Draft genome of the wheat A-genome progenitor *Triticum urartu*

Hong-Qing Ling[1]*, Shancen Zhao[2,3]*, Dongcheng Liu[1]*, Junyi Wang[1,2]*, Hua Sun[1]*, Chi Zhang[2]*, Huajie Fan[1], Dong Li[2], Lingli Dong[1], Yong Tao[2], Chuan Gao[2], Huilan Wu[1], Yiwen Li[1], Yan Cui[1], Xiaosen Guo[2], Shusong Zheng[1], Biao Wang[1], Kang Yu[1], Qinsi Liang[2], Wenlong Yang[1], Xueyuan Lou[1], Jie Chen[2], Mingji Feng[2], Jianbo Jian[2], Xiaofei Zhang[1], Guangbin Luo[1], Ying Jiang[2], Junjie Liu[2], Zhaobao Wang[2], Yuhui Sha[2], Bairu Zhang[1], Huajun Wu[4], Dingzhong Tang[1], Qianhua Shen[1], Pengya Xue[1], Shenhao Zou[1], Xiujie Wang[4], Xin Liu[1], Famin Wang[1], Yanping Yang[1], Xueli An[1], Zhenying Dong[1], Kunpu Zhang[1], Xiangqi Zhang[1], Ming-Cheng Luo[5], Jan Dvorak[5], Yiping Tong[1], Jian Wang[2], Huanming Yang[2], Zhensheng Li[1], Daowen Wang[1], Aimin Zhang[1] & Jun Wang[2,6,7]

Bread wheat (*Triticum aestivum*, AABBDD) is one of the most widely cultivated and consumed food crops in the world. However, the complex polyploid nature of its genome makes genetic and functional analyses extremely challenging. The A genome, as a basic genome of bread wheat and other polyploid wheats, for example, *T. turgidum* (AABB), *T. timopheevii* (AAGG) and *T. zhukovskyi* (AAGGA^{m}A^m), is central to wheat evolution, domestication and genetic improvement[1]. The progenitor species of the A genome is the diploid wild einkorn wheat *T. urartu*[2], which resembles cultivated wheat more extensively than do *Aegilops speltoides* (the ancestor of the B genome[3]) and *Ae. tauschii* (the donor of the D genome[4]), especially in the morphology and development of spike and seed. Here we present the generation, assembly and analysis of a whole-genome shotgun draft sequence of the *T. urartu* genome. We identified protein-coding gene models, performed genome structure analyses and assessed its utility for analysing agronomically important genes and for developing molecular markers. Our *T. urartu* genome assembly provides a diploid reference for analysis of polyploid wheat genomes and is a valuable resource for the genetic improvement of wheat.

Bread wheat is one of the most important food crops worldwide, and provides about 20% of the calories consumed by humans[5]. To accelerate wheat improvement, a substantial amount of research has been conducted on the genome. The International Wheat Genome Sequencing Consortium aims at flow-sorting and sequencing the individual chromosomes of bread wheat, and significant progress has been made with several chromosomes, for example 3B (ref. 6) and 4A (ref. 7). More recently, a whole-genome shotgun sequence analysis of bread wheat and its diploid relatives[8] has allocated more than 60% of the genes to the A, B and D genomes with more than 70% confidence. The sequence of diploid progenitor genomes will allow the complete and unambiguous assignment of their homeologous relationships.

We sequenced *T. urartu* accession G1812 (PI428198) using a whole-genome shotgun strategy on the Illumina HiSequation (2000) platform, and assembled the genome using SOAPdenovo (v. 1.05)[9] with 448.49 gigabases (Gb) of filtered high-quality sequence data (Supplementary Information). We estimated the genome size of *T. urartu* to be 4.94 Gb (Supplementary Information), which is consistent with previous reports of 4.8–5.7 Gb (refs 10, 11). The genome assembly reached 3.92 Gb with a contig N50 size (at which 50% of assembly was covered) of 3.42 kilobases (kb). After gap closure, the draft assembly was 4.66 Gb with a scaffold N50 length of 63.69 kb (Table 1 and Supplementary Information). The length of the contigs that contained intact or partial genes ranged from 200 base pairs (bp) to 65.8 kb, with an average length of 9.91 kb. The assembly was evaluated by comparisons with published bacterial artificial chromosome and expressed sequence tag (EST) sequences and by validation with PCR (Supplementary Information), and both indicated that the draft sequence had extensive genome coverage with high accuracy. The distribution of GC content in the *T. urartu* genome was comparable with those in the genomes of rice[12], maize[13], sorghum[14] and *Brachypodium distachyon*[15] (Supplementary Information).

