in *Ae. tauschii*; a 415 bp product amplified from the intact sequence was common to *Ae. tauschii*, SHW and bread wheat samples and two additional variants (430 and 454 bp) that were found only in *Ae. tauschii* and SHW (Table S7). A 297 bp product, associated with early flowering due to a ~ 2 kb deletion in the promoter, was only seen among bread wheat varieties, suggesting that this mutation is a post-domestication event. Bentley et al. (2011) previously reported that the photoperiod insensitive *Ppd-A1a* mutation also arose by mutation after domestication. The distribution of *Ppd-D1* alleles among the D-genome donors of the CIMMYT-SHW lines and *Ae. tauschii* accessions differs significantly (χ^2 test, $p < 0.0001$). The geographic distribution of *Ppd-D1* diversity reveals all three alleles present in accessions originating in the Caucasus and the southern shore of the Caspian Sea, while a 454 bp allele predominates in the eastern range of *Ae. tauschii* (Fig. S3). The *Ppd-D1* alleles with a 415 and 297 bp amplification products were seen among both groups of bread wheat varieties, though the 415 bp allele predominated in north western varieties and the 297 bp allele was most frequent among south eastern bread wheat lines.

Geographic distribution of diversity

Cross-referencing between the information supplied with the CIMMYT-SHW lines, passport data for *Ae. tauschii* accessions, and geographic databases enabled us to allocate geographic coordinates to 21 of the D-genome donors of the CIMMYT-SHW lines and to all but one of the genotyped *Ae. tauschii* accessions. These data allowed us to allocate climate proxies from the CRU CL 2.0 dataset, augmented by additional calculated data to each germplasm entry (Table S8), and to query the ArcGIS 10.0 (ESRI 2011) spatial datasets to describe the biome at each entry's point of origin (Table 3; Table S9). The range of biomes described at the collection points for the *Ae. tauschii* accessions and the D-genome donors of the CIMMYT-SHW lines differ significantly (χ^2 test, $p < 0.01$).

The STRUCTURE sub-populations revealed by SSR diversity at $k = 3$, sub-population II and III, have different geographic distributions (Fig. 2, upper map). Sub-population II is seen in both the eastern and western part of *Ae. tauschii*'s natural range, being widely distributed across the highland regions of central Asia and in the Caucasus. Sub-population III has a more limited distribution, predominating along the shore of the Caspian Sea and it is also seen

in the Caucasus. The sub-populations identified at $k = 7$ follow a similar pattern (Fig. 2, lower map) with sub-populations B and C, derived from sub-population II at $k = 3$, occurring together across the range. The four sub-populations (D-G) derived from sub-population III at $k = 3$, show a split in their distribution with D and E largely confined to the shore of the Caspian Sea and F is represented in the Caucasus. Sub-population G has a broad distribution. These differing geographic distributions are reflected in the assigned biomes for accessions from each sub-population (Table 1) and these differences are significant at $k = 3$ and $k = 7$ (χ^2 test, $p < 0.001$). Differences can also be seen in the mean elevation among the sub-populations at $k = 3$ (t test, $p < 0.001$) and at $k = 7$ between D and E and all other subpopulations (t test, $p < 0.01$).

Strategies for selecting germplasm

We have modelled selection strategies using differing criteria including selections intended to maximise genetic, environmental or geographic diversity, selections made for similarity to the climates in areas of interest to our breeding programmes and selections representing the ecology of the natural range of *Ae. tauschii*. We evaluate these selection strategies by examining the characteristics of each selection by comparison with randomly selected sets.

De novo selection to maximise geographic, genetic or environmental variation

Our initial strategy was to select *Ae. tauschii* accessions on the basis of distance based measures in the expectation of selecting accessions that maximise diversity. We tested the expectation that the three distance measures (geographic, genetic, and climatic) would be strongly correlated and found positive but low correlations in every case (genetic vs. geographic distance: $r = 0.07$, genetic vs. climatic distance: $r = 0.15$, climatic vs. geographic distance: $r = 0.35$, p value < 0.001 in each case). These results suggest that different accessions would be identified as representing novel diversity by using each distance measure in turn (Fig. 3).

De novo sampling to maximise coverage in genetic space, as may be expected, increased genetic diversity and showed bias in the representation of *Ppd-D1* diversity, STRUCTURE sub-population and sub-species diversity but was representative with regard to variance within the climate data. A de novo selection to maximise representation of geographic variation showed highly significant increases in the total variance of the climate data at the expense of genetic diversity. The second observation may be a consequence of distribution of genetic variability, with

Table 3 Organisation of sub-species, genetic and environmental diversity among D-genome donor of the CIMMYT-SHW and *Ae. tauschii* accessions

