ARTICLES
https://doi.org/10.1038/s41588-019-0382-2

nature genetics

Corrected: Publisher Correction

Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome

Fei He[1], Raj Pasam[2], Fan Shi[2], Surya Kant[2], Gabriel Keeble-Gagnere[2], Pippa Kay[2], Kerrie Forrest[2], Allan Fritz[3], Pierre Hucl[4], Krystalee Wiebe[4], Ron Knox[5], Richard Cuthbert[5], Curtis Pozniak[4], Alina Akhunova[1,6], Peter L. Morrell[7], John P. Davies[8], Steve R. Webb[8], German Spangenberg[2,9], Ben Hayes[2,10], Hans Daetwyler[2,9], Josquin Tibbits[2,9], Matthew Hayden[2,9]* and Eduard Akhunov[1]*

Introgression is a potential source of beneficial genetic diversity. The contribution of introgression to adaptive evolution and improvement of wheat as it was disseminated worldwide remains unknown. We used targeted re-sequencing of 890 diverse accessions of hexaploid and tetraploid wheat to identify wild-relative introgression. Introgression, and selection for improvement and environmental adaptation, each reduced deleterious allele burden. Introgression increased diversity genome wide and in regions harboring major agronomic genes, and contributed alleles explaining a substantial proportion of phenotypic variation. These results suggest that historic gene flow from wild relatives made a substantial contribution to the adaptive diversity of modern bread wheat.

Bread wheat is one of the major sources of calories and protein in the modern human diet. Its origin dates back to 8000 BC[1,2] and is attributed to a series of events that include: (1) the domestication of tetraploid wild emmer (*Triticum dicoccoides*) in the Fertile Crescent, which led to the origin of domesticated emmer wheat (*T. dicoccum*)[3,4], and (2) hybridization between free-threshing domesticated emmer with diploid goatgrass (*Aegilops tauschii*) from the southwestern Caspian region[5,6] to form hexaploid bread wheat (*T. aestivum*). Since its origin, bread wheat has spread across the world, reaching Europe around 6000 BC[7] and China around 2600 BC[8]. This dissemination required adaptation to new environmental conditions and agricultural practices quite distinct from those at the site of origin.

Wheat diversity has been affected by genetic bottlenecks caused by domestication and polyploidization[9,10]. A polyploidization-related bottleneck resulted in substantial loss of genetic diversity in the D genome of hexaploid wheat compared with its diploid ancestor[10]. A possible explanation for the lack of similar diversity reduction in the A and B genomes is gene flow from tetraploid wild emmer through pentaploid hybrids that can be produced from crosses between hexaploid and tetraploid wheat[11,12]. A study of introgression between sympatric populations of wild emmer and wheat at a single locus provides evidence for this hypothesis[13]. However, the scope of gene flow at the genome-wide level and its impact on adaptive evolution in wheat remained unexplored. The potential for adaptive introgression has been widely observed in plant populations[14–18]. Thus, identification of wild-relative introgression in the wheat genome can help to understand the role of ancestral diversity in defining the genetic diversity of modern wheat and evaluate its contribution to local adaptation.

Here, we used a reference wheat genome IWGSC RefSeq v.1.0 (ref. [19]) to generate a haplotype map on the basis of targeted re-sequencing[20] of 890 diverse wheat landraces and cultivars, and tetraploid wild and domesticated relatives to identify genomic regions showing the signals of introgression from wild emmer. By analyzing the distribution of SNPs relative to geography, historic environmental variables and improvement status, we sought to assess the contribution of introgression to local adaptation and crop improvement, and to evaluate the effects of these factors on deleterious allele burden in wheat. By partitioning genetic variance for major agronomic traits between the genomic regions with high and low incidence of introgression, we assessed the impact of historic gene flow on the phenotypic diversity of modern wheat.

Results

Population structure and genetic differentiation. The panel of geographically diverse wheat cultivars and landraces was re-sequenced using the sequence capture assay[20] (Fig. 1a,b and Supplementary Table 1), resulting in identification of about 7.3 million SNPs. A total of 3,573,809 filtered SNPs with minor allele frequency (MAF) ≥ 0.002 and missing genotype calls <25% were used in the study. This included 375,079 (10.5%) non-synonymous (nSNPs) and 49,233 (1.4%) potentially deleterious SNPs (dSNPs) that can negatively affect plant fitness[21,22] (Supplementary Table 2).

[1]Department of Plant Pathology, Kansas State University, Manhattan, KS, USA. [2]Agriculture Victoria, AgriBio, Centre for AgriBioscience, Bundoora, Victoria, Australia. [3]Department of Agronomy, Kansas State University, Manhattan, KS, USA. [4]Crop Development Centre, University of Saskatchewan, Saskatoon, Saskatchewan, Canada. [5]Swift Current Research and Development Centre, Swift Current, Saskatchewan, Canada. [6]Integrated Genomics Facility, Kansas State University, Manhattan, KS, USA. [7]Department of Agronomy and Plant Genetics, University of Minnesota, St Paul, MN, USA. [8]Corteva Agriscience, Agriculture Division of DowDuPont, Indianapolis, IN, USA. [9]School of Applied Systems Biology, La Trobe University, Bundoora, Victoria, Australia. [10]Present address: Queensland Alliance for Agriculture and Food Innovation, Centre for Animal Science, University of Queensland, St Lucia, Queensland, Australia. *e-mail: matthew.hayden@ecodev.vic.gov.au; eakhunov@ksu.edu

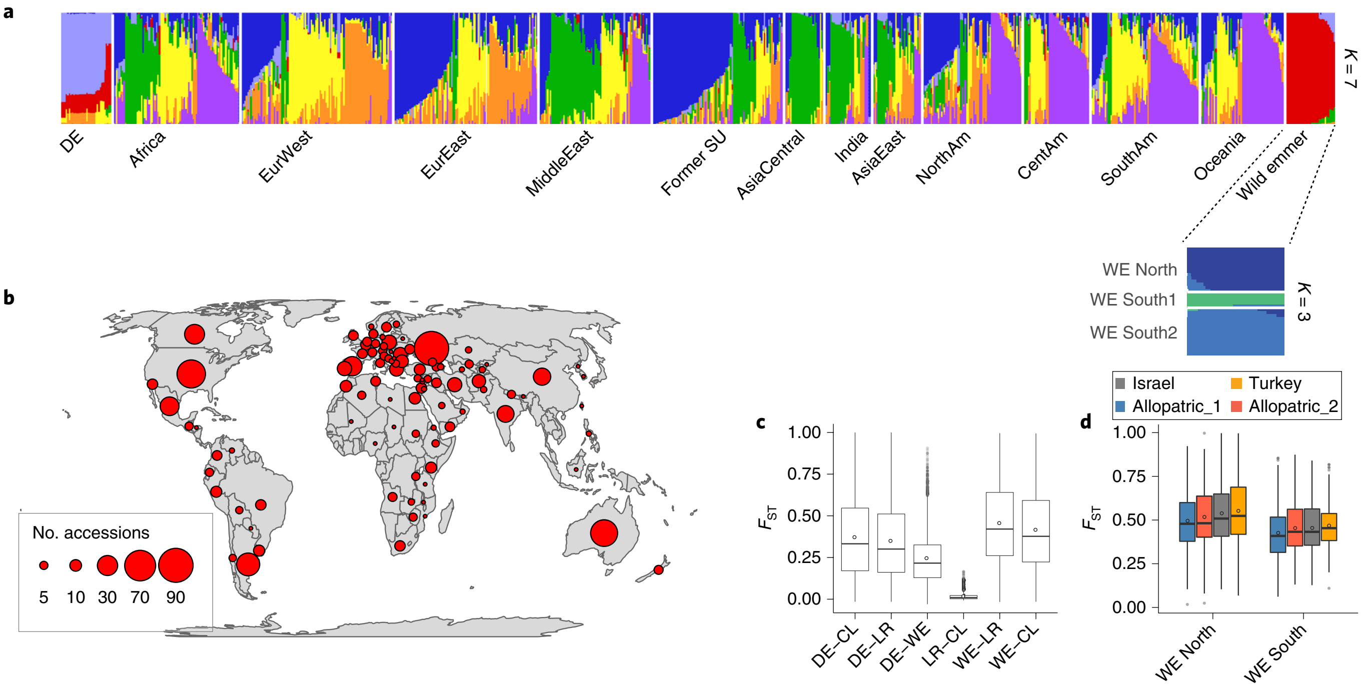

Fig. 1 | Population structure and genetic differentiation of wild and domesticated emmer, and bread wheat. a, Population structure of wheat populations combined with wild (WE) and domesticated emmer (DE) was inferred by assuming seven clusters (K) to illustrate admixture between wild and domesticated wheat. At $K=3$, wild emmer is separated into three subpopulations: North (Turkey, Iraq, Lebanon), South 1 (Central and Northern Israel) and South 2 (Southern Israel, Syria, Lebanon). **b**, Geographic distribution of wheat accessions by country of origin. Bubble sizes are proportional to the sample sizes. **c**,**d**, Box-plots were used to compare F_{ST} between populations in **c** and **d**. Box shows the median and interquartile ranges (IQR). The end of the top line is the maximum or the third quartile (Q) + 1.5× IQR. The end of the bottom line denotes either the minimum or the first Q − 1.5× IQR. The dots are either more than third Q + 1.5× IQR or less than first Q − 1.5× IQR. **c**, F_{ST} (2-Mb windows) among WE, DE, wheat cultivars (CL) and landraces (LR) (ANOVA $F=2{,}957$, d.f. = 5, $P=10^{-16}$). Means are shown by circles. All comparisons were significant at adjusted P value × 10^{-6}. **d**, Distribution of genetic differentiation estimates (F_{ST}) in 2-Mb windows between the sympatric and allopatric populations of wheat landraces and two wild emmer populations from the northern (Turkey) and southern (Israel, Syria) portions of the species' range. Allopatric_1 includes landraces from Azerbaijan, Armenia and Georgia in the putative region of bread wheat origin in Transcaucasia. Allopatric_2 includes landraces from Turkmenistan and Uzbekistan.

This SNP dataset was merged with the previously identified SNPs from wild and domesticated emmer[23] resulting in 348,372 SNPs from the A and B genomes. In the combined dataset, wild emmer shared 226,677, 255,615 and 249,246 SNPs with domesticated emmer, wheat cultivars and landraces, respectively.

