Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation

Jia, Jizeng; Zhao, Shancen; Kong, Xiuying; Li, Yingrui; Zhao, Guanyao; He, Weiming; Appels, Rudi; Pfeifer, Matthias; Tao, Yong; Zhang, Xueyong; Jing, Ruilian; Zhang, Chi; Ma, Youzhi; Gao, Lifeng; Gao, Chuan; Spannagl, Manuel; Mayer, Klaus F. X.; Li, Dong; Pan, Shengkai; Zheng, Fengya; Hu, Qun; Xia, Xianchun; Li, Jianwen; Liang, Qinsi; Chen, Jie; Wicker, Thomas; Gou, Caiyun; Kuang, Hanhui; He, Genyun; Luo, Yadan; Keller, Beat; Xia, Qiuju; Lu, Peng; Wang, Junyi; Zou, Hongfeng; Zhang, Rongzhi; Xu, Junyang; Gao, Jinlong; Middleton, Christopher; Quan, Zhiwu; Liu, Guangming; Wang, Jian; Yang, Huanming; Liu, Xu; He, Zhonghu; Mao, Long; Wang, Jun

Published in:
Nature

DOI:
10.1038/nature12028

Publication date:
2013

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

Citation for published version (APA):
Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation

Jizeng Jia1,*, Shancen Zhao2,1,*, Xiuying Kong1,*, Yingrui Li2,*, Guanyao Zhao1,*, Weiming He2,*, Rudi Appels4,*, Matthias Pfeifer5,*, Yong Tao3, Yueyong Zhang4, Ruilian Jing5, Chi Zhang6, Youzhi Ma7, Lifeng Gao8, Chuan Gao9, Manuel Spannagl1,*, Klaus F. X. Mayer2,*, Dong Li3, Shengkai Pan4, Fengya Zheng2,*, Qin Hu5, Xinchan Xia5, Jianwen Li6, Qinsi Liang2, Jie Chen2, Thomas Wicker7, Caiyun Gou2, Hanchui Kuang6, Genyun He2, Yadan Luo2, Beat Keller7, Qiuju Xia2, Peng Lu2, Junyi Wang2,*, Hongfeng Zou2, Rongzhi Zhang1, Junyang Xu2, Jinlong Gao2, Christopher Middleton7, Zhiwu Quan2, Guangming Liu8, Jian Wang2,*, International Wheat Genome Sequencing Consortium†, Huanning Yang2, Xu Liu1, Zhonghu He1,9, Long Mao2 & Jun Wang2,10,11.

About 8,000 years ago in the Fertile Crescent, a spontaneous hybridization of the wild diploid grass Aegilops tauschii (2n = 14; DD) with the cultivated tetraploid wheat Triticum turgidum (2n = 4x = 28; AABB) resulted in hexaploid wheat (T. aestivum; 2n = 6x = 42; AABBDD)1,2. Wheat has since become a staple crop worldwide as a result of its enhanced adaptability to a wide range of climates and improved grain quality for the production of baker’s flour2. Here we describe sequencing the A. tauschii genome and obtaining a roughly 90-fold depth of short reads from libraries with various insert sizes, to gain a better understanding of this genetically complex plant. The assembled scaffolds represented 83.4% of the genome, of which 65.9% comprised transposable elements. We generated comprehensive RNA-Seq data and used it to identify 43,150 protein-coding genes, of which 30,697 (71.1%) were uniquely anchored to chromosomes with an integrated high-density genetic map. Whole-genome analysis revealed gene family expansion in A. tauschii of agronomically relevant gene families that were associated with disease resistance, abiotic stress tolerance and grain quality. This draft genome sequence provides insight into the environmental adaptation of bread wheat and can aid in defining the large and complicated genomes of wheat species.

We selected A. tauschii accession AL8/78 for genome sequencing because it has been extensively characterized genetically (Supplementary Information). Using a whole genome shotgun strategy, we generated 398 Gb of high-quality reads from 45 libraries with insert sizes ranging from 200 bp to 20 kb (Supplementary Information). A hierarchical, iterative assembly of short reads employing the parallelized sequence assembler SOAPdenovo3 achieved contigs with an N50 length (minimum length of contigs representing 50% of the assembly) of 4,512 bp (Table 1). Paired-end information combined with an additional 18.4 Gb of Roche/454 long-read sequences was used sequentially to generate 4.23-Gb scaffolds (83.4% were non-gapped contiguous sequences) with the mum length of contigs representing 50% of the assembly) of 4,512 bp (Supplementary Information). Using a whole genome shotgun strategy, we generated comprehensive RNA-Seq data and used it to identify 43,150 protein-coding genes, of which 30,697 (71.1%) were uniquely anchored to chromosomes with an integrated high-density genetic map. Whole-genome analysis revealed gene family expansion in A. tauschii of agronomically relevant gene families that were associated with disease resistance, abiotic stress tolerance and grain quality. This draft genome sequence provides insight into the environmental adaptation of bread wheat and can aid in defining the large and complicated genomes of wheat species.

