ARTICLES
https://doi.org/10.1038/s41477-020-00845-2

nature plants

Check for updates

The role of gene flow and chromosomal instability in shaping the bread wheat genome

Alexandra M. Przewieslik-Allen[1]✉, Paul A. Wilkinson[1,2], Amanda J. Burridge[1], Mark O. Winfield[1], Xiaoyang Dai[1], Mark Beaumont[1], Julie King[3], Cai-yun Yang[3], Simon Griffiths[4], Luzie U. Wingen[4], Richard Horsnell[5], Alison R. Bentley[6], Peter Shewry[7], Gary L. A. Barker[1] and Keith J. Edwards[1]

Bread wheat (*Triticum aestivum*) is one of the world's most important crops; however, a low level of genetic diversity within commercial breeding accessions can significantly limit breeding potential. In contrast, wheat relatives exhibit considerable genetic variation and so potentially provide a valuable source of novel alleles for use in breeding new cultivars. Historically, gene flow between wheat and its relatives may have contributed novel alleles to the bread wheat pangenome. To assess the contribution made by wheat relatives to genetic diversity in bread wheat, we used markers based on single nucleotide polymorphisms to compare bread wheat accessions, created in the past 150 years, with 45 related species. We show that many bread wheat accessions share near-identical haplotype blocks with close relatives of wheat's diploid and tetraploid progenitors, while some show evidence of introgressions from more distant species and structural variation between accessions. Hence, introgressions and chromosomal rearrangements appear to have made a major contribution to genetic diversity in cultivar collections. As gene flow from relatives to bread wheat is an ongoing process, we assess the impact that introgressions might have on future breeding strategies.

Bread wheat (*Triticum aestivum*) is an allohexaploid (AABBDD) derived from the hybridization of the diploid *Aegilops tauschii* (DD) with the tetraploid *Triticum turgidum* (AABB) approximately 10,000 years ago[1,2]. The limited number of individuals involved in the initial hybridization, combined with subsequent inbreeding, means that bread wheat has reduced levels of diversity compared with its progenitors, particularly in the D genome[3,4]. This lack of genetic diversity limits the ability of breeders to develop cultivars able to respond to environmental challenges. Although elite wheat cultivars per se show limited genetic variation, their extended gene pool[5] (including landraces, progenitors and relatives belonging to the Triticeae tribe) provides a reservoir of alleles that may be exploited to increase diversity[6,7]. Until the past century, hexaploid bread wheat and its tetraploid relatives were commonly grown together (this practice still occurs in some regions of the world), and natural gene flow between them was possible; indeed, recent studies have demonstrated that gene flow from wild emmer (*T. turgidum* subsp. *dicoccoides*) has had a substantial effect on genome-wide single nucleotide polymorphism (SNP) diversity in wheat[8,9].

In breeding programmes, tetraploid species considered to be members of the primary gene pool of hexaploid bread wheat remain an important source of genetic diversity, and crosses with bread wheat are easily made[3,8]. Members of the secondary and tertiary gene pools have also contributed to diversity in the bread wheat genome. Although some selection of introgressed alleles may have been carried out by farmers and early breeders, the directed exploitation of introgressions started in the second half of the twentieth century when breeders recognized the value of relatives as sources of diversity for traits such as disease resistance. Recorded examples of such transfers include the translocation of the short arm of chromosome 1R from rye to the long arm of chromosome 1B in wheat, which occurred independently in several breeding programmes in the 1920s and is considered the most successful alien introduction into hexaploid bread wheat[10,11]. This introgression was initially intended to confer race-specific resistance to leaf, stem and stripe rust, as well as to powdery mildew[12], but the relatively rapid evolution of these fungal pathogens has meant that these resistances are no longer effective against new races in most environments[11,13]. Other recorded introgressions include the *A. ventricosa* 7D introgression conferring eyespot resistance and the uncharacterized *T. turgidum* subsp. *dicoccoides* introgressions in the wheat variety Robigus and its progeny[14–16]. Additionally, the creation of synthetic hexaploids through the hybridization of diploids and tetraploids has been an implementation of introgression in global breeding programmes to increase D-genome diversity[17,18].

Here we present a comprehensive study of gene flow between bread wheat and its relatives, based on a high-density (HD) array of molecular markers that work across multiple species in the *Triticum–Aegilops* complex[15]. In a genome-wide screen, the haplotypes of 358 bread wheat accessions are compared with those of 113 wheat relatives representing 44 species from the genera *Triticum*, *Aegilops*, *Amblyopyrum*, *Thinopyrum* and *Secale*. In so doing, we have produced an extensive catalogue of predicted introgressions from the primary, secondary and tertiary gene pools. To ensure the utility of the data, we have created interactive tools with which to interrogate and view the results. Our data show that the introgression of wheat relative DNA has been a common feature throughout the evolution and cultivation of bread wheat and represents a logical strategy to further increase the genetic diversity of this important crop.

[1]Life Sciences, University of Bristol, Bristol, UK. [2]Institute of Systems, Molecular & Integrative Biology, University of Liverpool, Liverpool, UK. [3]Plant Sciences Building, School of Biosciences, The University of Nottingham, Sutton Bonington, UK. [4]The John Innes Centre, Norwich, UK. [5]NIAB, Cambridge, UK. [6]International Maize and Wheat Improvement Center (CIMMYT), El Batán, Mexico. [7]Rothamsted Research, Harpenden, UK. ✉e-mail: A.Allen@bristol.ac.uk

Results

Array screening results. The 471 accessions screened on the wheat HD array[15] consisted of 358 hexaploid wheat lines and 113 wheat relatives. The call rates for hexaploid bread wheat accessions ranged from 95.5% to 99.3% with an average of 98.5%, while for wheat relatives the lowest call rate was 82.5% and the highest 98.5% with an average call rate of 93.5% (Supplementary Table 2). Over 80% (657,217) of the SNPs on the array were assigned an unambiguous physical position on the International Wheat Genome Sequencing Consortium (IWGSC) genome assembly with even distribution across the whole genome. There were 666,592 unambiguous alignments and 152,448 ambiguous alignments of a total of 819,040 (Supplementary Table 3). Unambiguously aligned SNPs to IWGSC v.0.4 assembly were evenly distributed across the chromosomes, although there was a drop in numbers around the centromeres, as we would expect with exome-derived SNPs[19] (Supplementary Fig. 1).

Diversity within heritage and modern hexaploid bread wheat collections. The bread wheat accessions screened on the wheat HD array[15] had registration dates ranging from 1790 to 2015 and include elite cultivars, heritage and landrace accessions (Supplementary Table 2). This set was divided into five collections on the basis of date of registration and breeding method (Collections 1–4, conventionally bred; Collection 5, hybridized novel synthetics). Regarding the genetic diversity, Collection 5 (the synthetics) exhibited the highest levels of diversity across all three genomes (Extended Data Fig. 1a). Of the conventionally bred collections (1–4), Collection 1 (1790–1930) had the highest level of genetic diversity, followed by Collection 3 (1966–1985), Collection 4 (1986–2015) and lastly Collection 2 (1930–1965). In these collections, the B genome had the highest level of genetic diversity, while the D genome had the lowest. In contrast, in the synthetic collection, diversity was highest in the D genome.

Variation within the genome was revealed by investigating the average genetic diversity separately for each chromosome (Extended Data Fig. 1b). Generally, the pattern for each chromosome reflected that of its base genome, but some chromosomes deviated from this: for example, chromosomes 1B and 2B in Collections 3 and 4 had higher average genetic diversity than the B genome in general, and chromosomes 7A and 4A in Collection 4 exhibited higher average genetic diversity. To determine which accessions were contributing to these patterns of increased diversity, principal component analysis (PCA) was performed for each chromosome separately. This revealed varied grouping patterns and mixtures of subgroups for each chromosome, indicating that ancestry and/or genome structure was different for each chromosome. In some cases, outlying samples were so distantly placed from all other samples that they effectively compressed the clusters of other accessions; these samples are listed in Supplementary Table 1 and included accessions from all four collections. These samples were removed from the relevant chromosome datasets, and the remaining samples were reclustered.

Repeating the PCA after the removal of extreme outliers better revealed subgroups within the collections (Fig. 1 and Supplementary Table 4). The cluster pattern was different for each chromosome, most likely indicating varying historical relatedness between accessions for different chromosomes. Some accessions were distinct and isolated in the PCA; others fell together in small subclusters. These subclusters tended to be of accessions belonging to the same collection; in particular, accessions from Collection 1 and Collection 4 formed distinct subclusters. For some chromosomes, such as chromosomes 6A and 6B, the PCA revealed a broad sweep of variation with the earlier landrace accessions (Collection 1) to the left of the *y* axis trending to the later collections (2 to 4) with positive values. An analysis of the coefficient of variation for each collection and each chromosome revealed regions of elevated diversity (Extended Data Fig. 2). Such regions of diversity were present in all collections or specific to one or more collections. For example, there was a large region of differentiation on the short arm of chromosome 1B in Collections 3 and 4, which can be attributed to the rye introgression present in some varieties in these collections. Interestingly, the D-genome chromosomes from Collection 5 (the synthetics) showed a striking elevation of diversity.

To analyse the patterns observed in the PCA plots and to identify the chromosomal regions responsible for the variation observed, genetic markers under selection were identified in each of the bread wheat collections (Fig. 2 and Supplementary Table 5)[20]. For each collection, unique markers or regions under selection were identified. The analysis revealed highly significant regions of selection in the landrace collection (1790–1930) on chromosomes 1D, 2D, 4A, 4D, 6B and 7B along with numerous other less significant regions. For Collection 2 (1931–1965), marker clusters across chromosomes 1D and 2B and at the telomeric ends of 4A showed significant signs of selection. For Collection 3 (1966–1986), regions on chromosomes 1B, 1D, 2D, 5A and 5D showed significant signs of selection. Finally, Collection 4 (1986–2015) had numerous tightly clustered regions with highly significant support for selection on chromosomes 1A, 1B, 1D, 2A, 2B, 2D, 3D, 4A, 4B, 5A, 5D and 6B. In some cases, these clusters of markers can be related to known genomic features such as translocations and introgressions within particular accessions. For example, the markers distinguishing the outlying cluster for chromosome 2A (Supplementary Table 5) were derived from the 2NS–2AS translocation from *Aegilops ventricosa*. On the basis of our results, we hypothesize that outlying samples in all plots contain substantial genome events such as deletions, inversions and translocations including introgressed segments from related *Triticale* species. Further smaller regions of genetic diversity are likely to represent similar regions maintained under selection within each collection.