Genetic assignment analysis with ADMIXTURE[24] in hexaploid wheat showed an optimal value of $K=11$ and high levels of interpopulation admixture consistent with previous findings[25,26] (Supplementary Note and Supplementary Fig. 1). To better illustrate admixture between domesticated and wild wheat, we performed joint analysis of hexaploid and tetraploid wheat samples at $K=7$, which separates wild and domesticated emmer from wheat and each other (Fig. 1a). At this value of K, on average, domesticated emmer and wheat accessions had 21.9% and 0.6% of ancestry assigned to wild emmer, respectively. The extent of genetic differentiation between wild emmer and wheat correlated negatively with the improvement status (landraces versus cultivars) (Fig. 1c). While this process might be largely driven by genetic drift and linked selection[27], it can also be influenced by gene flow between sympatric populations; a possibility consistent with higher F_{ST} between allopatric than sympatric populations of wild emmer and wheat landraces (Fig. 1d). Likewise, the three-population f_3 test[28], using the populations of wild emmer (Fig. 1a) and wheat as the sources and landraces as the target, identified significant Z scores less than −4.0, supporting gene flow from wild emmer to wheat (Supplementary Table 3).

In agreement with previous studies[25], we found that geographic proximity contributed more to genetic differentiation among the populations of landraces and cultivars than did improvement status (Supplementary Table 4 and Supplementary Fig. 2), suggesting that the local populations of landraces were broadly used to develop cultivars adapted to each major geographic region.

Genome-wide patterns of introgression from wild emmer. Three previously defined wild emmer subpopulations (North, South 1 and South 2 in Fig. 1a)[3] were used as sources to detect introgression into wheat landraces and cultivars. Four-taxon f_d statistic[29] (Fig. 2a), which estimates an excess of shared derived variants between two taxa, was calculated in 100-SNP windows across the A and B genomes of individual accessions as well as populations (Supplementary Tables 5–7). Introgressed regions were defined as the ninety-fifth percentile for outlier values of f_d statistic. We found more extensive wild emmer gene flow (two-tailed t-test, $t=10.7$, d.f. = 174,150, $P\leq 2.2\times 10^{-16}$) into landraces than cultivars (Fig. 2b and Supplementary Fig. 3), consistent with F_{ST} estimates (Fig. 1c). This conclusion was also validated by estimating the proportion of derived alleles shared only between wild emmer and wheat (referred to as derived wild emmer private (WEP) alleles in Supplementary Fig. 4). Consistent with the higher levels of gene flow from wild emmer, the proportion of WEP alleles was higher in landraces (two-tailed t-test, $t=7.7$, d.f. = 518, $P=5.8\times 10^{-14}$) than in cultivars (Supplementary Fig. 4).

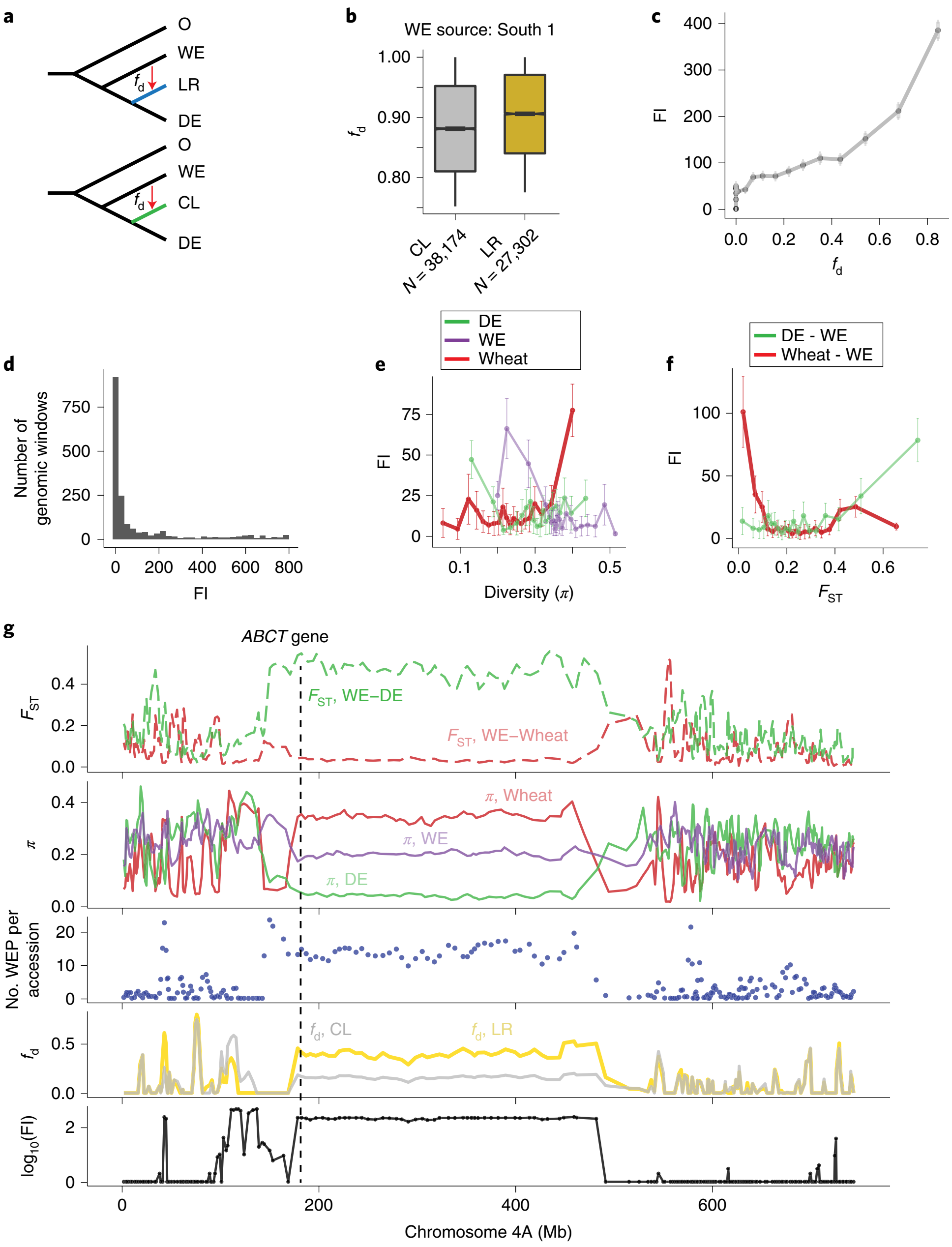

Fig. 2 | Identification of wild emmer introgression in the wheat genome. a, The f_d statistic was used to assess gene flow (red arrows) from wild emmer (WE) to wheat landraces (LR) or cultivars (CL). The multiple outgroup species (O) were used to infer the ancestral and derived alleles at SNP sites (Supplementary Note and Supplementary Table 7). **b**, Distribution of ninety-fifth percentile f_d outliers in wheat landraces and cultivars. Box shows the median and IQRs. The end of the top line is the maximum or the third quartile (Q) + 1.5× IQR. The end of the bottom line denotes either the minimum or the first Q − 1.5× IQR. **c**, FI correlates positively with f_d. **d**, Distribution of FI values in the entire wheat population. **e**, Relationship between FI in wheat and average pairwise diversity per SNP site (π) in the corresponding regions in the populations of wheat, wild (WE) and domesticated emmer (DE). **f**, Relationship between FI in wheat and F_{ST} in the corresponding genomic regions among wheat, wild and domesticated emmer. **g**, Distribution of FI from WE South 1 population into landraces, F_{ST}, average number of WEP alleles in 100-SNP window per landrace accession, population-based f_d statistic, and genetic diversity (π) along chromosome 4A. The location of the *ABCT* gene is shown by the vertical dashed line. Mean and standard error of FI was calculated for each five-percentile bin of the ranked f_d, π or F_{ST} values (**c**,**e**,**f**).

The frequency of introgression (FI) in a population can be affected by both the intensity of gene flow and selection acting on introgressed regions. The FI distribution indicates that most introgressed regions are rare in wheat, and that their frequency strongly correlates with the population-based f_d (mean Spearman rank correlation coefficient $r_s = 0.61 \pm 0.08$) (Fig. 2c,d and Supplementary Table 5). An increase in FI was accompanied by an increase in the genetic diversity of the corresponding genomic regions in the wheat populations (Fig. 2e). No such diversity increase was observed for the same genomic regions in wild and domesticated emmer, suggesting that this trend in wheat is associated with the introgression of wild emmer haplotypes rather than with the high levels of neutral variation in the corresponding regions of a common ancestor. Consistent with the predicted