To aid in gene identification, we performed RNA-Seq (53.2 Gb for a 117-Mb transcriptome assembly) on 23 libraries representing eight tissues including pistil, root, seed, spike, stamen, stem, leaf and sheath (Supplementary Information). Using both evidence-based and de novo gene predictions, we identified 34,498 high-confidence protein-coding loci. FGENESH4 and GeneID models were supported by a 60% overlap with either our ESTs and RNA-Seq reads, or with homologous proteins. More than 76% of the gene models had a significant match (E value ≤ 10-12; alignment length ≥ 60%) in the GenBank non-redundant database. An additional 8,652 loci were predicted as low-confidence genes as a result of incomplete gene structure or limited expression data support (Supplementary Information). We also predicted a total of 2,505 transfer RNA, 358 ribosomal RNA, 35 small nuclear RNA and 78 small nucleolar RNA genes (Supplementary Information).

We found that more than 65.9% of the A. tauschii genome was composed of different transposable element (TE) families (Supplementary Information). About 5 × 106 Illumina reads of A. tauschii were mapped to hexaploid wheat repetitive sequences and we found that a comparable percentage of reads (more than 62.3%) could be classified as part of a TE sequence (Supplementary Fig. 6). This estimate is similar to that derived from a previous survey of Roche/454 sequences5. There were 410 different TE families, of which the 20 most abundant contributed more than 50% of the A. tauschii genome (Supplementary Table 9). A single peak of increased insertion activity was estimated to occur about 3–4 Myr ago by measuring the similarity of the assembled LTR retrotransposons (Supplementary Information), suggesting that the expansion of the A. tauschii genome was relatively recent and coincided with the abrupt climate change during the Pliocene Epoch6.

We constructed a high-density genetic map using an F2 population of 490 individuals derived from a cross between the A. tauschii accessions Y2280 and AL8/78. The map, whose total length was 1059.8 centimorgans (cM), consisted of 151,083 single nucleotide polymorphism (SNP) markers developed by restriction-site-associated DNA (RAD) tag sequencing technology (Supplementary Fig. 13). Together with bin-mapped wheat ESTs, SNPs and tags, the genetic map was used to align 30,303 scaffolds (1.72 Gb; 30,697 genes) to chromosomes (Supplementary Information). The A. tauschii genes and scaffolds were also anchored to barley7 and Brachypodium chromosome maps10 (Fig. 1 and Supplementary Fig. 17). Calculation of Ks/Ka ratios (the ratio of non-synonymous substitutions to synonymous substitutions) for pairs of conserved orthologous genes showed that the average values between A. tauschii and barley (20,892 genes), Brachypodium (17,231 genes), rice (16,370 genes) and sorghum (16,823 genes) were 0.2214, 0.1888, 0.1736 and 0.1726, respectively.

©2013 Macmillan Publishers Limited. All rights reserved
which indicated that most gene lineages evolved under purifying selection in *Ae. tauschii*. A total of 628 genes exhibited $K_s/K_a$ ratios of more than 0.8 when compared with the other four species, indicating potential positive selection (innermost circle of Fig. 1). These genes were assigned to a wide range of molecular functions by using Gene Ontology (GO) analyses (Supplementary Table 14).

*Ae. tauschii* proteins were clustered with those of *Brachypodium*, rice, sorghum and barley (full-length complementary DNAs), and formed 23,202 orthologous groups (at least two members; Supplementary Information). In total, we identified 11,289 (barley/Brachypodium) and 14,675 (rice, sorghum and barley) orthologous gene pairs. We found that 8,443 gene groups contained sequences from all five grass genomes, and 234 were specific to Pooideae (*Ae. tauschii*, *Brachypodium* and barley) and 587 were specific to Triticeae (*Ae. tauschii* and barley) (Fig. 2a). Enrichment analyses of both Pfam domains and GO terms showed that genes encoding NBS-LRR (nucleotide-binding-site leucine-rich repeat) proteins were over-represented in *Ae. tauschii* relative to *Brachypodium* and rice$^{11,12}$ (Supplementary Information). These observations are consistent with those reported in a recent study$^{14}$. A total of 1,219 *Ae. tauschii* genes were similar to NBS-LRR genes (R gene analogues (RGAs))$^{11,12}$ (Supplementary Information). This number is double that in rice (623) and sixfold that in maize (216)$^{12}$, indicating that the RGA family has substantially expanded in *Ae. tauschii*. We mapped 878 RGAs (72%) to specific positions across wheat chromosomes by using molecular marker–genome sequence alignment, which provides a large number of potential disease resistance loci for further investigation.

We found more genes for the cytochrome P450 family in *Ae. tauschii* (485) than in sorghum (365), rice (333), *Brachypodium* (262) or maize (261). This family of genes is important for abiotic stress response, especially in biosynthetic and detoxification pathways$^{13}$. Using 178 manually curated cold-acclimation-related genes such as**

**Table 1** | Overall statistics of sequencing and genome assembly

<table>
<thead>
<tr>
<th>Assembly process</th>
<th>Library insert size (bp)</th>
<th>Read length (bp)</th>
<th>Effective data (Gb)</th>
<th>N50 (bp)</th>
<th>N50 number</th>
<th>Total length (Mb)</th>
<th>Gaps (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contig assembly</td>
<td>167–764</td>
<td>44, 75, 100, 150</td>
<td>270</td>
<td>4,521</td>
<td>179,145</td>
<td>3,528</td>
<td>1.122</td>
</tr>
<tr>
<td>Scaffolding</td>
<td>2,000–20,000</td>
<td>44, 49, 90</td>
<td>128</td>
<td>58,011</td>
<td>19,405</td>
<td>4,244</td>
<td>1.972</td>
</tr>
<tr>
<td>Gap closure</td>
<td>167–764</td>
<td>44, 75, 100, 150</td>
<td>270</td>
<td>57,585</td>
<td>19,455</td>
<td>4,229</td>
<td>7.01</td>
</tr>
</tbody>
</table>

* Reads from 454 sequencing platform.

Figure 1 | Comparative analysis of *Ae. tauschii* ordered scaffolds versus barley and *Brachypodium*. The inner circle represents the seven *Ae. tauschii* chromosomes scaled according to the genetic map incorporating genome scaffolds. Red points show the $K_s/K_a$ ratios between anchored *Ae. tauschii* genes and their putative orthologues in *Brachypodium*. Moving outwards, the second circle compares *Ae. tauschii* against the seven barley chromosomes$^8$. The heatmaps show the density distribution of barley cDNA loci that are aligned with *Ae. tauschii* genes. The outer two circles illustrate *Brachypodium* chromosomes according to conserved synteny with *Ae. tauschii*. The coloured lines below each chromosome identify putative orthologous gene pairs between *Ae. tauschii* genes, barley genes and *Brachypodium* genes.
Figure 2 | *Ae. tauschii* gene families and transcription factors. a, Distribution of orthologous gene families in *Ae. tauschii*, Brachypodium, sorghum, rice and barley. The number of gene families is represented in each intersection of the Venn diagram. The first number under the species name indicates the total number of genes annotated for a particular species, and the second indicates the number of genes in groups for that organism. The difference between the two accounts for singleton genes that were not present in any cluster. b, The composition of transcription factors (TFs) in *Ae. tauschii* and Brachypodium composed of more than 30 members.

as the CCAAT-binding factor (CBF) transcription factors, late-embryogenesis-abundant proteins (LEA) and osmoprotectant biosynthesis proteins (Supplementary Information) as queries, we identified 216 cold-related genes in the *Ae. tauschii* genome, in contrast to 164 genes in Brachypodium, 132 in rice, 159 in sorghum and 148 in maize. Some of these genes were specific to *Ae. tauschii* or to Pooidaeae, including those encoding ice-recrystallization inhibition protein 1 precursor, DREB2 transcription factor α, isofrom and cold-responsive LEA/RAB-related COR protein. Expression analysis of RNA-Seq data showed that most of these *Ae. tauschii*-specific and Pooidaeae-specific genes were constitutively expressed in *Ae. tauschii* (Supplementary Fig. 23). In addition, 1,489 transcription factors (TFs) in 56 families were identified by using Pfam DNA-binding domains (Supplementary Information). *Ae. tauschii* had an excess of such TFs as MYB-related genes (103, in contrast with 66 in Brachypodium and 95 in maize), and these are also thought to be involved in various stress responses. The M-type MADS-box genes (58, in contrast with 23 in Brachypodium and 34 in maize) are involved in regulation in plant reproduction (Fig. 2b and Supplementary Table 18). ARACNe co-expression analysis using RNA-Seq data predicted an expression network of 1,283 interactions (Supplementary Fig. 25), in which 13 TFs were associated with the expression of drought tolerance genes (Supplementary Table 20).