Genome annotation of the assembly was performed as described in Supplementary Information. About 66.88% of the *T. urartu* assembly was identified as repetitive elements, including long terminal repeat retrotransposons (49.07%), DNA transposons (9.77%) and unclassified elements (8.04%) (Supplementary Information). The proportion of repetitive DNA was lower than the roughly 80% previously reported[16], which is probably due to a decreased incorporation of repeat sequence reads into the assemblies.

To facilitate gene prediction, we generated a 116.65-megabase (Mb) transcriptome of *T. urartu* with 67.14 Gb of RNA-Seq data from eight different tissues and treatments using the HiSequation (2000) platform, and 49,935 assembled transcripts from six tissues using the Roche 454 sequencing platform (Supplementary Information). These data, together with publicly available ESTs from hexaploid wheat, and homologues from sequenced grass genomes[12–15,17], were used as evidence in gene prediction (Supplementary Information). In total, we predicted 34,879 protein-coding gene models. The average gene size was 3,207 bp, with a mean of 4.7 exons per gene, which was similar to that found for *B. distachyon* (5.2)[15] but slightly higher than that of rice (3.8)[12], maize (4.1)[13] and sorghum (4.3)[14]. In comparison with the 28,000 genes estimated for the A genome of hexaploid wheat[7], our gene set for *T. urartu* contained 6,800 more members, indicating a more complete representation of genes in our analysis. However, the different approaches used in this work and in a previous study[7], and the extensive loss of genes in the hexaploid A genome compared with its diploid progenitor[8], may also have contributed to this difference.

We also obtained 14,222,170 small RNA (sRNA) reads (18–30 bp) representing 4,369,970 unique sRNA tags. In total, 412 conserved and 24 new microRNAs (miRNAs) distributed into 116 families were identified. Comparison with the miRNAs of five monocots and five dicots showed that 73 miRNA families were specific to monocots, of which 23 were uniquely present in *T. urartu*. We predicted 244 target genes for these miRNAs and found that the target gene (TRIUR3_06170) of miRNA MIR5050 responded to cold treatment, which provides a

Table 1 | Sequencing and draft genome assembly statistics

Assembly step	DNA library insert size (bp)	Effective data (Mb)	N50 (bp)	Total length (Mb)	Ns† (Mb)	Number >2 kb
Contig construction	167–796	357,810	3,422	3,920	–	385,430
Scaffold construction	167–20,000	90,682	64,532	–	1,258	–
Gap closure*	167–796	357,810	63,687	4,660	738	81,689

* Final statistics of scaffolds after gap closure.
† The added N number during scaffolding.

new resource for investigating the regulation of cold adaptation through miRNA (Supplementary Information).

The gene families of *T. urartu* were compared with those of rice[12], maize[13], sorghum[14] and *B. distachyon*[15] using OrthoMCL[18] (Supplementary Information). We identified 24,339 families in the five grasses. Of these, 9,836 families, which contained 68,464 genes, were
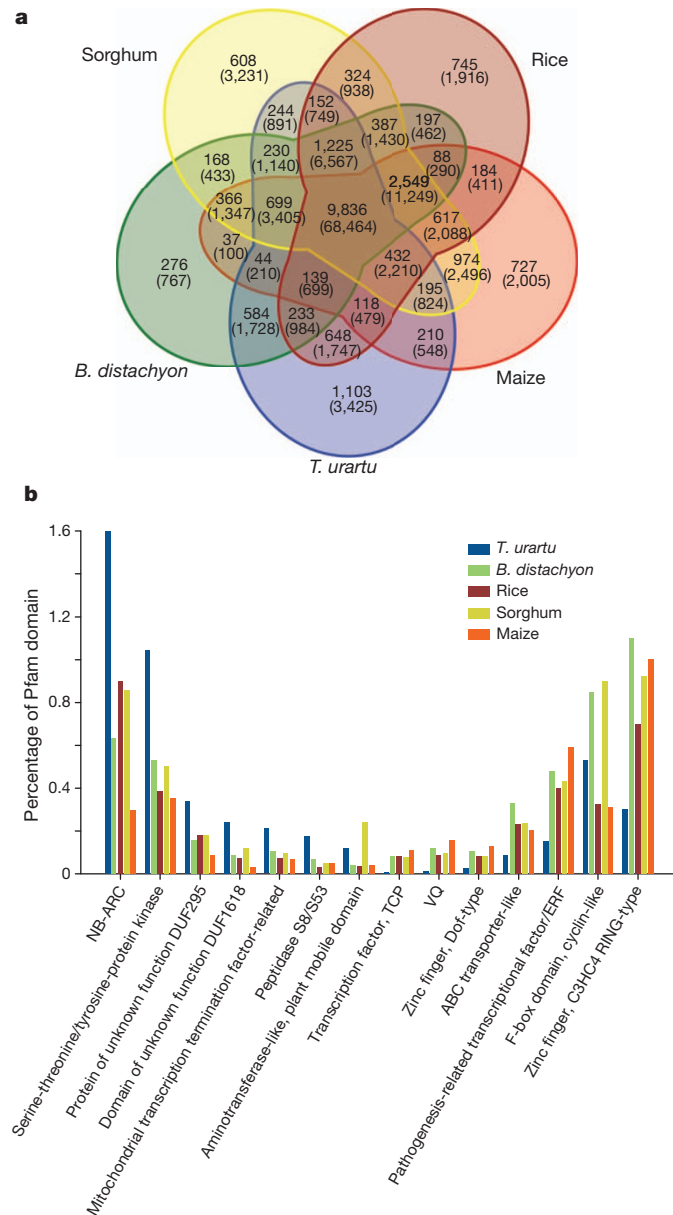
Figure 1 | Comparison of gene families and coding domains of *T. urartu* with rice, maize, sorghum and *B. distachyon*. a, Venn diagram illustrating shared and unique gene families (gene numbers in parenthesis) among the five grass species. **b**, Comparison of protein families with different Pfam domains among the five species. Fourteen Pfam domains that had significant differences ($P < 0.001$, Fisher's exact test) in their percentages between *T. urartu* and the other four grass species are shown. The percentage was calculated by dividing the number of each domain by the total gene set in a given species.

common to all five species. Another 1,103 families, containing 3,425 genes, were specific to *T. urartu* (Fig. 1a). GO analysis of the *T. urartu*-specific families revealed that 556, 230 and 841 genes were involved in biological processes, cellular compounds and molecular functions, respectively. In total, 2,067 Pfam domains were shared among the five species. Of these, 14 Pfam domains had differences in member numbers in *T. urartu* compared with the other four grasses (Fig. 1b). These included NB-ARC and serine–threonine/tyrosine-protein kinase domains that were markedly increased in *T. urartu*, and C3HC4 RING-type and pathogenesis-related transcriptional factor/ERF domains that were significantly decreased. However, determination of the significance and accuracy of these differences will require more detailed analysis.