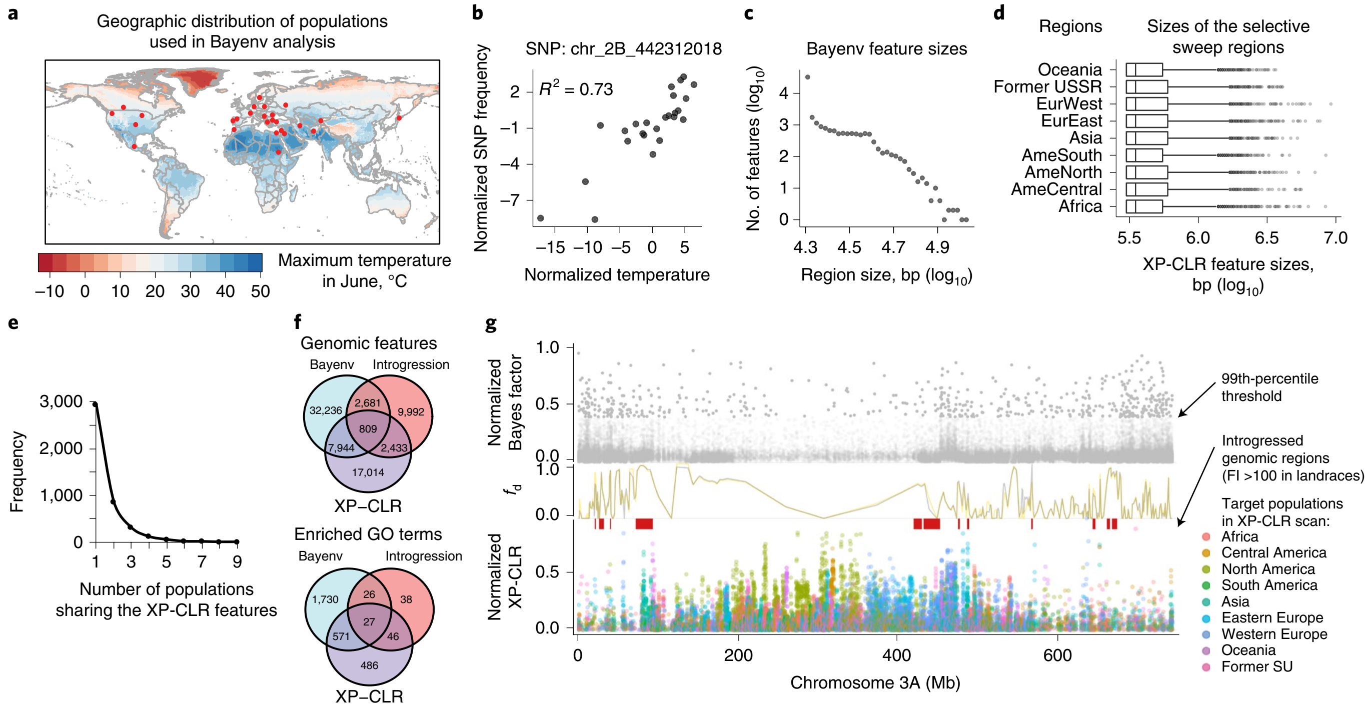

Fig. 3 | Distribution of introgressions, selective sweeps and regions showing environmental adaptation across the wheat genome. a, Geographic distribution of the 26 populations relative to one of the environmental factors, maximum temperature in June (tmax_6), which is shown as a heat map. **b**, Example of a SNP on chromosome 2B showing strong frequency correlation with a normalized tmax_6 factor. **c**, The size distribution of genomic regions (features) harboring the top 1% of Bayenv SNPs. The top 1% outliers of Bayenv analyses included ~78,000 SNPs, where each SNP was associated with an average of 12 out of 68 environmental and bioclimatic variables. **d**, The sizes of selective sweep regions identified by XP-CLR in the populations of cultivars using landraces as the reference population. Box shows the median and IQR. The end of the right line is the maximum or the third quartile (Q) + 1.5× IQR. The dots are more than third Q + 1.5× IQR. **e**, The number of populations sharing the same selective sweep regions identified by XP-CLR. **f**, Overlap between the genomic features (100-kb non-overlapping windows) identified using f_d, XP-CLR and Bayenv scans (top diagram). Number of enriched gene ontology (GO) terms overlapping among the regions identified using f_d, XP-CLR and Bayenv scans (bottom diagram). **g**, Distribution of normalized f_d, XP-CLR (on the scale from 0 to 1) and Bayenv statistics (on the scale from 0 to 1) along chromosome 3A. IGRs (FI > 100) are shown as red bars. XP-CLR statistics are shown for nine target populations.

effects of introgression[15,30], the FI increase resulted in reduced genetic differentiation (F_{ST}) and divergence (d_{xy}) between wheat and wild emmer, but not between wild and domesticated emmer (Fig. 2f and Supplementary Fig. 5).

To determine whether our f_d statistic analyses were consistent with the previously detected signal of wild emmer introgression[13], we examined the *ABCT* gene locus. The frequency of the *ABCT-A1b* allele from wild emmer was found to be higher in the wheat accessions from Europe than from Eastern Asia[13]. Consistent with this expectation, the frequency of WEP alleles at the *ABCT* locus was higher in the populations from Turkey and Europe than in the populations from India and East Asia (Supplementary Fig. 6). Using three wild emmer source populations, we showed that gene flow into wheat around the *ABCT* locus was from the South 1 population from central and northern Israel (Fig. 1a, Supplementary Fig. 6 and Supplementary Table 6). The lower level of wheat–wild emmer genetic differentiation ($F_{ST} = 0.35$) at the *ABCT* locus, compared with the genome-wide estimate ($F_{ST(\text{genome-wide})} = 0.38$), was also consistent with the expected effect of introgression. An analysis of all of chromosome 4A in our wheat population showed that the *ABCT* locus is located within a large region of introgression from wild emmer, which in addition to showing reduced differentiation from wild emmer ($F_{ST} = 0.32$ versus $F_{ST(\text{genome-wide})} = 0.38$) also has higher than average levels of genetic diversity in wheat ($\pi = 0.33$ versus $\pi_{\text{genome-wide}} = 0.18$). It is noteworthy that this region showed elevated F_{ST} between wild and domesticated emmer, and overlapped with a genomic region that was previously identified in a domestication selection scan[23] (Fig. 2g). Contrary to findings in maize that show gene flow from wild relatives is limited around domestication genes[31], we found evidence of introgression at the *ABCT* gene locus and domestication genes *BTR1-A* and *BTR1-B*[23,32]. No introgression was detected at domestication gene *Q* on chromosome 5A[33] (Supplementary Fig. 7).

The wild emmer source populations that contributed to gene flow differ across the wheat genome (Supplementary Table 6). It was lower in the A genome than in the B genome for WE North population (two-tailed Mann–Whitney *U*-test, sum of ranks (W) = 254,720, $N = 1{,}417$, $P = 3.3 \times 10^{-3}$), but it was higher in the A genome than the B genome for both WE South 1 (two-tailed Mann–Whitney *U*-test, $W = 265{,}680$, $N = 1{,}417$, $P = 2.6 \times 10^{-5}$) and South 2 (two-tailed Mann–Whitney *U*-test, $W = 254{,}720$, $N = 1{,}417$, $P = 9.8 \times 10^{-4}$) populations. If we define introgressed genomic regions (IGRs) as regions showing FI > 100 as a threshold, which is close to the FI value observed at the *ABCT* locus in landraces (Fig. 2g), on average, the IGRs would compose about 11.8% and 11.4% of genome per accession in landraces and cultivars (two-tailed *t*-test, $t = 3.5$, d.f. = 608.9, $P = 4.8 \times 10^{-4}$), respectively (Supplementary Fig. 8a,b and Supplementary Table 8). These results suggest that wheat experienced substantial levels of gene flow from its tetraploid wild relative. Considering our full wheat panel, we found that the total length of the IGRs from each of the three wild emmer source populations varied among chromosomes, with the largest number of IGRs found on chromosomes 1A, 4A, 4B, 5A and 6A (Supplementary Fig. 8c).

Patterns of introgression, selection and adaptive evolution. The spread of wheat from its center of origin in the Middle East to new geographic regions involved both agronomic and environmental adaptation. To identify genomic regions associated with local adaptation, we used the environmental association approach implemented in the program Bayenv[34] to study correlations between 49 environmental and 19 bioclimatic variables and allele frequencies in 26 wheat populations that were defined on the basis of geographic regions with relatively uniform climate (Figs. 3a,b and Supplementary Tables 1 and 9). By merging climate-associated SNPs located within 10 kilobases (kb) of each other, we have identified 43,670 genomic regions that, after extrapolating to the size of regions impacted by selection, cover about 988 megabases (Mb) of genome (Fig. 3c and Supplementary Table 10).

To identify genomic regions impacted by selection during improvement, we used the XP-CLR statistic[35], which was calculated by comparing each of the nine large regional populations of cultivars (Supplementary Table 1) with the reference population of landraces (Fig. 3d). We identified 4,316 genomic regions subjected to selection in at least one population (Supplementary Table 11) with the average size of the individual regions close to 100 kb (Fig. 3d). These genomic regions together span about 2.3 gigabases (Gb), indicating that improvement selection affected a substantial portion of the 17-Gb wheat genome. Cross-population comparisons showed that most selective sweep regions (2,947) do not overlap, and only three genomic regions were shared among all nine populations used in the comparison (Fig. 3e). Taken together, these results suggest that during the development of locally adapted cultivars, selection targeted unique genomic regions probably associated with region-specific agroecological factors. This possibility is supported by the 31% overlap between the regions identified in the XP-CLR and Bayenv scans (Fig. 3f).

We evaluated the contribution of introgression from wild emmer to wheat improvement and local adaptation. Genomic regions targeted by improvement selection overlapped with the 3,242 IGRs (20.4% of IGRs) (Fig. 3f,g and Supplementary Fig. 9). Nearly 8.0% of the Bayenv genomic windows associated with environmental adaptation overlapped with IGRs harboring 681 WEP alleles (Supplementary Fig. 4), suggesting that wild emmer gene flow may have contributed to local adaptation. A total of 809 IGRs (81 Mb) overlapped with the regions showing signatures of both environmental adaptation and improvement selection (Supplementary Table 12). Gene ontology terms enriched in the genomic regions overlapping among the introgression, XP-CLR and Bayenv scans were associated with biological processes targeted during crop improvement and development of locally adapted cultivars[36–39] (Fig. 3f, Supplementary Note and Supplementary Table 13).

Deleterious SNP alleles in the allopolyploid wheat genome. Deleterious alleles were shown to accumulate in genes that can affect complex phenotypic traits in maize, rice and barley[21,22,40]. To assess the effects of improvement selection, environmental adaptation and gene flow from the wild ancestor on mutation burden, we studied the distribution of non-synonymous, synonymous and deleterious SNPs (nSNPs, sSNPs and dSNPs, respectively) across the wheat genomes and among the regions identified in the selection and introgression scans.

The site frequency spectrum (SFS) for sSNPs was similar in the A and B genomes and significantly different from that of dSNPs (Kolmogorov–Smirnov test: $D_{\mathrm{Agenome}}=0.45$, $P=0.04$; $D_{\mathrm{Bgenome}}=0.5$, $P=0.01$; Fig. 4a). These patterns are consistent with the effect of purifying selection maintaining deleterious alleles at low frequency in both the A and B genomes of wheat. The SFS of dSNPs in the D genome was different from both the A and B genomes (Fig. 4a), but was not significantly different from the SFS of sSNPs in the D genome. The similarity of the SFS for dSNPs and sSNPs in the D genome is probably a result of reduced efficacy of selection and increased genetic drift owing to the genetic bottleneck caused by polyploidization during the origin of bread wheat[41].