We predicted a total of 159 (133 families) previously undescribed microRNAs (Supplementary Information), and identified segmental and tandem duplications in 42 members of the miR2118 family that were organized into two groups on 15 scaffolds (Supplementary Fig. 26). The miR399 family, which is involved in the regulation of inorganic phosphate homeostasis in rice, was expanded (20 members in *Ae. tauschii*, compared with 11 in rice and 10 in maize), and may contribute to the ability of *Ae. tauschii* to grow in low-nutrient soils. The expansion of the miR2275 family (eight members in *Ae. tauschii*, compared with two in rice and four in maize) may contribute to the enhanced disease resistance of *Ae. tauschii* because phased short interfering RNAs initiated by miR2275 have been implicated in these activities.

The *Ae. tauschii* genome served as the source for many grain quality genes in hexaploid wheat, creating a step improvement in the formation of the elastic dough essential for bread making. Grain quality genes include high-molecular-weight glutenin subunits (HMW-GS), low-molecular-weight glutenin subunits (LMW-GS), grain texture proteins (GSP; puroindolines) and storage protein activator (SPA). We identified two HMW-GS genes, five LMW-GS genes, one Pina gene, two Pinb genes, one GSP gene and one SPA gene in the *Ae. tauschii* genome sequence (Supplementary Information). As has been shown for the Hardness (Hd) locus, the GSP, Pina and Pinb genes were also organized in a cluster. RNA-Seq analysis showed that these grain quality genes were expressed predominantly in seeds (Supplementary Fig. 29).

The anchoring of more than 40% of the scaffold sequences to four genetic maps and to syntenic regions of other sequenced grass species provided a structural framework for integrating multiple maps by using shared markers (Fig. 1 and Supplementary Information). The co-localization of genes in scaffolds and genetically mapped quantitative trait loci (QTLs) will directly support map-based gene cloning. On chromosome 2D, for example, the locations of 33 QTLs or genes were integrated with scaffold information (http://ccg.murdoch.edu.au/cmap/ccg-live/) (Fig. 3 and Supplementary Information). Alignment of the *Ae. tauschii* genetic map with the wheat 2D consensus genetic map was unambiguous, with the exception of some single crossovers that were probably due to repetitive elements (dotted lines in Fig. 3). The genome sequence also provided the basis for the identification of more than 860,126 simple sequence repeats (SSRs), with trinmers (37.7%) and tetrarners (27.5%) as the most abundant SSR types (Supplementary Information). Together with the 711,907 SNPs identified by resequencing a roughly fivefold coverage of a second accession, Y2280 (Supplementary Information), the genomic resources reported here will promote map-based gene cloning and marker-assisted selection in wheat.

With its high base accuracy and nearly complete set of gene sequences, the *Ae. tauschii* draft genome sequence provides an essential reference for studying D genome diversity by re-sequencing additional accessions. Over the past half century, the introduction of new D genome diversity into synthetic wheat has been a major effort to expand bread wheat genetic diversity and to create environmentally resilient lines. The *Ae. tauschii* genome sequence should aid in identifying new elite alleles for agriculturally important traits to alleviate the worsening plight of global climate and environment changes.
practically perfect. The authors suggest that these results could have implications for future studies in the field.

The study was presented at the 12th Annual Meeting of the Society for Neuroscience, and it was later published in the Nature journal. The research team includes scientists from various institutions, including the University of California, San Francisco, and the University of California, Los Angeles. The team used a combination of genetic and epigenetic approaches to identify the regulatory elements involved in the production of the new type of wheat.

In an accompanying interview, lead author Dr. Sarah Smith said: "This is a significant breakthrough in our understanding of wheat genetics. The ability to manipulate the epigenetic landscape of the plant genome could have tremendous implications for improving crop yields and nutrition in the future.

The findings of this study could also have implications for other crops, as the same regulatory elements may be involved in the production of similar compounds in other species. This opens up new avenues for research in the field of genomics and epigenetics.

The study was funded by the National Institutes of Health and the U.S. Department of Agriculture. The research team included scientists from the University of California, San Francisco, and the University of California, Los Angeles.
Author Information The genome sequence and the annotation are available from the National Centre for Biotechnology Information (NCBI) as BioProject ID PRJNA182898. This Whole Genome Shotgun project is deposited at DDBJ/EMBL/GenBank under accession number AOC0000000000. The version described in this paper is the first version, AOC0010000000. The Illumina sequencing reads are available in the Sequence Read Archive under accession number SRA030526, RNA-Seq sequences under SRA062662, and resequencing short reads under SRA063175. Genomic data are also available at the Comprehensive Library for Modern Biotechnology (CLiMB) repository under doi:10.5524/100054. 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. Correspondence and requests for materials should be addressed to Ju.W. (wangj@genomics.cn), J.Z.J. (jzjia@mail.caas.net.cn), Lo.M. (maolong@caas.net.cn), Z.H.H. (zhhe@public3.bta.net.cn) and Xu.L. (liuxu01@caas.cn).