	<i>k</i> = 3			<i>k</i> = 7						Sub-species		<i>Ppd-D1</i> diversity		
	All	II	III	B	C	D	E	F	G	<i>strangulata</i>	<i>tauschii</i>	415	430	454
D-genome donor CIMMYT SHW														
Temperate Savannas grass and shrub-lands	3	1	2	0	1	0	2	0	0	1	2	1	0	2
Desert and Xeric shrublands	6	2	4	0	2	2	2	0	0	3	3	3	1	2
Temperate broadleaf and mixed forests	10	0	10	0	0	7	3	0	0	4	6	5	2	3
Temperate Conifer forests	2	0	2	0	0	0	2	0	0	0	2	0	0	2
Montane grasslands and shrublands	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<i>Ae. tauschii</i> accessions														
Temperate Savannas grass and shrub-lands	87	77	10	23	54	0	1	6	3	5	74	5	3	74
Desert and Xeric shrublands	82	68	14	11	57	2	0	8	4	17	64	6	6	65
Temperate broadleaf and mixed forests	39	14	25	6	8	8	5	10	2	18	20	17	3	16
Temperate Conifer forests	5	5	0	0	5	0	0	0	0	0	5	0	0	5
Montane grasslands and shrublands	40	32	8	8	24	0	0	2	6	4	35	2	4	33
<i>Ae. tauschii</i> summary														
Temperate Savannas grass and shrub-lands	90	78	12	23	55	0	3	6	3	6	76	6	3	76
Desert and Xeric shrublands	88	70	18	11	59	4	2	8	4	20	67	9	7	67
Temperate broadleaf and mixed forests	49	14	35	6	8	15	8	10	2	22	26	22	5	19
Temperate conifer forests	7	5	2	0	5	0	2	0	0	0	7	0	0	7
Montane grasslands and shrublands	40	32	8	8	24	0	0	2	6	4	35	2	4	33
D-genome donor CIMMYT SHW	48	12	36	0	12	19	17	0	0	9	15	19	6	23
<i>Ae. tauschii</i> accessions	254	197	58	48	149	10	6	27	15	44	199	31	16	195
<i>Ae. tauschii</i> summary	302	209	94	48	161	29	23	27	15	53	214	50	22	218
Sub-species														
<i>Ae. tauschii</i> ssp. <i>strangulata</i>	53	13	40	5	8	14	5	14	7	–	–	22	12	15
<i>Ae. tauschii</i> ssp. <i>tauschii</i>	214	175	39	39	136	6	12	13	8	–	–	18	8	179
<i>Ppd-D1</i> diversity														
415	50	1	49	0	1	17	7	21	4	22	18	–	–	–
430	22	0	22	0	0	10	0	5	7	12	8	–	–	–
454	218	200	18	44	155	2	14	0	2	15	179	–	–	–

The dissection of diversity within the biomes is based on a set of accessions where geographic coordinates were available. The dissection of all other diversity uses all genotyped accessions

greater variation seen in a relatively small region in the Caucasus and the southern Caspian compared with less variation with central Asia, a large region. Accessions selected to maximise representation of climate variation also showed highly significant increases in the total variance of the climate data but gave unbiased samples with respect to genetic diversity, distribution of sub-species, *Ppd-D1* alleles, STRUCTURE sub-populations and biome descriptions. A composite sample made up of the ‘top-slice’ of the sets selected by reference to genetic, geographic and climate diversity gave a balanced sample with respect to genetic diversity, distribution of sub-species, *Ppd-D1* alleles and STRUCTURE sub-populations, though climate variance was increased and gave a skewed representation of biome distributions.

Selection to complement variation within the CIMMYT-SHW

The CIMMYT-SHW deviated significantly when compared with randomly drawn samples, with increases in the total variance of climate and the χ^2 statistic for sub-species distributions, sub-population membership, *Ppd-D1* variation and biome origin. Sets chosen to complement the CIMMYT-SHW, while maximising genetic or geographic distances, were comparable with randomly drawn samples. A set chosen to complement the CIMMYT-SHW, while maximising climatic distance, showed a decrease in genetic diversity when compared to randomly drawn samples.

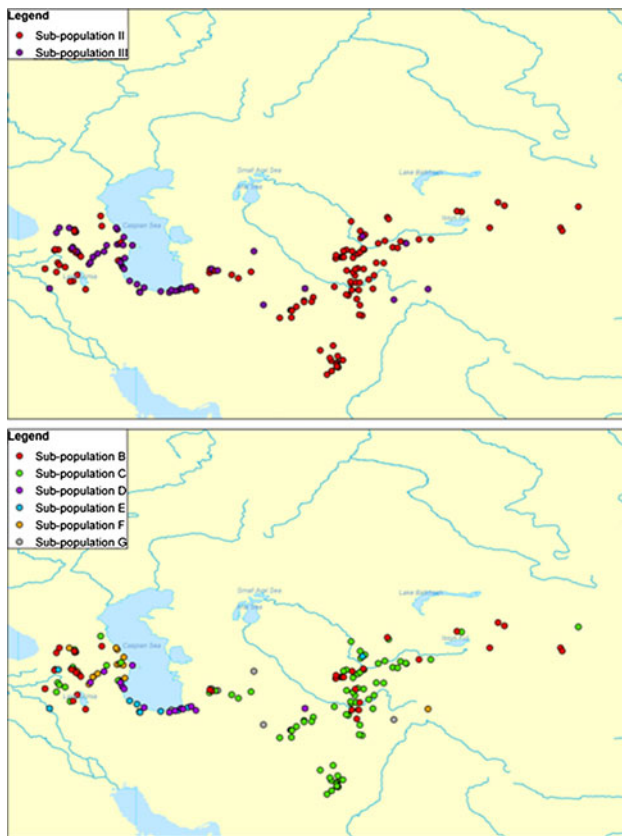


Fig. 2 Geographic distributions of sub-populations determined from D-genome specific SSR genotypes among the *Ae. tauschii* accessions at the optimum STRUCTURE solutions: number of sub-populations $k = 3$ (upper map) and $k = 7$ (lower map)