The average number of dSNPs per line varied among the wheat genomes, with the A, B and D genomes harboring 481, 548 and 297 dSNPs, respectively (Fig. 4b). However, the average dSNP/sSNP ratio was higher in the A genome (3.2×10^{-2}) than in the B (2.5×10^{-2}) and D (1.4×10^{-2}) genomes (Fig. 4c). This trend was partly associated with a 33.6% lower level of sSNP diversity in the A genome compared with that in the B genome (Supplementary Table 2)[20]. One of the factors contributing to higher dSNP and sSNP diversity in the B genome can be larger effective population size and outcrossing mating behavior of a diploid ancestral species closely related to *Aegilops speltoides*[42]. The proportions of the genome subjected to domestication selection, which tends to increase dSNP load[21], can also contribute to differences in dSNP enrichment between the A and B genomes. Consistent with this possibility, the size of selective sweep regions detected in domesticated emmer by using wild emmer[23] as the reference population was significantly higher (permutation-based $P=0.001$) in the A (1,237 windows covering 61.9 Mb) than in the B genome (784 windows covering 39.2 Mb).

In wheat, we found a strong negative correlation ($r^2=0.44$) between recombination rate and dSNP enrichment (Supplementary Fig. 10), consistent with findings made in other crops[21,22,40]. However, considering each genome separately, a strong correlation was found only in the A and B genomes, but not in the D genome (Fig. 4d and Supplementary Fig. 10)[41]. Although negative correlation between recombination and dSNPs in the A and B genomes is consistent with the patterns observed in diploid plants[22,40], and might suggest that polyploidy does not have an effect on dSNP load, we observed a tendency toward an increased number of dSNPs in genes duplicated due to polyploidy compared with that in single-copy genes (Supplementary Note and Supplementary Table 14). This trend implies that polyploidy probably resulted in relaxation of purifying selection in wheat.

Our results indicate that both selection and gene flow were important factors that affected the distribution of dSNPs across the wheat genome. We found evidence for the significant reduction of dSNP enrichment in cultivars compared with landraces (Fig. 4e and Supplementary Table 15). The reduction of mutation load in the XP-CLR outlier regions was consistent across all populations of cultivars from different geographic regions (Fig. 4e). Similarly, SNPs showing a strong correlation with the bioclimatic variables (Fig. 4f,g and Supplementary Fig. 11), or located within the regions of introgression, showed a significant reduction in deleterious mutation burden compared with other regions (Fig. 4h).

Effect of introgression on phenotypic variation. To evaluate the effects of gene flow, environmental adaptation and improvement selection on phenotypic variation, the diversity panel was phenotyped for grain filling period (GFP), harvest weight (HW), drought susceptibility index for harvest weight (HWS), heading date (HD) and plant height (PHT) traits. Among the trait-associated SNPs, 17.1%, 73.5% and 6.4% were located within the regions identified by the XP-CLR, Bayenv and introgression scans, respectively (Supplementary Table 16 and Supplementary Figs. 12 and 13a), indicating that some signals of environmental adaptation and improvement selection could be associated with known genes controlling adaptive and agronomic traits in wheat. Overlap of genome-wide association study signals with introgression suggests that variation contributed by gene flow may play a role in broadening phenotypic variation (Fig. 5a, Supplementary Table 17 and Supplementary Fig. 13a). At the loci controlling various agronomic traits, the average diversity in wheat lines showing no evidence of introgression ($\pi=6.46\times10^{-4}$) was increased by 41.5% ($\pi=9.14\times10^{-4}$) when lines with introgressions were included (Supplementary Table 17).

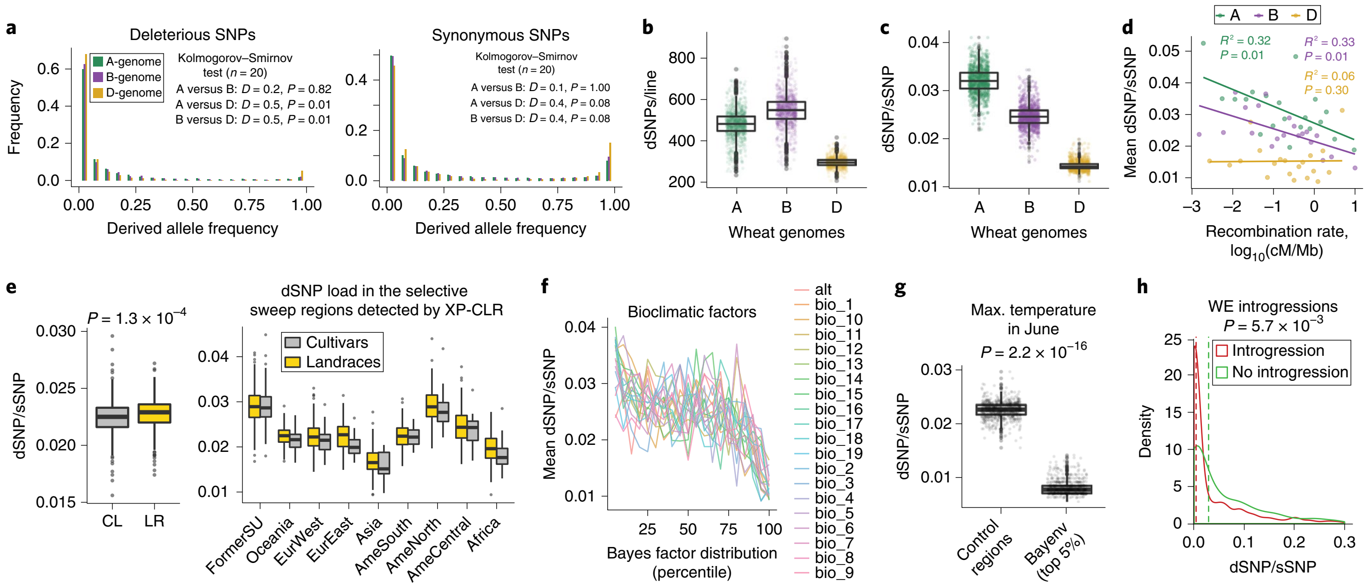

Fig. 4 | Distribution of dSNPs across the wheat genome. a, Derived SFS for dSNPs and sSNPs in the A, B and D genomes of wheat. Box-plots were used to compare dSNP load in **b**,**c**,**e**,**g**. Box shows the median and IQRs. The end of the top line is the maximum or the third quartile (Q) + 1.5× IQR. The end of the bottom line denotes either the minimum or the first Q − 1.5× IQR. The dots are either more than third Q + 1.5× IQR or less than first Q − 1.5× IQR. **b**, Number of observed dSNPs per wheat accession. **c**, dSNP/sSNP ratio in the different wheat genomes. **d**, Relationship between recombination rate and mean dSNP/sSNP ratio in each wheat genome. **e**, dSNP/sSNP ratio in wheat cultivars (CL) and landraces (LR) (left panel, two-tailed Mann–Whitney *U*-test, $W = 78{,}012$, $N = 743$, $P = 1.3 \times 10^{-4}$). dSNP/sSNP ratio in the selective sweep regions identified in cultivars compared with the same genomic regions in landraces (right panel). **f**, Relationship between dSNP/sSNP ratio and Bayes factors that reflect correlation between allele frequency and bioclimatic variables. **g**, dSNPs/sSNP ratio in top 5% of Bayenv SNPs correlating with the maximum temperature in June compared with genome-wide SNPs (two-tailed Mann–Whitney *U*-test, $N = 794$, $P < 2.2 \times 10^{-16}$). **h**, dSNP/sSNP ratio in the introgressed regions compared with regions without introgression (two-tailed Mann–Whitney *U*-test, $N = 285$, $P = 5.7 \times 10^{-3}$). Means are shown by dashed lines.

Using previously published RNA-seq data[43], we have evaluated the effect of introgression on the expression of duplicated homoeologous genes in the sequenced reference wheat cultivar Chinese Spring[19]. The proportion of genes showing an expression bias toward one of the three wheat genome homoeologs within the introgressed and non-introgressed genomic regions was mostly similar across 15 different wheat tissues (Supplementary Note and Supplementary Tables 18 and 19). At the genome-wide level, introgressed alleles of genes appear to have the same likelihood of showing biased gene expression as genes located within the non-introgressed regions. This lack of extensive expression bias in the introgressed regions can be associated with the necessity to maintain the correct balance of proteins encoded by duplicated genomes[44,45].

In addition, we have applied a variance-component method[46] to partition heritability for a range of traits using SNPs from the genomic regions with and without signals of wild emmer gene flow (Fig. 5b). SNPs private to wild emmer and wheat (WEP sites defined in Supplementary Fig. 4) and located within introgression were used to assess the contribution of wild emmer to trait variation in wheat. The genetic relationship matrices were constructed using WEP and non-WEP SNPs grouped into three sets on the basis of the derived allele frequency (DAF) in the population (Supplementary Table 20). For each DAF group, variance for each trait explained by SNPs within and outside introgression was estimated (Fig. 5b and Supplementary Fig. 13b). Compared with non-WEP SNPs, wild emmer SNPs with DAF < 0.1 explained most of the variation for harvest weight (up to 30.9%), drought susceptibility (up to 22.5%) and plant height (up to 35%). The proportion of variance explained by WEP SNPs with DAF ≥ 0.1 for the same traits was smaller compared with that explained by non-WEP SNPs. Overall, for SNPs with DAF < 0.1, the average proportion of phenotypic variance for all analyzed traits explained by WEP sites (18.7%) was close to that explained by non-WEP sites (21.8%) (Fig. 5b and Supplementary Fig. 13b). On average, the proportion of phenotypic variance explained by wild emmer SNPs declined with the increase in DAF.

Discussion

Our study reveals that the genome-wide SNP diversity in wheat was strongly influenced by gene flow from its tetraploid wild ancestor[13]. Regions of introgression in the wheat A and B genomes showed increased levels of genetic diversity and reduced genetic differentiation from wild emmer. Both patterns were consistent with wild-relative introgression into wheat, which offset the effects of polyploidization and domestication bottlenecks on diversity in these genomes[10].

Introgression frequency tended to be relatively low in wheat populations and was distributed non-uniformly among the genomes and chromosomes. In agreement with findings from previous studies, limited gene flow was found around the domestication-related genes, or regions showing evidence of a domestication-selective sweep[23,31–33]. However, while it appears that selection against introgression around one of the primary domestication genes, *Q*, was not affected by the merger of the domesticated AB genome progenitor with the wild D genome ancestor during the origin of bread wheat, some regions overlapping with the selective sweeps found in domesticated emmer[23] or domestication genes[23] showed evidence of introgression. It is possible that these regions did not contribute to domestication traits or that their contribution is modulated by the D genome, which would suggest a more complex domestication trait architecture in hexaploid wheat[47,48].