To investigate the outliers detected in our initial PCA screen, nine accessions were subjected to genomic in situ hybridization (GISH) as a standard technique to identify large deletions and chromosome abnormalities (Supplementary Table 1). The GISH analysis revealed large-scale genomic rearrangements and deletions in the majority of accessions analysed and confirmed in many cases that these are most likely responsible for the cluster patterns seen in the PCA (Supplementary Table 1, Fig. 3 and Supplementary Fig. 2). This analysis also revealed the presence of smaller (so far undocumented) chromosomal rearrangements: translocations, deletions or introgressions for lines Russet, SoGood, Diablo, Atou and Gondola. Some accessions demonstrated variation in chromosome numbers between plants (Watkins 816) and even within plants (CIMCOG 59). In four cases, a deletion or rearrangement was not detected on the relevant chromosome via GISH (for example, cultivar Diablo chromosome 2B). Here we hypothesize that the cluster pattern was caused by substantial sequence variation over a large region, potentially due to an introgressed region from a wheat relative, as seen with the documented 2B *T. timopheevii* introgression on lines Cook and Sunvale. We believe that unknown events requiring further investigation (Supplementary Table 1) could represent a deletion, an introgression or a chromosomal rearrangement such as a large-scale gene conversion or translocation.

To further investigate the link between patterns of diversity observed in the accessions and genomic events such as introgressions and deletions, we screened individual accessions for copy number variation (CNV) across the genome and the presence of introgressed regions from wheat relatives. To detect deletions and rearrangements in the genome, all accessions were subjected to CNV analysis[21]. Using this procedure, we were able to highlight SNP loci associated with the large deletions detected via GISH and also revealed further, smaller genome events in all lines. We have

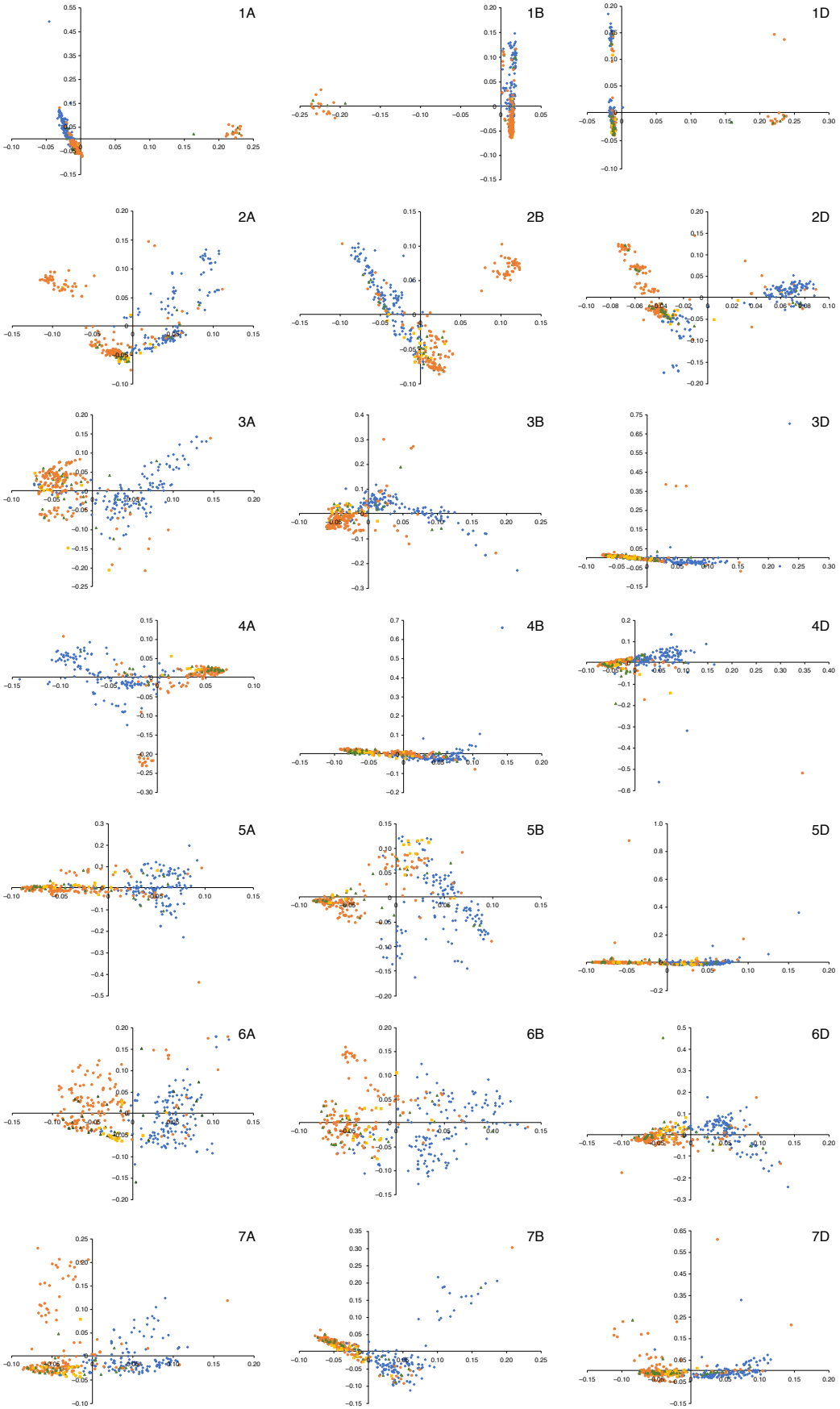


Fig. 1 | PCA plots of hexaploid bread wheat accessions for each chromosome. In all plots, PC1 is plotted along the *x* axis, and PC2 is plotted along the *y* axis. The samples are coloured on the basis of accession release date: blue, Collection 1 (1790–1930); yellow, Collection 2 (1931–1965); green, Collection 3 (1966–1985); orange, Collection 4 (1986–2015).

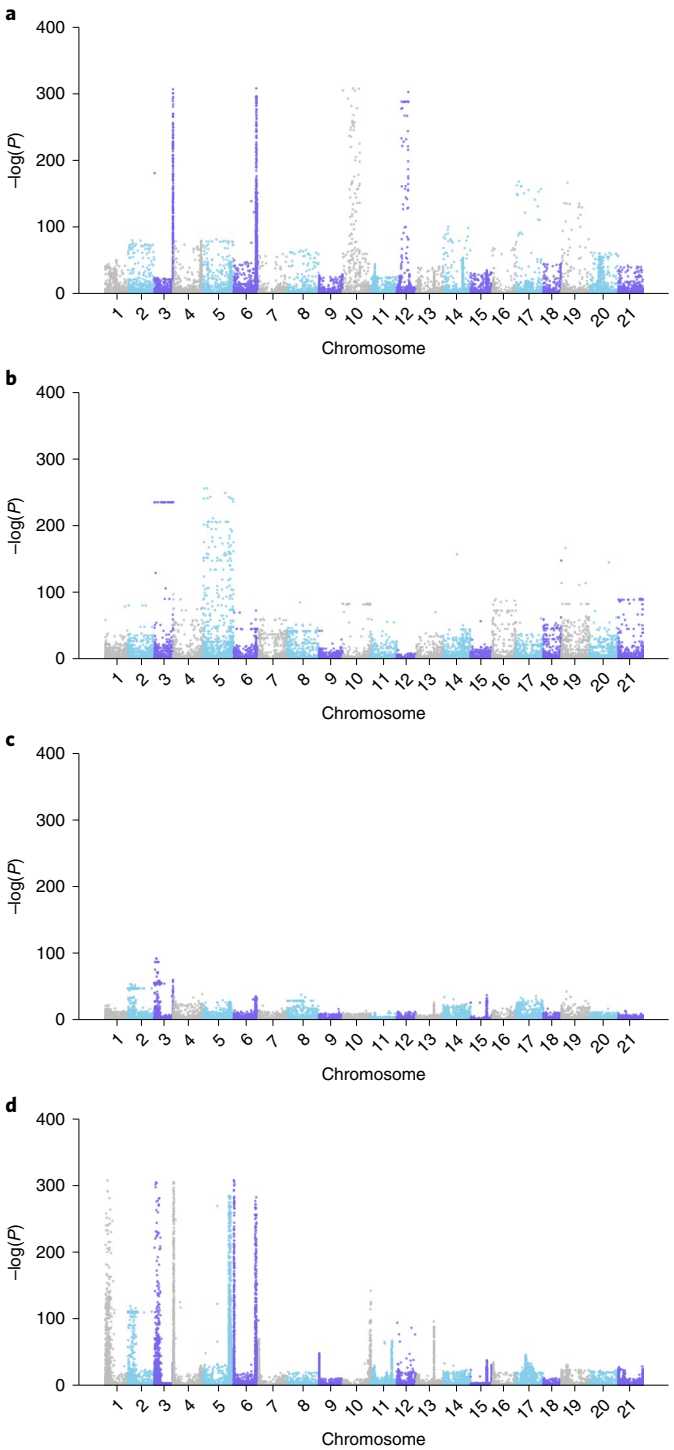


Fig. 2 | Manhattan plots of $-\log_{10}$-transformed *P* values assigned to each SNP marker derived from the Mahalanobis distance test statistic calculated via PCAdapt. **a**, Landraces, 1790–1930 (Collection 1). **b**, Accessions, 1931–1965 (Collection 2). **c**, Accessions, 1966–1985 (Collection 3). **d**, Accessions, 1986–2015 (Collection 4). The peaks indicate genomic regions potentially under selection. The analysis was performed separately for each collection. Chromosomes were labelled 1–21 in the order: 1A, 1B, 1D, 2A, ..., 7D.

shown previously that regions where a reduced hybridization signal is observed may be indicative of a deletion or an introgression from a wheat relative[21]. Of the accessions screened in this study, all lines had multiple copy number losses and gains detected, ranging in size from 17 base pairs (bp) to 44.1 mega base pairs (Mbp; Supplementary Table 6). The elevation and reduction of copy numbers can be associated with translocations, deletions and introgressions and has been included in our introgression discovery tool (described below).

Detection of introgressions. To identify predicted introgressions from wheat relatives, we screened 113 relative accessions with the Axiom HD arrays, in addition to the hexaploid bread wheat accessions described above. These wheat relatives are considered to be the main source species for introgressed alleles, but they may not represent a complete list. Previous examination of lines carrying the known introgressions from rye (1RS) or *A. ventricosa* confirmed the validity of using these markers[15]. Regions of similarity were identified and assigned a physical map position in the genome, giving both the location and the size of predicted introgressions (Supplementary Table 7). Using a new pipeline to detect introgressions and bespoke tools to visualize them, we identified numerous introgressions; some of these have been reported previously, but many have not. The extent of introgression and the size of individual introgressed regions varied considerably between accessions and individual chromosomes (Fig. 4).