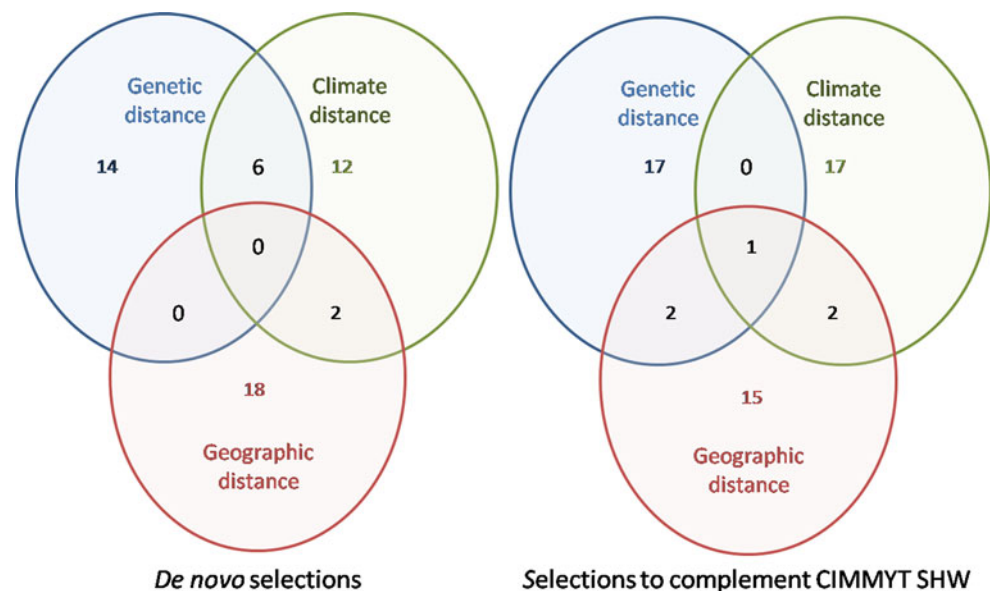
Targeted selection for similarity to the climates in areas of interest to our breeding programmes

When we implemented an alternative strategy which targeted accessions originating in those locations with a climate similar to our target breeding environments [Danube basin (Osijek and Novi Sad) and north western Europe (Cambridge and Tours)] we found increases in genetic diversity and a highly significant decreases in the total variance of the climate data. High values for the χ^2 statistic showed over representation of the ‘Temperate Broadleaf and Mixed Forests’ biome, sub-population III in both sets and a highly significant bias in sampling sub-species, STRUCTURE sub-populations and Ppd-D1 diversity in the set selected for its similarity to north western Europe (Cambridge and Tours).

Selecting under-represented germplasm

The results presented above suggest a strategy for sampling *Ae. tauschii* accessions for inclusion in a breeding programme where groups under-represented in the CIMMYT-SHW should be sampled preferentially in the expectation this will bring novel functional diversity into play. These under-represented groups include *Ae. tauschii* accessions classified as *ssp. tauschii*, belonging to sub-population II with the 454 bp *Ppd-D1* allele, and those assigned to biomes described as ‘Temperate Grasslands, Savannas and Shrublands’ and ‘Montane Grasslands and Shrublands’.

Fig. 3 Low correlations between the three distance measures (geographic, genetic, and climatic) result in different selections when samples are drawn to maximise each facet of diversity



Selecting germplasm by reference to the biome at its point of origin

Selections taken from accessions three of the six biomes represented within *Ae. tauschii*'s natural range ('Desert and Xeric Shrubland', 'Temperate Grasslands, Savannas and Shrublands' and 'Montane Grasslands and Shrublands') showed unbiased sampling with respect to genetic diversity, distribution of sub-species, *Ppd-D1* alleles and STRUCTURE sub-populations. Samples drawn from the 'Desert and Xeric Shrubland' and 'Temperate Grasslands, Savannas and Shrublands' biomes show a decrease in climate variance while accessions originating in the 'Montane Grasslands and Shrublands' biome show increased variation among climate measures. When accessions were chosen from among those originating in the 'Temperate Broadleaf and Mixed Forests' biome, high values for the χ^2 statistic showed highly significant bias in sampling the distribution of sub-species, *Ppd-D1* alleles and STRUCTURE sub-populations. Sub-species *strangulata*, the 415 bp allele at *Ppd-D1*, sub-population III were each over represented among the selected set and genetic diversity was elevated. There were too few accessions originating in the 'Temperate Conifer Forests' (7) or 'Mediterranean Forests, Woodland and Scrub' (1) biomes to conduct comparable random sampling.

Targeting selection according to *Ppd-D1* diversity:

Selections taken from accessions harbouring *Ppd-D1* 415 and 430 bp alleles showed unbiased sampling with respect to genetic diversity and climatic diversity the distribution of sub-species is skewed in favour of *ssp. strangulata*, STRUCTURE sub-population III is over-represented and, among the biomes, representation of 'Temperate Grasslands' is reduced while representation of 'Temperate Broadleaf Forests' is increased. The selection taken from accessions harbouring *Ppd-D1* 454 bp allele shows reduced genetic diversity but shows unbiased sampling in all other respects.