Deleterious SNPs were postulated to have a negative effect on crop performance and their elimination was proposed as a possible breeding strategy[49]. Our results show that dSNP burden in wheat was reduced by gene flow, breeding and environmental adaptation. Selection associated with both environmental adaptation and wheat improvement appears to have been effective at purging deleterious

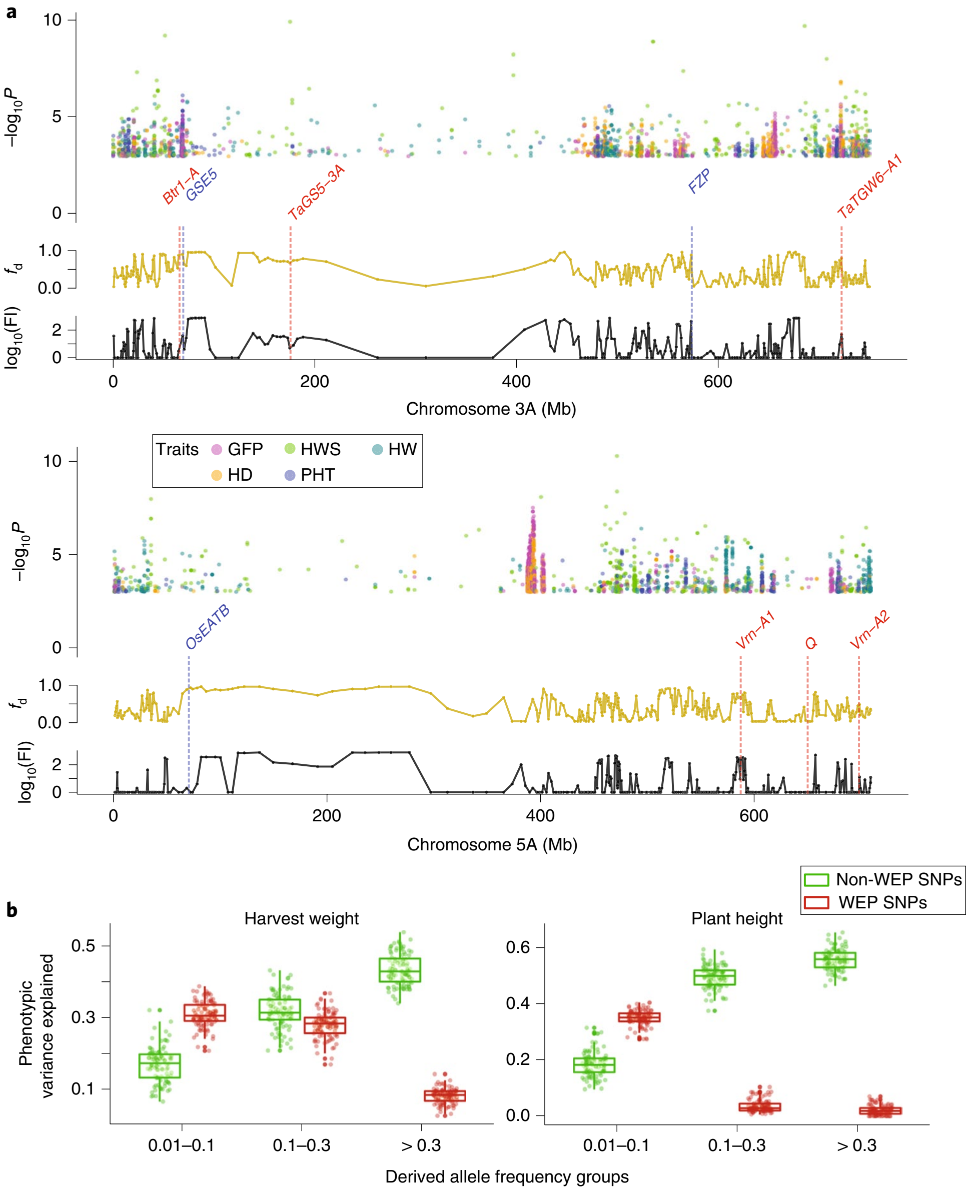

Fig. 5 | SNPs from the introgressed regions explain a large proportion of phenotypic variance in wheat. a, Distribution of association *P* values for various phenotypic traits (only SNPs showing $P < 10^{-3}$ are shown), population-based f_d for wheat landraces and FI in entire wheat population along wheat chromosomes 3A and 5A. The locations of wheat genes (red) and rice gene orthologs (blue) are shown by vertical dashed lines. Traits: GFP, grain filling period; HW, harvest weight; HWS, drought susceptibility index for harvest weight; HD, heading date; PHT, plant height. **b**, Partitioning the heritability for harvest weight and plant height explained by WEP and non-WEP SNPs grouped on the basis of derived allele frequency ((0.01–0.1), (0.1–0.3), (0.3–1.0)) into three groups. Box shows the median and IQRs. The end of top line is the maximum or the third quartile (Q) + 1.5× IQR. The end of bottom line denotes either the minimum or the first Q − 1.5× IQR. The dots are either more than third Q + 1.5× IQR or less than first Q − 1.5× IQR.

alleles from wheat populations. Although domestication and a founder event were shown to increase deleterious allele load in maize[16], this effect was mostly associated with a genetic bottleneck rather than with linkage to positively selected alleles. The lack of a strong bottleneck during the transition from landraces to cultivars in wheat[25] probably facilitated effective selection against deleterious alleles. Gene flow further reduced mutation load, a trend that was also observed in maize and teosinte[16,31], and was possibly linked with a larger effective population size of wild emmer capable of effective removal of deleterious alleles.

Deleterious effects of mutations can potentially be masked in polyploids such as wheat[20,50], which is consistent with our finding that single-copy genes tend to carry fewer dSNPs than genes duplicated due to polyploidy. However, in spite of polyploidy, the frequency and patterns of dSNP distribution in the A and B genomes were similar to those observed in diploid crops[21,22]. Most dSNPs were rare, and dSNP density in the A and B genomes is correlated negatively with recombination, consistent with the effective removal of deleterious alleles in the highly recombining regions. A severe polyploidization bottleneck in the D genome accompanying the origin of wheat appears to have increased genetic drift and reduced the efficiency of selection against dSNPs in the high recombining regions[21,41]. A similar uncoupling between recombination and dSNP load was found in cassava and is attributed to decreased efficiency of selection due to clonal propagation[51]. Our results indicate that the effects of dSNPs in wheat are only partially masked by polyploidy and sufficiently large to be selected against. Increased hybrid wheat performance shown in recent studies[52] is consistent with this possibility, suggesting that, in addition to polyploidy, the negative effects of dSNPs can be reduced in the heterozygous state. Thus, the effects of purifying selection on deleterious mutations are evident not only in paleopolyploid plants, where the strength of selection acting on retained duplicates and their single-copy orthologs in diploids[53] was similar, but also in young polyploids.

The overlap between the signals of gene flow, genome-wide association studies, improvement and environmental adaptation is suggestive of adaptive introgression. Using the variance-component method[46] we show that introgression contributed to wheat phenotypic diversity for HW, PHT, GFP and HD traits. Less common alleles (DAF < 0.1) introduced from wild emmer explained a substantial proportion of phenotypic variance for harvest weight (up to 30.9%), drought susceptibility (up to 22.5%) and plant height (up to 35%) traits. The proportion of phenotypic variance explained by introgressed variants declined with an increase in population frequency, suggestive of either negative selection against introgression allowing only small-effect-size alleles to reach high frequency in population, or introgression being adaptive only in specific geographically constrained habitats.

Our results provide evidence that historic gene flow from wild emmer played an important role in shaping the agronomic phenotypes in modern wheat and probably broadened its adaptive potential. These findings have important implications for the future of wheat breeding. A detailed map of genome-wide introgression developed in our study can guide targeted deployment of wild-relative diversity in wheat-breeding programs. These efforts, besides introducing novel adaptive alleles into cultivars and broadening phenotypic diversity available for selection, hold great potential to reduce the deleterious mutation burden in the wheat genome, further accelerating breeding.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-019-0382-2.

Received: 2 August 2018; Accepted: 26 February 2019;
Published online: 1 May 2019