Introgression from the primary gene pool. A large percentage of hexaploid wheat accessions (47–179, or 14.5–55.1% of the lines studied, depending on the chromosome studied) showed evidence of introgression from tetraploid species (Supplementary Table 7). The sizes of these introgressions varied from 19,626 bp (0.002% of chromosome 2B) to 213 Mbp (30.0% of chromosome 7A). Collections 2, 3 and 4 contained fewer and smaller introgressions than Collection 1. The average introgression size was larger in the B genome of Collection 4 (80.7 Mbp) than in the A genome (36.5 Mbp). Conversely, in Collection 1, the average introgression size was higher in the A genome (83.1 Mbp) than in the B genome (65.0 Mbp). Collection 5 (synthetics) had larger areas of similarity to *T. turgidum* subsp. *dicoccoides* than to any other tetraploid because this species was the donor of the AB genomes to most modern synthetic wheats[3].

While gene flow from tetraploids to hexaploids is known to have occurred, much less is known about the transfer of genes to the D genome. To examine the relationship between the D genome of hexaploid wheat accessions and multiple *A. tauschii* accessions, we generated a similarity matrix based on D-genome markers (Supplementary Table 8). *A. tauschii* accessions split into two groups; the first included *A. tauschii* subsp. *tauschii* and *A. tauschii* subsp. *strangulata*, while the second consisted of *A. tauschii* subsp. *tauschii*. All bread wheat accessions were more similar to group 1 *A. tauschii* (average, 0.72) than to group 2 accessions (average, 0.36). The *A. tauschii* accession with the highest similarity to the D genome of hexaploid wheat was Ent 088, an *A. tauschii* subsp. *strangulata* accession collected in Iran and allocated to the first *A. tauschii* group, consistent with previous reports[22]. We further examined this relationship by identifying genomic regions of similarity between bread wheat and *A. tauschii* accessions. A specific *A. tauschii* region was assigned to between 5 and 31 hexaploid accessions (Supplementary Table 7). The sizes of these regions of similarity varied from 83,636 bp (1.1% of chromosome 7D) to 84 Mbp (16.4% of chromosome 4D) with an average of 4.9 Mbp. The reduced frequency and size of these regions in the D genome suggests that, in contrast to the A and B genomes, relatively little gene flow has occurred. The synthetic collection had substantially larger regions of similarity (with an average total region size of 74 Mbp compared with 1.4 Mbp in Collections 1–4) to the accessions of the second group of *A. tauschii*, reflecting the source of D-genome donors used in the creation of these lines.

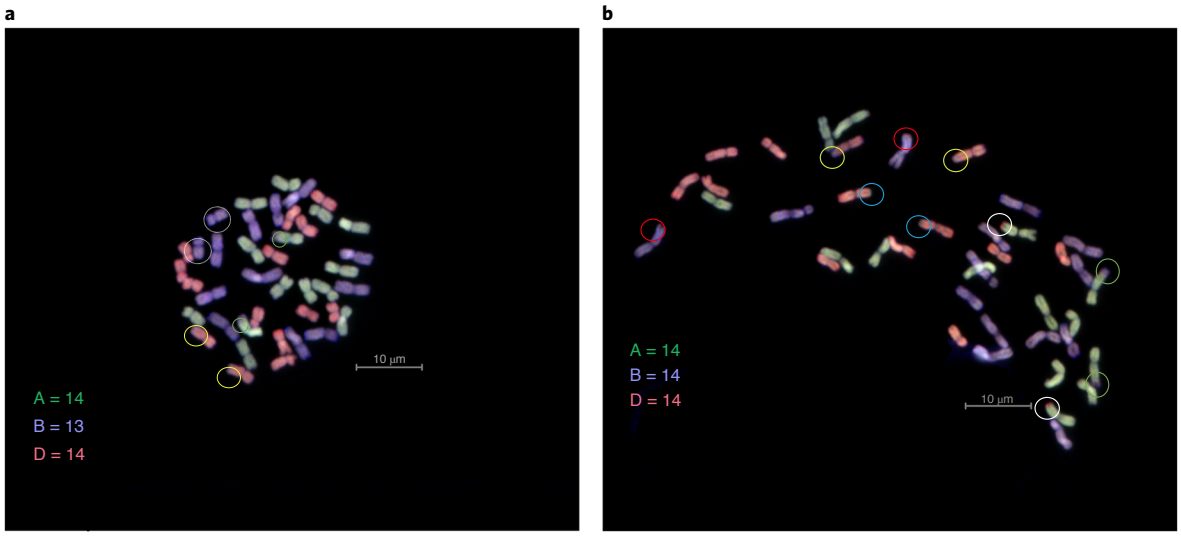


Fig. 3 | Detection of translocations/introgressions in accessions Russet and Diablo using GISH. a, Russet is aneuploid with only 13 B chromosomes. The B genome also has two metacentric chromosomes (dotted circles) that are smaller than expected. Two A and two D chromosomes have telomeric translocations or introgressions (yellow circles). **b**, Diablo is euploid but has translocations/introgressions on five pairs of chromosomes: one pair has a translocation from D (white circles), one pair of B chromosomes has an apparent telomeric translocation from the D genome (red circles), one pair of D chromosomes has a telomeric translocation from the A genome (blue circles) and one pair has distal introgression from the B genome (yellow circles). The green circles indicate the relatively widespread 4A/5A/7B translocation in both varieties. For each line, the experiment was repeated twice independently using separate plants.

Finally, we examined areas of similarity between bread wheat (*T. aestivum* subsp. *aestivum*) collections and other hexaploid *T. aestivum* subspecies. Regions of similarity varied from 197 kbp in length (0.03% of chromosome 2D) to 557 Mbp (74.8% of chromosome 4A). The oldest landraces and heritage cultivars had the most regions of similarity to *T. aestivum* subspecies, and the majority of these were detected in the A and B genomes. The most common and largest regions of similarity were detected in Indian shot wheat (*T. aestivum* subsp. *sphaerococcum*), followed by club wheat (*T. aestivum* subsp. *compactum*).

Introgressions from the secondary and tertiary gene pools. Several breeding programmes have employed wide crosses in wheat to introduce novel alleles from members of its secondary and tertiary gene pools. However, the species and accessions used for such crosses are often not documented. Here, we attempt to identify such introgressed regions and assign sources from within the secondary and tertiary gene pools. There are distinct differences in the patterns of introgressions between the wheat collections from different periods (Fig. 5). Collections 3 and 4 consistently had the largest average total introgression size detected for all secondary and tertiary gene pools. A trend was observed of increased hexaploid and tetraploid gene flow in accessions bred pre-1960 to the introduction of more exotic introgressions in post-1960 accessions. This reflects the change in the breeders' use of the gene pool over the 60 years since the Green Revolution[23]. To examine the introgressions in detail, the data have been summarized for each accession for all relatives and displayed as a heat map alongside CNV data to help to identify which relative might be the likely donor. To examine the sizes and locations of the introgressions, the data can be viewed as a Circos plot where it is viewed alongside CNV data (Fig. 4). The introgression data confirm the GISH results and support the conclusion of an introgression where no deletion or other rearrangement was detected by GISH for the accessions Cook, Sunvale and Diablo.

The predicted introgression data were compared with regions identified as contributing significantly to the diversity within a collection. Regions with consecutive SNPs with highly significant support for being under selection ($P < 0.001$; Fig. 2) were found to be correlated with a predicted deletion or introgression identified in the collection (Table 1), suggesting that these genome features have a significant impact on the genetic diversity of a collection. In Collection 4, these included the 1BL–1RS rye translocation, the 2NS–2AS *A. ventricosa* introgression and the *Triticum timopheevii* introgression on 2B. In Collection 1, a large *T. aestivum* subsp. *macha* introgression on 4A was detected, common to 52 heritage and landrace accessions originating from Europe, the Middle East, Asia and Australia. A screen of five publicly available Watkins × Paragon quantitative trait locus (QTL) datasets revealed seven examples where QTLs for agronomic traits were located in genomic regions predicted to have introgressions in the Watkins parent in our data (http://wisplandracepillar.jic.ac.uk/results_resources.htm; Supplementary Table 9).

A case study was conducted into the important cultivar Robigus from Collection 4, a cultivar known to have *T. turgidum* subsp. *dicoccoides* in its breeding background. Six different *T. dicoccoides* accessions were screened as part of this study, and predicted introgressions from each in Robigus were mapped onto the genome. The predicted introgressions ranged from 18 Mbp to 24.5 Mbp on the A and B genomes (Supplementary Table 10). One predicted *T. dicoccoides* introgression on 4A located between 732 Mbp and 743 Mbp and containing seven SNPs was highlighted as under selection in Collection 4 by the PCAdapt analysis and was exclusive to Robigus and its progeny (14 of the accessions included in this study). One accession, Player, appeared to have an intermediate haplotype for this region, with heterozygous genotypes for the loci. We further examined this region using exome-captured sequence data and identified 308 polymorphic SNPs between seven varieties (Robigus, Chinese Spring, Avalon, Cadenza, Rialto and Savannah from a previously published study[24]; Player from this study). Of these SNPs, 139 (45%) had a genotype unique to Robigus (compared between 1% and 10% for the other six varieties examined). Of the 308 SNPs in the region, Player exhibited a match to Robigus or a heterozygous genotype at 55.8% of the loci. In four of the regions

highlighted by PCAdapt, there was evidence in a small number of accessions for the inheritance of smaller or partial introgressions, suggesting that the segment has been broken down or recombined during the breeding process (Table 1).

We further investigated the bread wheat accession Player, as it was predicted to contain exotic introgressions from a range of sources. This accession was analysed by exome capture and Illumina sequencing to study regions of correspondence between the two platforms. From the exome capture data, potential introgressions were identified as regions with low coverage when BLASTed against the Chinese Spring sequence. In general, coverage was greater at distal regions of the chromosomes, reflecting the higher gene density towards the telomeres. Chromosomes 2A and 2B showed distinct drops in coverage where predicted introgressions from *A. ventricosa* and *T. timopheevii*, respectively, were thought to be located (Fig. 6). However, the predicted introgression on 5A from *T. aestivum* subsp. *compactum* is not reflected by a drop in sequence coverage. This is because the donor species is a close relative and the sequences map efficiently onto the wheat reference. For more exotic donors, a drop in sequence coverage is detected, as the sequence variation prevents the sequences from being mapped. This pattern is also reflected in the CNV analysis, where introgressions from closer relatives (in the primary or secondary gene pool) do not result in a detected loss of copy number, while regions from exotic (tertiary gene pool) relatives do. Figure 6 shows the alignment of the exome capture data with predicted introgressions and diversity measures, highlighting the impact that introgressions from wheat relatives have had on this modern-day wheat cultivar.

