Journal of Genetics and Genomics xxx (2016) 1–11

Contents lists available at ScienceDirect

Journal of Genetics and Genomics

Journal homepage: www.journals.elsevier.com/journal-of-genetics-and-genomics/

Original research

Sequencing and comparative analysis of *Aegilops tauschii* chromosome arm 3DS reveal rapid evolution of Triticeae genomes

Jingzhong Xie [a, 1], Naxin Huo [b, c, 1], Shenghui Zhou [a, 1], Yi Wang [b, 1], Guanghao Guo [a], Karin R. Deal [c], Shuhong Ouyang [a], Yong Liang [a], Zhenzhong Wang [a], Lichan Xiao [c], Tingting Zhu [c], Tiezhu Hu [b], Vijay Tiwari [d], Jianwei Zhang [e], Hongxia Li [b], Zhongfu Ni [a], Yingyin Yao [a], Huiru Peng [a], Shengli Zhang [b], Olin D. Anderson [b], Patrick E. McGuire [c], Jan Dvorak [c, *], Ming-Cheng Luo [c, *], Zhiyong Liu [a, *], Yong Q. Gu [b, *], Qixin Sun [a, *]

[a] State Key Laboratory for Agrobiotechnology, China Agricultural University, Beijing 100193, China
[b] USDA-ARS West Regional Research Center, Albany, CA 94710, USA
[c] Department of Plant Sciences, University of California at Davis, Davis, CA 95616, USA
[d] Department of Plant Pathology, Kansas State University, Manhattan, KS 66506, USA
[e] Arizona Genomics Institute, School of Plant Science, University of Arizona, Tucson, AZ 85721, USA

ARTICLE INFO

Article history:
Received 19 August 2016
Received in revised form
26 September 2016
Accepted 27 September 2016
Available online xxx

Keywords:
Aegilops tauschii
Genome sequencing
Sequence assembly
Comparative genomics
Grass evolution

ABSTRACT

Bread wheat (*Triticum aestivum*, AABBDD) is an allohexaploid species derived from two rounds of interspecific hybridizations. A high-quality genome sequence assembly of diploid *Aegilops tauschii*, the donor of the wheat D genome, will provide a useful platform to study polyploid wheat evolution. A combined approach of BAC pooling and next-generation sequencing technology was employed to sequence the minimum tiling path (MTP) of 3176 BAC clones from the short arm of *Ae. tauschii* chromosome 3 (At3DS). The final assembly of 135 super-scaffolds with an N50 of 4.2 Mb was used to build a 247-Mb pseudomolecule with a total of 2222 predicted protein-coding genes. Compared with the orthologous regions of rice, *Brachypodium*, and sorghum, At3DS contains 38.67% more genes. In comparison to At3DS, the short arm sequence of wheat chromosome 3B (Ta3BS) is 95-Mb large in size, which is primarily due to the expansion of the non-centromeric region, suggesting that transposable element (TE) bursts in Ta3B likely occurred there. Also, the size increase is accompanied by a proportional increase in gene number in Ta3BS. We found that in the sequence of short arm of wheat chromosome 3D (Ta3DS), there was only less than 0.27% gene loss compared to At3DS. Our study reveals divergent evolution of grass genomes and provides new insights into sequence changes in the polyploid wheat genome.

1. Introduction

Globally, common (bread) wheat (*Triticum aestivum* L.) is one of the primary staple crops, supplying nearly one-fifth of the calories consumed by humans. The current growth rate of the human population is exceeding the rate of food production, and closing the gap will demand more efficient breeding methodologies for all major crops, including wheat (FAO, 2016).

Bread wheat is a hexaploid species ($2n = 6x = 42$, genome AABBDD), which originated by two hybridization steps. The first step was the hybridization of *T. urartu* ($2n = 2x = 14$, AA) with a species closely related to *Aegilops speltoides* ($2n = 2x = 14$, SS) (Dvorak and Zhang, 1990; Dvorak et al., 1993). This event gave rise to wild emmer wheat (*T. turgidum* ssp. *dicoccoides*, $2n = 4x = 28$, AABB). The second step, which occurred 8000 to 10,000 years ago (Nesbitt and Samuels, 1996), was the hybridization of an unknown

* Corresponding authors.
E-mail addresses: jdvorak@ucdavis.edu (J. Dvorak), mcluo@ucdavis.edu (M.-C. Luo), zyliu@genetics.ac.cn (Z. Liu), yong.gu@ars.usda.gov (Y.Q. Gu), qxsun@cau.edu.cn (Q. Sun).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jgg.2016.09.005

domesticated tetraploid wheat (*T. turgidum*) with *Ae. tauschii* ($2n = 2x = 14$, DD) (Kihara, 1944; Mc and Sears, 1946; Dvorak et al., 2012).

One of the factors limiting progress in wheat breeding is the low genetic diversity in the wheat D genome (Akhunov et al., 2010). *Ae. tauschii*, the wild donor of the wheat D genome, is a geographically widespread and genetically diverse species (Wang et al., 2013) and is a logical source for broadening genetic diversity in the wheat D genome (Warburton et al., 2006). A number of *Ae. tauschii* genes for resistance to various diseases, especially for rust resistance, such as *Lr21*, *Lr22*, *Lr32*, *Lr39*, *Lr42*, *Sr33* and *Sr45* (McIntosh et al., 2013), have been successfully incorporated into wheat. Genetic improvement in wheat would be greatly facilitated by the availability of a high-quality reference genome sequence for *Ae. tauschii*, an important wheat relative.

An attempt to construct a whole-genome shotgun (WGS) sequence of the *Ae. tauschii* (accession AL8/78) genome has been reported (Jia et al., 2013), but the fragmented nature and largely unordered scaffolds of the sequence limit its utility. The *Ae. tauschii* genome is highly repetitive (Jia et al., 2013) and of large size (estimated to be between 4 and 5 Gb) (Rees and Walters, 1965; Arumuganathan and Earle, 1991), which is the primary cause of the failure to produce a high-quality reference sequence by the WGS sequencing approach.

Ae. tauschii, along with wheat, barley and several hundred of other species, belongs to the tribe Triticeae of the grass family (Love, 1984). A large genome size and high content of repeated sequences are attributes of all species in the tribe, including wheat, and have been the principal obstacles to genome sequencing in this economically important plant group. To overcome these obstacles, alternative sequencing approaches have been pursued, such as the construction of BAC-based physical maps followed by ordered-clone genome sequencing along the minimum tiling path (MTP), complexity reduction through sequencing of flow-sorted chromosomes (International Wheat Genome Sequencing Consortium, 2014), and optical BioNano genome mapping (Hastie et al., 2013). A reference sequence of the common wheat chromosome 3B, the largest chromosome in the wheat genome, was successfully assembled by sequencing BAC pools along the MTP of the 3B physical map, using a combination of Roche 454 and Illumina sequencing technologies coupled with extensive BAC end sequencing and ordering of super-scaffolds on a high-density genetic map (Choulet et al., 2014). This success demonstrated that the ordered-clone sequencing approach can generate a single-chromosome reference-quality sequence in Triticeae, using a flow-sorted chromosome library, thus reducing the complexity of the entire genome. A relevant question is whether this approach can be scaled up to sequence an entire Triticeae genome without resorting to the complexity reduction by chromosome sorting.

Here we sequenced the short arm of *Ae. tauschii* chromosome 3D (henceforth At3DS) with the goal of assessing the use of the ordered clone sequencing approach to generate a high-quality reference sequence for an entire *Ae. tauschii* genome. The ordered-clone sequencing of the At3DS arm employed a genome-wide MTP across the entire physical map of the *Ae. tauschii* genome (Luo et al., 2013). We evaluated synteny of the At3DS pseudomolecule with those of other sequenced grass genomes, such as *Brachypodium*, rice, and sorghum, as well as with the homoeologous wheat chromosome arm 3BS. In addition, the high-quality At3DS sequence assembly allowed us for the first time to evaluate the sequence polymorphism between the D genomes of polyploid wheat and its diploid ancestor on a large chromosomal scale.

2. Results

2.1. Sequence and assembly of the Ae. tauschii 3DS pseudomolecule

In this study, 3176 MTP clones from 365 BAC contigs across the *Ae. tauschii* 3DS physical map (Safar et al., 2010; Luo et al., 2013) were sequenced and assembled using a hybrid strategy as detailed in Materials and Methods. After merging assembled sequences based on MTP BAC order, we obtained a final assembly of 689 scaffolds with an N50 of 766 kb and a total length of 292,922,430 bp. A BioNano map of *Ae. tauschii* was used to align, order, and orient the scaffolds to generate 135 super-scaffolds with an N50 of 4.2 Mb using a technique previously described by Hastie et al. (2013). Several map resources, including two high-resolution genetic maps (Jia et al., 2013; Luo et al., 2013) and one Radiation-Hybrid (RH) map (Kumar et al., 2015) for *Ae. tauschii* were used to build an At3DS pseudomolecule. The final pseudomolecule was 247,348,992 bp long; 77 scaffolds with a total length of 22,279,739 bp remained unanchored. These unanchored scaffolds accounted for 8% of the total At3DS sequences (Table S1). To assess the completeness of the sequence, we assumed that At3DS had a size and gene content equivalent to the wheat 3DS arm (Ta3DS). Comparison between the At3DS assembly and the chromosome arm size of Ta3DS (Luo et al., 2010) gave an estimate that the At3DS MTP BAC-based sequence assembly covered about 90% of the At3DS arm sequence.

2.2. Transposable elements

Transposable elements (TEs) represented 81.18% of the At3DS pseudomolecule (Table 1). This TE proportion was similar to that in the survey sequences of *Ae. tauschii* chromosome 5D (82.8%) (Akpinar et al., 2015), *T. aestivum* cv. Chinese Spring chromosome 3B (85.5%) (Choulet et al., 2014) and global Chinese Spring chromosome arm survey sequence assemblies (81.5%) (International Wheat Genome Sequencing Consortium, 2014), respectively. Over 80% of the At3DS TEs belonged to three superfamilies: *CACTA*, *Gypsy*, and *Copia*. More than two thirds of the TEs were

Table 1
Features of protein coding genes and TEs of At3DS.