Targeting selection by sub-species

The selection made within *ssp. strangulata* shows an increase in genetic diversity and the distribution of *Ppd-D1* diversity is skewed in favour of the 415 and 430 bp alleles, STRUCTURE sub-populations III is over-represented and, among the biomes, representation of 'Temperate Broadleaf Forests' is increased. The selection targeting *ssp. tauschii* showed unbiased sampling apart from among the biomes where representation of 'Temperate Grasslands, Savannas and Shrublands' is increased.

Using STRUCTURE sub-populations as a guide to sampling

Genetic diversity is decreased in the selection made within STRUCTURE sub-population II and the distribution of sub species diversity is skewed in favour *ssp. tauschii*. The STRUCTURE sub-populations III selection shows increased genetic diversity and, among the biomes, representation of 'Temperate Broadleaf Forests' was increased, the proportion of *ssp. strangulata* was increased and the frequency of the *Ppd-D1* 454 bp allele was decreased.

How do selections made using SSR or KASPar data differ?

De novo selections within the reduced set of accessions that had both SSR and KASPar data available were made to maximise genetic diversity. A set selected to maximise KASPar diversity showed a highly significant increase in SSR diversity while a set selected to maximise SSR diversity showed a significant increase in KASPar diversity as shown by *H* and PIC values, while the heterozygosity showed a highly significant increase. The selections made using each data set had only three accessions in common. The selections differed in their diversity with respect to their sub-species, *Ppd-D1* and STRUCTURE sub-population diversity and in their representation of the biomes but none of the differences were significant. When all other selection methods were applied to this reduced set of accessions the outcomes were broadly similar to the results of selection applied above.

Selection summary

We have shown that alternative sampling strategies result in selections with different characteristics. Using one set of measures, targeting genetic diversity or *Ppd-D1* diversity, for example, to make selections has consequences when other measures are considered. Clearly, selection to maximise one facet of diversity may produce unbalanced samples with respect to other aspects of diversity. Conversely, a more balanced sample resulted from 'top-slicing' each of the three sets selected to maximise genetic, geographic and climate diversity. The value of targeted sampling by climate or biome remains unproven though, as the CIMMYT synthetic programme matures, data is being collected that could determine the merit of such an approach.

Discussion

The variation within wild *Ae. tauschii* populations is an important resource for biologists and plant breeders

introducing novel D-genome diversity via wheat re-synthesis. The process of selecting among this exotic material for a plant breeding programme represents a considerable commitment of resources. The methods described here integrate passport data available for wild accessions with climate and ecological information from publically available sources and a limited set of genotype data. Genetic distances were used to guide selection for a new re-synthesis programme and then reviewed using the integrated data to ensure a balanced selection.

Organisation of diversity

Genetic variation within and between *Ae. tauschii* populations has been recently described by AFLP (Mizuno et al. 2010), DArT markers (Sohail et al. 2012), and SNP (Wang et al. 2013). In the current study, division of *Ae. tauschii* accessions into two sub-populations (II and III at $k = 3$) is similar to the distribution of lineages described by these authors and those previously described by Pestsova et al. (2000). Our finding that morphological diversity, described by the division between ssp. *tauschii* and ssp. *strangulata* is not clearly congruent with genetic division (Table 3) is in line with the findings of these authors. While *Ae. tauschii* ssp. *tauschii* predominates among the accessions of sub-population II it is not exclusive, both sub-species are seen in equal proportions within sub-population III. This remains the case when considering the sub-populations at $k = 7$, where both sub-species are seen among all sub-populations, though ssp. *tauschii* predominates among the accessions of sub-populations B and C. The low correlation between sub-species distributions and D-genome wide genetic diversity is mirrored by the relationship between sub-species and *Ppd-D1* diversity. Looking at the distribution of sub-species among the biomes, sub-species *tauschii* predominates in four out of five biomes, with the two sub-species seen in similar proportions within the 48 accessions originating in the ‘Temperate Broadleaf and Mixed Forests’ biome.

Genetic diversity shows a greater degree of organisation, both in the relationship between the whole genome population structure and the *Ppd-D1* diversity and in the relationship between the genetic diversity across each biome (Table 3). Within the sub-populations defined at $k = 3$, the 454 bp allele overwhelmingly predominates at *Ppd-D1* in sub-population II. Three *Ppd-D1* alleles were identified within sub-population III, including the 415 bp allele at a frequency of 0.55. Of these *Ppd-D1* alleles, only the 415 bp allele is found in bread wheat, suggesting that the D-genome progenitor of modern bread wheat belonged to sub-population III.

The Kullback–Leibler distances between sub-populations calculated by STRUCTURE offer further support to

the D-genome progenitor of bread wheat having its origin in sub-population III at $k = 3$ (Supplementary Table S16, Fig. S15). The same distance measures at $k = 7$ differentiate between the sub-populations and suggest sub-populations D or E as the D-genome donor. Eastern Turkey, Transcaucasia and the south eastern shores of the Caspian Sea in Iran have been suggested as the probable geographic location of the bread wheat foundation polyploidisation event (Dvorak et al. 1998b; Giles and Brown 2006; Matsuoka 2011; Wang et al. 2013). The distribution of the sub-populations at $k = 7$ shows a mix of members of sub-populations B, C, D, E, and F in eastern Turkey and Transcaucasia while members of sub-population D and E alone are found in the region around the southern shores of the Caspian Sea. This geographic distribution of sub-populations offers support to the Caspian region of Iran as the location of the bread wheat D-genome. The similarity between the north west European and south east European bread wheat D-genomes shows that domestication through polyploidisation captured a limited portion of *Ae. tauschii* diversity, suggesting that polyploidisation events were infrequent, supporting Giles and Brown (2006) suggestion that the bread wheat D-genome has a diphyletic rather than a multiphyletic origin.