References

1. Nesbitt, M. & Samuel, D. From staple crop to extinction? The archaeology and history of the hulled wheats. in *Proc. 1st Int. Workshop Hulled Wheats* (eds Padulosi, S. et al.) 41–100 (Italy International Plant Genetic Resources Institute, 1996).
2. Tanno, K.-I. & Willcox, G. How fast was wild wheat domesticated? *Science* **311**, 1886 (2006).
3. Luo, M.-C. et al. The structure of wild and domesticated emmer wheat populations, gene flow between them, and the site of emmer domestication. *Theor. Appl. Genet.* **114**, 947–959 (2007).
4. Ozkan, H., Willcox, G., Graner, A., Salamini, F. & Kilian, B. Geographic distribution and domestication of wild emmer wheat (*Triticum dicoccoides*). *Genet. Resour. Crop Evol.* **58**, 11–53 (2011).
5. Kihara, H. Discovery of the DD-analyser, one of the ancestors of *Triticum vulgare*. *Agric. Hortic.* **19**, 889–890 (1944).
6. Dvorak, J., Luo, M. C., Yang, Z. L. & Zhang, H. B. The structure of the *Aegilops tauschii* genepool and the evolution of hexaploid wheat. *Theor. Appl. Genet.* **97**, 657–670 (1998).
7. Smith, O. et al. Sedimentary DNA from a submerged site reveals wheat in the British Isles 8000 years ago. *Science* **347**, 998–1001 (2014).
8. Long, T. et al. The early history of wheat in China from ^{14}C dating and Bayesian chronological modelling. *Nat. Plants* **4**, 272–279 (2018).
9. Haudry, a et al. Grinding up wheat: a massive loss of nucleotide diversity since domestication. *Mol. Biol. Evol.* **24**, 1506–1517 (2007).
10. Akhunov, E. D. et al. Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. *BMC Genomics* **11**, 702 (2010).
11. Kerber, E. R. Wheat: reconstitution of the tetraploid component (AABB) of hexaploids. *Science* **143**, 253–255 (1964).
12. Dvorak, J., Luo, M. & Akhunov, E. D. N. I. Vavilov' s theory of centres of diversity in the light of current understanding of wheat diversity, domestication and evolution. *Czech. J. Genet. Plant Breed.* **47**, 1–8 (2011).
13. Dvorak, J., Akhunov, E. D., Akhunov, A. R., Deal, K. R. & Luo, M.-C. Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from wild tetraploid wheat to hexaploid wheat. *Mol. Biol. Evol.* **23**, 1386–1396 (2006).
14. Salojärvi, J. et al. Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nat. Genet.* **49**, 904–912 (2017).
15. Rendón-anaya, M. et al. Genomic history of the origin and domestication of common bean unveils its closest sister species. *Genome Biol.* **18**, 1–17 (2017).
16. Wang, L. et al. The interplay of demography and selection during maize domestication and expansion. *Genome Biol.* **18**, 1–13 (2017).
17. Hardigan, M. A. et al. Genome diversity of tuber-bearing *Solanum* uncovers complex evolutionary history and targets of domestication in the cultivated potato. *Proc. Natl Acad. Sci. USA* **114**, E9999–E10008 (2017).
18. Hübner, S. et al. Islands and streams: clusters and gene flow in wild barley populations from the Levant. *Mol. Ecol.* **21**, 1115–1129 (2012).
19. International Wheat Genome Sequencing Consortium. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar7191 (2018).
20. Jordan, K. et al. A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. *Genome Biol.* **16**, 48 (2015).
21. Liu, Q., Zhou, Y., Morrell, P. L., Gaut, B. S. & Ge, S. Deleterious variants in Asian rice and the potential cost of domestication. *Mol. Biol. Evol.* **34**, 908–924 (2017).
22. Mezmouk, S. & Ross-Ibarra, J. The pattern and distribution of deleterious mutations in maize. *G3 (Bethesda)* **4**, 163–171 (2014).
23. Avni, R. et al. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* **97**, 93–97 (2017).
24. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
25. Cavanagh, C. R. et al. Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proc. Natl Acad. Sci. USA* **110**, 8057–8062 (2013).
26. Wang, S. et al. Characterization of polyploid wheat genomic diversity using a high-density 90,000 single nucleotide polymorphism array. *Plant Biotechnol. J.* **12**, 787–796 (2014).
27. Poets, A. M. et al. The effects of both recent and long-term selection and genetic drift are readily evident in North American barley breeding populations. *G3 (Bethesda)* **6**, 609–622 (2016).
28. Patterson, N. et al. Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
29. Martin, S. H., Davey, J. W. & Jiggins, C. D. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* **32**, 244–257 (2015).
30. Smith, J. & Kronforst, M. R. Do *Heliconius* butterfly species exchange mimicry alleles? *Biol. Lett.* **9**, 1–4 (2013).
31. Hufford, M. B. et al. The genomic signature of crop-wild introgression in maize. *PLoS Genet.* **9**, e1003477 (2013).
32. Nave, M., Avni, R., Ben-Zvi, B., Hale, I. & Distelfeld, A. QTLs for uniform grain dimensions and germination selected during wheat domestication are co-located on chromosome 4B. *Theor. Appl. Genet.* **129**, 1303–1315 (2016).
33. Simons, K. J. et al. Molecular characterization of the major wheat domestication gene Q. *Genetics* **172**, 547–555 (2006).
34. Günther, T. & Coop, G. Robust identification of local adaptation from allele frequencies. *Genetics* **195**, 205–220 (2013).
35. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).
36. Kant, S., Bi, Y. & Rothstein, S. J. Understanding plant response to nitrogen limitation for the improvement of crop nitrogen use efficiency. *J. Exp. Bot.* **62**, 1499–1509 (2011).
37. Forde, B. G. Glutamate signalling in roots. *J. Exp. Bot.* **65**, 779–787 (2014).
38. Lu, G. et al. Application of T-DNA activation tagging to identify glutamate receptor-like genes that enhance drought tolerance in plants. *Plant Cell Rep.* **33**, 617–631 (2014).
39. Kiba, T., Krapp, A. & Science, R. Plant nitrogen acquisition under low availability: regulation of uptake and root architecture. *Plant Cell Physiol.* **57**, 707–714 (2016).
40. Kono, T. J. Y. et al. The role of deleterious substitutions in crop genomes. *Mol. Biol. Evol.* **33**, 2307–2317 (2016).
41. Jordan, K. W. et al. The genetic architecture of genome-wide recombination rate variation in allopolyploid wheat revealed by nested association mapping. *Plant J.* **95**, 1039–1054 (2018).
42. Kilian, B. et al. Independent wheat B and G genome origins in outcrossing *Aegilops* progenitor haplotypes. *Mol. Biol. Evol.* **24**, 217–227 (2007).
43. Choulet, F. et al. Structural and functional partitioning of bread wheat chromosome 3B. *Science* **345**, 1249721 (2014).
44. Akhunova, A. R., Matniyazov, R. T., Liang, H. & Akhunov, E. D. Homoeolog-specific transcriptional bias in allopolyploid wheat. *BMC Genomics* **11**, 1–16 (2010).

45. Veitia, R. A., Bottani, S. & Birchler, J. A. Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. *Trends Genet.* **24**, 390–397 (2008).
46. Gusev, A. et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
47. Peleg, Z., Fahima, T., Korol, A. B., Abbo, S. & Saranga, Y. Genetic analysis of wheat domestication and evolution under domestication. *J. Exp. Bot.* **62**, 5051–5061 (2011).
48. Stitzer, M. C. & Ross-Ibarra, J. Maize domestication and gene interaction. *New Phytol.* **220**, 395–408 (2018).
49. Morrell, P. L., Buckler, E. S. & Ross-Ibarra, J. Crop genomics: advances and applications. *Nat. Rev. Genet.* **13**, 85–96 (2011).
50. Krasileva, K. V. et al. Uncovering hidden variation in polyploid wheat. *Proc. Natl Acad. Sci. USA* **114**, 913–E921 (2017).
51. Ramu, P. et al. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat. Genet.* **49**, 959–963 (2017).
52. Zhao, Y. et al. Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. *Proc. Natl Acad. Sci. USA* **112**, 15624–15629 (2015).
53. Hao, Y. et al. Patterns of population variation in two paleopolyploid eudicot lineages suggest that dosage-based selection on homeologs is long-lived. *Genome Biol. Evol.* **10**, 999–1011 (2018).

Acknowledgements

This project was supported by the Agriculture and Food Research Initiative Competitive Grants 2017-67007-25939 (Wheat-CAP) and 2016-67013-24473 from the USDA National Institute of Food and Agriculture, and grants from the Bill and Melinda Gates Foundation and Kansas Wheat Commission. Exome sequencing of Canadian wheat cultivars was supported through the Canadian *Triticum* Applied Genomics grant funded by Genome Canada, Genome Prairie, Saskatchewan Ministry of Agriculture, and the Western Grains Research Foundation. P.L.M. was supported by grant IOS-1339393 from the US National Science Foundation. Corteva Agriscience, Agriculture Division of DowDuPont provided financial support through collaboration with Agriculture Victoria Services enabling the development of the SNP dataset and technologies used in this manuscript. The authors would like to thank International Wheat Genome Sequencing Consortium for providing access to wheat genome sequence under Toronto agreement, D. Andresen for assistance with the computing resources of the KSU Beocat cluster funded by NSF grant ACI-144054 and K. Jordan for valuable suggestions and editing the manuscript.

Author contributions

F.H. led the bioinformatic and statistical analyses of data and helped to draft the first version of the manuscript. R.P. led phenotypic analyses. F.S. contributed to genomic analyses. S.K. was responsible for field trials and phenotype collection. G.K.-G. contributed to bioinformatic analyses of data. P.K. was responsible for exome sequencing of most wheat lines. K.F. was responsible for exome sequencing and 90K SNP data analyses. A.F. contributed to generating wild emmer exome capture data. P.H., K.W., R.K., R.C. and C.P. generated and contributed exome sequencing data for wild and domesticated emmer, and Canadian wheat cultivars. A.A. contributed to exome capture of wild and domesticated emmer, and wheat. P.L.M. contributed to data interpretation and manuscript writing. C.P., J.P.D., S.R.W. and G.S. contributed to project design. B.H., H.D. and J.T. contributed to project coordination and data analyses. M.H. provided project leadership, coordinated data collection and next generation sequencing (NGS) data analyses, and contributed to manuscript writing. E.A. conceived the idea, coordinated data collection and NGS data analyses and data interpretation, and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0382-2.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.H. or E.A.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Wheat accessions. The exome-sequenced hexaploid wheat accessions were selected from a worldwide population of landraces and cultivars obtained from the USDA National Small Grains Collection and the Australian Grains Genebank (formerly the Australian Winter Cereal Collection). The panel included 3,990 accessions from 106 countries that were genotyped using the 90K iSelect assay[26]. The MCG method[54] was used to select accessions for exome sequencing. This method uses a genetic relationship matrix calculated from genotype data to iteratively chose an arbitrary number of accessions (in the case of this study 20% of the accessions) that collectively explain the largest proportion of variance in genetic relationships among the whole set.

DNA extraction and sequence capture. Genomic DNA was extracted from leaf tissue for each accession using the Agencourt DNAdvance Genomic DNA Isolation Kit (Beckman Coulter) and subjected to sequence capture using the NimbleGen SeqCap EZ wheat whole-genome assay[20]. In brief, 1 μg of DNA was fragmented with the Covaris S2 instrument to obtain an average fragment length of 300 bp. Shorter fragments were removed using Agencourt AMPure XP beads (Beckman Coulter). The KAPA Library Preparation Kit (Kapa Biosystems) and adapters supplied in the NimbleGen SeqCap EZ Reagent Kit Plus v2 (Roche) were used to prepare libraries for the capture assay, according to the KAPA protocol (excluding steps 8.1–8.22). The quality and yield of each sample library were assessed using an Agilent 2200 TapeStation (Agilent Technologies), and sequence capture was then performed according to the Roche protocol. Each library was sequenced on an Illumina HiSeq2000 instrument (Illumina) to generate about 30 million reads per accession.

SNP calling. GYDLE (GYDLE Inc.) software was used to quality filter the raw sequence reads (phred score ≥ 20; read length ≥ 50 bp) and align them to sequences used in the exome capture design[20]. These on-target reads were then realigned to the IWGSC RefSeq v.1.0 (ref. [19]) using GYDLE that conducts exhaustive search to detect uniquely mapped reads. Because the level of divergence among wheat homoeologs is 1–2%, consistent assignment to the correct wheat genome was possible for nearly all reads. Multi-mapping reads as well as genomic regions showing high depth of read coverage were excluded. SNP discovery and genotype calling were performed using the 'find snp' function of GYDLE.