Feature	Pseudomolecule	Unanchored
Protein coding genes		
No. of loci	2222	166
Mean/median no. of exons	4.3/3	3.2/2
Mean/median gene size (bp)	3821/2856	3644/2511
Mean/median transcripts length (bp)	1884/1097	1810/1014
Retrotransposons		
Copia	13.83%	12.48%
Gypsy	34.08%	35.49%
Athila	0.03%	0.02%
TRIM	0.00%	0.01%
Unclassified LTR-Retrotransposon	14.89%	14.64%
LINE	0.26%	0.23%
SINE	0.00%	0.01%
Total retrotransposons	63.09%	62.88%
DNA transposons		
CACTA	17.96%	20.43%
hAT	0.00%	0.02%
Mutator	0.08%	0.11%
Tc1/Mariner	0.01%	0.01%
PIF/Harbinger	0.03%	0.03%
Unclassified class II with TIRs	0.00%	0.01%
MITE	0.01%	0.01%
Helitron	0.02%	0.02%
Total DNA transposons	18.11%	20.64%
Total TEs	81.18%	83.50%

Please cite this article in press as: Xie, J., et al., Sequencing and comparative analysis of *Aegilops tauschii* chromosome arm 3DS reveal rapid evolution of Triticeae genomes, Journal of Genetics and Genomics (2016), http://dx.doi.org/10.1016/j.jgg.2016.09.005

retrotransposons, with the *Gypsy* LTR superfamily being the most abundant (34.08%). The TE content and profile of superfamilies in At3DS were similar to those in the wheat homoeologous chromosome arms 3AS, 3BS, and 3DS (Fig. 1). The percentages of TEs representing the homoeologous chromosome arm Os1S of rice (*Oryza sativa*), Sb3S of sorghum (*Sorghum bicolor*), and the distal portion of the short arm of *Brachypodium distachyon* chromosome 2 (Bd2S) were all lower than that in At3DS, reflecting their smaller genome sizes. The proportion of *CACTA* superfamily represented approximately 17% in the investigated Triticeae chromosome arms but only about 1% in the homoeologous chromosome regions in the three small grass genomes. Fig. 1 shows that the increase in genome size in Triticeae compared to the three representative small genome grass species occurred primarily by increase in the abundance of LTR retroelements and *CACTA* DNA elements.

We estimated the length of the At3DS centromeric region by plotting the density of the centromere-specific repeat families *Cereba* and *Quinta* along the At3DS pseudomolecule (Fig. S1). A region of 77 Mb from 170 to 247 Mb can be defined as a putative pericentromeric-centromeric region. Since this 77-Mb sequence is only from the short-arm region of the 3D centromere, the size of the entire centromeric region of *Ae. tauschii* chromosome 3D is therefore comparable to the size of the 122-Mb centromeric region (265 Mb–387 Mb) of wheat 3B chromosome (Choulet et al., 2014), suggesting that they may have similar centromeric structures with regard to size and repeat element composition. We used the middle of the centromeric region to define the short arm of wheat chromosome 3B (Ta3BS). Most of the expansion in Ta3BS appears to be in the non-centromeric region of the chromosome, as the Ta3BS non-centromeric region (265 Mb) is 95 Mb larger than that of At3DS (170 Mb). Our results are consistent with the observation that the total centromere size is positively correlated with the genome size among species, but is independent of chromosome size within the species (Zhang and Dawe, 2012).

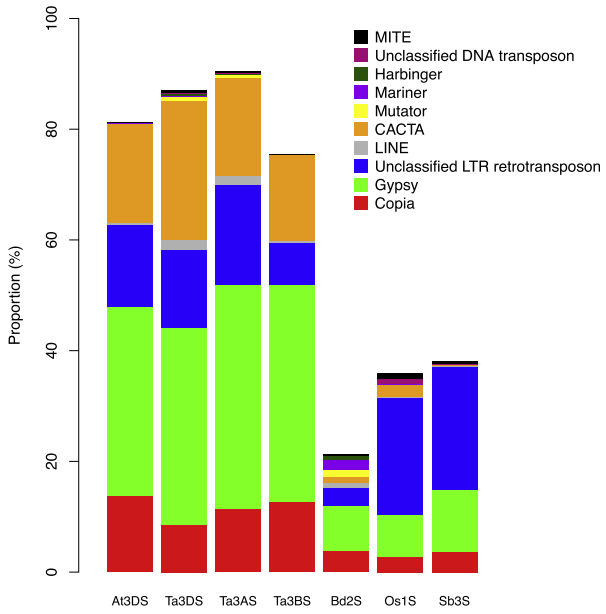

Fig. 1. Comparison of repetitive sequence types between homologous grass chromosomes. Proportions of ten different types of repetitive sequences that account for at least one percent of at least one of the seven chromosomes were plotted in different colors for At3DS and relevant homologous chromosome arms, including *T. aestivum* chromosome arm 3BS (Ta3BS), part of *Brachypodium* chromosome 2 short arm (Bd2S), rice chromosome 1 (Os1S), and sorghum chromosome 3 (Sb3S).

2.3. Gene content

The genes in the At3DS sequence were annotated with the reference-based TriAnnot pipeline (Leroy et al., 2012), which had previously been used for gene prediction in the wheat 3B chromosome. A total of 2222 protein-coding genes, including 2124 high confidence (HC) genes and 98 low confidence (LC) genes, were predicted and annotated on the At3DS pseudomolecule. Among them, 1823 (85.96%) HC genes were supported by reference proteins with more than 70% coverage (Table 1). Expression data from 8 tissues at different developmental stages were used to validate the annotated genes (Jia et al., 2013). A total of 1873 (88.18%) of the 2124 HC genes and 51 (52.04%) of the 98 LC genes were expressed in at least one of the eight tissues. These results indicate that most of our gene predictions are reliable. The average length of the At3DS genes was 3821 bp, generating transcripts 1884 bp long. The genes had on average 4.3 exons (Table 1). This is similar to the average exon numbers reported in some other annotated genome sequences: Ta3B (4.2), *Triticum urartu* (4.7), *B. distachyon* (5.5), rice (4.7) and sorghum (4.7) (International Rice Genome Sequencing Project, 2005; Paterson et al., 2009; International Brachypodium Initiative, 2010; Ling et al., 2013; Choulet et al., 2014). Additionally, 166 genes were annotated on the unanchored scaffolds, accounting for ~7% of the total number of 2388 genes (2222 plus 166) on At3DS.

2.4. Relationship among TE density, gene density, and recombination rate

TE density increased from 54% to nearly 94% towards the centromeric region along At3DS, with a relatively rapid increase at the distal 60 Mb region (Fig. 2A). The *Gypsy* LTR superfamily was the most abundant and exhibited a similar distribution to the total TE density ($r = 0.79$, P-value < 0.01), suggesting that the *Gypsy* LTR superfamily may be a major cause of variation in TE density along At3DS (Fig. S2). The proportion of the *Copia* LTR superfamily was negatively associated with that of the *Gypsy* LTR superfamily ($r = -0.61$, P-value < 0.01), while no significant correlation between the *CACTA* TIR superfamily and the two LTR superfamilies was observed. *CACTA* elements appear to be more evenly distributed in the sequence.

Gene density was uneven along At3DS and decreased from telomere to centromere (Fig. 2B). The highest gene density was in the distal 10 Mb region, where it was 25.3 genes per Mb on average. The gene density declined very sharply over the distal 60 Mb region, and then decreased at a relatively moderate rate to 2.5 genes per Mb in the centromeric region (about one tenth of the gene density of the telomeric region). This is consistent with the telomere-centromere gene density decline found on all seven *Ae. tauschii* chromosomes (Luo et al., 2013) and also on *T. aestivum* chromosome 3B (Choulet et al., 2014). We also confirmed the absence of large gene islands on At3DS.

Using the high-density SNP linkage map (Luo et al., 2013), the recombination rate of At3DS was estimated to be 0.33 cM/Mb on average, which is nearly identical to the average recombination rate of 0.32 cM/Mb across the entire *Ae. tauschii* genome (Luo et al., 2013). In the distal 60 Mb region, the recombination rate dropped precipitously from 3.12 cM/Mb to 0.2 cM/Mb, and was then maintained at a nearly constant low level in the centromeric region (nearly 0 cM/Mb) (Fig. 2C), indicating that most recombination take place in the distal region. Consequently, the first 60-Mb region accounts for 90% of the genetic length of the At3DS arm. The recombination rate is positively correlated with the gene density ($r = 0.89$, P-value < 0.01), but negatively correlated with TE density ($r = -0.89$, P-value < 0.01) along the chromosome arm (Fig. 2).

Please cite this article in press as: Xie, J., et al., Sequencing and comparative analysis of *Aegilops tauschii* chromosome arm 3DS reveal rapid evolution of Triticeae genomes, Journal of Genetics and Genomics (2016), http://dx.doi.org/10.1016/j.jgg.2016.09.005