Marker selection

In a comparison of SSR and SNP diversity in a panel of maize lines (Hamblin et al. 2007), SSRs offered better performance in clustering varieties into sub-populations and better resolution when calculating genetic distances. Ascertainment bias is a widely recognised problem and, as the SNP we used were discovered in a panel of bread-wheat varieties, bias may be a constraint in using the KASPar data for a crop wild relative. Our observation that SNP genetic diversity within sub-population II is very low relative to sub-population III, where sub-population II is less closely related to bread wheat, suggests that these KASPar markers are subject to ascertainment bias in our accession set. Ascertainment bias and lower resolution power among the SNP may account for the different selections made when using either SNP or SSRs. If genetic diversity is to be used as a guide to making selections from among crop wild relatives, care must be taken to ensure that the makers used are free of ascertainment bias and offer sufficient resolution to fully describe the structure of genetic diversity.

Does ecology influence geographic distributions of sub-populations?

The distribution of D-genome STRUCTURE sub-populations at $k = 3$ across the biomes of central Asia shows sub-

population III predominates in the ‘Temperate Broadleaf and Mixed Forests’ biome while sub-population II accounts for the majority of accessions originating within all other biomes. A similar distribution is seen at $k = 7$ with 79 and 53 % of accessions from sub-populations D and E, respectively. The distribution pattern of *Ppd-D1* alleles is similar with the 415 bp allele appearing as the most frequent allele within the ‘Temperate Broadleaf and Mixed Forests’ biome whereas the 454 bp allele is the majority allele in all other biomes. The distribution of sub-populations and *Ppd-D1* alleles within the biomes suggests that, at least in part, adaptation to environmental variation has played a role in the establishment of population substructure in *Ae. tauschii*.

These observations may also have significance in selecting accessions for deployment in plant breeding programmes. Many of the major wheat growing regions of the world; the northern European Plain, the Danube Basin, the American mid west, eastern China and south eastern Australia are classified as belonging to the ‘Temperate Broadleaf and Mixed Forests’ biome. Our data is consistent with the D-genome progenitor of bread wheat originating within the ‘Temperate Broadleaf and Mixed Forests’ biome and, as a consequence, it may have harboured adaptive variation that offers enhanced fitness within the environments included within that biome. Should that be the case, adaptation to environments within the ‘Temperate Broadleaf and Mixed Forests’ biome rather than adaptation to the more continental climates of central Asia (Feuillet et al. 2008) may have been a key factor allowing hexaploid wheat to flourish across such a broad range. This suggests sampling within the pool of accessions originating within the ‘Temperate Broadleaf and Mixed Forests’ biome as a strategy for plant breeders seeking adaptation to temperate environments. An alternative would be to consider the detail of the climate data collated for the point of origin for each accession. Using these data, it is apparent that there are some differences between the climate prevailing at the point of origin of sub-populations D and E compared to that for the other populations (Figs. S4–14). The minimum temperatures and humidity are higher, the precipitation is higher and the moisture deficit is lower during the later months in the year and the temperature is higher and frost frequency lower from autumn through to spring. Taken together, these and other differences result in many of the members of sub-populations D and E occupying a distinct region within climate space (Fig. 4). When locations representing the wheat growing regions of the Danube basin (Osijek and Novi Sad) and north western Europe (Cambridge and Tours) are plotted in the same space it can be seen that members of sub-populations D and E are among the accessions occupying a similar region within climate space (Fig. 4).

Using simple data sets in order to select germplasm for a breeding programme

In this analysis, we have selected simple data sets including high resolution climate data and ecological information in the form of biome data from publically available sources, genotypes from screening germplasm with 15 nuclear microsatellite markers and an indel marker associated with the photoperiod response gene *Ppd-D1*. In our case, the population structures we found are well supported by previous workers and recent publications; if this had not been the case, the robustness of genetic structures should be tested by permutation within the marker set. We have used the passport data (sub-species information), biome data, sub-populations calculated from the microsatellite data and the allele score at *Ppd-D1* to group accessions. The groupings obtained from our analyses have shown that the D-genome donors of the CIMMYT-SHW lines over represent some sub-groups within the *Ae. tauschii* accessions available in world wide germplasm collections and we have suggested sampling methods that could be used to address those imbalances. We have compared the climate and biome information available for the *Ae. tauschii* accessions and for the wheat growing regions targeted in our breeding programme and seen that the D-genome donors selected for the re-synthesis of the CIMMYT-SHW are those that might be predicted to harbour variation that would confer enhanced fitness within the temperate regions