Discussion

The Axiom 820k wheat HD SNP array has been used for diversity screens of a large collection of wheat lines and to develop pipelines to extract information on CNV and predicted introgressions from wheat relatives[15]. The genome events detected by the data have been independently investigated by GISH and exome sequencing. A summary of the workflow is illustrated in Supplementary Fig. 3. The bread wheat accessions examined in this study varied by year of registration, country of origin and growth habit, providing a global snapshot of genetic diversity present in different collections during different temporal periods of wheat breeding and development. As has previously been observed, a higher level of diversity was observed for all cultivar collections on the A and B genomes than on the D genome, reflecting the genetic bottleneck imposed during the formation of hexaploid wheat[4]. The most modern collection, comprising synthetic hexaploid wheats, exhibited the opposite pattern, with higher diversity detected on the D genome than on the A and B genomes. This is because the D-genome progenitors were purposely chosen to be as diverse as possible during the creation of these novel wheats, and this reflects the potential genetic diversity existing in wild populations of *Aegilops tauschii*[23,25]. The contribution of the increased diversity to the gene pool of historic and modern breeding material is demonstrated by the striking increase in variation plotted in Extended Data Fig. 2.

Except for recently created synthetic wheats, genetic diversity was highest in the collection of bread wheat accessions dating from 1790 to 1930. Diversity dropped in accessions from 1931 to 1965 but then increased in the period after the Green Revolution with the advent of modern breeding before levelling off again. There was substantial variation observed between chromosomes, genomes and bread wheat accessions. PCA plots revealed different cluster patterns for each chromosome for each of the collections. Plots with little structure and differentiation may represent chromosomes conserved from landraces and taken through into modern breeding lines. Some chromosomes, such as 6A and 6B, exhibit diffuse cluster patterns typical of whole-chromosomal consistent variation. However, most chromosomes had unique and tighter cluster patterns for the accessions screened. Outliers were investigated cytologically via GISH and were found to have large, whole-chromosome or chromosome arm deletions and introgressions. CNV analysis confirmed the large deletions visualized by GISH and additionally highlighted many smaller deletions and rearrangements in these accessions. The polyploid nature of the wheat genome makes such structural alterations viable, and our findings support a high degree of chromosomal variability in our set of accessions dating from the past 200 years.

For samples located in subclusters, we have demonstrated that the cluster patterns are likely to be caused by substantial sequence variation over a large region, potentially due to an introgressed region from a wheat relative, as seen with the documented rye 1B.1R translocation and the 2B *T. timopheevii* introgression on lines Cook and Sunvale. Our introgression analysis was used to further investigate the nature of undocumented genome events detected via GISH and CNV, to produce a catalogue of predicted introgressions and to develop interactive online tools for visualizing and interrogating the data. The SNP data and introgression pipeline generated were accurate at predicting introgressions even where a close relative was the donor—a result that might be missed if looking at sequence coverage alone.

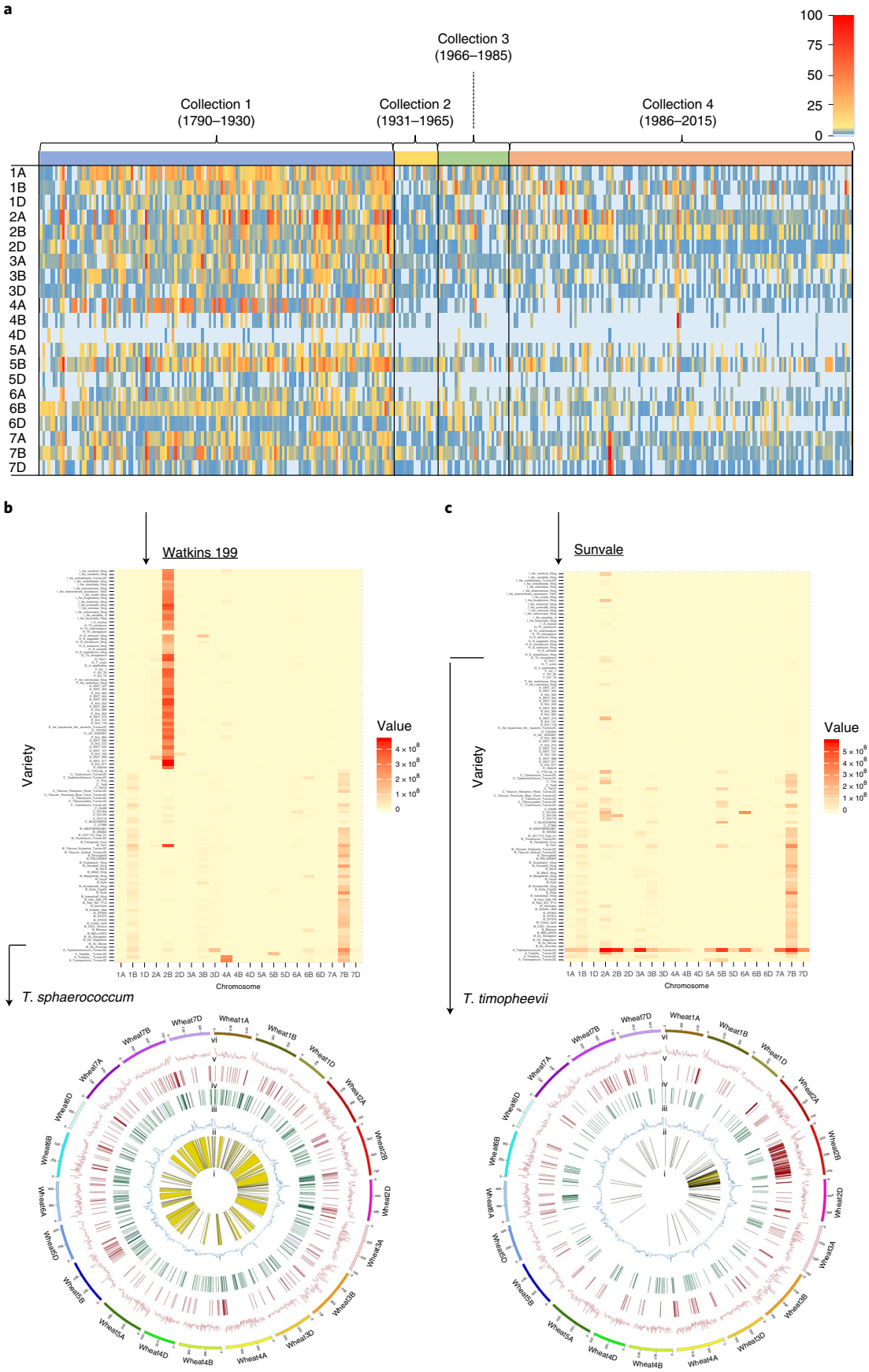
We have demonstrated that the older bread wheat accessions, in contrast to modern cultivars, have more introgressions from primary relatives (*T. aestivum* and *T. turgidum* subspecies). As has been previously suggested, these findings provide evidence for interspecies crossing occurring historically between hexaploid bread wheat and tetraploid and hexaploid relatives, resulting in introgressed material being incorporated in the A and B genomes of bread wheat[3,8,9]. This is likely to account for the higher level of diversity in the A and B genomes than in the D genome, where introgressions are thought to be few due to the lack of suitable donors[25,26]. The predicted introgressions unique to Collection 1 may represent areas of diversity contained in the primary gene pool but absent from current elite bread wheat germplasm and hence may represent potential breeding targets. Overall, Collections 2, 3 and 4 contained fewer and smaller primary gene pool introgressions than Collection 1. This suggests either that the ancestors chosen in the breeding of modern elites were those with fewer introgressions from related tetraploid species or that breeder-driven selection pressure has subsequently removed many tetraploid-derived alleles.

In contrast, modern cultivars (from 1986 onwards) have a higher proportion of predicted introgressions from more distant relatives (in the secondary and tertiary gene pools) than historical cultivars and landraces. While historical crossing may have resulted in 'accidental' introgressions due to different species being cultivated alongside each other, the incorporation of introgressed regions from secondary and tertiary gene pool relatives is a result of more targeted crossing for a specific purpose or trait. To achieve this, breeders have reached back to before polyploidization and hybridization

Fig. 4 | Summary of the introgression data. a, Heat map of the maximum introgression size detected on each chromosome for each accession. The accessions are ordered horizontally on the basis of date of release. Introgression size is shaded according to size as a percentage of the chromosome. **b**,**c**, Introgression plotter images selected accessions from Collection 1 (Watkins 199) (**b**) and Collection 4 (Sunvale) (**c**). The heat map plots the total introgression size for each comparison, with relatives ordered vertically by relatedness to bread wheat and chromosomes ordered horizontally. Circos plots: i, introgressed regions; ii, SNP density for each wheat chromosome; iii, CNV gain; iv, CNV loss; v, minor allele frequency; vi, wheat chromosomes (500 = 500 Mbp).

events to wild relatives to introduce specific regions. Some of these introgressions are well documented, but for others, records have not been well maintained or linked to specific varieties.