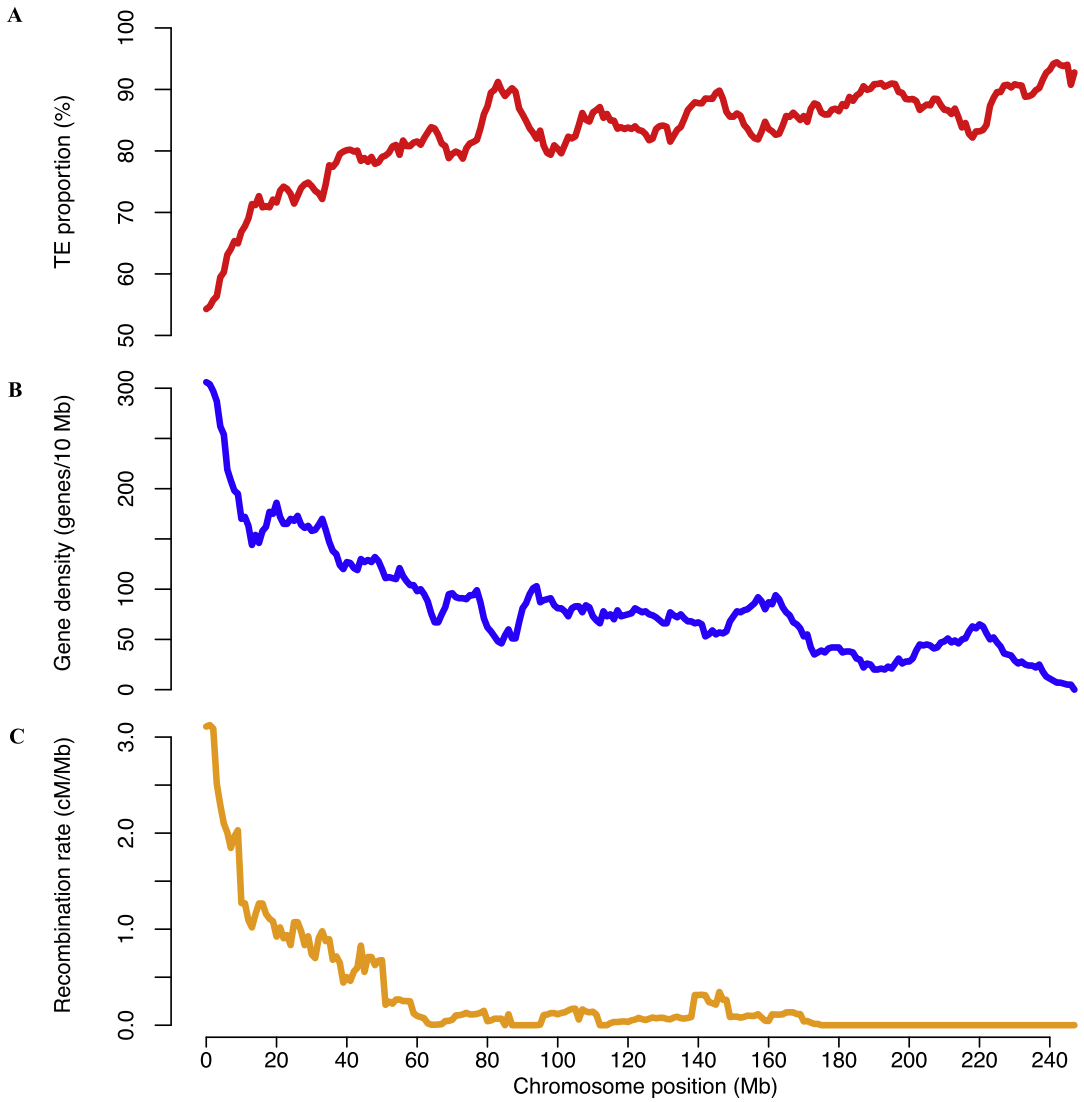

Fig. 2. Distributions of TE density, gene density, and recombination rate along At3DS. TE density (**A**), gene density (**B**), and recombination rate (**C**) are all counted in a sliding window of 10 Mb with a step of 1 Mb, and were plotted along the chromosome.

2.5. Synteny

At3DS is syntenic with Os1S, Sb3S, and the distal part of Bd2S (Luo et al., 2013), and the borders of the syntenic regions were identified using genes as anchors (Table S2). To make the result comparable with these grass genomes, we applied a filtration process similar to that described by Glover et al. (2015) to identify functional core gene sets by discarding possible mispredicted genes and transposon genes. The core gene sets consisted of 1929, 1357, 1465, and 1352 genes located in the syntenic region of At3D, Bd2S, Os1S, and Sb3S, respectively (Fig. 3). The regions of the Bd2S, Os1S, and Sb3S pseudomolecules orthologous to At3DS had similar numbers of core genes, an average of 1391 genes, but At3DS contained 538 (38.67%) additional genes.

To assess synteny, we defined syntenic genes as orthologous core genes found in orthologous (homoeologous) regions in at least two of the three compared species. If an orthologous core gene was not found in a syntenic position in any species, we considered it non-syntenic. A total 1046, 1231, 1311, and 1242 syntenic genes and 883 (45.77%), 126 (9.28%), 154 (10.51%), and 110 (8.13%) non-syntenic genes were found in At3DS, Bd2S, Os1S and Sb3S, respectively (Fig. 3). *Ae. tauschii* had about 220 fewer syntenic genes than the average syntenic gene number of the others. The smaller number of syntenic genes could potentially be caused by incomplete coverage of the entire homoeologous region, about 90%, but this result should be taken with due caution. However, synteny assessment clearly showed that At3DS had a large number of non-syntenic genes, representing 45.77% of the total At3DS core genes, which is about 36.47% more than that of *Brachypodium*, rice and sorghum. The relatively high proportion of non-syntenic genes has been reported in Ta3B as well (Choulet et al., 2014).

Non-syntenic genes could be caused by transposition/translocation of genes from elsewhere in the genome. To assess this possibility, the genic sequences of non-syntenic At3DS genes were aligned to the D genome of the IWGSC survey sequence assembly of Chinese Spring, excluding the short arm of 3D (International Wheat Genome Sequencing Consortium, 2014). We found that 420 (48%) non-syntenic At3DS genes have at least one copy on chromosome regions elsewhere in the wheat D genome, suggesting that nearly half of the non-syntenic At3DS genes resulted from inter-chromosomal gene duplications. Inter-chromosomal gene duplication that gives rise to non-syntenic genes often occurs at the single gene level, which has been reported previously (Glover et al., 2015). Of these non-syntenic genes, we found that 21% were either

Please cite this article in press as: Xie, J., et al., Sequencing and comparative analysis of *Aegilops tauschii* chromosome arm 3DS reveal rapid evolution of Triticeae genomes, Journal of Genetics and Genomics (2016), http://dx.doi.org/10.1016/j.jgg.2016.09.005

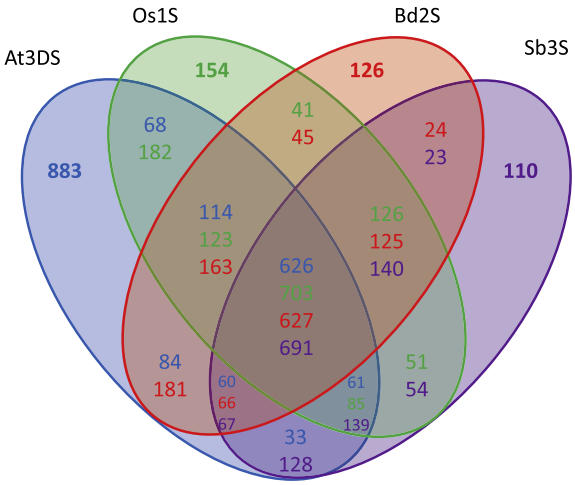

Fig. 3. Venn diagram displaying the number of genes conserved in grasses. Four core gene sets, including *Aegilops tauschii* chromosome 3 short arm (At3DS, blue), rice chromosome 1 short arm (Os1S, green), part of *Brachypodium* chromosome 2 short arm (Bd2S, red) and sorghum chromosome 3 short arm (Sb3S, purple), were compared to each other. The number of homologous genes in each intersection for the gene sets from different species is shown accordingly in different colors. The number of non-syntenic genes is indicated in bold for each gene set.

further duplicated tandemly or dispersed within the sequenced At3DS region, suggesting the intra-chromosomal duplication has further increased the total number of non-syntenic genes.

The rate of duplication/fixation of non-syntenic genes in each genome can be estimated using the divergence time among grass species (International Brachypodium Initiative, 2010; Massa et al., 2011; Glover et al., 2015). We calculated that the rates of non-syntenic gene duplication/fixation for *Brachypodium*, rice and sorghum were similar, ranging from 1.4×10^{-3} to 2.4×10^{-3} locus^{-1} MYA^{-1}, which is nearly identical to those reported by Massa et al. (2011). However, the rate for *Ae. tauschii* is 11.7×10^{-3} locus^{-1} MYA^{-1}, nearly five-fold higher than that in the other grass species, suggesting a fast evolution of the *Ae. tauschii* genome with regard to the duplication/fixation rate of non-syntenic genes.

2.6. Synteny between At3DS and Ta3BS

The reference sequence of the wheat 3B chromosome (Choulet et al., 2014) makes it possible for the first time to perform a detailed comparison of synteny between two Triticeae genomes, albeit based on a single chromosome arm. The TriAnnot gene annotation pipeline predicted a total of 3101 genes in Ta3BS as compared to 2222 genes in At3DS, a difference of 879 genes. Of the At3DS genes, 1715 (77.18%) were homologous to 2287 (73.75%) genes on Ta3BS. The extra 572 genes in Ta3BS are likely derived from gene duplications. Since synteny analysis can be complicated by tandem gene duplications, gene transpositions, and chromosome rearrangements, we used collinearity, a more specific form of synteny that requires conservation of gene order, to reveal evolutionary changes (Fig. 4). Collinearity analysis identified 1296 collinear gene pairs between At3DS and Ta3BS, indicating that 58.32% of the At3DS genes were collinear with 41.79% of the Ta3BS genes. In other words, at least 18.86% of the At3DS genes and 31.96% of the Ta3BS genes have been involved in sequence rearrangement events, resulting in changes of gene positions.

Additionally, we examined the distribution of the non-collinear genes (41.68% of the At3DS genes and 58.21% of the Ta3BS genes) along the chromosome and found that non-collinear gene numbers decline gradually from telomere to centromere in both At3DS and Ta3BS (Fig. 4). The number of non-collinear genes is positively associated with recombination rate ($r = 0.85$, P-value < 0.01) and gene density ($r = 0.94$, P-value < 0.01) in At3DS, suggesting a possible relationship between recombination and collinearity. Moreover, no large-scale structural rearrangements were found between At3DS and Ta3BS, except for a few small-scale rearrangements in the centromeric region. These rearrangements are likely caused by inversions and translocations, or by assembly errors in the centromeric regions in the pseudomolecules (Fig. 4). Meanwhile, although Ta3BS has 879 more annotated genes than At3DS, the average gene densities in these two arms are similar, 8.9 and 9.0 genes per Mb for Ta3BS and At3DS, respectively, due to the fact that the Ta3BS arm is longer than the At3DS arm. This fact implies that the expansion of the Ta3BS arm was in proportion to the increase in total gene number.

2.7. Gene deletions in the wheat 3DS arm

A whole genome duplication caused by polyploidization generates fabric for the evolution of new genes (Ohno, 1970) and turns on several genomic processes. One of them is gene loss. The At3DS sequence provides a reference for the assessment of the process on genome sequence level. To analyze gene loss in the wheat 3DS arm, the predicted Coding Sequence (CDS) of each gene in At3DS was aligned to the genomic DNA sequences of Ta3DS using the splice site alignment program Gmap (Wu and Nacu, 2010) to find the wheat D-genome homologs. *Ae. tauschii* CDSs that did not align well with the Ta3DS genes were reanalyzed using a local alignment algorithm to identify potential structure variations between homologs, employing a manual examination to validate the alignment.

In total, 2076 (93% of the total annotated genes) of the At3DS genes had homologs on Ta3DS. To determine if the remaining 146 At3DS genes were truly absent from Ta3DS and not just missing sequences in the assembly, the D-genome specific primer pairs for the 146 genes were designed based on the *Ae. tauschii* sequences and PCR was performed using genomic DNA of *Ae. tauschii* accession AL8/78, Chinese Spring, and Chinese Spring nullisomic-tetrasomic lines N3A-T3D, N3B-T3D and N3D-T3A. Only six of the 146 (0.27%) gene sequences could be amplified in *Ae. tauschii* but not in Chinese Spring and nullisomic-tetrasomics, indicating that these six genes were absent from the Ta3DS. Out of the 146 genes, two gene sequences were amplified in *Ae. tauschii*, Chinese Spring, N3A-T3D and N3B-T3D but not in N3D-T3A, showing that they were located on Ta3D and the putative loss was a false positive result. The other 138 gene loss predictions were not confirmed by PCR assay. The likely explanations were gene duplications in Ta3D or sequence polymorphism. The missing 0.27% of the At3DS genes represented an upper limit of the percentage of genes that may have been deleted from the single wheat lineage, which Chinese Spring represents. This result is close to the 0.17% of deleted genes in the wheat D genome estimated earlier employing several hundred *T. aestivum* accessions (Dvorak et al., 2004).