SNP imputation and filtering. The Beagle program (beagle.21Jan17.6cc.jar) was used for SNP imputation with the following parameters: 'overlap = 500, window = 5,000, ne = 12,000' (ref. [55]). The effective population size was based on the previous estimates[56]. The accuracy for imputation was estimated to be higher than 90% in both GT and GL modes.

Genotype calls with a genotype probability less than 0.8 were set as missing data. Sites with > 75% missing data and > 3% heterozygote calls were removed. For Bayenv and XP-CLR analyses, we used 1.39 million SNPs spaced at least 100 bp apart from each other (thinned using PLINK, --bp-space 100) and MAF > 0.01 in the wheat landraces and cultivars. The VCF file can be downloaded at (http://wheatgenomics.plantpath.ksu.edu/1000EC).

The 90K Infinium iSelect SNP array[26] was used to evaluate the SNP genotyping error rate of SNP from the exome capture. The flanking sequences of SNP on the 90K array were aligned to IWGSC RefSeq v.1.0 using the following parameters: 'coverage > 95%, identity > 97%, e-value < 1×10^{-10}'. SNPs with multiple mapped locations were removed. Among the 23,577 90K SNPs mapped to the reference genome, ~10,000 were shared between the 90K array and the exome capture data. Concordance rate between the 90K and exome capture genotype calls was >0.98, suggesting that the exome capture and imputation procedures applied in our study generated high quality SNP data.

Merging with published emmer wheat SNP data. To analyze introgression between wild emmer and hexaploid wheat, our data were merged with the recently published SNP dataset for wild and domesticated emmer[23]. For each emmer SNP, the flanking sequences of 200 bp on the emmer reference genome were aligned to the IWGSC RefSeq v.1.0 genome using BLAT. The following parameters were used to define a blat hit, 'alignment coverage > 95%, identities > 97%, e-value < 1×10^{-10}'. A total of 1.26 million published emmer wheat SNPs have been uniquely mapped to the IWGSC RefSeq v.1.0 genome[19]. A total of 348,372 merged SNPs that were also polymorphic in our dataset were used for further analyses. The merged VCF file can be downloaded at (http://wheatgenomics.plantpath.ksu.edu/1000EC).

Using generated SNP data, all accessions were clustered using the phylo program from the VCF-kit (https://vcf-kit.readthedocs.io/en/latest/). On the basis of the patterns of clustering inconsistent with the clustering of wheat accessions having the same improvement status (wild, domesticated, landrace, cultivar), we have removed from analyses three wheat landraces (PI 345355, PI 131592, PI 534284) and two accessions of wild and domesticated emmer (PI 467000, PI 415152).

Inference of ancestral allelic states at SNP sites. We assessed the probability of ancestral versus derived allelic states using the maximum likelihood method described in ref. [57], which permits the use of multiple outgroup species. For inferring ancestral states of SNPs in the A and B genomes we have used *Aegilops tauschii*[58], D genome of hexaploid wheat[19], *Triticum urartu*[59], *Triticum monococcum*, *Aegilops sharonensis*, *A. speltoides*, *Thinopyrum elongatum* and *Hordeum vulgare*[60] as outgroups. Sequences were obtained from public databases: URGI (https://wheat-urgi.versailles.inra.fr/), Ensembl Plant (plants.ensembl.org), and NCBI. In addition, intergenomic comparisons of homoeologous genes were used to supplement these analyses; for example, sequences of genes from the A genome of wheat and wild emmer can be used as an outgroup to infer ancestral states in the B genome, and vice versa. Taken together, these species span a broad range of divergence from the wheat A and B genomes, and provide a powerful framework for the accurate assessment of the ancestral allelic states. The ancestral states were inferred for 213,528 SNP sites (Supplementary Table 7 and Supplementary Note).

SNP annotation and genetic load. The SNPeff program was used to annotate SNPs[61] using both high ('HC') and low ('LC') quality gene models from the IWGSC RefSeq v.1.0 genome[19]. SNPs with the high impact (stop_gain/lost, start_gain/lost, splice_donor/acceptor_variant) were considered deleterious. The dSNP/sSNP and nSNP/sSNP ratios were used to assess the enrichment of dSNPs and nSNPs relative to neutral background. A total of 346,146 sSNPs, 15,895 dSNPs and 390,661 nSNPs were used in the analyses.

Population structure and diversity statistics. The ADMIXTURE program was used for genetic assignment[24] using a subset of 17,656 SNPs. This subset was selected by applying the following criteria: (1) SNPs with linkage disequilibrium (LD) above 0.4 were removed using Plink '--indep-pairwise 1000 10 0.4', and (2) SNPs with MAF ≥ 0.001 and located more than 1,000 bp apart were retained. Ten independent runs of ADMIXTURE with different random seeds were performed and summarized using CLUMPP[62]. The R package pophelper was used to generate the ancestry barplots.

We used PLINK1.9 and VCFtools v0.1.16 for the calculation of principal components, and other basic diversity statistics. A sliding window analyses of genetic differentiation (F_{ST}), divergence (d_{xy}) and SNP diversity (π) were performed using ABBABABAwindows.py (https://github.com/simonhmartin/genomics_general)[29]. This script uses only variable sites for calculating the diversity statistics.

Detection of selective sweeps by XP-CLR. The XP-CLR statistic[35] was used to identify selective sweeps associated with modern wheat improvement. A population of wheat landraces from Eurasia was used as reference in comparison with nine test populations of cultivars from nine geographic areas (defined in Supplementary Table 1). For analysis, the genetic location of each SNP was interpolated using the R function 'approx' (method = 'linear', rule = 1). XP-CLR was run with the grid size of 50 kb, the window size of 1 cM, the maximum number of SNPs within a window of 200 and the correlation levels as 0.95. We considered the top 1% outliers of chromosome-level test statistic from each population as the genomic segments under selective sweep.

Detecting local adaptation using BAYENV. Association between local environment and SNP frequency was identified using BAYENV 2.0 (ref. [34]). Data for 49 environmental and 19 bioclimatic variables were obtained from the WorldClim database (http://www.worldclim.org/) for 26 wheat populations defined on the basis of geography (Supplementary Table 1). To control for population structure, a randomly selected set of 20,000 SNPs was used to estimate the covariance matrix with 1,000,000 iterations. The association between the 1.39 million SNPs and the 68 environmental and bioclimatic variables was tested using 1,000,000 iterations for each SNP. The median value of the Bayes factor calculated for each SNP using data from ten independent Bayenv runs. Top 1% cutoff value of the median Bayes factor was used to select locally adaptive SNPs for each of the 68 environmental variables. In total, about 78,000 SNPs were associated with at least one environmental or bioclimatic variable. SNPs located within a 10 kb region were merged to estimate the overlap among genomic regions detected by the XP-CLR, Bayenv and introgression scans.

Detection of introgression from wild emmer. A four-taxon f_d statistic was used to identify the genomic segments introgressed from wild emmer wheat[29]. Multiple outgroup species (O) were used to infer the ancestral (A) and derived (B) SNP allelic states in the populations of wild emmer (WE or P3), domesticated emmer (DE or P1), wheat cultivars (CL or P2) and landraces (LR or P2). Inference of ancestral state is described above. Three previously identified[3] WE source populations were used as P3: North, South 1, and South 2 (defined in Fig. 1a). Without gene flow, the ABBA and BABA allele configurations in the four-taxa tree (((P1, P2),P3),O), should be equally frequent; gene flow between WE and CL or LR would result in an excess of ABBA relative to BABA that can be detected using the f_d statistic[29]. The f_d statistic was calculated in sliding windows of 100 SNPs with a step of 50 SNPs. Windows with less than three informative SNPs (neither 'ABBA' nor 'BABA') were ignored. Windows with negative values of Patterson's *D* statistic[63] (closely related to the f_d statistic) and $f_d > 1$ were ignored. For analyses of introgression, we used a set of 35 accessions of DE and 33 accessions of WE from our previously published study[23].

The f_d statistic for 4,264 genomic windows was calculated for both individual accessions and populations. Three sets of f_d estimates were obtained using three WE source populations. Within each set, the ninety-fifth percentile outliers of f_d distribution were used to detect regions introgressed from wild emmer. The ninety-fifth percentile f_d thresholds for datasets generated using WE source populations

North, South 1 and South 2 were 0.91, 0.78 and 0.47, respectively. The FI within each genomic region in wheat population was estimated by counting overlapping introgressed regions among accessions. To define the IGRs in populations, we have used the FI > 100, which is close to FI at the previously identified introgression around the *ABCT* gene locus on chromosome 4A[13] (genomic window used on chromosome 4A span region 174,725,311–185,745,837 bp). The proportion of IGRs in a single accession was estimated by summing up all regions defined as IGRs at the population level.

To compare overlap among the genomic regions identified by the XP-CLR, Bayenv and f_d analyses, test statistics values obtained using each method were assigned to 100-kb non-overlapping genomic windows. The bedmap tool[64] was used to find the overlap between the datasets.

Gene ontology analysis. In addition to gene ontology data released with the IWGSC RefSeq v.1.0 genome[19], we used blast2go[65] to generate detailed functional annotations of both HC and LC gene sets from the reference genome. The RefSeq v.1.0 genome annotation was combined with our blast2go analyses resulting in 142,346 gene ontology-annotated genes. Fisher's exact test was used to determine the significance of the enrichment for every gene ontology term, followed by the Benjamini–Hochberg correction[66]. A list of gene ontology terms with adjusted $P < 0.005$ was obtained for genes located within the regions identified by the XP-CLR, Bayenv and f_d statistic methods. For the gene ontology enrichment analyses, the selective sweep regions identified in all nine regional populations (top 1% XP-CLR outliers) were combined into a non-redundant set. For gene ontology analyses of genomic regions identified by Bayenv, we used genes that were closest to the top 1% SNPs associated with the environmental variables. This list of gene ontology terms was submitted to Revigo[67] to generate summaries.