The detected introgressions varied in size, with some varieties showing large and widespread areas of introgressed material from specific relatives. For example, Watkins 199, collected in India,



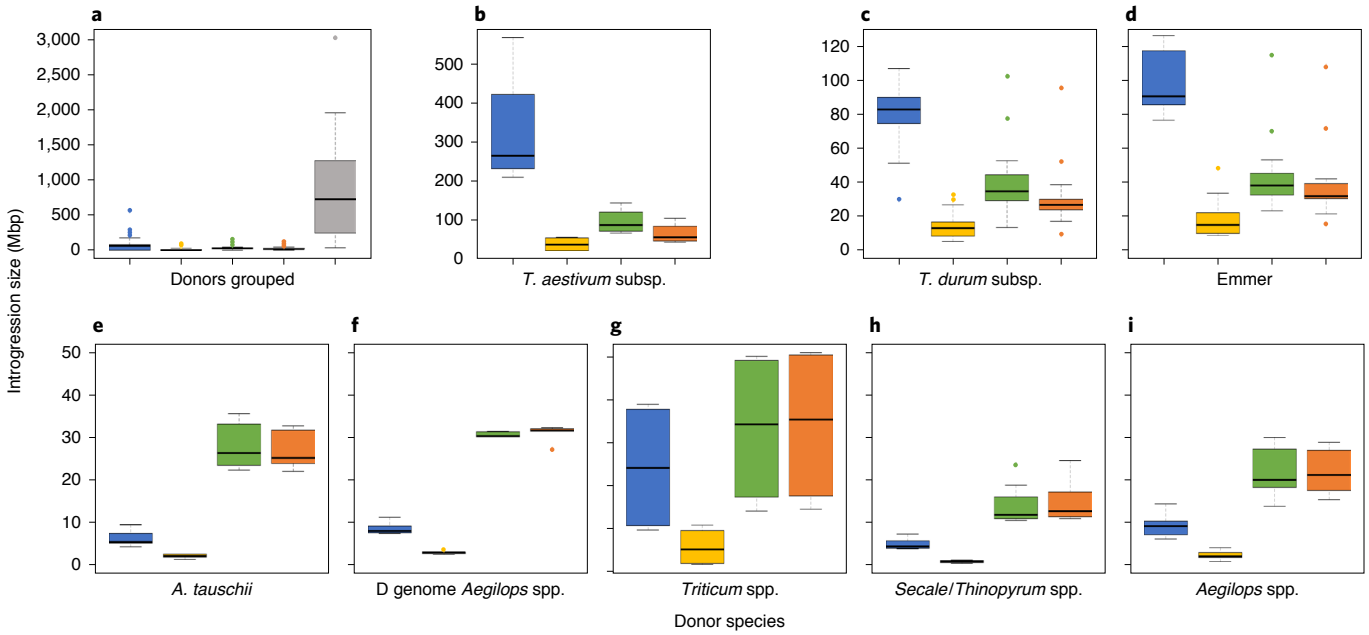


Fig. 5 | Average total introgression size for wheat collections based on the predicted donor wheat relative. a, Average total introgression size including Collection 5 (synthetics) for all wheat relative donors. **b–i**, Average total introgression size without Collection 5, divided into different donor classifications. The collections are coloured on the basis of accession release date: blue, Collection 1 (1790–1930); yellow, Collection 2 (1931–1965); green, Collection 3 (1966–1985); orange, Collection 4 (1986–2015); grey, Collection 5 (novel synthetics). $n = 358$ biologically independent samples split into 147, 18, 29, 141 and 23 for Collections 1–5, respectively.

Table 1 | Regions with consecutive SNPs with highly significant support for being under selection ($P < 0.001$) using the Mahalanobis distance test statistic calculated by PCAdapt, implementing and associated genome features detected via CNV and introgression analysis

Collection	Chromosome	Start position (bp)	End position (bp)	Predicted number of genes	Number of significant SNPs	Associated genome feature	Number of lines
1	1D	484402678	495229142	235	183	Deletion	25 (+6 partial)
1	2D	461537459	627113794	2,483	475	Deletion	39 (+4 partial)
1	4A	11844358	448254938	1,265	53	*T. macha* introgression	52
1	4D	122193919	302688837	1,257	42	Unknown	52
2	1D	5357381	494553729	5,904	302	Deletion (Tet1)	1
2	4A	5464991	721412913	1,732	80	Unknown	1
4	1A	4046985	236720937	1,175	118	Unknown	16
4	1B	1431485	621636781	4,001	631	*S. cereale* introgression	15
4	1D	5517400	162065293	1,661	69	Unknown	16
4	2A	254840	191878486	1,534	201	*A. ventricosa* introgression	51 (+3 partial)
4	2B	667805430	742524112	630	336	*T. timopheevii* introgression	48
4	2D	274484	22818447	2,745	134	Unknown	50
4	2D	558802310	608200664	740	203	Unknown	47
4	3A	12479443	13413554	34	24	*T. dicoccum* introgression	29
4	4A	732512426	742549832	139	6	*T. dicoccoides* introgression	14 (+2 partial)
4	5A	461492802	470170989	76	87	Unknown	17

The exact *P* values for all significant SNPs are provided in Supplementary Table 5.

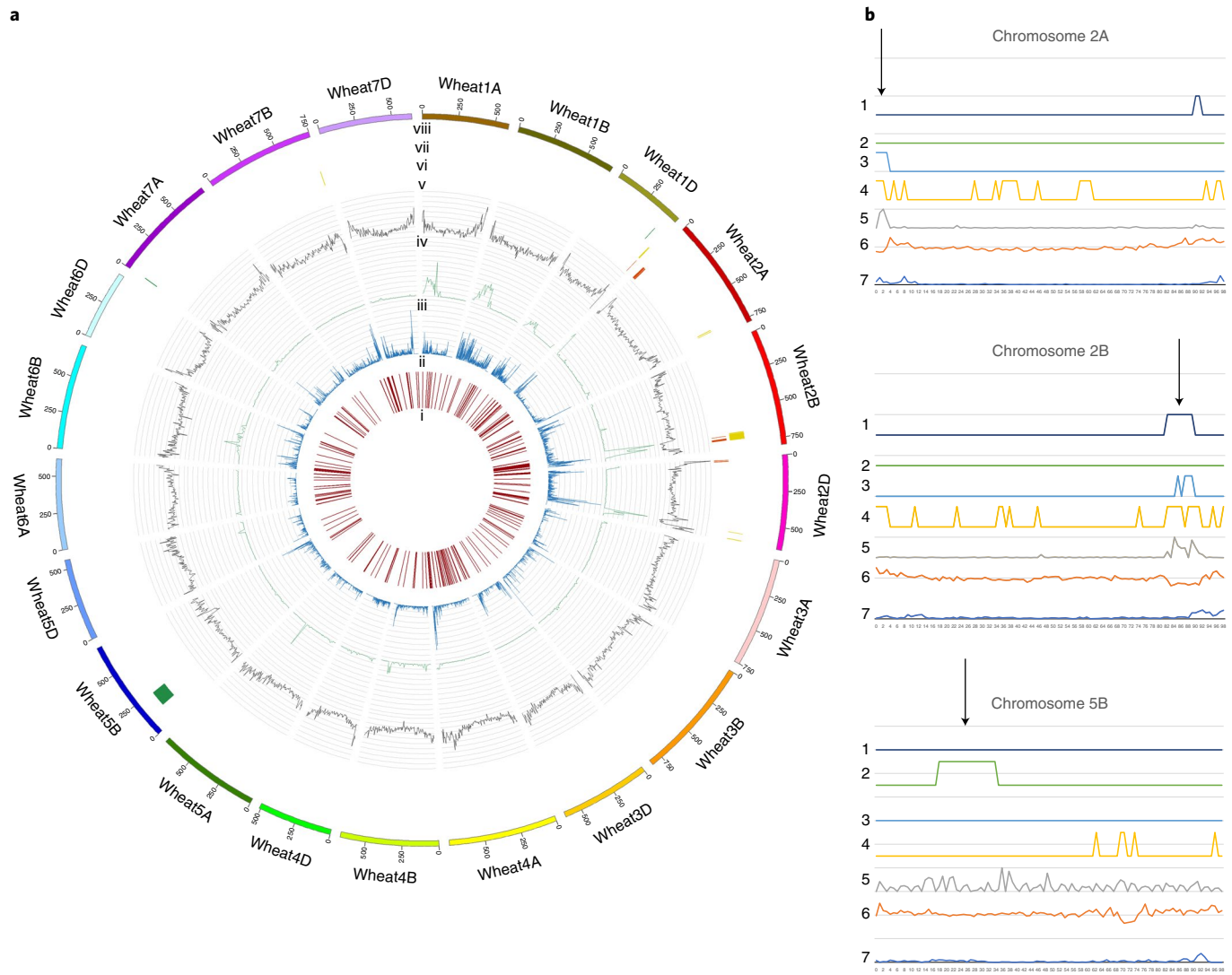


Fig. 6 | Alignment of the exome capture data with predicted introgressions and diversity measures. a, Circos plot of predicted introgressed regions, signatures of selection and CNV across the genome for the variety Player. i, CNV loss; ii, Collection 4 s.d. data (the s.d. values are used to summarize the data from all varieties in a collection, with peaks in the values representing regions where introgression scores vary because a subset of varieties have an introgression not shared by all varieties examined); iii, $-\log_{10}$-transformed *P* values assigned to each SNP marker derived from the Mahalanobis distance test statistic calculated via PCAdapt for Collection 4; iv, Player exome capture sequence coverage; v, predicted *Aegilops ventricosa* introgressions; vi, predicted *Triticum timopheevii* introgressions; vii, predicted *Triticum aestivum* subsp. *compactum* introgressions; viii, ideogram of wheat chromosomes with scales added (100 equals 100Mbp). **b**, Individual plots for chromosomes 2A, 2B and 5B were made using the Circos data (**a**), rescaled to percentage of the maximum on both axes. The arrows indicate the locations of predicted introgressions. (1) *T. timopheevii* introgression; (2) *T. aestivum* subsp. *compactum* introgression; (3) *A. ventricosa* introgression; (4) CNV loss; (5) $-\log_{10}$-transformed *P* values assigned to each SNP marker derived from the Mahalanobis distance test statistic calculated via PCAdapt for Collection 4; (6) exome capture sequence coverage; (7) Collection 4 s.d. data.

exhibited extensive similarity (39.0% of the genome) to *T. aestivum* subsp. *sphaerococcum* across most chromosomes (chromosomes 2A, 2B, 3A, 5B, 6A, 7A, 7B and 7D in particular had over 50% similarity). Another heritage line, April Bearded (dating from 1838), had large areas of similarity to *T. aestivum* subsp. *compactum*, particularly on chromosome 4A, where the predicted introgression extended across 76.2% of the chromosome. This 4A introgression appears widespread in the heritage/landrace collection of accessions and is responsible for the selection signature seen in Collection 1 detected by PCAdapt. Other regions, highlighted by the PCAdapt genome scan, were investigated in detail and linked to genome structural abnormalities such as introgressions and deletions (Table 1). These included relatively common introgressions in Collection 4 from *A. ventricosa* (24.5% of chromosome 2A) and *T. timopheevii* (9.3% of chromosome 2B), which were present in 51 and 48 accessions, respectively. These important introgressions were used to bring multiple disease-resistance genes into the wheat genome. The 2NS/2AS translocation[27] derived from *A. ventricosa* carries Lr37 for leaf rust resistance, Sr38 for resistance to stem rust caused by *Puccinia graminis tritici* and Yr17 for resistance to stripe rust, *P. striiformis*[28,29]. *T. timopheevii* was the source of the stem rust resistance gene Sr36 and the closely linked powdery mildew resistance gene, Pm6, which was transferred to wheat chromosome 2B[30]. In some cases, such as the 2A *A. ventricosa* introgression, there is evidence of some variation in the sizes of introgressions among accessions, indicating that the region has been 'eroded' or recombined during the breeding process (Table 1). As with the GISH data, the observation of multiple undocumented

introgressions was also reflected in the introgression data for many accessions.