2.8. Single nucleotide polymorphism and indels

Single nucleotide polymorphism (SNP) and indel polymorphism between the *Ae. tauschii* and hexaploid wheat D genomes were examined to infer the SNP and indel distribution. A total of 254,682 SNPs were identified between homologous gene pairs between At3DS and Ta3DS, with 66,395 SNPs in promoters, 94,940 SNPs in 5′ UTRs, 14,882 SNPs in CDS, 50,104 SNPs in introns, and 28,361 SNPs in 3′ UTRs (Fig. 5A). The alignment of annotated *Ae. tauschii* genes with those of Ta3DS identified a total of 56,020 indels in genic regions, with 14,965 indels in promoters, 18,054 indels in 5′ UTRs, 6667 indels in CDS, 10,354 indels in introns, and 5980 indels in 3′

Please cite this article in press as: Xie, J., et al., Sequencing and comparative analysis of *Aegilops tauschii* chromosome arm 3DS reveal rapid evolution of Triticeae genomes, Journal of Genetics and Genomics (2016), http://dx.doi.org/10.1016/j.jgg.2016.09.005

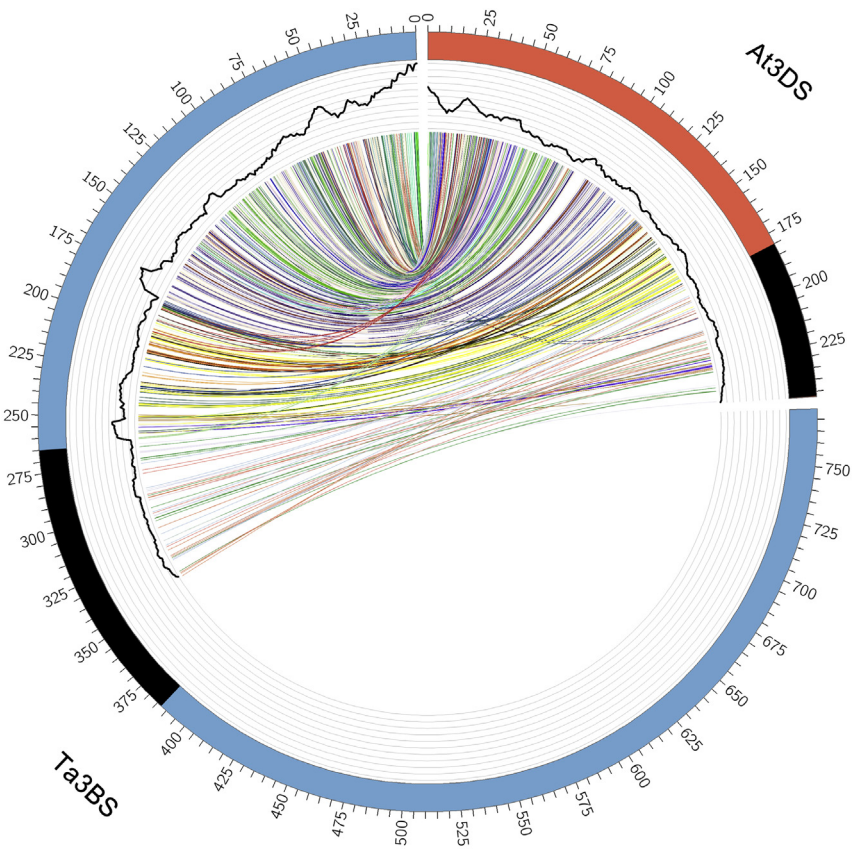

Fig. 4. Collinear and non-collinear genes between At3DS and Ta3BS. The 1296 collinear gene pairs were identified between At3DS and Ta3BS, and are linked by lines in different colors. The black lines in the inner circle indicate the number of non-collinear genes along the chromosomes. The non-collinear gene number was calculated in a 10 Mb window with an increment of 1 Mb. The outer circle shows the chromosomes, with centromeric region in black and the non-centromeric region displayed in blue (Ta3B) or red (At3DS). The scales for each region are in Mb.

UTRs (Fig. 5B). The exon lengths were nearly identical between homologous gene pairs in At3DS and Ta3DS, while the lengths of introns were variable (Fig. S3).

Within genes, CDSs are the most conserved regions, followed by 5′ UTR, 3′ UTR, introns, and promoter regions (Fig. 5C). The analysis showed that the median CDS sequence identities between At3DS and Ta3DS are 99.9%. Similar median sequence identities for 5′ UTRs, 3′ UTRs, and introns were observed (99.8%). Promoter regions seem to experience a faster sequence differentiation, with a median sequence identity of 99.1%. Overall, the sequence polymorphism in the genic regions (including promoter, 5′ UTR, CDS, introns, and 3′ UTR) between At3DS and Ta3DS is less than one SNP per kb on average. Additionally, we observed that the CDS identities increased from telomere to centromere along At3DS, with lower CDS identities in the telomeric regions and higher identities in the centromeric region, except for a slight decline in the pericentromeric region (Fig. 6). In the CDS regions, we found 8361 (57%) synonymous and 6521 (43%) nonsynonymous SNPs, resulting in a ratio of nonsynonymous to synonymous substitutions of 0.78 and a corresponding overall K_a/K_s value of 0.33, suggesting that the genomic regions have been under purifying selection.

3. Discussion

3.1. Sequencing strategy

The ordered-clone sequencing of the At3DS arm resulted in a pseudomolecule 247 Mb long with the superscaffold N50 is equal to 4.2 Mb. The pseudomolecule covered ~90% of the At3DS arm sequence and contained 2124 high confidence protein-coding genes. The successful construction of the pseudomolecule was primarily made possible by use of the *Ae. tauschii* physical map, in which 84.2% of the MTP was anchored on a high-resolution genetic map (Luo et al., 2013). A radiation hybrid map of the *Ae. tauschii* genome (Kumar et al., 2015), another high-density genetic map (Chapman et al., 2015), BioNano genome map (Hastie et al., 2013; Stankova et al., 2016), and the conserved collinearity with *Brachypodium* were employed during manual editing of superscaffolds and the construction of the pseudomolecule.

The ordered-clone sequencing approach was previously successfully used to sequence wheat chromosome 3B (Choulet et al., 2014). Because sequencing of chromosome 3B was based on flow-sorted chromosome BAC libraries, it was a poor predictor of problems that will be encountered if this technology is scaled up to the sequencing of an entire Triticeae genome. The successful construction of the At3DS pseudomolecule provides evidence that it will be feasible to sequence the entire *Ae. tauschii* genome by our approach and this approach could be used for the sequencing of any other Triticeae genome if a prerequisite physical map comparable in quality to that of *Ae. tauschii* exists. The same technology that was used for the development of the *Ae. tauschii* physical map can also be used for the development of a global physical map of hexaploid wheat (Luo et al., 2010) indicating that genomes of polyploid Triticeae species, including wheat, can be sequenced using the genome-wide ordered-clone approach.

Since the At3DS assembly accounted for ~90% of the At3DS arm, it is likely that the other whole-genome sequences constructed with this approach will have about the same coverage. The

Please cite this article in press as: Xie, J., et al., Sequencing and comparative analysis of *Aegilops tauschii* chromosome arm 3DS reveal rapid evolution of Triticeae genomes, Journal of Genetics and Genomics (2016), http://dx.doi.org/10.1016/j.jgg.2016.09.005

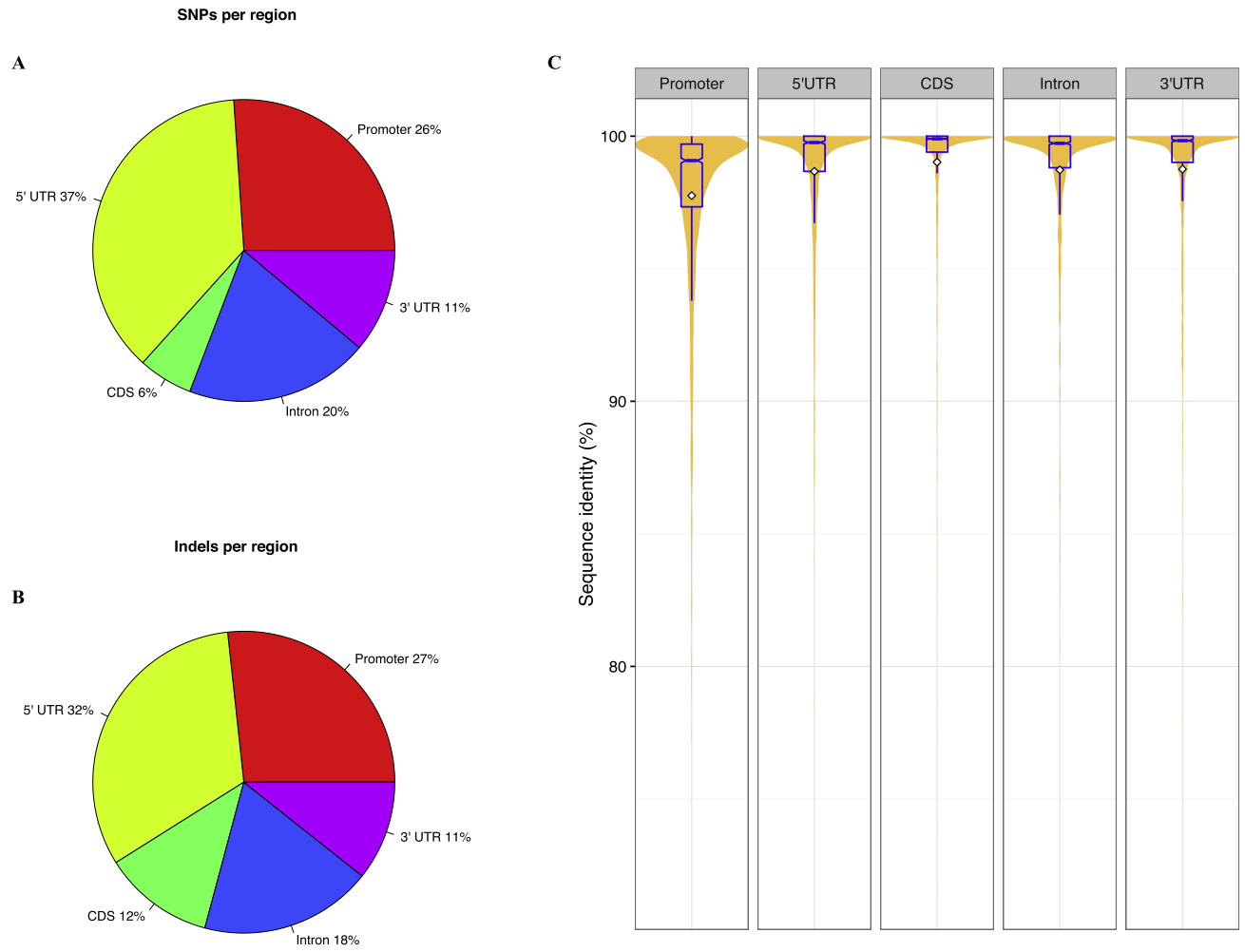

Fig. 5. SNPs, indels, and sequence identities between At3DS and Ta3DS. SNPs (**A**) and Indels (**B**) in orthologous gene pairs between At3DS and Ta3DS were summarized per region in the pie chart with their percentages labeled. Gene regions are shown in different colors. Violin plots (**C**) in orange were used to compare sequence identity distributions. Narrow box plots were overlaid to show the quartiles. The white dots show the means.

combination of 454 sequencing technology that was used to sequence the At3DS arm and other NGS platforms could improve the sequence assembly and increase the contig and scaffold lengths. However, no NGS platform can close gaps caused by sequences missing from the BAC libraries on which the physical map is based. To close gaps and bring the sequence coverage to the vicinity of 100% may require the deployment of a hybrid sequencing approach incorporating a WGS sequence into the final genome sequence assembly.

3.2. Recombination rate and TE and gene distribution

The At3DS pseudomolecule contained 81.18% TEs. LTR retrotransposons, primarily the *Gypsy* superfamily, were the major TE component of the At3DS pseudomolecule. The *Gypsy* superfamily was also the most abundant TE superfamily in wheat homoeologous chromosome arms 3AS and 3BS, and also in the much smaller *B. distachyon*, rice and sorghum genomes (International Rice Genome Sequencing Project, 2005; Paterson et al., 2009; International Brachypodium Initiative, 2010).

The abundance of the *Gypsy* superfamily was the primary factor responsible for the TE telomere-centromere gradient along the At3DS arm. TE density increases from distal region, where TEs accounted for about 54% of the sequence to the centromeric region where it accounted for 94% of the sequence. TE distribution was negatively correlated with gene density ($r = -0.96$, P-value < 0.01) and recombination rate ($r = -0.88$, P-value < 0.01). The negative correlation between gene density and overall TE density is intuitive since the presence of one excludes the presence of the other. However, the correlations of TE density and gene density with recombination rate are also obvious. In addition to the overall gene density (Akhunov et al., 2003) and TE density (Choulet et al., 2010), other variables, such as the densities of single-copy loci (Akhunov et al., 2003), multigene loci (Akhunov et al., 2003), non-collinear genes (Luo et al., 2013), synteny (Akhunov et al., 2003; Luo et al., 2009), disease resistance genes (Luo et al., 2013), gene deletions and duplications (Dvorak et al., 2004; Dvorak and Akhunov, 2005), RFLP (Dvorak et al., 1998), and SNP (Wang et al., 2014), have been shown to be positively or negatively correlated with recombination rate in wheat or its close relatives, including *Ae. tauschii*. The fact that recombination is the independent variable in so many genomic relationships suggests that it may play the central role in shaping the structure and evolution of Triticeae genomes.

3.3. TEs and gene number

The genetic and physical mapping of the *Ae. tauschii* genome revealed that the *Ae. tauschii* genome has accumulated a large number of structural changes, and it was suggested that the large amount of repeated sequences and their precipitous rate of

Please cite this article in press as: Xie, J., et al., Sequencing and comparative analysis of *Aegilops tauschii* chromosome arm 3DS reveal rapid evolution of Triticeae genomes, Journal of Genetics and Genomics (2016), http://dx.doi.org/10.1016/j.jgg.2016.09.005

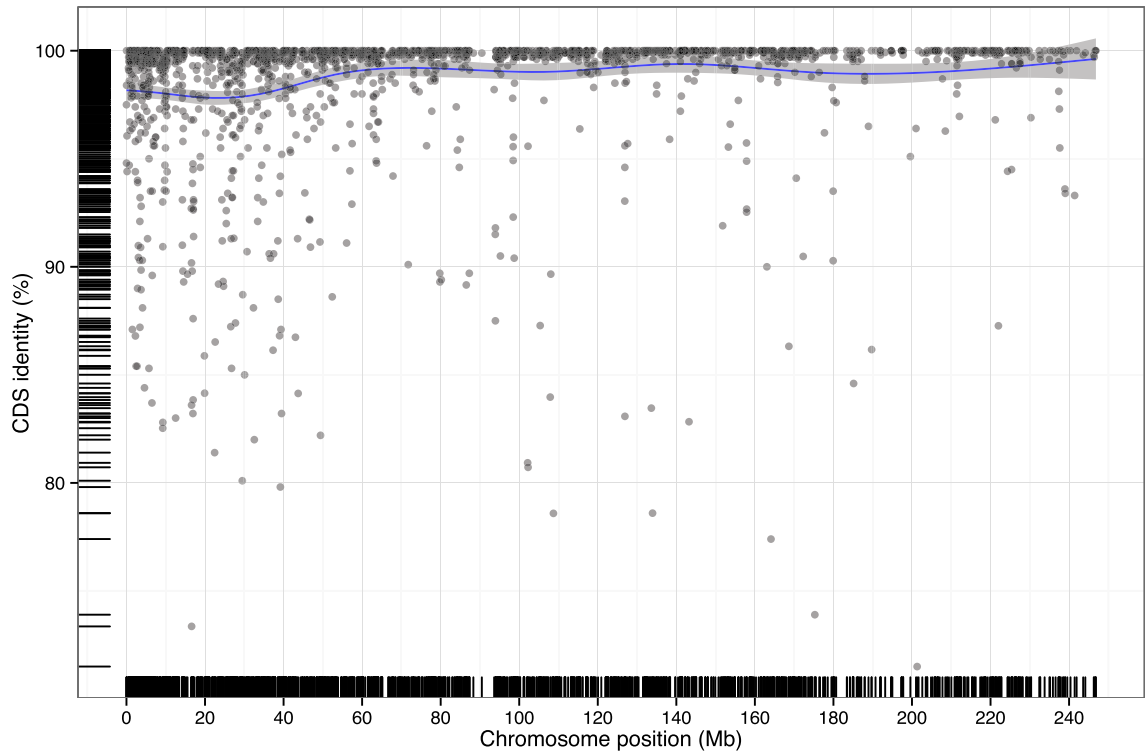

Fig. 6. Chromosomal distribution of CDS identities between At3DS and Ta3DS. The CDS identities of homologous gene pairs between At3DS and Ta3DS are represented in black dots along the At3DS pseudomolecule sequence. The loess function in ggplot2 was used to draw the regression curve. The Y axis indicates the percent identity of each CDS.

turnover (Dubcovsky and Dvorak, 2007) resulted in faster rate of genome evolution in the *Ae. tauschii* linage compared to linages of grass species with small genomes and small TE contents (Luo et al., 2009). It was also suggested that one of the consequences of this genomic instability was the increase in the number of genes in the large *Ae. tasuchii* genome compared to small grass genomes (Massa et al., 2011). Manual annotation of the 9.7 Mb of sequence sampled from a number of regions of the *Ae. tauschii* genome estimated that 26% of the annotated genes were not syntenic. The annotation of the entire At3DS arm estimated the percentage of non-syntenic genes to be 45% of the total gene number. Both estimates are higher than the 10% of non-syntenic genes found in small grass genomes. If the suggested relationship between genome size and gene number in grasses were true, 3BS arm should contain more genes and more non-collinear genes than the 3DS arm since the 3BS arm is about 30% larger than the 3DS arm (Dvorak et al., 1984). We found that the non-collinear genes account for 59% of the Ta3BS genes, which is consistent with this prediction.

Also consistent with this prediction is the number of genes annotated on the two chromosome arms, 3101 on the 3BS and 2222 on At3DS. Although both gene number estimates were generated with the same annotation pipeline, caution is nevertheless needed in taking the difference at its face value. Of the At3DS protein-coding genes 88% were supported by mRNA evidence but only 71% of genes annotated on the 3BS arm pseudomolecule were supported by mRNA evidence. Lower mRNA support for genes annotated on the 3BS arm suggests that some of the genes annotated on the 3BS arm could be pseudogenes, and that the difference in the number of genes between the two arms may not be as great as the data appear to suggest.

3.4. Gene loss from the wheat D genome and gene divergence

Polyploidy duplicates the gene content of a genome. The resulting whole-genome duplication (WGD) is then slowly reversed by functional divergence of duplicated genes (Ohno, 1970) or their loss. The latter process leads to “meropolyploidy”, a situation in which a portion of the genome remains polyploid but a portion is diploid due to loss of duplicated genes (Langham et al., 2004). Hexaploid wheat is an outstanding system to study this process because of the availability of the ancestral genomes makes it possible to quantify the loss of wheat genes. Using restriction fragment polymorphism (RFLP) in wheat and wheat ancestors, the rates of gene loss from the three wheat genomes were estimated to be 2.11%, 5.18% and 0.17% of wheat A, B, and D-genome single-copy genes, respectively, and the rates were highly dependent on the recombination rate along the chromosomes (Dvorak et al., 2004). The At3DS pseudomolecule provided a reference for the assessment of the process at the genome sequence level. The pseudomolecule was compared to the Ta3DS survey sequence and the putative gene deletions in the Ta3DS arm were validated by PCR. Since *Ae. tauschii* accession AL8/78 is not the exact source of the wheat D genome (Wang et al., 2014) and an unknown percentage of the absent genes may have actually been deleted at the diploid level where they existed as deletion polymorphism and then introduced or introgressed into hexaploid wheat, an estimate based on simple genome sequence comparisons represents an upper limit estimate. We showed that at most 0.27% of the genes present in the At3DS pseudomolecule may have been lost from the D genome of hexaploid wheat since the origin of hexaploid wheat about 8000 years ago. This estimate of gene loss from the wheat D genome is remarkably close to the 0.17% loss of D-genome single-copy genes based on RFLP (Dvorak et al., 2004). As noted previously (Dvorak et al., 2004) these estimates of gene loss are a stark contrast to 15% of the total DNA lost within a few generations from the D genome in artificially synthetized wheat (Feldman and Levy, 2009). Nascent amphiploids including synthetic wheat are intrinsically unstable (Zhang et al., 2013b) due to complex meiotic processes

Please cite this article in press as: Xie, J., et al., Sequencing and comparative analysis of *Aegilops tauschii* chromosome arm 3DS reveal rapid evolution of Triticeae genomes, Journal of Genetics and Genomics (2016), http://dx.doi.org/10.1016/j.jgg.2016.09.005

accompanying meiotic restitution (Oleszczuk and Lukaszewski, 2014) and meiotic stability in subsequent generations. The great discrepancy between the level of DNA lost reported for an artificially produced hexaploid wheat and that observed in natural hexaploid wheat is probably caused by purifying natural selection that has stabilized the wheat meiotic system but also underscores the danger in making evolutionary inferences from short-term experimental data and a limited number of observations.

The comparison of gene sequences in the 3DS arm of *Ae. tauschii* accession AL8/78 and Chinese Spring 3DS arm showed that genes were highly similar and shared with greater than 99% sequence identity. However, not all gene regions were equally conserved. Coding sequences were the most conserved (99.7% identity) whereas the promoter regions were the least conserved (99.1%). In coding sequences, SNPs occurred mostly in the third codon positions, amounting to 57% of the 14,882 SNPs detected in coding sequences. The overall K_a/K_s value for homologous genes from Ae3DS and Ta3DS was 0.33, suggesting purifying selection acting on them. Since most, if not all, of this SNP may have originated at the diploid level, the observed K_a/K_s ratio may be the consequence of natural selection in *Ae. tauschii,* not in wheat. It would therefore be a mistake to take the K_a/K_s value as evidence of purifying selection acting on wheat genes until SNPs that originated during polyploid wheat evolution are unequivocally identified.

In contrast to other polyploids (Zhang et al., 2015), neither gene loss nor gene divergence indicates that the wheat D genome experienced a severe “genome shock” in transition from the tetraploid to hexaploid level. This is consistent with minor epigenetic adjustments of the D genome (Zhao et al., 2011) and suggests that changes in gene expression and small RNA-mediated dynamic regulation of homologs may have played a primary role in hexaploid wheat evolution and adaptation (Li et al., 2014).

4. Materials and methods

4.1. BAC DNA isolation, library construction, and sequencing

A total of 3176 MTP clones on the *Ae. tauschii* physical map were selected for sequencing (Luo et al., 2013). 2312 MTP BACs were derived from 143 FPC BAC contigs that were genetically anchored on the *Ae. tauschii* chromosome 3D short arm, and 864 MTP BACs came from 224 FPC BAC contigs which were co-assembled with the *T. aestivum* cv. Chinese spring 3DS physical map (Luo et al., 2010) and considered to be part of the *Ae. tauschii* chromosome 3D short arm. The MTP clones belonging to the same BAC contig were re-arrayed in alpha-numerical order. Equal numbers of cells from eight BAC clone cultures were mixed, and the BAC pool DNA was isolated with the Qiagen Midiprep kit. The BAC pool DNA was then sheared to a peak size of around 1400 base pairs with the Covaris, according to the Roche Rapid Library Method Manual for the GS FLX + Series. As for mate-pair sequencing library preparation, 5 g of mate-pair pool DNA was used to construct 3 kb paired-end libraries respectively and sequenced using the Roche GS FLX Titanium XL Chemistry protocol.

Shotgun libraries were constructed following a modified Roche Rapid Library Preparation Method Manual for GS FLX + Series (Lennon et al., 2010). Briefly, 1 μg DNA of each BAC pool was end-repaired and ligated with one of the 24 molecular identifier (MID) adapters (GS FLX Titanium Rapid Library MID Adaptors MID 1–24), and quantified again by a Quant-iT™ PicoGreen® assay. In general, DNAs from 20 to 24 adapter-ligated BAC pools were proportionally pooled before constructing 454 sequencing libraries with the Rapid Library Preparation kit for sequencing using XL + Chemistry Series according to the manufacturer's protocol. The BAC end sequences (BES) were sequenced following the protocol described by Huo et al. (2008).

4.2. Sequence assembly, manual curation, and pseudomolecule construction

Sequence reads were assembled using the Roche 454 gsAssembler V2.6 package (Margulies et al., 2005) with parameters requiring at least 96% identity within a minimal overlap of 50 bp. Assembled contigs were scaffolded by adding 3–5 kb mate-pair reads using the Consed package (Gordon and Green, 2013). BAC ends sequences (BES) were aligned to scaffolds using Consed (Gordon and Green, 2013) to identify the BAC end position and corresponding BAC sequences in each scaffold. All BAC sequences from the same BAC contig were merged using a golden path (AGP) builder pipeline (Genome Puzzle Master) (Zhang et al., 2016) The assembly was manually curated if the sequences conflicted with either the information from physical map or the BES position.

To further anchor, order, and orient large-scaffolds on the physical map for a pseudomolecule, 5773 genetically mapped SNP markers on 3DS from Jia et al. (2013) and 473 genetically mapped SNP markers from Luo et al. (2013) were aligned to BAC contig sequences using the short reads aligner BWA (Li and Durbin, 2010). All the contiguous sequences were *in silico* digested, then aligned to 345 contigs, totaling 349 Mb and having an N50 of 1.46 Mb, of the *Ae. tauschii* genome map to assist with the ordering, orientating and anchoring of FPC contigs. All above maps were merged to generate a consensus map using the software ALLMAPS (Tang et al., 2015) with default parameters except for different weights. The large-scaffolds near the centromere on the physical map with low recombination rates were ordered based on the radiation map (Kumar et al., 2015), BioNano optical map, and synteny with *Brachypodium* genome. Finally, the sequences of super-scaffolds were concatenated to form the pseudomolecule following the order and orientation in the consensus map by inserting 100 N′s among sequences of the same BAC contig, and 200 N′s between different BAC contigs, using in-house perl scripts.

4.3. Annotation of repeats and coding genes

Sequences and annotation files of all species in this study were downloaded from Ensembl Plant database release 25 (Kersey et al., 2014). Repeat DNA analysis was performed using the software RepeatMasker (Smit et al., 2015) with the repeat database of MIPS v9.3 (Nussbaumer et al., 2013) and default parameters. The total lengths and proportion of each TE family was obtained by adding up the length of each annotated feature of that specific family.

The TriAnnot gene prediction pipeline version 4.3 (Leroy et al., 2012) was applied to annotate the repeat-masked sequences. The annotation workflow combines different external evidences, including annotated proteins of closely related species, consisting of *T. urartu*, *Ae. tauschii,* barley, *Brachypodium*, rice, sorghum, maize, bread wheat chromosome 3B, as well as publicly available fl-cDNAs, ESTs, and RNA-Seq assembly of *T. aestivum*. The predicted protein-coding genes with >40% reference sequence coverage were classified as high confidence (HC) genes, and the remains were considered as low confidence (LC) genes.

4.4. Identification of syntenic genes, non-syntenic genes, and collinearity blocks

A process similar to that as described by Glover et al. (2015) was applied to filter genes in At3DS, Ta3BS, Bd, Os, and Sb to generate core gene sets for comparative analyses. In this process, TE relevant genes were discarded according to the annotations in each genome, and only the longest isoform for each gene was kept. Genes with

Please cite this article in press as: Xie, J., et al., Sequencing and comparative analysis of *Aegilops tauschii* chromosome arm 3DS reveal rapid evolution of Triticeae genomes, Journal of Genetics and Genomics (2016), http://dx.doi.org/10.1016/j.jgg.2016.09.005

unknown homolog (BLASTP hit with ≥35% amino acid identity, ≤1e−5 *e*-value and ≥35% query coverage) in the existing protein database were excluded as well (Camacho et al., 2009). Syntenic genes were defined as genes having their best BLAST hits (*e*-value <1e−5) located on at least one of the syntenic chromosome arms in At3DS, Ta3BS, Bd, Os or Sb. If all the best BLAST hits in At3DS, Ta3BS, Bd, Os, or Sb were found on non-orthologous chromosomes, the genes were considered non-syntenic.

Collinearity between the genes located on chromosomes At3DS and Ta3BS was analyzed by identifying collinearity blocks using the program MCScanX (Wang et al., 2012). All the amino acid sequences of the gene sets of each species were used in an all-by-all BLASTp comparison (using *e*-value cutoff at 1e−10 and description top 5). Collinearity and the distribution of non-collinear genes at chromosomes was visualized using the Circos software (Krzywinski et al., 2009).

4.5. Indels and sequence nucleotide polymorphism analysis

The transcripts of annotated *Ae. tauschii* genes were aligned to the wheat IWGSC genomic assembly (International Wheat Genome Sequencing Consortium, 2014) using spliced alignment software Gmap with default parameters (Wu and Nacu, 2010). After filtering low-quality alignments such as those due to mapping quality <20 or misalignments due to high copy number genes, the alignment results were used for calling SNPs and indels in coding and non-coding region between At3DS and Ta3DS genes. Promoter sequences were defined as the upstream 2 kb sequence beyond 5′ UTR. Sequence alignments of promoters were conducted by BLAST for each orthologous gene pairs. SNPs, indels and sequence identities were parsed and calculated from the tabular output of Gmap and BLAST using in-house perl and shell scripts.

4.6. Synonymous (K_s) and non-synonymous (K_a) substitution rates analysis

For calculation of K_a/K_s between homologous gene pairs of At3DS and Ta3DS, the CDS were aligned to each other and the K_a and K_s rates were computed using the software gKaKs (Zhang et al., 2013a) using the default parameters. To calculate the K_a/K_s ratio, genes with K_s value equal to 0 were excluded.

Acknowledgements

The authors would like to thank James Thomson and Toni Mohr for their critical reading and providing comments to improve the manuscript. The work was supported by funding from the National Natural Science Foundation of China (Nos. 31290210, 31210103902), the Unites States National Science Foundation grant (No. IOS 1238231), the USDA-Agricultural Research Service CRIS project (No. 5325-21000-019), and the Ministry of Education of China (111 project).

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jgg.2016.09.005.

References

Akhunov, E.D., Akhunova, A.R., Anderson, O.D., Anderson, J.A., Blake, N., Clegg, M.T., Coleman-Derr, D., Conley, E.J., Crossman, C.C., Deal, K.R., Dubcovsky, J., Gill, B.S., Gu, Y.Q., Hadam, J., Heo, H., Huo, N., Lazo, G.R., Luo, M.C., Ma, Y.Q., Matthews, D.E., McGuire, P.E., Morrell, P.L., Qualset, C.O., Renfro, J., Tabanao, D., Talbert, L.E., Tian, C., Toleno, D.M., Warburton, M.L., You, F.M., Zhang, W., Dvorak, J., 2010. Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. BMC Genomics 11, 702.

Akhunov, E.D., Akhunova, A.R., Linkiewicz, A.M., Dubcovsky, J., Hummel, D., Lazo, G., Chao, S., Anderson, O.D., David, J., Qi, L., Echalier, B., Gill, B.S., Miftahudin, Gustafson, J.P., La Rota, M., Sorrells, M.E., Zhang, D., Nguyen, H.T., Kalavacharla, V., Hossain, K., Kianian, S.F., Peng, J., Lapitan, N.L., Wennerlind, E.J., Nduati, V., Anderson, J.A., Sidhu, D., Gill, K.S., McGuire, P.E., Qualset, C.O., Dvorak, J., 2003. Synteny perturbations between wheat homoeologous chromosomes caused by locus duplications and deletions correlate with recombination rates. Proc. Natl. Acad. Sci. U. S. A. 100, 10836–10841.

Akpinar, B.A., Lucas, S.J., Vrana, J., Dolezel, J., Budak, H., 2015. Sequencing chromosome 5D of *Aegilops tauschii* and comparison with its allopolyploid descendant bread wheat (*Triticum aestivum*). Plant Biotechnol. J. 13, 740–752.

Arumuganathan, K., Earle, E.D., 1991. Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep. 9, 208–218.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421.

Chapman, J.A., Mascher, M., Buluc, A., Barry, K., Georganas, E., Session, A., Strnadova, V., Jenkins, J., Sehgal, S., Oliker, L., Schmutz, J., Yelick, K.A., Scholz, U., Waugh, R., Poland, J.A., Muehlbauer, G.J., Stein, N., Rokhsar, D.S., 2015. A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. Genome Biol. 16, 26.

Choulet, F., Alberti, A., Theil, S., Glover, N., Barbe, V., Daron, J., Pingault, L., Sourdille, P., Couloux, A., Paux, E., Leroy, P., Mangenot, S., Guilhot, N., Le Gouis, J., Balfourier, F., Alaux, M., Jamilloux, V., Poulain, J., Durand, C., Bellec, A., Gaspin, C., Safar, J., Dolezel, J., Rogers, J., Vandepoele, K., Aury, J.M., Mayer, K., Berges, H., Quesneville, H., Wincker, P., Feuillet, C., 2014. Structural and functional partitioning of bread wheat chromosome 3B. Science 345, 1249721.

Choulet, F., Wicker, T., Rustenholz, C., Paux, E., Salse, J., Leroy, P., Schlub, S., Le Paslier, M.C., Magdelenat, G., Gonthier, C., Couloux, A., Budak, H., Breen, J., Pumphrey, M., Liu, S., Kong, X., Jia, J., Gut, M., Brunel, D., Anderson, J.A., Gill, B.S., Appels, R., Keller, B., Feuillet, C., 2010. Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. Plant Cell 22, 1686–1701.

Dubcovsky, J., Dvorak, J., 2007. Genome plasticity a key factor in the success of polyploid wheat under domestication. Science 316, 1862–1866.

Dvorak, J., Akhunov, E.D., 2005. Tempos of gene locus deletions and duplications and their relationship to recombination rate during diploid and polyploid evolution in the *Aegilops-Triticum* alliance. Genetics 171, 323–332.

Dvorak, J., Deal, K.R., Luo, M.C., You, F.M., von Borstel, K., Dehghani, H., 2012. The origin of spelt and free-threshing hexaploid wheat. J. Hered. 103, 426–441.

Dvorak, J., Luo, M.C., Yang, Z.L., 1998. Restriction fragment length polymorphism and divergence in the genomic regions of high and low recombination in self-fertilizing and cross-fertilizing aegilops species. Genetics 148, 423–434.

Dvorak, J., Mcguire, P.E., Mendlinger, S., 1984. Inferred chromosome morphology of the ancestral genome of *Triticum*. Plant Syst. Evol. 144, 209–220.

Dvorak, J., Terlizzi, P., Zhang, H.B., Resta, P., 1993. The evolution of polyploid wheats: identification of the A genome donor species. Genome 36, 21–31.

Dvorak, J., Yang, Z.L., You, F.M., Luo, M.C., 2004. Deletion polymorphism in wheat chromosome regions with contrasting recombination rates. Genetics 168, 1665–1675.

Dvorak, J., Zhang, H.B., 1990. Variation in repeated nucleotide sequences sheds light on the phylogeny of the wheat B and G genomes. Proc. Natl. Acad. Sci. U. S. A. 87, 9640–9644.

FAO, 2016. Cereals and us: time to renew an ancient bond. In: Save and Grow in Practice: Maize, Rice, Wheat, Chapter 1, pp. 3–10.

Feldman, M., Levy, A.A., 2009. Genome evolution in allopolyploid wheat-a revolutionary reprogramming followed by gradual changes. J. Genet. Genomics 36, 511–518.

Glover, N.M., Daron, J., Pingault, L., Vandepoele, K., Paux, E., Feuillet, C., Choulet, F., 2015. Small-scale gene duplications played a major role in the recent evolution of wheat chromosome 3B. Genome Biol. 16, 188.

Gordon, D., Green, P., 2013. Consed: a graphical editor for next-generation sequencing. Bioinformatics 29, 2936–2937.

Hastie, A.R., Dong, L., Smith, A., Finklestein, J., Lam, E.T., Huo, N., Cao, H., Kwok, P.Y., Deal, K.R., Dvorak, J., Luo, M.C., Gu, Y., Xiao, M., 2013. Rapid genome mapping in nanochannel arrays for highly complete and accurate de novo sequence assembly of the complex *Aegilops tauschii* genome. PLoS One 8, e55864.

Huo, N., Lazo, G.R., Vogel, J.P., You, F.M., Ma, Y., Hayden, D.M., Coleman-Derr, D., Hill, T.A., Dvorak, J., Anderson, O.D., Luo, M.C., Gu, Y.Q., 2008. The nuclear genome of *Brachypodium distachyon*: analysis of BAC end sequences. Funct. Integr. Genomics 8, 135–147.

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

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

International Wheat Genome Sequencing Consortium, 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 345, 1251788.

Jia, J., Zhao, S., Kong, X., Li, Y., Zhao, G., He, W., Appels, R., Pfeifer, M., Tao, Y., Zhang, X., Jing, R., Zhang, C., Ma, Y., Gao, L., Gao, C., Spannagl, M., Mayer, K.F., Li, D., Pan, S., Zheng, F., Hu, Q., Xia, X., Li, J., Liang, Q., Chen, J., Wicker, T., Gou, C., Kuang, H., He, G., Luo, Y., Keller, B., Xia, Q., Lu, P., Wang, J., Zou, H., Zhang, R., Xu, J., Gao, J., Middleton, C., Quan, Z., Liu, G., Wang, J., International Wheat Genome Sequencing C., Yang, H., Liu, X., He, Z., Mao, L., Wang, J., 2013. *Aegilops*

tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. Nature 496, 91–95.

Kersey, P.J., Allen, J.E., Christensen, M., Davis, P., Falin, L.J., Grabmueller, C., Hughes, D.S., Humphrey, J., Kerhornou, A., Khobova, J., Langridge, N., McDowall, M.D., Maheswari, U., Maslen, G., Nuhn, M., Ong, C.K., Paulini, M., Pedro, H., Toneva, I., Tuli, M.A., Walts, B., Williams, G., Wilson, D., Youens-Clark, K., Monaco, M.K., Stein, J., Wei, X., Ware, D., Bolser, D.M., Howe, K.L., Kulesha, E., Lawson, D., Staines, D.M., 2014. Ensembl Genomes 2013: scaling up access to genome-wide data. Nucleic Acids Res. 42, D546–D552.

Kihara, H., 1944. Discovery of the DD-analyser, one of the ancestors of *Triticum vulgare*. Agri. Hort. 19, 13–14.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., Marra, M.A., 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19, 1639–1645.

Kumar, A., Seetan, R., Mergoum, M., Tiwari, V.K., Iqbal, M.J., Wang, Y., Al-Azzam, O., Simkova, H., Luo, M.C., Dvorak, J., Gu, Y.Q., Denton, A., Kilian, A., Lazo, G.R., Kianian, S.F., 2015. Radiation hybrid maps of the D-genome of *Aegilops tauschii* and their application in sequence assembly of large and complex plant genomes. BMC Genomics 16, 800.

Langham, R.J., Walsh, J., Dunn, M., Ko, C., Goff, S.A., Freeling, M., 2004. Genomic duplication, fractionation and the origin of regulatory novelty. Genetics 166, 935–945.

Lennon, N.J., Lintner, R.E., Anderson, S., Alvarez, P., Barry, A., Brockman, W., Daza, R., Erlich, R.L., Giannoukos, G., Green, L., Hollinger, A., Hoover, C.A., Jaffe, D.B., Juhn, F., McCarthy, D., Perrin, D., Ponchner, K., Powers, T.L., Rizzolo, K., Robbins, D., Ryan, E., Russ, C., Sparrow, T., Stalker, J., Steelman, S., Weiand, M., Zimmer, A., Henn, M.R., Nusbaum, C., Nicol, R., 2010. A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454. Genome Biol. 11, R15.

Leroy, P., Guilhot, N., Sakai, H., Bernard, A., Choulet, F., Theil, S., Reboux, S., Amano, N., Flutre, T., Pelegrin, C., Ohyanagi, H., Seidel, M., Giacomoni, F., Reichstadt, M., Alaux, M., Gicquello, E., Legeai, F., Cerutti, L., Numa, H., Tanaka, T., Mayer, K., Itoh, T., Quesneville, H., Feuillet, C., 2012. TriAnnot: a versatile and high performance pipeline for the automated annotation of plant genomes. Front. Plant. Sci. 3, 5.

Li, A., Liu, D., Wu, J., Zhao, X., Hao, M., Geng, S., Yan, J., Jiang, X., Zhang, L., Wu, J., Yin, L., Zhang, R., Wu, L., Zheng, Y., Mao, L., 2014. mRNA and small RNA transcriptomes reveal insights into dynamic homoeolog regulation of allopolyploid heterosis in nascent hexaploid wheat. Plant Cell 26, 1878–1900.

Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589–595.

Ling, H.Q., Zhao, S., Liu, D., Wang, J., Sun, H., Zhang, C., Fan, H., Li, D., Dong, L., Tao, Y., Gao, C., Wu, H., Li, Y., Cui, Y., Guo, X., Zheng, S., Wang, B., Yu, K., Liang, Q., Yang, W., Lou, X., Chen, J., Feng, M., Jian, J., Zhang, X., Luo, G., Jiang, Y., Liu, J., Wang, Z., Sha, Y., Zhang, B., Wu, H., Tang, D., Shen, Q., Xue, P., Zou, S., Wang, X., Liu, X., Wang, F., Yang, Y., An, X., Dong, Z., Zhang, K., Zhang, X., Luo, M.C., Dvorak, J., Tong, Y., Wang, J., Yang, H., Li, Z., Wang, D., Zhang, A., Wang, J., 2013. Draft genome of the wheat A-genome progenitor *Triticum urartu*. Nature 496, 87–90.

Love, A., 1984. Conspectus of the Triticeae. Feddes Repert. 95, 425–521.

Luo, M.C., Deal, K.R., Akhunov, E.D., Akhunova, A.R., Anderson, O.D., Anderson, J.A., Blake, N., Clegg, M.T., Coleman-Derr, D., Conley, E.J., Crossman, C.C., Dubcovsky, J., Gill, B.S., Gu, Y.Q., Hadam, J., Heo, H.Y., Huo, N., Lazo, G., Ma, Y., Matthews, D.E., McGuire, P.E., Morrell, P.L., Qualset, C.O., Renfro, J., Tabanao, D., Talbert, L.E., Tian, C., Toleno, D.M., Warburton, M.L., You, F.M., Zhang, W., Dvorak, J., 2009. Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in Triticeae. Proc. Natl. Acad. Sci. U. S. A. 106, 15780–15785.

Luo, M.C., Gu, Y.Q., You, F.M., Deal, K.R., Ma, Y., Hu, Y., Huo, N., Wang, Y., Wang, J., Chen, S., Jorgensen, C.M., Zhang, Y., McGuire, P.E., Pasternak, S., Stein, J.C., Ware, D., Kramer, M., McCombie, W.R., Kianian, S.F., Martis, M.M., Mayer, K.F., Sehgal, S.K., Li, W., Gill, B.S., Bevan, M.W., Simkova, H., Dolezel, J., Weining, S., Lazo, G.R., Anderson, O.D., Dvorak, J., 2013. A 4-gigabase physical map unlocks the structure and evolution of the complex genome of *Aegilops tauschii*, the wheat D-genome progenitor. Proc. Natl. Acad. Sci. U. S. A. 110, 7940–7945.

Luo, M.C., Ma, Y., You, F.M., Anderson, O.D., Kopecky, D., Simkova, H., Safar, J., Dolezel, J., Gill, B., McGuire, P.E., Dvorak, J., 2010. Feasibility of physical map construction from fingerprinted bacterial artificial chromosome libraries of polyploid plant species. BMC Genomics 11, 122.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., Rothberg, J.M., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380.

Massa, A.N., Wanjugi, H., Deal, K.R., O'Brien, K., You, F.M., Maiti, R., Chan, A.P., Gu, Y.Q., Luo, M.C., Anderson, O.D., Rabinowicz, P.D., Dvorak, J., Devos, K.M., 2011. Gene space dynamics during the evolution of *Aegilops tauschii, Brachypodium distachyon, Oryza sativa*, and *Sorghum bicolor* genomes. Mol. Biol. Evol. 28, 2537–2547.

Mc, F.E., Sears, E.R., 1946. The origin of *Triticum spelta* and its free-threshing hexaploid relatives. J. Hered. 37, 107–116.

McIntosh, R.A., Yamazaki, Y., Dubcovsky, J., Rogers, W.J., Morris, C., Appels, R., Xia, X.C., 2013. Catalogue of gene symbols for wheat. In: 12th International Wheat Genetic Symposium, Yokohama, Japan.

Nesbitt, M., Samuels, D.C., 1996. From stable crop to extinction? The archaeology and history of the hulled wheats. In: Padulosi, S., Hammer, K.,J.,H. (Eds.), Hulled Wheats. International Plant Genetic Resources Institute, Rome, pp. 41–100.

Nussbaumer, T., Martis, M.M., Roessner, S.K., Pfeifer, M., Bader, K.C., Sharma, S., Gundlach, H., Spannagl, M., 2013. MIPS PlantsDB: a database framework for comparative plant genome research. Nucleic Acids Res. 41, D1144–D1151.

Ohno, S., 1970. Evolution by Gene Duplication. Springer, Berlin Heidelberg.

Oleszczuk, S., Lukaszewski, A.J., 2014. The origin of unusual chromosome constitutions among newly formed allopolyploids. Am. J. Bot. 101, 318–326.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., Schmutz, J., Spannagl, M., Tang, H., Wang, X., Wicker, T., Bharti, A.K., Chapman, J., Feltus, F.A., Gowik, U., Grigoriev, I.V., Lyons, E., Maher, C.A., Martis, M., Narechania, A., Otillar, R.P., Penning, B.W., Salamov, A.A., Wang, Y., Zhang, L., Carpita, N.C., Freeling, M., Gingle, A.R., Hash, C.T., Keller, B., Klein, P., Kresovich, S., McCann, M.C., Ming, R., Peterson, D.G., Mehboob ur, R., Ware, D., Westhoff, P., Mayer, K.F., Messing, J., Rokhsar, D.S., 2009. The *Sorghum bicolor* genome and the diversification of grasses. Nature 457, 551–556.

Rees, H., Walters, M.R., 1965. Nuclear DNA and evolution of wheat. Heredity 20, 73–82.

Safar, J., Simkova, H., Kubalakova, M., Cihalikova, J., Suchankova, P., Bartos, J., Dolezel, J., 2010. Development of chromosome-specific BAC resources for genomics of bread wheat. Cytogenet. Genome Res. 129, 211–223.

Smit, A., Hubley, R., Green, P., 2015. RepeatMasker Open-4.0. http://www.repeatmasker.org/.

Stankova, H., Hastie, A.R., Chan, S., Vrana, J., Tulpova, Z., Kubalakova, M., Visendi, P., Hayashi, S., Luo, M., Batley, J., Edwards, D., Dolezel, J., Simkova, H., 2016. BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. Plant Biotechnol. J. 14, 1523–1531.

Tang, H., Zhang, X., Miao, C., Zhang, J., Ming, R., Schnable, J.C., Schnable, P.S., Lyons, E., Lu, J., 2015. ALLMAPS: robust scaffold ordering based on multiple maps. Genome Biol. 16, 3.

Wang, J., Luo, M.C., Chen, Z., You, F.M., Wei, Y., Zheng, Y., Dvorak, J., 2013. *Aegilops tauschii* single nucleotide polymorphisms shed light on the origins of wheat D-genome genetic diversity and pinpoint the geographic origin of hexaploid wheat. New Phytol. 198, 925–937.

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B.E., Maccaferri, M., Salvi, S., Milner, S.G., Cattivelli, L., Mastrangelo, A.M., Whan, A., Stephen, S., Barker, G., Wieseke, R., Plieske, J., International Wheat Genome Sequencing, C., Lillemo, M., Mather, D., Appels, R., Dolferus, R., Brown-Guedira, G., Korol, A., Akhunova, A.R., Feuillet, C., Salse, J., Morgante, M., Pozniak, C., Luo, M.C., Dvorak, J., Morell, M., Dubcovsky, J., Ganal, M., Tuberosa, R., Lawley, C., Mikoulitch, I., Cavanagh, C., Edwards, K.J., Hayden, M., Akhunov, E., 2014. Characterization of polyploid wheat genomic diversity using a high-density 90,000 single nucleotide polymorphism array. Plant Biotechnol. J. 12, 787–796.

Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H., Marler, B., Guo, H., Kissinger, J.C., Paterson, A.H., 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40, e49.

Warburton, M.L., Crossa, J., Franco, J., Kazi, M., Trethowan, R., Rajaram, S., Pfeiffer, W., Zhang, P., Dreisigacker, S., van Ginkel, M., 2006. Bringing wild relatives back into the family: recovering genetic diversity in CIMMYT improved wheat germplasm. Euphytica 149, 289–301.

Wu, T.D., Nacu, S., 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26, 873–881.

Zhang, C., Wang, J., Long, M., Fan, C., 2013a. gKaKs: the pipeline for genome-level Ka/Ks calculation. Bioinformatics 29, 645–646.

Zhang, H., Bian, Y., Gou, X., Dong, Y., Rustgi, S., Zhang, B., Xu, C., Li, N., Qi, B., Han, F., von Wettstein, D., Liu, B., 2013b. Intrinsic karyotype stability and gene copy number variations may have laid the foundation for tetraploid wheat formation. Proc. Natl. Acad. Sci. U. S. A. 110, 19466–19471.

Zhang, H., Dawe, R.K., 2012. Total centromere size and genome size are strongly correlated in ten grass species. Chromosome Res. 20, 403–412.

Zhang, J., Kudrna, D., Mu, T., Li, W., Copetti, D., Yu, Y., Goicoechea, J.L., Lei, Y., Wing, R.A., 2016. Genome puzzle master (GPM): an integrated pipeline for building and editing pseudomolecules from fragmented sequences. Bioinformatics 32, 3058–3064.

Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., Zhang, J., Saski, C.A., Scheffler, B.E., Stelly, D.M., Hulse-Kemp, A.M., Wan, Q., Liu, B., Liu, C., Wang, S., Pan, M., Wang, Y., Wang, D., Ye, W., Chang, L., Zhang, W., Song, Q., Kirkbride, R.C., Chen, X., Dennis, E., Llewellyn, D.J., Peterson, D.G., Thaxton, P., Jones, D.C., Wang, Q., Xu, X., Zhang, H., Wu, H., Zhou, L., Mei, G., Chen, S., Tian, Y., Xiang, D., Li, X., Ding, J., Zuo, Q., Tao, L., Liu, Y., Li, J., Lin, Y., Hui, Y., Cao, Z., Cai, C., Zhu, X., Jiang, Z., Zhou, B., Guo, W., Li, R., Chen, Z.J., 2015. Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. Nat. Biotechnol. 33, 531–537.

Zhao, N., Zhu, B., Li, M., Wang, L., Xu, L., Zhang, H., Zheng, S., Qi, B., Han, F., Liu, B., 2011. Extensive and heritable epigenetic remodeling and genetic stability accompany allohexaploidization of wheat. Genetics 188, 499–510.

Please cite this article in press as: Xie, J., et al., Sequencing and comparative analysis of *Aegilops tauschii* chromosome arm 3DS reveal rapid evolution of Triticeae genomes, Journal of Genetics and Genomics (2016), http://dx.doi.org/10.1016/j.jgg.2016.09.005