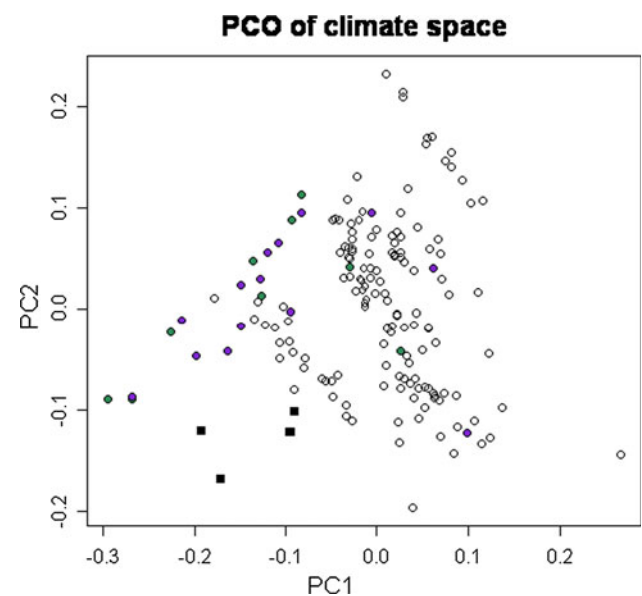


Fig. 4 Representation of climate variation at the geographic origin of *Ae. tauschii* accessions by principal component ordination. The climate at the origin of members of sub-populations D and E (at $k = 7$, plotted in green and purple, respectively) can be seen in relation to the climate at four locations chosen to represent European climates (black squares)

of Europe. Whether these selections would be optimal in other regions is a question that remains to be addressed.

The genetic distances among the sub-populations derived from the microsatellite data and the *Ppd-D1* genetic data suggest the original polyploidisation event giving rise to hexaploid bread wheat resulted from a hybridisation between tetraploid wheat and members of a sub-population of *Ae. tauschii* with a limited geographic and ecological range. This finding has implications for the history of early agriculture and may suggest the region of northern Iran on the coast of the Caspian Sea as a target for investigations to address the geography and chronology of the origins of bread wheat. Placing the origin of hexaploid wheat evolution in the part of *Ae. tauschii*'s range that is most similar ecologically and climatically to temperate Europe has implications for plant breeders. Breeders working to create varieties to be grown within temperate Europe might conclude that their search within wild progenitor germplasm should be confined to the same geographic and ecological range that we suggest the D-genome donor of bread wheat originated. Conversely, breeders working in other environments may use the methods described here to guide their search towards *Ae. tauschii* accessions originating in very different regions and environments. Against the background of climate change the methods described here could be used to select D-genome donors tailored to offer adaptation to a range of environments.

Concluding remarks

It should be possible to implement these methods in a range of species, using a combination of passport data, climate or ecological data and economical genetic data in any programme attempting to incorporate exotic germplasm. The selection of germplasm from exotic sources poses a major challenge. Plant breeders have to choose from collections that number in the hundreds or thousands. Where accessions are available from a broad geographic range, heterogeneity within the environment may result in pockets of genetic sub-structure that merit attention for their functional diversity, while broad regions with less environmental variation may result in widespread sub-populations where variation is simply the expression of dispersal and isolation by distance.

These approaches require an evaluation to assess which, if any of the sampling strategy discussed offers useful prediction of the breeding value of *Ae. tauschii* in a wheat re-synthesis programme. Since 2000, an international nursery of SHW derived from crosses of *T. turgidum* cultivars and *Ae. tauschii* accessions in the wide crosses working collection at CIMMYT have been deployed in

wheat breeding projects across the globe. Each breeding project will have introgressed SHW into locally adapted wheat germplasm and made selections on the progeny in subsequent generations. Each centre will have made selections in their local situation under conditions that could be described eco-geographically in terms of the biome or by climate data. It would be possible to quantify the breeding value of *Ae. tauschii* accessions in each environment by identifying the *Ae. tauschii* accessions in the pedigrees of selected lines. This would allow an exploration of the relationship between breeding value and climate data and biome description, sub-population membership, *Ppd-D1* diversity and sub-species membership among the *Ae. tauschii* accessions. Should clear relationships emerge from this meta-analysis we would be able to select those *Ae. tauschii* accessions most likely to offer novel diversity tailored to the environment in any regional breeding programme.

Acknowledgments The authors gratefully acknowledge the NIAB Trust and the UK Biotechnology and Biological Sciences Research Council (grants BB/E006868/1, BB/I002561/1, BB/I004596/1, BB/J004588/1) for funding this work. The following germplasm collections provided seed and passport data for the *Ae. tauschii* accessions, USDA-ARS Small Grains Collection, University of Tokyo, IPK Gatersleben, ICARDA, Kansas State University, John Innes Centre Germplasm Resources Unit and the Vavilov Institute. The authors wish to thank Jon Raupp of the Kansas State University, Steve Reader and Mike Ambrose of the John Innes Centre.

References

- Allen AM, Barker GLA, Berry ST, Coghill JA, Gwilliam R, Kirby S, Robinson P, Brenchley RC, D'Amore R, McKenzie N, Hall A, Bevan M, Hall N, Edwards KJ (2011) Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (*Triticum aestivum* L.). *Plant Biotechnol J* 9:1086–1099
- Allen AM, Barker GLA, Wilkinson P, Burrige A, Winfield M, Coghill J, Uauy C, Griffiths S, Jack P, Berry S, Werner P, Melichar JPE, McDougall J, Gwilliam R, Robinson P, Edwards KJ (2013) Discovery and development of exome-based, co-dominant single nucleotide polymorphism markers in hexaploid wheat (*Triticum aestivum* L.). *Plant Biotechnol J* 11(3):279–295. doi:10.1111/pbi.12009
- Beales J, Turner A, Griffiths S, Snape JW, Laurie DA (2007) A pseudo-response regulator is misexpressed in the photoperiod insensitive Ppd-D1a mutant of wheat (*Triticum aestivum* L.). *Theor Appl Genet* 115:721–733
- Bentley AR, Turner AS, Gosman N, Leigh FJ, Maccaferri M, Dreisigacker S, Greenland A, Laurie DA (2011) Frequency of photoperiod-insensitive Ppd-A1a alleles in tetraploid, hexaploid and synthetic hexaploid wheat germplasm. *Plant Breed* 130: 10–15
- Bhullar NK, Street K, Mackay M, Yahiaoui N, Kellera B (2009) Unlocking wheat genetic resources for the molecular identification of previously undescribed functional alleles at the Pm3 resistance locus. *Proc Natl Acad Sci USA* 106:9519–9524

- Brevis JC, Morris CF, Manthey F, Dubcovsky J (2010) Effect of the grain protein content locus *Gpc-B1* on bread and pasta quality. *J Cereal Sci* 51:357–365
- Campana MG, Hunt HV, Jones H, White J (2010) CorrSieve: software for summarizing and evaluating Structure output. *Mol Ecol Resour* 11:340–352
- Dubcovsky J, Dvorak J (2007) Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* 316:1862–1866
- Dudnikov AJ (2011) Chloroplast DNA non-coding sequences variation in *Aegilops tauschii* Coss.: evolutionary history of the species. *Genet Resour Crop Evol* 59:683–699
- Dvorak J, Zhang HK (1992) Reconstruction of the phylogeny of the genus *Triticum* from variation in repeated nucleotide sequences. *Theor Appl Genet* 84:419–429
- Dvorak J, Luo MC, Yang ZL, Zhang HB (1998a) The structure of the *Aegilops tauschii* genepool and the evolution of hexaploid wheat. *Theor Appl Genet* 97:657–670
- Dvorak J, Luo M-C, Yang Z-L (1998b) Genetic evidence on the origin of *Triticum aestivum* L. In: Damania AB, Valkoun J, Willcox G, Qualset CO (eds) The origins of agriculture and crop domestication (The Harlan Symposium). ICARDA, Aleppo, pp 235–251
- Endresen DTF (2010) Predictive association between trait data and ecogeographic data for Nordic Barley Landraces. *Crop Sci* 50:2418–2430
- Endresen DTF, Street K, Mackay M, Bari A, De Pauw E (2011) Predictive association between biotic stress traits and ecogeographic data for wheat and barley landraces. *Crop Sci* 51:2036–2055
- ESRI (2011) ArcGIS Desktop: Release 10. Environmental Systems Research Institute, Redlands
- Eujayl I, Sorrells M, Baum M, Wolters P, Powell W (2002) Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. *Theor Appl Genet* 104:399–407
- Feldman M, Lupton FGH, Miller TE (1995) Wheats. In: Smartt J, Simmonds NW (eds) Evolution of Crops. Longman Scientific, London, pp 184–192
- Feuillet C, Langridge P, Waugh R (2008) Cereal breeding takes a walk on the wild side. *Trends Genet* 24:24–32
- Fulton TM, Chunwongse J, Tanksley SD (1995) Microprep protocol for extraction of DNA from tomato and other herbaceous plants. *Plant Mol Biol Rep* 13:207–209
- Giles RJ, Brown TA (2006) *GluDy* allele variations in *Aegilops tauschii* and *Triticum aestivum*: implications for the origins of hexaploid wheats. *Theor Appl Genet* 112:1563–1572
- Gupta PK, Balyan HS, Edwards KJ, Isaac P, Korzun V, Röder M, Gautier M-F, Joudrier P, Schlatter AR, Dubcovsky J et al (2002) Genetic mapping of 66 new microsatellite (SSR) loci in bread wheat. *Theor Appl Genet* 105:413–422
- Hamblin MT, Warburton ML, Buckler ES (2007) Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. *PLoS One* 2:e1367. doi:10.1371/journal.pone.0001367
- Heywood V (2011) Introductory and Background Material. In: Hunter D, Heywood V (eds) Crop Wild Relatives: A Manual of in situ Conservation. London and Washington Earthscan in Association with Biodiversity International, pp 3–30. ISBN 9781849711791
- Hijmans RJ, Jacobs M, Bamberg JB, Spooner DM (2003) Frost tolerance in wild potato species: assessing the predictivity of taxonomic, geographic, and ecological factors. *Euphytica* 130:47–59
- Hübner S, Höffken M, Oren E, Haseneyer G, Stein N, Graner A, Schmid K, Fridman E (2009) Strong correlation of wild barley (*Hordeum spontaneum*) population and precipitation variation. *Mol Ecol* 18:1523–1536
- IUCN (1974) Uvardy MDF. A classification of the biogeographical provinces of the world. IUCN occasional paper No 18, International Union for Conservation of Nature, Gland, Switzerland
- Jiang J, Friebe B, Gill BS (1994) Recent advances in alien gene transfer in wheat. *Euphytica* 73:199–212
- Jones H, Lister DL, Bower MA, Leigh FJ, Smith LM, Jones MK (2008) Approaches and constraints of using existing landrace and extant plant material to understand agricultural spread in prehistory. *Plant Genet Resour* 6:98–112
- Jones H, Cíván P, Cockram J, Leigh FJ, Smith LMJ, Jones MK, Charles MP, Molina-Cano J-L, Powell W, Jones G, Brown TA (2011) Evolutionary history of barley cultivation in Europe revealed by genetic analysis of extant landraces. *BMC Evol Biol* 11:320. <http://www.biomedcentral.com/1471-2148/11/320>
- Kilian B, Graner A (2012) NGS technologies for analyzing germplasm diversity in genebanks. *Briefings Funct Genomics*. doi:10.1093/bfpg/blr046
- Kishii M, Delgado R, Rosas V, Cortes A, Cano S, Sanchez J, Mujeeb-Kazi A (2007) Exploitation of Genetic Resources through Wide Crosses. Proceedings International Symposium on wheat Yield Potential CIMMYT, Mexico
- Knaggs P, Ambrose MJ, Reader S, Miller TE (2000) Morphological characterization and evaluation of the subdivision of *Aegilops tauschii* Coss. *Wheat Inform Serv* 91:15–19
- Liu K, Muse SV (2005) Powermarker: integrated analysis environment for genetic marker data. *Bioinformatics* 21:2128–2129
- Matsuoka Y (2011) Evolution of polyploid triticum wheats under cultivation: the role of domestication, natural hybridization and allopolyploid speciation in their diversification. *Plant Cell Physiol* 52:750–764
- Maxted N, Kell SP (2009) Establishment of a global network for the in situ conservation of crop wild relatives: status and needs. FAO Commission on Genetic Resources for Food and Agriculture, Rome
- Maxted N, White K, Valkoun J, Konopka J, Hargreaves S (2008) Towards a conservation strategy for *Aegilops* species. *Plant Genetic Resour Charact Utilization* 6:126–141
- Mizuno N, Yamasaki M, Matsuoka Y, Kawahara T, Takumi S (2010) Population structure of wild wheat D-genome progenitor *Aegilops tauschii* Coss: implications for intraspecific lineage diversification and evolution of common wheat. *Mol Ecol* 19:999–1013
- Moragues M, Comadran J, Waugh R, Milne I, Flavell AJ, Russell JR (2010) Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. *Theor Appl Genet* 120:1525–1534
- Mujeeb-Kazi A, Gul A, Ahmad I, Farooq M, Rizwan S, Bux H, Iftikhar S, Asad S, Delgado R (2007) *Aegilops tauschii*, as a spot blotch (*Cochliobolus sativus*) resistance source for bread wheat improvement. *Pak J Bot* 39:1207–1216
- New M, Lister D, Hulme M, Makin I (2002) A high-resolution data set of surface climate over global land areas. *Clim Res* 21:1–25
- Peeters JP, Wilkes HG, Galwey NW (1990) The use of ecogeographical data in the exploitation of variation from gene banks. *Theor Appl Genet* 80:110–112
- Pestsova E, Korzun V, Goncharov NP, Hammer K, Ganal MW, Röder MS (2000) Microsatellite analysis of *Aegilops tauschii* germplasm. *Theor Appl Genet* 101:100–106
- Prada D (2009) Molecular population genetics and agronomic alleles in seed banks: searching for a needle in a haystack? *J Exp Bot* 60:2541–2552
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959
- Reynolds M, Trethowan RM (2007) Physiological interventions in breeding for adaptation to abiotic stress. In: Spiertz JHJ, Struik