The introgression analysis demonstrated that the oldest accessions in this study (Collection 1) contain widespread and varied introgressed material and promise to be a source of novel alleles potentially useful in enriching the current breeding gene pool. The overlap between QTLs for agronomic traits and predicted introgressions in the Watkins accessions potentially reveals phenotypes associated with novel introgressed alleles. One important modern breeding line, Robigus, is widely regarded to have *T. dicoccoides* in its pedigree[14]. Robigus has recently been named the most influential UK wheat variety from the past 100 years by NIAB and is found in 50% of the recommended list varieties since 2014. An investigation into introgressions detected in this important cultivar revealed multiple predicted *T. dicoccoides* introgressions, including 10.0 Mbp and 21.3 Mbp on chromosomes 4A and 5B, respectively. The 4A introgression was exclusive to Robigus and its progeny in our data, with one line, Player, exhibiting heterozygous genotypes in this region. Player is a French variety from the cross H00143/Inoui, but we were unable to ascertain whether Robigus was present further back in the pedigree. The case studies of both Robigus and Player demonstrate how these data and tools may be used to investigate both genomic regions and specific varieties of interest. When combined with other resources such as the IWGSC genome assembly and QTL data, it is clear how far the field of wheat genomics has advanced over the past decade, and we are now in a position to relate knowledge at the DNA sequence level to phenotypic features.

Our data show that introgressions from wheat relatives have shaped the genetic diversity present in bread wheat cultivars both historically and in modern elite breeding lines. Specific areas of increased diversity in bread wheat collections were aligned with regions exhibiting elevated historical selection within a collection and further associated with genome features such as deletions and wheat relative introgressions. Every accession screened had multiple genome events detected in the form of introgressions, deletions and rearrangements, suggesting that no single wheat variety can be used to represent the wheat genome and why it is important to study wheat's pangenome[31]. Such an observation has important ramifications for wheat breeders, as it suggests that by using a relatively narrow selection of lines in breeding programmes, they are more likely to lose segments and hence diversity. GISH analysis revealed a further important factor: of the nine accessions analysed, two lines showed variation between the two plants screened, with one of these showing substantial variation in cells from the same plant (Supplementary Fig. 2). Overall, this indicates that the hexaploid wheat genome remains a highly dynamic structure, which continues to evolve, both by the introduction of novel genetic material from a range of relatives and by genome-specific rearrangements, including deletions and translocations typical of a polyploid. Our detailed analysis of 471 wheat genomes shows that both gene flow from wheat relatives and chromosomal instability have shaped and continue to shape the bread wheat genome. Bread wheat is one of the major sources of calories for humankind. It is therefore important that wheat breeders have access to this knowledge so that they can fully utilize the plasticity of the genomes to develop new high-yielding and resilient varieties.

Methods

Plant material. The accessions grown for DNA extraction (listed in Supplementary Table 2) were obtained from the Germplasm Resources Unit (https://www.seedstor.ac.uk/) and NIAB (synthetic lines); they were grown in peat-based soil in pots and maintained in a glasshouse at 15–25 °C with a 16-h-light, 8-h-dark cycle. Leaf tissue was harvested from six-week-old plants, immediately frozen on liquid nitrogen and then stored at 20 °C before nucleic acid extraction. Genomic DNA was prepared from the leaf tissue using a phenol–chloroform extraction method[32]. The genomic DNA samples were treated with RNase-A (New England Biolabs UK) according to the manufacturer's instructions and purified using the QiaQuick PCR purification kit (QIAGEN).

Genotyping. The Axiom Wheat HD Genotyping Arrays (Affymetrix UK) were used to genotype 471 samples using the Affymetrix GeneTitan system according to the procedure described by Affymetrix (Axiom 2.0 Assay Manual Workflow User Guide Rev3). These arrays contain 819,571 SNPs obtained from genic sequences derived via targeted capture resequencing of numerous wheat lines[15]. Allele calling was carried out using the Affymetrix proprietary software packages Affymetrix Power Tools (Release 1.15.0), following the Axiom Best Practices Genotyping Workflow (http://media.affymetrix.com/support/downloads/manuals/axiom_genotyping_solution_analysis_guide.pdf) and using the thresholds established for optimized genotype calling for this array[15].

Exome capture. Genomic DNA from leaf tissue (14 days after germination) was extracted, RNase treated and purified[32]. A total volume of 55 µl was sheared to an average of 300 bp using a E220 Focused-ultrasonicator (Covaris). The SeqCap EZ HyperCap Workflow User's Guide (v.2.0) was used with the following modifications. The starting material was increased to 2 µg of DNA. The A-tailing reaction was changed to 20 °C for 30 minutes, followed by 65 °C for 30 minutes. The size selection of the precapture libraries was replaced with a 0.9 bead:sample ratio. The precapture amplification was changed to nine cycles followed by immediate clean-up. The COT human DNA was replaced with 1 µl of Developer Reagent Plant Capture Enhancer (NimbleGen) per 100 ng of DNA. Exome capture was performed using Gene Capture v.1, 4000026820 and Promoter Capture v.1, 4000030160 wheat capture probes[33]. The gene and promoter capture probes were not lyophilized, but the captures were performed separately. For the capture wash, the first Wash Buffer I and both Stringent Wash Buffer steps used buffer preheated to 57 °C. The fragment size distribution throughout was determined by TapeStation (Agilent) analysis. Capture-probe-enriched sequencing libraries for the gene and promoter capture probes were sequenced separately at the Bristol Genomics Facility using NextSeq 500 and NextSeq500 2 × 150 bp Mid-Output v.2.5 kit (Illumina). A final library concentration of 0.8 pM was used with a 5% PhiX control library.

The Illumina paired-end reads were trimmed for adapter sequences and by quality scores using the FASTQ preprocessor fastp (v.0.20.0), and both trimmed and raw reads were assessed using the quality-control tool FastQC (v.0.11.4) to ascertain the effect of the preprocessing step. After trimming, 333,952,552 (92.1%) reads remained, and these were aligned to the Chinese Spring wheat genome (IWGSC v.1.0) using the Burrows–Wheeler aligner bwa (v.0.7.7-r441) using the bwa-mem algorithm. The number of reads that could be mapped to the genome was 333,709,069 (99.9%), and the number that mapped to genes (that is, on-target reads) was 75,489,075 (22.6%), with an average depth of coverage for gene sequences of 681× (median, 399×) and an average breadth of coverage for gene sequences of 92.6%. Read coverage was calculated at each position in the genome using bedtools (v.2.25.0) using the coverage option. Coverage graphs were then plotted for each chromosome using the ggplot2 package (v1.0.0) in R (v.3.2.5) using a bin size of 5,000,000 bp.

Introgression identification. Regions of similarity were identified and assigned a physical map position in the genome, giving both the location and the size of the predicted introgression. This data table is available at www.cerealsdb.uk.net/cerealgenomics/Introgression_table_280617.tdt.

SNP markers were assigned a physical map position by BLAST searching the probe sequences to the IWGSC whole genome assembly v.0.4 (https://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies).

The genotyping data were initially processed to identify SNP markers with a minor allele present at a frequency of 0.2 or below in the hexaploid lines, which was also shared with a progenitor line. This threshold was selected to identify genome calls that were effectively present in 20% or less of the screened hexaploids to focus on less widespread regions. The next part of the introgression prediction pipeline takes the predicted scores for each chromosome from the previous step, uses a sliding window of ten SNPs along the chromosome and calculates the average similarity scores for consecutive ten-SNP windows for each of the varieties, to identify predicted regions of introgression along each chromosome. A summary file was then generated for each variety to identify regions of introgression for each chromosome. The summary file reports the chromosomal location and size of predicted introgressions above 0.4 for each wheat–relative comparison. The analysis pipeline achieved this by looking for consecutive SNPs that had a relative match of 0.4 or higher, and if the next SNP along did not have a relative match of 0.4 or greater, the preceding SNP was classified as a small introgression. A threshold of 0.4 was chosen following empirical testing on the basis of the efficient identification of known introgressed regions from a variety of donors. If successive SNPs all had a relative match of 0.4 or greater, then the distance between the first and last SNP with a match ≥0.4 was calculated to identify the approximate size of the introgression. For example, if five consecutive SNPs had a similarity score greater than 0.4, the distance between the first and the fifth SNPs would be used to calculate the size of the introgression. In this way, it was possible to identify potential introgressions along each chromosome for every combination of hexaploid wheat variety and wheat relative/progenitor species. The tools for

viewing introgressions are available on the CerealsDB website (https://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/search_introgressions.php); all Circos plots were constructed using the CerealsDB Circos tool[34]. In accordance with the philosophy of data sharing described by Moore[6], this data table is available at www.cerealsdb.uk.net/cerealgenomics/Introgression_table_280617.tdt, and the scripts are available via GitHub (https://github.com/pr0kary0te/). Similarity matrices were created using Flapjack v.1.20.10.07 (ref. [35]).

CNV analysis. CNV analysis was performed using the Affymetrix CNV Tool software (v.1.1). CEL files from the Axiom Wheat HD Genotyping Array were processed using Affymetrix Power Tools as described above. The annotation file was generated using the Affymetrix Annotation Converter (v.1.0). The Affymetrix CNV tool calculates the copy number at each SNP position for each sample using CEL intensity files and comparing them with the reference set. The $\log_2$ ratio at a marker is computed by dividing the intensity of the marker by the median intensity of that marker in the chosen reference set in log space. The reference set represents the normal copy number state for each marker and is created from the entire group of individuals genotyped, assuming that for each marker the vast majority of individuals on the plate are expected to have a normal copy number status. Events were defined as copy number gain and copy number loss using the segmentation algorithm in Nexus Copy Number v.9.0.

PCA plots. A distance matrix was generated from the genotype scores using R package SNPRelate (v.3.12)[36]. The proportions of variance for the first six eigenvalues were as follows: 8.66, 6.58, 5.85, 5.10, 3.88 and 3.68. The first two eigenvalues, accounting for over 15% of the variance, were plotted as a PCA plot.

PCAdapt. The R package PCAdapt (v.3.0)[20] was used to perform genome scans to detect genes under selection on the basis of population genomic data. The Mahalanobis distance was computed for each SNP to detect outliers for which the vector of z scores does not follow the distribution of the main bulk of points. The Mahalanobis distances were transformed into P values for multiple hypothesis testing[20]. The $-\log_{10}$-transformed P values of individual markers were plotted along the chromosome as Manhattan plots for each collection.

