Analysis of the bread wheat genome using whole-genome shotgun sequencing

Rachel Brenchley1, Manuel Spannagl2, Matthias Pfeifer2, Gary L. A. Barker3, Rosalinda D’Amore1, Alexandra M. Allen1, Neil McKenzie4, Melissa Kramer5, Arnaud Kerhornou6, Dan Bolser6, Suzanne Kay1, Darren Waite4, Martin Trick4, Ian Bancroft4, Yong Gu7, Naxin Huo7, Ming-Cheng Luo8, Sunish Sehgal9, Bikram Gill9, Sharyar Kianian10, Olin Anderson7, Paul Kersey6, Jan Dvorak8, W. Richard McCombie5, Anthony Hall1, Klaus F. X. Mayer2, Keith J. Edwards3, Michael W. Bevan4 & Neil Hall1

Bread wheat (Triticum aestivum) is a globally important crop, accounting for 20 per cent of the calories consumed by humans. Major efforts are underway worldwide to increase wheat production by extending genetic diversity and analysing key traits, and genomic resources can accelerate progress. But so far the very large size and polyploid complexity of the bread wheat genome have been substantial barriers to genome analysis. Here we report the sequencing of its large, 17-gigabase-pair, hexaploid genome using 454 pyrosequencing, and comparison of this with the sequences of diploid ancestral and progenitor genomes. We identified between 94,000 and 96,000 genes, and assigned two-thirds to the three component genomes (A, B and D) of hexaploid wheat. High-resolution synteny maps identified many small disruptions to conserved gene order. We show that the hexaploid genome is highly dynamic, with significant loss of gene family members on polyploidization and domestication, and an abundance of gene fragments. Several classes of genes involved in energy harvesting, metabolism and growth are among expanded gene families that could be associated with crop productivity. Our analyses, coupled with the identification of extensive genetic variation, provide a resource for accelerating gene discovery and improving this major crop.

With a global output of 681 million tonnes in 2011,1 bread wheat accounts for 20% of the calories consumed by humans2 and is an important source of protein, vitamins and minerals. It originated from hybridization between cultivated tetraploid emmer wheat (AABB, Triticum dicoccoides) and diploid goat grass (DD, Aegilops tauschii) approximately 8,000 years ago.3 Bread wheat cultivation and domestication has been directly associated with the spread of agriculture and settled societies, and it is now one of the most widely cultivated crops owing to its high yields and nutritional and processing qualities. The three diploid progenitor genomes, AA from Triticum urartu, BB from a species that is unknown but which may be of the section Sitopsis (to which Aegilops speltoides belongs), and DD from Ae. tauschii, radiated from a common Triticeae ancestor between 2.5 and 4.5 million years ago.4,5 Nucleotide diversity in the AABB and DD genomes is substantially reduced compared with ancestral populations, indicating a major diversity bottleneck on the transition to cultivated lines.6

Grass genomes show extensive long-range conservation of gene order7–9. Nevertheless, they are highly dynamic owing to the activities of repeats that contribute to tremendous variation in genome size10, changes in local gene order and pseudogene formation, particularly in larger genomes such as those of maize11 and wheat12. From analysis of BAC contigs on chromosome 3B, the 17-gigabase-pair (Gb) genome was estimated to be composed of approximately 80% repeats, primarily retroelements, with a gene density of between 1 per 87 kilobase pairs and 1 per 184 kilobase pairs.13 Despite both the substantial knowledge gained of the wheat genome from these studies and the central importance of the wheat crop, a comprehensive genome-wide analysis of gene content has yet to be conducted owing to its large size, repeat content and polyploid complexity.

We have analysed a low-coverage, long-read (454) shotgun sequence of the hexaploid wheat genome using gene sequences from diverse grasses. From this, we created assemblies of wheat genes in an orthologous gene family framework, used diploid wheat relatives to classify homeologous relationships, and defined a genome-wide catalogue of single nucleotide polymorphisms (SNPs) in the A, B and D genomes. These analyses provide a foundation for genetic and genomic analysis of this key crop.

Sequence analysis

The wheat variety Chinese Spring (CS42) was selected for sequencing because of its wide use in genome studies14,15. Purified nuclear DNA was sequenced using Roche 454 pyrosequencing technology (GS FLX Titanium and GS FLX + platforms) to generate 85 Gb of sequence (220 million reads), corresponding to approximately fivefold coverage on the basis of an estimated genome size of 17 Gb. Supplementary Table 1 shows that 79% of the reads had matches to the Triticeae Repeat Sequence Database, and most hit retrotransposons, consistent with previous studies.16 To identify A-, B- and D-genome-derived gene assemblies in the hexaploid sequences, we used illumina sequence assemblies of Triticum monococcum, related to the A-genome donor, Ae. speltoides complementary DNA (cDNA) assemblies and 454 sequences from the D-genome donor Ae. tauschii, respectively. The SOLiD platform was used to generate additional sequence of CS42 and three commercial wheat varieties to increase the accuracy of homoeologous SNP identification. Data sets are summarized in Table 1 and Supplementary Table 2, and SNP

1 Centre for Genome Research, University of Liverpool, Liverpool L69 7ZB, UK. 2 MIPS/IBIS, Helmholtz- Zentrum München, 85764 Neuherberg, Germany. 3 School of Biological Sciences, University of Bristol, Bristol BS8 1UG, UK. 4 John Innes Centre, Norwich NR4 7UH, UK. 5 Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA. 6 European Bioinformatics Institute, Hinxton CB10 1SD, UK. 7 USDA Western Regional Laboratory, Albany, California 94710, USA. 8 Department of Plant Sciences, University of California, Davis, California 95616, USA. 9 Department of Plant Pathology, Kansas State University, Manhattan, Kansas 66506, USA. 10 Department of Plant Sciences, North Dakota State University, Fargo, North Dakota 58018-6050, USA.
identification methods are described in Supplementary Information, section 5.2.

**Sequence assembly**

An orthologous group assembly (Supplementary Table 3) was created by clustering 454 reads by sequence similarity to orthologous grass gene sequences, and separate assembly of the clusters at high stringency using Newbler (Supplementary Information, section 2). The orthologous genes were derived from rice, sorghum, *Brachypodium* and barley full-length cDNAs by OrthoMCL clustering. This generated 20,496 orthologous groups (Supplementary Table 4 and Supplementary Fig. 1). The gene model with highest similarity to wheat (termed the orthologous group representative (OGR)) was selected from each orthologous group by stringent BLASTX comparison to a low-copy-number genome assembly (LCG) made by filtering out repetitive sequences and assembling the remaining low-copy-number sequences de novo (Supplementary Table 3). The assemblies are described in Table 2. Nearly 90% of the metabolic genes in *Arabidopsis* matched OGRs, and the 20,051 OGRs matched 92% of publicly available wheat chromosomes 1A, 1B and 1D (Supplementary Table 6).

Table 1 | Sequence sources used for analysis

| Genome | Platform | Size of data set | Reference |
|--------|----------|-----------------|-----------|
| *T. aestivum* (CS42) genomic DNA | 454 GS FLX Titanium/454 GS FLX+ | 85 Gb | EBI study: ERP000319 Ref. 12 |
| *T. aestivum* (CS42) genomic DNA from sorted chromosomes 1A, 1B and 1D | 454 GS FLX Titanium | 1A: 287 Mb 1B: 392 Mb 1D: 375 Mb | EBI study: ERP001493 |
| *T. aestivum* (CS42) genomic DNA | SOLiD 3/SOLiD 4 | 15.2 billion reads | EBI study: ERP001415 NCBI archive: SRP004490.3 |
| *T. aestivum* (CS42) cDNA | 454 GS FLX Titanium/454 GS FLX+ | 1.6 Gb | EBI study: ERP001415 NCBI archive: SRP004490.3 |
| *T. aestivum* cDNA | SOLiD 4 | 80–100-fold coverage | M.-C.L. et al., submitted |
| Ae. tauschii genomic DNA | 454 GS FLX Titanium | 12.8 Gb | J. Dvorak, unpublished observations |
| Ae. tauschii genomic DNA | SOLiD 4 | | |

EBI, European Bioinformatics Institute; NCBI, US National Center for Biotechnology Information.

With multiple gene copies. In contrast, the low depth coverage by the LCG assembly suggested that gene family numbers were collapsed. The number of wheat assemblies for each OGR was calculated to determine gene copy numbers (Supplementary Table 7). Figure 1c shows that most OGRs had between one and five distinct wheat gene assemblies, with a peak of two genes.

The A, B and Ae. tauschii (D) genome3,13,20,21 have been estimated to contain approximately 28,000, 38,000 and 36,000 genes, respectively. We estimated the number of genes in the hexaploid wheat genome to range between 94,000 and 96,000 (Supplementary Information, section 2.10). This is reasonably consistent with estimates based on wheat chromosome sequences13. Comparing our transcriptome assembly (Supplementary Information, sections 2.8 and 2.9) and wheat harvEST with the wheat OGRs showed that 76% and, respectively, 65% were expressed under the conditions used for RNA isolation. Similar results were found in barley, rice and maize3, indicating that the assemblies are bona fide wheat genes.

We identified many insertions and/or translocations of blocks of genome and from a triplicated rice gene set, with the introduction of hexaploid sequences. Maize reads covered their OGRs to a median depth of approximately five, consistent with fivefold coverage.

Simulated maize and triplicated rice 454 reads were used to optimize assembly parameters. Assembly at 99% minimum sequence identity (m.i.) using 40-bp overlap length predicted gene family sizes most accurately (Supplementary Figs 3–6). Wheat 454 reads were preprocessed (Supplementary Table 5) and assembled using 99% m.i. (Supplementary Tables 6 and 7) to create the orthologous group assembly. Figure 1b shows that the depth of coverage of the orthologous group assembly followed a similar pattern to maize, consistent with fivefold coverage. Simulated maize and triplicated rice 454 reads were used to optimize assembly parameters. Assembly at 99% minimum sequence identity (m.i.) using 40-bp overlap length predicted gene family sizes most accurately (Supplementary Figs 3–6). Wheat 454 reads were preprocessed (Supplementary Table 5) and assembled using 99% m.i. (Supplementary Tables 6 and 7) to create the orthologous group assembly. Figure 1b shows that the depth of coverage of the orthologous group assembly followed a similar pattern to maize, consistent with fivefold coverage.

**Genome change in polyploid wheat**

We determined the influence of polyploidy on gene content in hexaploid wheat by defining the sizes of gene families in hexaploid wheat and the diploid progenitor Ae. tauschii from the copy number of genes.
for each OGR, which were then paired with the gene family size of the OGR in sequenced diploid grasses (Supplementary Information, section 2.6). The mean family size was 1.4 members. Supplementary Fig. 8 shows relationships between wheat and diploid orthologous gene family across the full scale of orthologous gene family sizes. This approach accurately reconstructed gene family sizes in simulated maize and ‘hexaploid’ rice genomes (Figs 3a, b), although larger gene family sizes tended to be underestimated. Figure 3c, d shows the relationships between Ae. tauschii and wheat genes. Single-member gene families in hexaploid wheat and Ae. tauschii were maintained to a similar extent as those seen in sequenced diploid grasses, consistent with Southern blot analyses of single-copy genes27. Using the D genome as a diploid reference, we calculated the Triticeae hexaploid/diploid gene family size ratio to be between 2.5:1 and 2.7:1, derived from the geometric mean (2.5:1) and the slopes of the blue line and the red line (2.7:1) in Fig. 3e. Comparing this with the expected hexaploid/diploid ratio of 3:1 indicates the loss of between 10,000 and 16,000 genes in hexaploid wheat compared with the three diploid progenitors (Supplementary Information, section 2.10). This is consistent with earlier studies of gene loss in newly synthesized wheat polyploids28 and the erosion of genetic diversity during wheat domestication2.

Despite this overall trend of gene family size reduction, gene families with fewer or more members than expected were identified in Ae. tauschii and hexaploid wheat, as shown by green dots (more members) and brown dots (fewer members) in Fig. 3c (Ae. tauschii) and Fig. 3d (hexaploid wheat). Supplementary Tables 10–12 show the over- and under-represented functional categories of protein. Most of the over-represented categories in expanded gene families are common to wheat and Ae. tauschii: these include ribosome proteins, components of photosystem II, storage proteins, transposon-related proteins, cytochrome P450s, NB-ARC domain proteins involved in defence responses, proteins related to pollen allergens and F-box proteins. Five of the eleven families encoding hydrogen ion transmembrane transporters were significantly more numerous in Ae. tauschii than in wheat. Analysis of gene families (Supplementary Fig. 9) showed that they encode different subunits of ATPases. We speculate that they may provide proton gradients to support Na⁺ exclusion in Ae. tauschii29 and the accumulation of minerals in other Aegilops species30.

**Pseudogene analysis**

Several classes of plant DNA transposons31,32 and retroelements33 create and amplify gene fragments, disrupt genes and create pseudogenes, which can influence gene expression through epigenetic mechanisms34. We identified a set of almost 233,000 gene fragments that mapped to the same regions of their OGRs, forming ‘stacks’ that were sufficiently divergent not to assemble into their cognate gene assemblies (Fig. 4a). Two classes were identified: those containing Pfam domains and those aligning with non-Pfam domains of OGRs. Nearly 30% of the OGRs had associated gene fragments (Supplementary Table 13) that most frequently covered between 5 and 15% of the OGR length (Fig. 4b). Figure 4c shows that the alignment identities of gene fragments against their OGRs were substantially lower than the identities of cognate regions within wheat gene assemblies. Supplementary Fig. 10 shows the distribution of stacks along genes and the ratio of non-synonymous to synonymous substitutions (Ks/Ka) along the genes. Pfam domains found in stacks were enriched for zinc-finger motifs in mutator transposons (Supplementary Table 14), consistent

---

**Figure 1** | Coverage of OGRs by wheat 454 sequence reads and simulated 454 reads from rice and maize. a. Coverage of OGRs by repeat-masked wheat 454 sequence reads (black line), wheat LCG (black dashed line) and the orthologous group assembly (blue line), together with rice genes (red line) and maize simulated reads (green line). b. Median coverage depth over protein-coding regions of OGRs (amino terminus = 0; carboxy terminus = 100). The colour coding is the same as in a, except simulated hexaploid reads from rice (red line) were used. c. Distribution of wheat gene copy numbers from the orthologous group assembly.

**Figure 2** | Alignment of wheat 454 reads, SNPs and genetic maps to the B. distachyon genome. The inner circle represent gene order on the five Brachypodium chromosomes (Bd1–Bd5). Track 1 illustrates conservation between wheat 454 reads and Brachypodium genes, shown as a window of genes present in wheat. Tracks 2–4 show SNP density (the mean number of SNPs per gene in a window of 20 genes) in the A (track 2), B (track 3) and D (track 4) genomes of wheat. Tracks 5–7 show wheat synteny with Brachypodium for the A (track 5), B (track 6) and D (track 7) genomes. Genetic markers (shown in darker colours) are colour-coded by wheat chromosome. Gaps between markers are filled in to show synteny (lighter colours).
with their role in pseudogene formation. F-box, protein kinase and NB-ARC domains, which are found in the most rapidly evolving gene families in plants, are also over-represented.

**Determining homeologous relationships of gene assemblies**

We classified gene assemblies as A-, B- or D-genome-derived according to sequence similarity to Illumina sequence assemblies from *T. monococcum*, cDNA assemblies from *Ae. speltoides* and, respectively, 454 reads from *Ae. tauschii* by applying a support vector machine learning approach (Supplementary Section 5, Supplementary Figs 11 and 12, and Supplementary Tables 15–18). Supplementary machine learning approach confirmed a subset of A-, B- and D-genome SNPs showed high precision (Fig. 13) shows that 66% of the gene assemblies were classified with high overall precision (70%) and recall into the A genome (28.3%), the B genome (29.2%) and the D genome (33.8%). The other 9% of classified assemblies have stop codons. The others 34% with low classification probabilities are likely to be very similar homeologues. The othes 34% with low classification probabilities are likely to be very similar homeologues.

Figure 3 | Gene family sizes in orthologous assemblies of hexaploid wheat, *Ae. tauschii*, simulated maize and hexaploid rice. The boxes and whiskers indicate the number of genes in diploid gene families of different sizes. The black lines represent expected gene family sizes, and the red lines show the gene family sizes determined from the orthologous group assembly, derived by polynomial regression fit. Only gene families with up to ten members are shown. A, Maize gene family sizes predicted from orthologous assembly of simulated 454 reads. B, Rice gene family sizes predicted from orthologous assembly of simulated 454 reads derived from triplicated rice genes. C, *Aegilops tauschii* gene family sizes obtained from orthologous assembly of repeat-masked 454 reads. Expanded gene families are shown as green dots. D, Wheat gene family sizes in the orthologous group assembly. E, Amalgamation of wheat and *Ae. tauschii* gene copy numbers. The black line shows the expected gene copy numbers for wheat and *Ae. tauschii*. The red line shows the regression fit for wheat, and the blue line shows the regression fit for *Ae. tauschii*. The grey zone between these lines estimates the extent of gene loss in hexaploid wheat. For each family size, the left-hand boxes represent hexaploid wheat and the right-hand boxes represent *Ae. tauschii*.

Conclusions

Using whole-genome 454 sequencing, we assembled gene sequences representing an essentially complete gene set, and a significant number were assigned to the A, B or D genome. Although the assemblies are fragmentary, they form a powerful framework for identifying genes, accelerating further genome sequencing and facilitating genome-scale analyses. The identification of over 132,000 SNPs in A, B and D genes facilitates analysis of quantitative trait loci and association studies of pseudogenes. The blue and red lines represent stacks with and without protein domains, respectively. C, Protein identity between subassemblies forming stacks of gene fragments. The blue and red lines represent stacks with and without protein domains, respectively, and the black line represents subassemblies forming genes.
traits. Comparison with the sequences of diploid progenitors and relatives showed pronounced reductions in the size of large gene families in wheat despite the relatively recent formation of the hexaploid (Fig. 3e), consistent with smaller-scale analyses.\(^{24,38}\) The scale of gene loss in hexaploid wheat compared with maize\(^{46}\) and *Brassica rapa*\(^{47}\) is significantly smaller, possibly as a result of its relatively recent origin and the absence of intergenome recombination\(^{48}\). Nevertheless, gene loss in wheat could be rapid, as shown in the newly created allopolyploid *Tragopogon miscellus*\(^{41}\). Most functional classes show equal gene loss in the three genomes, but families of transcription factors showed a clear tendency to be retained as functional genes in all three genomes. These may maintain transcriptional networks in each genome and contribute to non-additive gene expression\(^{42}\) and genome plasticity. In contrast to the overall loss of gene family members, several classes of gene families with predicted roles in defence, nutritional content, energy metabolism and growth have increased sizes in the Triticeae lineage, possibly as a result of selection during domestication.

Major efforts are underway to improve wheat productivity by increasing genetic diversity in breeding materials and through genetic analysis of traits\(^{33}\). The genomic resources that we have developed promise to accelerate progress by facilitating the identification of useful variation in genes of wheat landraces and progenitor species, and by providing genomic landmarks to guide progeny selection. Analysis of complex polygenic traits such as yield and nutrient use efficiency will also be accelerated, contributing to sustainable increases in wheat crop production.

**METHODS SUMMARY**

A single-seed descent line of *T. aestivum* landrace Chinese Spring was sequenced, because it is widely used for cytogenetic analysis\(^{44}\) and physical mapping\(^{4}\). *Triticum monococcum* accession 4342-96 is a community standard line for targeting induced local lesions in genomes, physical mapping and genetic analysis; and *Ae. tauschii* ssp. *sativum* accession AL87/78, which is used for physical and genetic mapping, was sequenced using 454 technology.

Sequence for the *T. aestivum* wheat gene assembly was generated using Roche 454 pyrosequencing on the GS FLX Titanium and GS FLX+ platforms. Additional sequence read data sets for *T. aestivum*, *T. monococcum* and *Ae. tauschii* were generated using three platforms, Illumina, 454 and SOLiD, to analyse homologous sequences and SNPs (a list of all data sets is in Supplementary Table 2).

Orthologous groups were created from rice, sorghum and *Brachypodium distachyon* genome sequences and physically mapped using three platforms, Illumina, 454 and SOLiD, to analyse homeologous regions of orthologous sequences and SNPs (a list of all data sets is in Supplementary Information). Sequence assembly and functional annotation were performed using the software package MapMan 3.0. Sequence assemblies, annotated gene sequences and their relationships are available for download from the International Rice Genome Sequencing Project, A map-based sequence of the rice genome. Nature 436, 793–800 (2005).

7. Moore, G., Devos, K. M., Wang, Z. & Gale, M. D. Cereal genome evolution. Grasses, lines up and form a circle. *Curr. Biol.*, 5, 737–739 (1995).

8. Paterson, A. H. et al. The Sorghum bicolor genome and the diversification of grasses. *Nature* 457, 551–556 (2009).

9. The International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* 463, 763–768 (2010).

10. Smith, D. B. & Flavell, R. B. Characterisation of the wheat genome by association genetics. *Genome* 50, 959–969 (2007).

11. Baroucl, R. S. et al. Exceptional diversity, non-random distribution, and rapid evolution of retrotransposons in the B73 maize genome. *PLoS Genet.* 6, e1000732 (2010).

12. Wicker, T. et al. Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. *Plant Cell* 23, 1706–1718 (2011).

13. Choulet, F. et al. Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell* 22, 1686–1700 (2011).

14. Gill, B. S. et al. A workshop report on wheat genome sequencing: International Genome Research on Wheat Consortium. *Genetics* 168, 1087–1096 (2004).

15. Jiao, X. et al. A physical map of the 1-gigabase wheat chromosome 3B. *Science* 292, 101–104 (2008).

16. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* 436, 793–800 (2005).

17. Liu, L., Stoeckert, C. J., Jr & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189 (2003).

18. Mochida, K., Yoshida, T., Sakurai, T., Ogihara, Y. & Shinozaki, K. TrifLDDB: a database of clustered full-length coding sequences from *Triticaceae* with applications to comparative grass genomics. *Plant Physiol.* 150, 1135–1146 (2009).

19. Richter, D. C., Ott, F., Auch, A. F., Schmid, R. & Huson, D. H. MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS ONE* 3, e3377 (2008).

20. Hernandez, P. et al. Next-generation sequencing and syntonic integration of flow-sorted arms of wheat chromosome 4A4 exposes the chromosome structure and gene content. *Plant J.* 69, 377–386 (2012).

21. Massa, A. N. et al. Gene space dynamics during the evolution of *Aegilops tauschii*, *Brachypodium distachyon*, *Oryza sativa*, and *Sorghum bicolor* genomes. *Mol. Biol. Evol.* 28, 2537–2547 (2011).

22. The International Barley Genome Sequencing Consortium. A physical, genetic, and functional sequence assembly of the barley genome. *Nature* doi:10.1038/nature11543 (this issue).

23. Schnable, P. S. et al. The B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112–1115 (2009).

24. Lee, E. K. et al. A functional phylogenetic view of the seed plants. *PLoS Genet.* 7, e1002111 (2011).

25. Allen, A. M. et al. Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid wheat (*Triticum aestivum* L.). *Plant Biotechnol. J.* 9, 1086–1099 (2011).

26. Salse, J. et al. Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* 20, 1124–1134 (2008).

27. Qi, L. L. et al. A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics* 168, 701–712 (2004).

28. Ozkan, H., Levy, A. A. & Feldman, M. Allopolyploid-induced rapid genome evolution in the wheat (*Aegilops-Triticum*) group. *Plant Cell* 13, 1735–1747 (2001).

29. Shavrukov, Y., Langridge, P. & Tester, M. Salinity tolerance and sodium exclusion in *Triticum*. *Breed. Sci.* 59, 671–678 (2009).

30. Wang, S., Yin, L., Tanaka, K., Tanaka, H. & Tsujimoto, H. Wheat-Aegilops chromosome addition lines showing high iron and zinc contents in grains. *Breed. Sci.* 61, 189–195 (2011).

31. Jian, X., Bao, Z., Zhang, X., Eddy, S. R. & Wessler, S. R. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* 431, 569–573 (2004).

32. Morgante, M. et al. Gene duplication and exon shuffling by heitlin-re like transposons generate intraspecies diversity in maize. *Nature Genet.* 37, 997–1002 (2005).

33. Jin, Y. K. & Bennetzen, J. L. Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the *B1* retroelement of maize. *Curr. Biol.* 6, 1177–1186 (1994).

34. Lippman, Z. et al. Rapid, repeated, and clustered loss of duplicate genes in *Brachypodium* syntenic-based genome browser. *http://plantsensembl.org/brachypodium_distantly.html*.

Received 4 March; accepted 1 October 2012.

1. United States Department of Agriculture. *World Agricultural Supply and Demand Estimates*. Report No. WASDE-511; http://usda01.library.cornell.edu/usda/ current/wasde/wasde-10-11-2012.pdf (2012).

2. Food and Agriculture Organisation of the United Nations. http://faostat.fao.org/default.aspx?lang=en (2011).

3. Neubert, M. & Samuel, D. in Hulled Wheats (eds Padulosi, S., Hammer, K. & Helleur, J.) 41–100 (Proc. 1st Internat. Workshop Hulled Wheats, International Plant Genetic Resources Institute, 1996).

4. 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 provinces evidence for gene flow from wild tetraploid wheat to hexaploid wheat. *Mol. Biol. Evol.* 23, 1386–1396 (2006).

5. Salse, J. & et al. New insights into the origin of the B genome of hexaploid wheat: evolutionary relationships at the SPA genomic region with the S genome of the diploid relative *Aegilops speltoides*. *BMC Genomics* 9, 555 (2008).

6. Haudry, A. et al. Gridding up wheat: a massive loss of nucleotide diversity since domestication. *Mol. Biol. Evol.* 24, 1506–1517 (2007).
Pumphrey, M., Bai, J., Laudencia-Chingcuanco, D., Anderson, O. & Gill, B. S. Nonadditive expression of homoeologous genes is established upon polyploidization in hexaploid wheat. *Genetics* **181**, 1147–1157 (2009).

Tester, M. & Langridge, P. Breeding technologies to increase crop production in a changing world. *Science* **327**, 818–822 (2010).

Sears, E. R. in *Chromosome Manipulation and Plant Genetics* (eds Riley, R. & Lewis, K. R.) 22–45 (Oliver and Boyd, 1966).

Supplementary Information is available in the online version of the paper.

Acknowledgements DNA sequence was generated by The University of Liverpool Centre for Genomic Research (United Kingdom), 454 Life Sciences (United States), The Cold Spring Harbor Woodbury Genome Centre (United States) and The Genome Analysis Centre (United Kingdom). This work was supported by UK Biological and Biotechnological Sciences Research Council (BBSRC) grants BB/G012865, BB/G013985/1 and BB/G013004/1, to K.J.E., M.W.B. and N.Hall; a Wolfson Merit Award from the Royal Society, to N.Hall; BBSRC Strategic Programme grant BB/J004588/1 (GRO), to M.W.B.; EC TriticeaeGenome grant number 212019, to K.F.X.M. and M.W.B.; The TRITEX Project of the Plant20130 Initiative of the German Ministry of Education and Research grant number 0315954C, to K.F.X.M.; EC Transplant Grant 283496, to K.F.X.M. and P.K., a BBSRC Career Development Fellowship BB/H022333/1, to A.H.; US NSF grants IOS-1032105 and DBI-0923128, to W.R.M.; USDA-NIFA grant 2008-35300-04588, to B.G.; and US NSF grants DBI-0701916, to J.D., and DBI-0822100, to S. Kianian.

**Author Contributions** R.B., M.S., M.P., G.L.A.B. and R.D. are joint first authors. K.J.E., M.W.B., N. Hall and A.H. designed the project; W.R.M., M.K., M.T., I.B., J.D., M.-C.L., O.A., S. Kianian, N. Huo, B.G. and S.S. provided data and advice; R.D., N.M. and S. Kay conducted experiments; K.F.X.M., N. Hall and M.W.B. planned and conducted analyses; and R.B., M.S., M.P., G.L.A.B., A.M.A., D.B., D.W., P.K. and A.H. carried out analyses. K.J.E., A.H., W.R.M. and R.B. contributed to the text and M.W.B., N. Hall and K.F.X.M. wrote the manuscript. All authors commented on the manuscript.

**Author Information** Sequence assemblies have been submitted to the European Nucleotide Archive under project accession number PRJEB568. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.W.B. (michael.bevan@jic.ac.uk), K.F.X.M. (k.mayer@helmholtz-muenchen.de), N. Hall (Neil.Hall@liverpool.ac.uk), A.H. (Anthony.Hall@liverpool.ac.uk), or K.J.E. (k.j.edwards@bristol.ac.uk). This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported licence. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-sa/3.0/