Draft genome of the wheat A-genome progenitor Triticum urartu

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

Published in:
Nature

DOI:
10.1038/nature11997

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Draft genome of the wheat A-genome progenitor
*Triticum urartu*

Hong-Qing Ling1,*, Shancen Zhao2,1,*, Dongcheng Liu1,*, Junyi Wang1,2,*, Hua Sun1,*, Chi Zhang2,*, Huaiej Fan1, Dong Li1, Lingli Dong1, Yong Tao2, Chuhan Gao2, Huilin Wu1, Yiwen Li1, Yan Cui1, Xiaosen Guo3, Shusong Zheng1, Biao Wang1, Kang Yu1, Qinxi Liang1, Wenlong Yang1, Xueyuan Lou1, Jie Chen3, Mingji Feng4, Jianbo Jian2, Mingjiang Zhang1, Yingjiang2, Junjie Liu3, Zhaobao Wang5, Yuhui Sha6, Bairu Zhang1, Huajun Wu1, Difang Zhang1, Qianhua Shen1, Pengya Xue1, Shenhao Zou1, Xiujie Wang3, Xin Liu3, Famin Wang1, Yanping Yang1, Xueli An1, Zhenying Dong1, Kunpu Zhang1, Xiangqi Zhang1, Ming-Cheng Luo1, Jan Dvorak1, Yiping Tong1, Jian Wang1, Huaming Yang2, Zhensheng Li3, Daowen Wang1, Aimin Zhang1 & Jun Wang2,*

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* (AAGGAm), is central to wheat evolution, domestication and genetic improvement1. The progenitor species of the A genome is the diploid wild einkorn wheat *T. urartu*, which resembles cultivated wheat more extensively than do *Aegilops speltoides* (the ancestor of the B genome2) and *Ae. tauschii* (the donor of the D genome3), especially in the morphology and development of spike and seed. Here we present the generation, assembly and analysis of a whole-genome shotgun 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 humans4. 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 relatives8 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 homoeologous relationships.

We sequenced *T. urartu* accession G1812 (PI428198) using a whole-genome shotgun strategy on the Illumina HiSeqation (2000) platform, and assembled the genome using SOAPdenovo (v. 1.05)9 with 488.49 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 genome size of *T. urartu* was predicted 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 rice12, maize3, sorghum14 and *Brachypodium distachyon*11 (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 reported16, 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 genomes12–15,17, 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)11 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 wheat2, 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 study7, and the extensive loss of genes in the hexaploid A genome compared with its diploid progenitor8, 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
new resource for investigating the regulation of cold adaptation through miRNA (Supplementary Information).

The gene families of T. urartu were compared with those of rice12, maize13, sorghum14 and B. distachyon15 using OrthoMCL18 (Supplementary Information). We identified 24,339 families in the five grasses. Of these, 9,836 families, which contained 68,464 genes, were 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 resistance14, 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 data17, 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 ESTs20 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 previously15.

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 rice11. We found two haplotypes (H1 and H2) for TaGASR7 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

### Table 1 | Sequencing and draft genome assembly statistics

<table>
<thead>
<tr>
<th>Assembly step</th>
<th>DNA library insert size (bp)</th>
<th>Effective data (Mb)</th>
<th>N50 (bp)</th>
<th>Total length (Mb)</th>
<th>Ns (Mb)</th>
<th>Number &gt;2 kb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contig construction</td>
<td>167–796</td>
<td>357,810</td>
<td>3,422</td>
<td>3,920</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Scaffold construction</td>
<td>167–20,000</td>
<td>90,682</td>
<td>64,532</td>
<td>-</td>
<td>1,258</td>
<td>-</td>
</tr>
<tr>
<td>Gap closure*</td>
<td>167–796</td>
<td>357,810</td>
<td>63,687</td>
<td>4,660</td>
<td>738</td>
<td>81,689</td>
</tr>
</tbody>
</table>

*Final statistics of scaffolds after gap closure.
†The added N number during scaffolding.
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) 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 located in the 15-cM region containing MQTL_6; and six scaffolds with an overall length of 529,604 bp that were distributed in five separate scaffolds (totaling 1,092 kb). The total intergenic space in the displayed T. urartu region is more than 20-fold that in the corresponding Bd1 region.

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 (totaling 1,092 kb). The total intergenic space in the displayed T. urartu region is more than 20-fold that in the corresponding Bd1 region.

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 (TuGASR7). 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, 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 HiSeq2000 platform. These data were used to assemble the draft genome sequence with the use of the SOAPdenovo 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. Protein-coding genes were predicted by using an *ab initio* approach, sequence similarity search and RNA-Seq data to build reliable gene models. Detailed methodology descriptions are given in Supplementary Information.

Received 17 March 2012; accepted 8 February 2013.

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

**Acknowledgements** We thank L. Goodman for assistance in editing the manuscript, and M. Bevan and Y.-B. Xue for critical reading of the manuscript. This work was supported by grants from the Ministry of Science and Technology of China (the special fund for the State Key Laboratory of Plant Cell and Chromosome Engineering, 2010DFB33540, 2011CB100304, 2011AA100104 and 2009CB18300).


**Author Information** This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under accession number A0T00000000. Sequence assemblies and all short-read data are under project accession numbers SRA030525 (genomic short reads), SRA046608 (sequencing reads), PRJNA182347 (assembly and annotation) and SRA064213 (RNA-Seq). The version described in this paper is the first version, A0T01000000. Genomic data are also available at the Comprehensive Library for Modern Biotechnology (CLMB) repository. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper.

This work is licensed under a Creative Commons Attribution-NonCommercial-Share Alike 3.0 Unported licence. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-sa/3.0

©2013 Macmillan Publishers Limited. All rights reserved