Cytogenetic analysis. The protocol for GISH was as described in ref. [37] with some modifications. Genomic DNA was isolated using a CTAB method[37] from young leaves of the three putative diploid progenitors of bread wheat: *T. urartu* (A genome), *A. speltoides* (B genome) and *A. tauschii* (D genome). The genomic DNA of *T. urartu* was labelled by nick translation with Chroma Tide Alexa Fluor 488-5-dUTP (Invitrogen; C11397). The genomic DNA of *A. tauschii* was labelled with Alexa Fluor 594-5-dUTP (Invitrogen; C11400). The genomic DNA of *A. speltoides* was fragmented to 300–500 bp in boiling water. Roots from each germinated introgression line were excised and treated with nitrous oxide gas at 10 bar for 2 h. The treated roots were fixed in 90% acetic acid for 10 min and then washed three times in water on ice. The root tip was dissected and digested in 20 µl of 1% pectolyase Y23 and 2% cellulase Onozuka R-10 (Yakult Pharmaceutical) solution for 50 min at 37 °C and then washed three times in 70% ethanol. The root tips were crushed in 70% ethanol, and the cells were collected by centrifugation at 3,000 *g* for 1 min, briefly dried and then resuspended in 30–40 µl of 100% acetic acid before being placed on ice. The cell suspension was dropped onto glass slides (6–7 µl per slide) in a moist box and dried slowly under cover. The slide was probed with labelled DNAs of *T. urartu* (100 ng) and *A. tauschii* (200 ng) and fragmented DNA of *A. speltoides* (5,000 ng) as blocker in the ratio 1:2:50 to detect the AABBDD genomes of wheat. All slides were counterstained with DAPI and analysed using a Zeiss Axio Imager.Z2 upright epifluorescence microscope (Carl Zeiss) with a MetaSystems Coolcube 1 m CCD camera, and image analysis was carried out using Metafer4 (v.4) and ISIS software (v.5.8.5, Metasystems).

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

Data availability

The genotype data that support the findings of this study are available in the European Variation Archive (EVA) with the identifier PRJEB29561. Source data are provided with this paper.

Code availability

The custom PERL scripts described in this study are available via GitHub (https://github.com/pr0kary0te/).

Received: 3 August 2020; Accepted: 18 December 2020;
Published online: 1 February 2021

References

1. Dubcovsky, J. & Dvorak, J. Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* **316**, 1862–1866 (2007).
2. Shewry, P. R. Wheat. *J. Exp. Bot.* **60**, 1537–1553 (2009).
3. Pont, C. et al. Tracing the ancestry of modern bread wheats. *Nat. Genet.* **51**, 905–911 (2019).
4. Haudry, A. et al. Grinding up wheat: a massive loss of nucleotide diversity since domestication. *Mol. Biol. Evol.* **24**, 1506–1517 (2007).
5. Harlan, J. R. & de Wet, J. M. Toward a rational classification of cultivated plants. *Taxon* **20**, 509–517 (1971).
6. Moore, G. Strategic pre-breeding for wheat improvement. *Nat. Plants* **1**, 15018 (2015).
7. He, F. et al. Molecular cytogenetic identification of a wheat–*Thinopyrum ponticum* translocation line resistant to powdery mildew. *J. Genet.* **96**, 165–169 (2017).
8. Cheng, H. et al. Frequent intra- and inter-species introgression shapes the landscape of genetic variation in bread wheat. *Genome Biol.* **20**, 136 (2019).
9. He, F. et al. Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. *Nat. Genet.* **51**, 896–904 (2019).
10. Molnár-Láng, M. et al. in *Genomics of Plant Genetic Resources* (eds Tuberosa, R. et al.) 255–283 (Springer, 2014).
11. Worland, A. J. & Snape, J. W. in *The World Wheat Book* (eds Bonjean, A. P. & Angus, W. J.) 59–100 (Lavoisier, 2001).
12. Villareal, R. L., Rajaram, S., Mujeebkazi, A. & Deltoro, E. The effect of chromosome 1B/1R translocation on the yield potential of certain spring wheats (*Triticum aestivum*). *Plant Breed.* **106**, 77–81 (1991).
13. Schlegel, R. & Meinel, A. A quantitative trait locus (QTL) on chromosome arm 1RS of rye and its effect on yield performance of hexaploid wheat. *Cereal Res. Commun.* **22**, 7–13 (1994).
14. Gardner, K. A., Wittern, L. M. & Mackay, I. J. A highly recombined, high-density, eight-founder wheat MAGIC map reveals extensive segregation distortion and genomic locations of introgression segments. *Plant Biotechnol. J.* **14**, 1406–1417 (2016).
15. Winfield, M. O. et al. High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotechnol. J.* **14**, 1195–1206 (2016).
16. Gale, M. D. et al. An α-amylase gene from *Aegilops ventricosa* transferred to bread wheat together with a factor for eyespot resistance. *Heredity* **52**, 431–435 (1984).
17. Jafarzadeh, J. et al. Breeding value of primary synthetic wheat genotypes for grain yield. *PLoS ONE* **11**, e0162860 (2016).
18. Hao, M. et al. A breeding strategy targeting the secondary gene pool of bread wheat: introgression from a synthetic hexaploid wheat. *Theor. Appl. Genet.* **132**, 2285–2294 (2019).
19. Lange, T. M. et al. In silico quality assessment of SNPs—a case study on the Axiom® Wheat genotyping arrays. *Curr. Plant Biol.* **21**, 100140 (2020).
20. Luu, K., Bazin, E. & Blum, M. G. pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Mol. Ecol. Resour.* **17**, 67–77 (2017).
21. Allen, A. M. et al. Characterization of a Wheat Breeders' Array suitable for high throughput SNP genotyping of global accessions of hexaploid bread wheat (*Triticum aestivum*). *Plant Biotechnol. J.* **15**, 390–401 (2016).
22. Jones, H. et al. Strategy for exploiting exotic germplasm using genetic, morphological, and environmental diversity: the *Aegilops tauschii* Coss. example. *Theor. Appl. Genet.* **126**, 1793–1808 (2013).
23. Fradgley, N. et al. A large-scale pedigree resource of wheat reveals evidence for adaptation and selection by breeders. *PLoS Biol.* **17**, e3000071 (2019).
24. Winfield, M. O. et al. Targeted re-sequencing of the allohexaploid wheat exome. *Plant Biotechnol. J.* **10**, 733–742 (2012).
25. Akhunov, E. D. et al. Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. *BMC Genomics* **11**, 702 (2010).
26. Jordan, J. D. et al. A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. *Genome Biol.* **16**, 48 (2015).
27. Fang, T. et al. Stripe rust resistance in the wheat cultivar Jagger is due to Yr17 and a novel resistance gene. *Crop Sci.* **51**, 2455–2465 (2011).
28. Bulos, M. et al. Occurrence of the rust resistance gene *Lr37* from *Aegilops ventricosa* in Argentine cultivars of wheat. *Electron. J. Biotechnol.* **9**, 580–586 (2006).
29. Xue, S. et al. Mapping of leaf rust resistance genes and molecular characterization of the 2NS/2AS translocation in the wheat cultivar Jagger. *G3 (Bethesda)* **8**, 2059–2065 (2018).
30. Allard, R. W. & Shands, R. G. Inheritance of resistance to stem rust and powdery mildew in cytologically stable spring wheats derived from *Triticum timopheevi*. *Phytopathology* **44**, 266–274 (1954).
31. Montenegro, J. D. et al. The pangenome of hexaploid bread wheat. *Plant J.* **90**, 1007–1013 (2017).
32. Burridge, A. J. et al. in *Wheat Biotechnology: Methods and Protocols* (Bhalla, P. L. & Singh, M. B.) 293–306 (Humana Press, 2017).
33. Gardiner, L. J. et al. Integrating genomic resources to present full gene and putative promoter capture probe sets for bread wheat. *GigaScience* **8**, giz018 (2019).

34. Wilkinson, P. A. et al. CerealsDB—new tools for the analysis of the wheat genome: update 2020. *Database* **2020**, 1–13 (2020).
35. Milne, I. et al. Flapjack—graphical genotype visualization. *Bioinformatics* **26**, 3133–3134 (2010).
36. Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
37. Kato, A., Lamb, J. C. & Birchler, J. A. Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize. *Proc. Natl Acad. Sci. USA* **101**, 13554–13559 (2004).

Acknowledgements

We thank the Bristol Genomics Facility for the Illumina sequencing data and the Germplasm Resource Unit (GRU) for providing many of the accessions used in this paper. We thank the Biotechnology and Biological Sciences Research Council, UK, for funding this work (award nos BB/N021061/1 and BBS/E/J/000PR9781).

Author contributions

A.M.P.-A., P.A.W., A.J.B., M.O.W., G.L.A.B. and K.J.E. conceived and planned the experiments. S.G., L.U.W., R.H., A.R.B. and P.S. provided the plant material for the analysis. A.J.B., J.K. and C.Y. carried out the lab experiments. A.M.P.-A., P.A.W., M.O.W., X.D., M.B., G.L.A.B. and K.J.E. planned and carried out the computational analyses. A.M.P.-A. took the lead in writing the manuscript. All authors provided critical feedback and helped interpret the data and shape the research, analysis and manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41477-020-00845-2.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41477-020-00845-2.

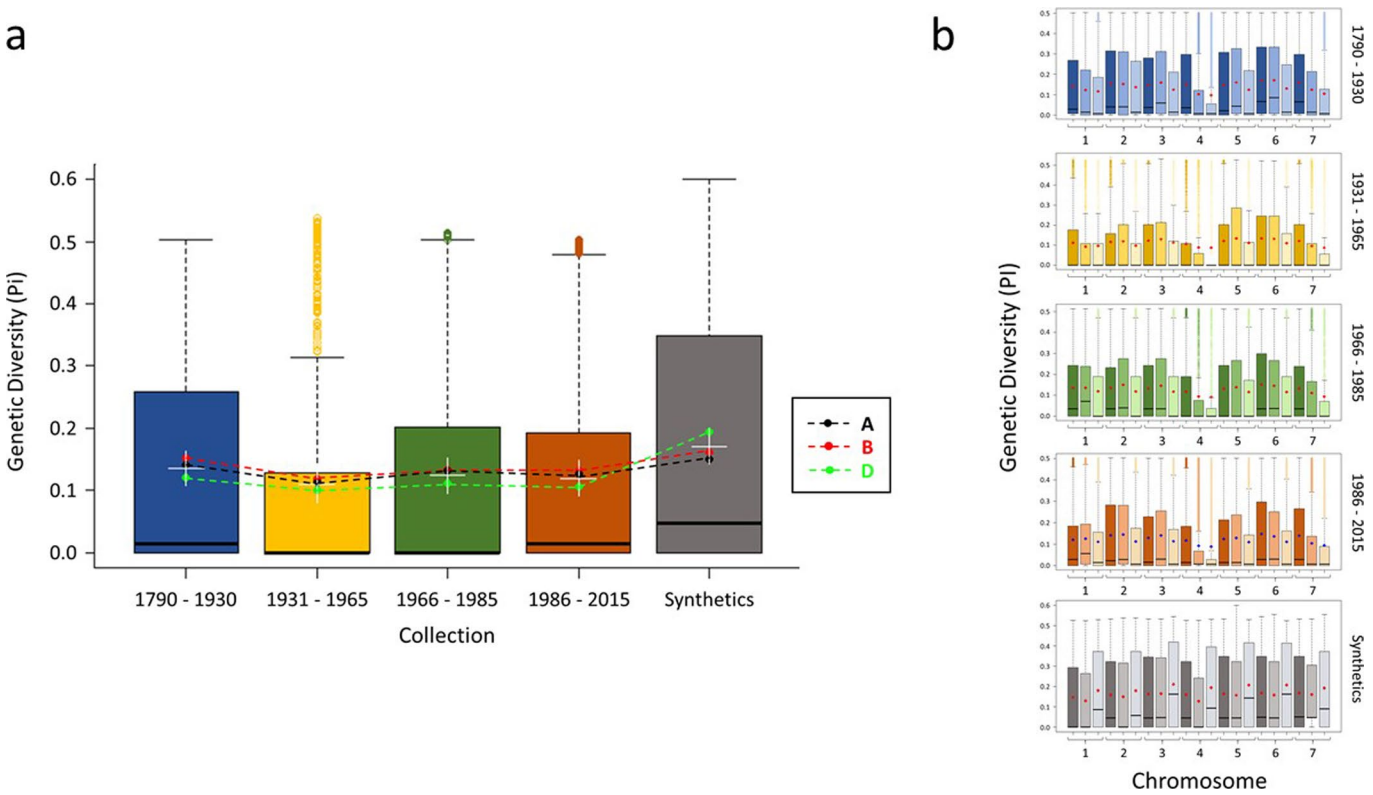
Correspondence and requests for materials should be addressed to A.M.P.-A.

Peer review information *Nature Plants* thanks Rudi Appels, Agnieszka Aleksandra Golicz and Isobel Parkin for their contribution to the peer review of this work.

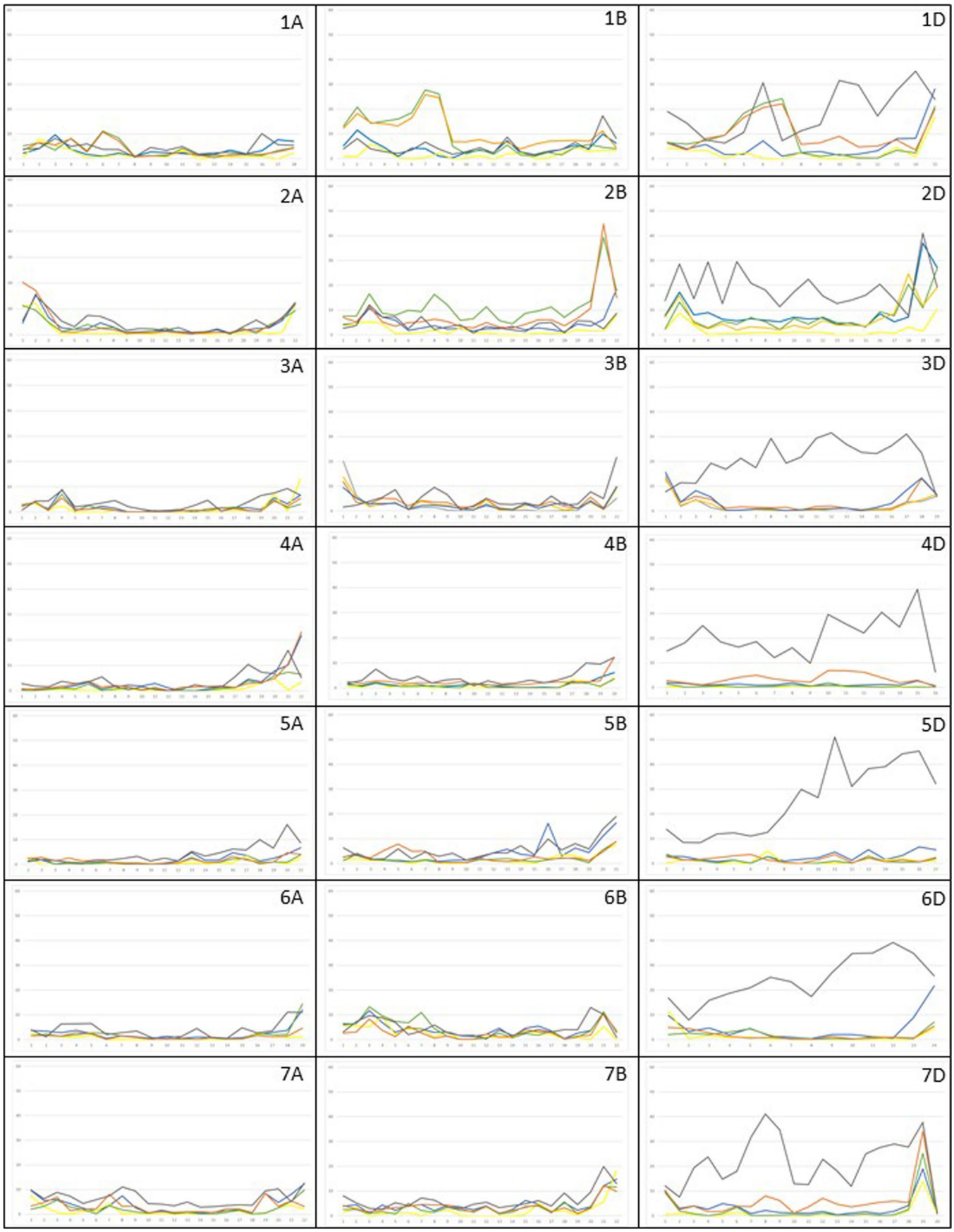
Reprints and permissions information is available at www.nature.com/reprints.

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

© The Author(s), under exclusive licence to Springer Nature Limited 2021



Extended Data Fig. 1 | Genetic diversity calculated for five collections of T. aestivum, averaged across genomes (**a**) and chromosomes (**b**). The mean is displayed as a white cross (**a**) or red dots (**b**). For each collection chromosomes are ordered 1A, 1B, 1D, … 7D. A two-sided Kruskal-Wallis (non-parametric ANOVA) showed that that mean diversity values were significantly different among the three genomes ($H = 10{,}270$, $n = 3{,}376{,}940$, d.f. = 2, $p = 0.000$) and among the five populations ($H = 41{,}383$, $n = 3{,}376{,}940$, d.f. = 4, $p = 0.000$). Post-hoc Mann-Whitney two-sided tests with Bonferroni correction also showed that all pairwise comparisons between the three genomes were significant as were all comparisons between the five populations with $p = 0.000$ in all cases.



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | The standard deviation (STDEV) of introgression scores for each chromosome. Chromosomes are divided into approximately 50 Mb bins (x-axis) and separate lines are shown for each bread wheat collection based upon accession release date: blue, Collection 1 (1790-1930); yellow, Collection 2 (1931-1965); green, Collection 3 (1966-1985); orange, Collection 4 (1986-2015); grey, Collection 5 (novel synthetics). The STDEV values are used to summarise data from all varieties, with peaks in the values representing regions where introgression scores vary because a sub-set of varieties have an introgression not shared by all varieties examined. For example, the clear peaks in Chromosome 1B arise because a sub-set of varieties have the 1BL/1RS introgression.

nature research

Corresponding author(s): Alexandra Przewieslik-Allen

Last updated by author(s): Dec 11, 2020

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 our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the 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
☐	☒	A statement on 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 statistical parameters 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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Affymetrix Power Tools (Release 1.15.0) was used to extract genotype calls from Affymetrix GeneTitan CEL files. For exome sequence data the illumina paired end reads were trimmed for adapter sequences and by quality scores using the FASTQ pre-processor fastp (version 0.20.0) and both trimmed and raw reads were assessed using the quality control tool FastQC (version 0.11.4) to ascertain the effect of the pre-processing step. GISH image analysis was carried out using Metafer4 (version 4) and ISIS software (version 5.8.5).
Data analysis	Sequence reads were aligned using the Burrows-Wheeler aligner bwa (version 0.7.7-r441). Read coverage was calculated using bedtools (version 2.25.0). Custom PERL scripts used to identify introgressions are available via GitHub (https://github.com/pr0kary0te/). Copy number variation (CNV) analysis was performed using the Affymetrix CNV Tool software (version 1.1) and visualised in Biodiscovery Nexus Copy Number (version 9.0). The R packages SNPRelate (version 3.12) and PCAdapt (version 3.0) were used for PCA analyses. Flapjack (1.20.10.07) was used to generate similarity matrices.

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 and reviewers. 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

The genotype data that support the findings of this study are available in the European Variation Archive (EVA) with the identifier PRJEB29561

(https://www.ebi.ac.uk/eva/?eva-study=PRJEB29561). The data has also been made available through the cerealsDB database (www.cerealsdb.uk.net/cerealgenomics/Introgression_table_280617.tdt). SNP markers were assigned a physical map position by BLAST searching the probe sequences to the International Wheat Genome Sequencing Consortium (IWGSC) whole genome assembly v0.4 (https://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies). Germplasm can be searched and obtained from the Germplasm Resources Unit (https://www.seedstor.ac.uk/).

Field-specific reporting

Please select the one below that is 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/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	Sample sizes were selected based on being a representative subset of the larger landrace, heritage and elite cultivar germplasm collections determined from diversity screens. All NIAB-produced synthetic wheats available were included. Wheat relative accessions were selected from the germplasm collection available to us, with multiple accession where possible.
Data exclusions	No data were excluded from the analyses.
Replication	Replicates of 10 samples were included for quality control purposes with >99% agreement.
Randomization	Samples were randomised across multiple genotyping arrays. Accessions were organised into relevant groups based upon date of registration.
Blinding	Investigators were blinded during data collection and analyses by keeping the sample group designation and description anonymous.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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