- PC, Van Laar HH (eds) Scale and complexity in plant systems research, gene-plant-crop relations. Springer, The Netherlands
- Reynolds M, Dreccer F, Trethowan R (2007) Drought-adaptive traits derived from wheat wild relatives and landraces. *J Exp Bot* 58:177–186
- Röder MS, Korzun V, Wendehake K, Plaschke J, Tixier M-H, Leroy P, Ganal MW (1998) A microsatellite map of wheat. *Genetics* 149:2007–2023
- Sakamoto Y, Ishiguro M, Kitagawa G (1986) Akaike information criterion statistics. D. Reidel Publishing Company, Dordrecht
- Sohail Q, Shehzad T, Kilian A, Eltayeb AE, Tanaka H, Tsujimoto H (2012) Development of diversity array technology (DArT) markers for assessment of population structure and diversity in *Aegilops tauschii*. *Breed Sci* 62:38–45. doi:[10.1270/jsbbs.62.38](https://doi.org/10.1270/jsbbs.62.38)
- Song QJ, Fickus EW, Cregan PB (2002) Characteristics of trinucleotide markers in wheat. *Theor Appl Genet* 104:286–293
- Song QJ, Shi JR, Singh S, Fickus EW, Costa JM, Lewis J, Gill BS, Ward R, Cregan PB (2005) Development and mapping of microsatellite (SSR) markers in wheat. *Theor Appl Genet* 110:550–560
- Trethowan RM, Mujeeb-Kazi A (2008) Novel germplasm resources for improving environmental stress tolerance of hexaploid wheat. *Crop Sci* 48:1255–1265
- Vavilov NI (1957) World resources of cereals, leguminous seed crops and flax and their utilisation in plant breeding. The Academy of Sciences of the USSR. Moscow translated and published by the National Science Foundation, Washington (1960)
- Wang J, Luo MC, Chen Z, You FM, Wei Y, Zheng Y, Dvorak J (2013) *Aegilops tauschii* single nucleotide polymorphisms shed light on the origins of wheat D-genome genetic diversity and pinpoint the geographic origin of hexaploid wheat. *New Phytol.* doi:[10.1111/nph.12164](https://doi.org/10.1111/nph.12164)