Given that NB-ARC domain proteins function mainly in disease resistance[19], we analysed the genes encoding R proteins in the *T. urartu* genome and identified 593 such genes, which were more abundant than in *B. distachyon* (197), rice (460), maize (106) and sorghum (211) (Supplementary Information). In contrast with barley genome data[17], the ratio of NBS-LRR type genes in *T. urartu* (1.21%) was also substantially higher than that in barley (0.73%). These analyses indicate that there was a specific expansion of *R* genes in the *T. urartu* genome.
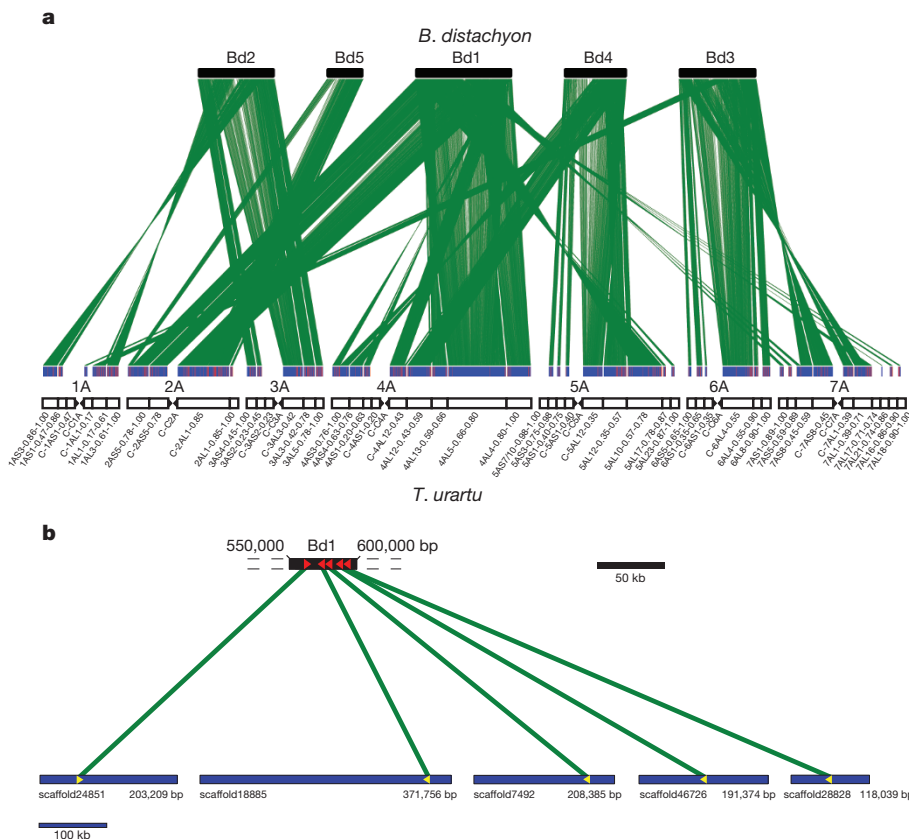
The scaffolds and gene models of *T. urartu* were assigned to chromosomes by using genetically mapped bread wheat ESTs[20] as queries to search for homologous sequences in the *T. urartu* assembly (Supplementary Information). A total of 8,715 scaffolds, harbouring 14,578 genes (41.8% of the total predicted genes) were mapped to 45 chromosomal regions of the wheat A genome. Syntenic alignments between the *T. urartu* and *B. distachyon* genomes were constructed by using a set of 14,578 orthologous genes (Fig. 2a). These gene-based alignments conform, and supply more details, to the broad framework of genome synteny between wheat and *B. distachyon* proposed previously[15].

The 4.94-Gb *T. urartu* genome is more than 18 times larger than the 272-Mb genome of *B. distachyon*. Given that the average gene size of *T. urartu* is similar to that of *B. distachyon*, and the predicted gene number (34,879) is only about 1.37-fold that of *B. distachyon* (25,532), the larger genome size of *T. urartu* might be due to increased intergenic spaces. We therefore compared the intergenic space of the syntenic blocks between *T. urartu* and *B. distachyon* (Supplementary Information). About 21% of *T. urartu* genes had similarly sized intergenic spaces to those in *B. distachyon*, but most *T. urartu* genes were separated by a greatly increased intergenic space enriched in *Gypsy* and *Copia* retrotransposons, and were present in separate scaffolds (Fig. 2b). This provides the genome sequence-scale evidence for the role of repeat expansion in genome size enlargement during the evolution of the tribe Triticeae.

We next demonstrated the utility of the *T. urartu* draft genome sequence for finding agronomically important genes through identifying the *T. urartu* homologue of *OsGASR7*, a gibberellin-regulated gene that controls grain length in rice[21]. We found two haplotypes (H1 and H2) for *TuGASR7* in 92 diverse *T. urartu* accessions collected from different regions. H1 was significantly associated with greater values of grain length and grain weight (Fig. 3 and Supplementary Information). We also found natural variation of *TaGASR7* in bread wheat, with the elite variant associated with improved yield-related traits, suggesting that *TaGASR7* is of use for the improvement of wheat yield (Supplementary Information).

The *T. urartu* assembly also served as a rich resource for the development of genetic markers for molecular breeding through genomic

Figure 2 | Synteny analysis between *T. urartu* and *B. distachyon*, and genome expansion in *T. urartu*.
**a**, Syntenic relationships between the seven hypothetical chromosomes of *T. urartu* (1A–7A) and the five chromosomes of *B. distachyon* (Bd1–Bd5). The deletion bin maps of bread wheat 1A–7A chromosomes are noted at the bottom. **b**, Comparison of intergenic spaces in one representative syntenic block from *T. urartu* and *B. distachyon*. The 50-kb block from *B. distachyon* chromosome 1 (Bd1) contains five genes (indicated by red arrowheads). The *T. urartu* orthologues (yellow arrowheads) of the five genes are distributed in five separate scaffolds (totalling 1,092 kb). The total intergenic space in the displayed *T. urartu* region is more than 20-fold that in the corresponding Bd1 region.

selection. We identified 739,534 insertion-site-based polymorphism (ISBP) markers and 166,309 simple sequence repeats (SSRs) (Supplementary Information). PCR validation showed that 94.5% of the SSRs and 87% of the ISBP markers gave the expected products, and that

33.61% of the SSRs and 10.19% of the ISBP markers were specific to the A genome. Moreover, 28.7% of the SSR loci were polymorphic in bread wheat (Supplementary Information). To enable the identification of single nucleotide polymorphisms (SNPs), we re-sequenced another *T. urartu* accession (DV2138) and obtained 78.6 Gb of high-quality data. Comparison of the genome data between the two *T. urartu* accessions (G1812 versus DV2138) allowed the discovery of 2,989,540 SNPs, which will be useful for the future development of SNP markers (Supplementary Information).

Previous studies have revealed that more than half of the 60 meta-quantitative trait loci (meta-QTLs)[22,23] related to wheat yield traits are present in the A genome of bread wheat, and three meta-QTLs (MQTL_5, 6 and 7) are located on chromosome 5A (ref. 22). We therefore searched the *T. urartu* scaffolds using available markers around the three meta-QTLs (Supplementary Information). We found ten scaffolds with a total length of 772,014 bp that were distributed in the 14-centimorgan (cM) region of MQTL_5; nine scaffolds with a combined length of 783,140 bp were located in the 15-cM region containing MQTL_6; and six scaffolds with an overall length of 529,604 bp were assigned to the 9-cM region harbouring MQTL_7. The sequence information of these scaffolds will expedite the development of more polymorphic markers within the three meta-QTL regions and facilitate the identification of their corresponding genes.
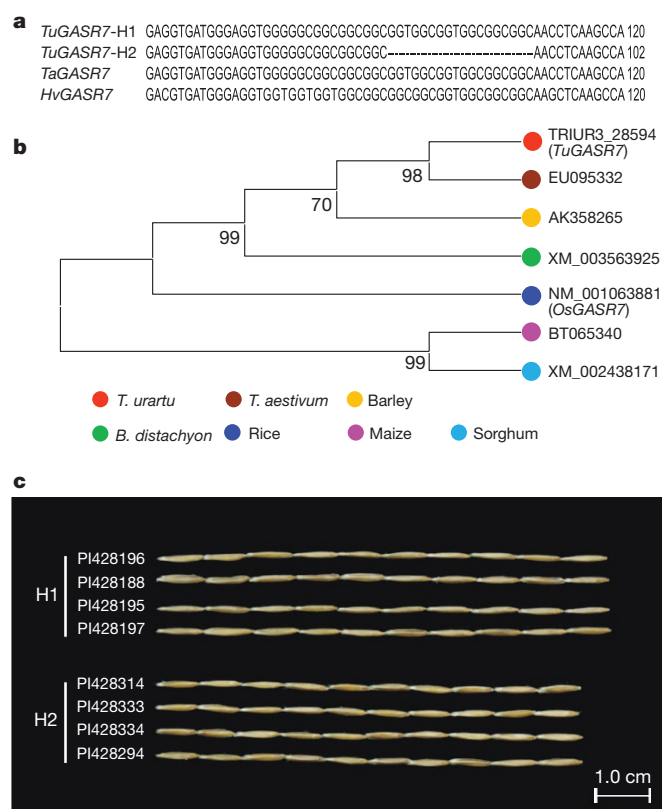
Figure 3 | Association analysis of the *TuGASR7* gene. a, Indel differences between the *TuGASR7* haplotypes H1 and H2 compared with their corresponding regions in the *GASR7* homologues from barley (*Hordeum vulgare*; *HvGASR7*) and *T. aestivum* (*TaGASR7*). **b**, Phylogenetic relationships among the homologous genes of rice *GASR7* from different grass species. TRIUR3_28594 is the predicted homologue of *GASR7* in *T. urartu*. **c**, Typical grain lengths for the two detected *TuGASR7* haplotypes, H1 (top panel) and H2 (bottom panel). For each haplotype, the seeds from four representative *T. urartu* accessions were photographed. To illustrate the difference in grain length more clearly, ten seeds were randomly selected from each of the compared accessions and aligned closely for the photograph.

Our *T. urartu* draft genome sequence provides new insights into the A genome that is shared by many polyploid wheat species. The large set of gene models (34,879) and abundant genetic markers anchored in sequence scaffolds, together with the emerging genomic resources from bread wheat[8], promise to accelerate deeper and more systematic genomic and breeding studies of bread wheat that are required to meet the future challenges of food security and sustainable agriculture.

METHODS SUMMARY

The genome of *T. urartu* accession G1812 was sequenced on the Illumina HiSequation (2000) platform. These data were used to assemble the draft genome sequence with the use of the SOAPdenovo[9] software. RNA-Seq data were generated on the same platform and Roche 454 for genome annotation and transcriptome analysis. Repeat sequences were identified through sequence similarity at the nucleotide and protein levels[24]. Protein-coding genes were predicted by using an *ab initio* approach, sequence similarity search and RNA-Seq data to build reliable gene models[25]. Detailed methodology descriptions are given in Supplementary Information.

1. Peng, J. H., Sun, D. F. & Nevo, E. Domestication evolution, genetics and genomics in wheat. *Mol. Breed.* **28,** 281–301 (2011).
2. Dvorak, J., di Terlizzi, P., Zhang, H. B. & Resta, P. The evolution of polyploid wheats: identification of the A genome donor species. *Genome* **36,** 21–31 (1993).
3. Dvorak, J. & Zhang, H. B. Variation in repeated nucleotide sequences sheds light on the phylogeny of the wheat B and G genomes. *Proc. Natl Acad. Sci. USA* **87,** 9640–9644 (1990).
4. McFadden, E. S. & Sears, E. R. The origin of *Triticum spelta* and its free-threshing hexaploid relatives. *J. Hered.* **37,** 81–89 (1946).
5. Food and Agriculture Organisation of the United Nations. FAOSTAT. http://faostat.fao.org/site/339/default.aspx (2011).
6. Paux, E. *et al.* A physical map of the 1-gigabase bread wheat chromosome 3B. *Science* **322,** 101–104 (2008).
7. Hernandez, P. *et al.* Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *Plant J.* **69,** 377–386 (2012).
8. Brenchley, R. *et al.* Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491,** 705–710 (2012).
9. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20,** 265–272 (2010).
10. Bennett, M. D. & Smith, J. B. Nuclear DNA amounts in angiosperms. *Phil. Trans. R. Soc. Lond. B* **274,** 227–274 (1976).
11. Ozkan, H., Tuna, M., Kilian, B., Mori, N. & Ohta, S. Genome size variation in diploid and tetraploid wild wheats. *AoB Plants* **2010,** 1–11 (2010).
12. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436,** 793–800 (2005).
13. Schnable, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326,** 1112–1115 (2009).
14. Paterson, A. H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457,** 551–556 (2009).
15. Vogel, J. P. *et al.* Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463,** 763–768 (2010).
16. Choulet, F. *et al.* Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell* **22,** 1686–1701 (2010).
17. The International Barley Genome Sequencing Consortium. A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491,** 711–716 (2012).
18. Li, L., Stoeckert, C. J., Jr & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13,** 2178–2189 (2003).
19. Collier, S. M. & Moffett, P. NB-LRRs work a 'bait and switch' on pathogens. *Trends Plant Sci.* **14,** 521–529 (2009).
20. 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).
21. Huang, X. *et al.* Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nature Genet.* **44,** 32–39 (2012).
22. Gegas, V. C. *et al.* A genetic framework for grain size and shape variation in wheat. *Plant Cell* **22,** 1046–1056 (2010).
23. Zhang, L. Y. *et al.* Genomic distribution of quantitative trait loci for yield and yield-related traits in common wheat. *J. Integr. Plant Biol.* **52,** 996–1007 (2010).
24. Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protocols Bioinformat.* **25,** 4.10.1–4.10.14 (2004).
25. Elsik, C. G. *et al.* Creating a honey bee consensus gene set. *Genome Biol.* **8,** R13 (2007).
26. Ling, H.-Q, *et al.* Genomic data from *Triticum urartu*—the progenitor of wheat A genome. *GigaScience* http://dx.doi.org/10.5524/100050 (2013).