Field trials and phenotypic observations. Field trials were conducted for two consecutive years under rainfed and irrigated conditions at the Agriculture Victoria departmental research station located at Horsham, Victoria, Australia. Horsham is located in a medium rainfall zone, with average annual rainfall of 400 mm and a temperate climate. Locations of field trials: (1) rainfed trial 2014: latitude 36° 45' 3.97" S, longitude 142° 6' 57.51" E; (2) irrigated trial 2014: latitude 36° 44' 38.29" S, longitude 142° 6' 12.40" E; (3) rainfed trial 2015: latitude 36° 44' 14.77" S, longitude 142° 6' 50.79" E; (4) irrigated trial 2015: latitude 36° 44' 26.36" S, longitude 142° 6' 6.17" E. Each wheat accession was sown in triplicated 4.5 m single rows in a randomized block design in each of the rainfed and irrigated trials, with a seed-to-seed density of 3.6 cm and row-to-row spacing of 65 cm. The trials were managed using best practice for weed and disease control. Heading date was recorded as the date on which 50% of the heads in the experimental row had fully emerged from the culms. Physiological maturity was the date on which 95% of the plants in the plot had senesced. Plant height was measured from the ground to the tip of the spike (excluding awns).

The estimated means (best linear unbiased estimates) were obtained using a model with fixed genotype effects and all other effects as random in an individual year. The traits were analyzed separately for each year and environment. The trait values from the rainfed (RF) and irrigated (I) trials were used to calculate the stress susceptibility index (SSI)[68]. For each trait, the year and environment were added as a suffix to the trait name. Descriptions for the traits included for analysis are: days to heading in 2014 (HD14_I); days to heading in 2015 (HD15_I); plant height in 2014 (PHT14_I); plant height in 2015 (PHT15_I); grain filling period in 2014 (GFP14_I); grain filling period in 2015 (GFP15_I); harvest weight of grain in 2014 (HW14_I); harvest weight of grain in 2015 (HW15_I); SSI for harvest weight in 2014 (HW14_S); and SSI for harvest weight in 2015 (HW14_S). For the latter, the higher the SSI value the lower the tolerance of the accession to the stress. Accessions that have an SSI value close to zero perform equally well in both the rainfed and irrigated trials.

The stress susceptibility index/drought susceptibility index was calculated according to ref. [68] using the formula SSI = (1 − TRF/TIR)/1 − (μRF/μIR), where TRF is the trait value under rainfed conditions, TIR is the trait value under irrigated conditions, μRF is the mean of the trait across all genotypes under rainfed conditions and μIR the mean of the trait across all genotypes under irrigated conditions.

Estimation of the proportion of phenotypic variance explained by introgression. Genome-wide association mapping was performed using the mixed linear model implemented in GCTA[69]. A total of 2.5 million SNPs with MAF > 0.01 were tested for marker-trait association using best linear unbiased estimates for each trait. The first three principal components were used for controlling the population structure. For estimating kinship coefficients and principal componenet analysis, we used a set of 233,059 SNPs selected using the following criteria: MAF > 0.01, LD (r^2) < 0.8 (plink --indep-pairwise, window size = 10 Mb, step = 10 SNP sites), and keep at most 1 SNP within 1,000 bp. The models' type I error (false positive) rate was estimated using a random set of 10,000 genome-wide distributed SNPs by plotting observed and expected *P* values (Supplementary Note and Supplementary Fig. 12).

The GCTA-GREML method[69] was used to partition the phenotypic variation for a range of traits into components explained by SNPs from the introgressed and non-introgressed genomic regions. For the regions of introgression we have selected 11,032 WEP sites (MAF > 0.01) that are private to wild emmer when compared with domesticated emmer (defined in Supplementary Fig. 4). A total of 58,795 non-WEP SNPs (MAF > 0.01) from the regions without the signals of wild emmer gene flow were used as control. To account for the effect of allele frequency on the estimates of heritability, we have grouped WEP and non-WEP SNPs into three classes depending on the frequency of the derived allele (DAF): 0.01–0.1, 0.1–0.3, >0.3 (Supplementary Table 20). Within each frequency class, we have estimated the proportion of genetic variation for different traits explained by WEP and non-WEP SNPs. For each DAF class, WEP and non-WEP SNPs were used to build two genetic relationship matrices using the following parameter, 'make-grm-inbred, autosome-num 30'. Then, the phenotypic variance for each phenotype explained by these two SNP sets was estimated using 'mgrm'. Because the number of non-WEP SNPs was greater than the number of WEP SNPs, the variance partitioning was repeated 100 times, each time randomly sampling the same number of non-WEP SNPs as the number of WEP SNPs. The proportions of phenotypic variance for each trait, *V*(G)/*V*(p), were extracted from each of the 100 calculations.

Statistical tests. Analysis of variance (ANOVA) tests followed by Tukey's test were performed to compare window-based F_{ST} estimates among populations. Fisher's exact test was used to determine the significance of the enrichment for every gene ontology term, followed by the Benjamini–Hochberg correction. Comparison between two group means was performed using two-tailed *t*-test and two-tailed Mann–Whitney *U*-test. The SFS for SNPs was compared using the Kolmogorov–Smirnov test.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data have been deposited in the European Variation Archive (EVA) under project PRJEB31218 and NCBI SRA under project PRJNA517692, and are available for viewing and download from http://wheatgenomics.plantpath.ksu.edu/1000EC.

References

54. Yu, X., Woolliams, J. A. & Meuwissen, T. H. E. Prioritizing animals for dense genotyping in order to impute missing genotypes of sparsely genotyped animals. *Genet. Sel. Evol.* **46**, 1–8 (2014).
55. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).
56. Thuillet, A.-C., Bataillon, T., Poirier, S., Santoni, S. & David, J. L. Estimation of long-term effective population sizes through the history of durum wheat using microsatellite data. *Genetics* **169**, 1589–1599 (2005).
57. Keightley, P. D. & Jackson, B. C. Inferring the probability of the derived vs. the ancestral allelic state at a polymorphic site. *Genetics* **209**, 897–906 (2018).
58. Luo, M.-C. et al. Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. *Nature* **551**, 498–502 (2017).
59. Ling, H.-Q. et al. Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature* **496**, 87–90 (2013).
60. Mascher, M. et al. A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**, 427–433 (2017).
61. De Baets, G. et al. SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Res.* **40**, D935–D939 (2012).
62. Jakobsson, M. & Rosenberg, N. A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801–1806 (2007).
63. Green, R. E. et al. A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
64. Neph, S. et al. BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920 (2012).
65. Conesa, A. & Götz, S. Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics* **2008**, 619832 (2008).
66. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B* **57**, 289–300 (1995).
67. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**, e21800 (2011).
68. Fischer, R. A. & Maurer, R. Drought resistance in spring wheat cultivars: I. Grain yield responses. *Aust. J. Agric. Res* **29**, 897–912 (1978).
69. Yang, J. et al. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519–525 (2011).

nature research

Corresponding author(s): Eduard Akhunov, Matthew Hayden

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
□	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
□	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	no software was used
Data analysis	Bedops version 2.4.24 Beagle version 4.0 SNPeff version 4.3 ADMIXTURE version 1.3.0 CLUMPP version 1.1.2 PopHelper version 2.2.6 GYDLE version 1.0 GCTA version 1.9.1 BAYENV2 XP-CLR version 1.0 PLINK version 1.9 VCFtool version 0.1.14 phylo in VCF-kit, version 0.1.3 ABBABABAwindows.py and popgenWindows.py from https://github.com/simonhmartin/genomics_general

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data is deposited to European Variation Archive (EVA) under project PRJEB31218 and NCBI SRA under project PRJNA517692, and available for viewing and download from the link: http://wheatgenomics.plantpath.ksu.edu/1000EC

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

Targeted capture re-sequencing of 890 accessions of hexaploid wheat cultivars and landraces, and tetraploid wild and domesticated emmer wheat. Diversity data was used to study selective sweeps associated with adaptation and and gene flow from the wild relative. The effects of improvement selection, adaptation and gene flow on deleterious allele load was investigated. The impact of gene flow on phenotypic variation in wheat was investigated by conducting GWAS for number of agronomic traits and by partitioning genetic variance between introgression hotspots and deserts of the wheat genome.

Research sample

Sample of accessions was selected to represent maximum genetic and geographic diversity of hexaploid bread wheat (Triticum aestivum).

Sampling strategy

For this purpose, 3990 accessions from 106 countries that were genotyped using the Infinium iSelect 90K SNP wheat bead chip array. The MCG method was used to select accessions for exome sequencing. This method uses a genetic relationship matrix calculated from genotype data to iteratively chose an arbitrary number of accessions (in the case of this study 20% of the accessions) that collectively explain the largest proportion of variance in genetic relationships among the whole set.

Data collection

Wheat seeds are obtained from the USDA National Small Grains Collection and the Australian Grains Genebank (formally the Australian Winter Cereal Collection). Seeds were grown by selfing for two generations to remove possible residual heterozygosity.

Timing and spatial scale

Not applicable

Data exclusions

Using generated SNP data, all accessions were clustered using the phylo program from the VCF-kit. Based on the patterns of clustering inconsistent with the clustering of wheat accessions having the same improvement status (wild, domesticated, landrace, cultivar), we have removed from analyses three wheat landraces (PI 345355, PI 131592, PI 534284), and two accessions of wild and domesticated emmer (PI 467000, PI415152).

Reproducibility

Not directly applicable to our study. The study focused on characterizing existing patterns of genetic diversity to detect targets of selection and evidence of gene flow.

Randomization

Field trials were conducted for two consecutive years under rainfed and irrigated conditions at the Agriculture Victoria departmental research station located at Horsham, Victoria, Australia. Each wheat accession was sown in triplicated 4.5 m single rows in a randomized block design in each of the rainfed and irrigated trials, with a seed-to-seed density of 3.6 cm and row-to-row spacing of 65 cm.

Blinding

Not applicable. Our study is not focused on detecting the outcome of biological process but rather on characterizing the existing patterns of genetic diversity.

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions

Field trials were conducted for two consecutive years under rainfed and irrigated conditions at the Agriculture Victoria departmental research station located at Horsham, Victoria, Australia. Horsham is located in a medium rainfall zone, with average annual rainfall of 400 mm and a temperate climate.

Location

Locations of field trials: 1) rainfed trial 2014: latitude- 36°45'3.97"S, longitude- 142° 6'57.51"E; 2) irrigated trial 2014: latitude-

Location	36°44'38.29"S, longitude- 142° 6'12.40"E; 3) rainfed trial 2015: latitude- 36°44'14.77"S, longitude- 142° 6'50.79"E; 4) irrigated trial 2015: latitude- 36°44'26.36"S, longitude- 142° 6'6.17"E.
Access and import/export	Not applicable
Disturbance	Not applicable

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study	
☒	□	Unique biological materials
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
☒	□	Human research participants

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging