Plant Biotechnology Journal

Plant Biotechnology Journal (2011) **9**, pp. 768–775 doi: 10.1111/j.1467-7652.2010.00587.x

Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS

Paul J. Berkman[1], Adam Skarshewski[1], Michał T. Lorenc[1], Kaitao Lai[1], Chris Duran[1], Edmund Y.S. Ling[1], Jiri Stiller[1], Lars Smits[1], Michael Imelfort[1], Sahana Manoli[1], Megan McKenzie[1], Marie Kubaláková[2], Hana Šimková[2], Jacqueline Batley[1], Delphine Fleury[3], Jaroslav Doležel[2] and David Edwards[1]*

[1]*School of Land, Crop and Food Sciences and Australian Centre for Plant Functional Genomics, University of Queensland, Brisbane, QLD, Australia*
[2]*Centre of the Region Haná for Biotechnological and Agricultural Research, Institute of Experimental Botany, Olomouc, Czech Republic*
[3]*Australian Centre for Plant Functional Genomics, University of Adelaide, Glen Osmond, Australia*

Received 7 October 2010;
revised 24 November 2010;
accepted 25 November 2010.
**Correspondence* (fax +61 (0) 7 3365 1176;
email Dave.Edwards@uq.edu.au)

Keywords: wheat genome sequence, chromosome 7, genome assembly, gene annotation.

Summary

The genome of bread wheat (*Triticum aestivum*) is predicted to be greater than 16 Gbp in size and consist predominantly of repetitive elements, making the sequencing and assembly of this genome a major challenge. We have reduced genome sequence complexity by isolating chromosome arm 7DS and applied second-generation technology and appropriate algorithmic analysis to sequence and assemble low copy and genic regions of this chromosome arm. The assembly represents approximately 40% of the chromosome arm and all known 7DS genes. Comparison of the 7DS assembly with the sequenced genomes of rice (*Oryza sativa)* and *Brachypodium distachyon* identified large regions of conservation. The syntenic relationship between wheat, *B. distachyon* and *O. sativa*, along with available genetic mapping data, has been used to produce an annotated draft 7DS syntenic build, which is publicly available at http://www.wheatgenome.info. Our results suggest that the sequencing of isolated chromosome arms can provide valuable information of the gene content of wheat and is a step towards whole-genome sequencing and variation discovery in this important crop.

Introduction

Genome sequence information greatly assists our understanding of crop growth and development as the sequence content of the genes, or specifically, the allelic variants of the sequenced genes are responsible for almost all of the heritable differences between crop varieties. Initial analysis of the genetic variation in wheat applies genetic markers such as isozymes, restriction fragment length polymorphisms (RFLPs), simple sequence repeats and, more recently, single nucleotide polymorphisms (SNPs). The latter two markers are sequence based and require either specific sequencing for their identification or mining of existing sequence data. The availability of a genome sequence for wheat would greatly assist wheat improvement, through the discovery of molecular genetic markers, their placement on a physical map and an understanding of which genes underlie genetically mapped traits.

The size of the wheat genome (*Triticum aestivum*) is approximately 16 000 Mbp/1C, much larger than related cereal genomes such as barley (*Hordeum vulgare*, 5000 Mbp/1C), rye (*Secale cereale*, 9100 Mbp/1C) and oat (*Avena sativa*, 11 000 Mbp/1C) (Bennett and Leitch, 1995). Both the size and hexaploid nature of the wheat genome create significant problems in elucidating its DNA sequence. The traditional method applied for sequencing large eukaryotic genomes involves the production of an overlapping tiling path of large genomic fragments maintained within bacterial hosts in a bacterial artificial chromosome (BAC) vector. The BAC clones are then each shotgun-sequenced, resulting in many reads that are assembled to produce the sequence of the genomic fragment within the BAC. The whole genome may then be reassembled from these sequenced BACs based on overlaps. This approach led to the first complete plant genome sequences for Arabidopsis and rice (*Oryza sativa*) (Arabidopsis Genome I, 2000; Matsumoto *et al.*, 2005) and has been used to sequence the large genome of maize (Schnable *et al.*, 2009). As sequencing technology and bioinformatics techniques advance, an increasing number of genomes are being sequenced using whole-genome shotgun methods (Imelfort and Edwards, 2009; Imelfort *et al.*, 2009a).

Second-generation sequencing describes platforms that produce large amounts of short DNA sequence reads of length typically between 25 and 500 bp. Illumina produce one of the leading technologies for second-generation sequencing. These systems use reversible terminator chemistry, and the Illumina HiSeq 2000 can generate more than 200 thousand million bases of usable data per run. The short length of the sequence reads creates a challenge for genome assembly; however, the application of paired read technology, where sequence reads are produced in pairs with a known orientation and approximate distance between them, has greatly increased the use of this type of data for genome assembly (Imelfort and Edwards, 2009).

Illumina technology has now been applied for the sequencing of several crop species including cucumber (Huang *et al.*, 2009) and *Brassica* species (http://www.uq.edu.au/news/index.html?article=20010). Given the size and complexity of the hexaploid wheat genome, it has previously been unfeasible to attempt whole-genome shotgun sequencing. Using current technologies, it is unlikely that a complete wheat genome, including repeats, would be assembled by this method. However, by focussing on the non-repetitive regions of the genome and dissecting the genome into smaller parts by isolating single chromosome arms using flow cytometric sorting (Doležel *et al.*, 2004), an initial draft assembly may be produced. As sequencing technology continues to improve, with longer read lengths, greater accuracy and increased output, combined with advances in bioinformatics tools to assemble these data, complex cereal genome sequencing may become a realistic ambition.

We have tested the utility of this strategy in wheat by shotgun-sequencing isolated chromosome arm 7DS, which represents 2.25% of the wheat genome. Sequences are assembled into contigs which cover approximately 40% of the chromosome arm, predominantly the non-repetitive regions, and contain all known 7DS genes. This assembly provides a framework for SNP genetic marker discovery, the association of traits with underlying genes and a model for complete genome sequencing of this important crop.

Results

In total, eight lanes of Illumina GAIIx data were produced across six runs, generating 154 225 186 paired reads of 75, 76, or 100 bp and an insert size of 320 bp, totalling 13.05 Gbp. With a predicted molecular size for 7DS of 381 Mbp (Šafář *et al.*, 2010), the coverage was estimated to be 34x. All data have been submitted to the GenBank short-read archive (SRA025100.1).

The 7DS reads were compared with the genomes of *B. distachyon* and *O. sativa* to validate the sequence data and define the syntenic regions. Coverage varied across the chromosomes (Figure 1). Three regions with a high read density were identified on *B. distachyon* chromosomes one (between 39 and 51 Mbp) and three (between 8 and 20 Mbp and 39 and 44 Mbp). Similar analysis of the *O. sativa* genome identified regions on chromosome six (between 1 and 12 Mbp) and chromosome eight (between 23 Mbp and the telomere). 7DS read coverage co-located with genes, with reads predominantly matching gene exons (Figure 2). Additional regions of local high coverage were observed, associated with conserved repetitive elements and low-complexity sequences such as simple sequence repeats (data not shown).

Following data pre-processing, the 7DS sequences were assembled using Velvet (Zerbino and Birney, 2008). The assembly contained 571 038 contigs, with an N50 of 1159 bp and maximum contig length of 32 648 bp. The total assembly length was 153 653 984 bp, approximately 40% of the predicted size of this chromosome arm (Šafář *et al.*, 2010).

Comparison of the wheat contigs with 65 sequence-based *Aegilops tauschii* genetic markers which had been mapped to the short arm of chromosome 7 (Luo *et al.*, 2009) identified 60 correspondences. Five markers that had been genetically mapped to *Ae. tauschii* 7S did not share significant sequence identity with any 7DS contig. However, in contrast to the majority of mapped markers, these five were all mapped as RFLPs and may demonstrate a lower sequence identity than SNP-based marker assays.

In addition to the 65 *Ae. tauschii* genetic markers, we identified cDNAs in the GrainGenes database (Matthews *et al.*, 2003; Carollo *et al.*, 2005), which had been mapped by hybridization with DNA from deletion wheat lines, missing defined chromosomal regions (Hossain *et al.*, 2004). Comparison of these sequences with our 7DS assembly identified matching contigs for 315 out of the 354 7DS-mapped cDNAs (88.5%). To assess the accuracy of bin mapping, we compared all 354 cDNA sequences with the complete *B. distachyon* genome. Of the 315 cDNAs that were found in the 7DS assembly, 279 had a match in the *B. distachyon* genome, 199 (71%) of which matched the syntenic regions, defined as *B. distachyon* chromosome 1 between 35 and 55 Mbp, and *B. distachyon* chromosome 3 between 8 and 20 Mbp as well as between 39 and 45 Mbp. In contrast, none of the remaining 39 cDNAs matched a *B. distachyon* syntenic region, while 32 of these 39 matched non-syntenic regions of *B. distachyon*, suggesting that they may be erroneously bin-mapped to 7DS. This is consistent with the expected error rate for bin-mapping cDNAs in wheat (M. Sorrells personal communication). A large number of 7AS, 7BS and 4AL mapped genes also matched the 7DS assembly (Supplementary data S1), reflecting gene similarity between the 7S arms, and the translocation between 4AL and 7BS (Devos *et al.*, 1995).

The assembled contigs were compared with the *B. distachyon* genome using reciprocal best BLAST (RBB). The 7DS contigs matched a total of 1488 *B. distachyon* genes, 1022 (69%) of which were in the predicted syntenic regions (636 in Bd1 35–55 Mbp, 103 in Bd3 8–20 Mbp and 283 in Bd3 39–45 Mbp). An additional 466 7DS contigs had RBB hits

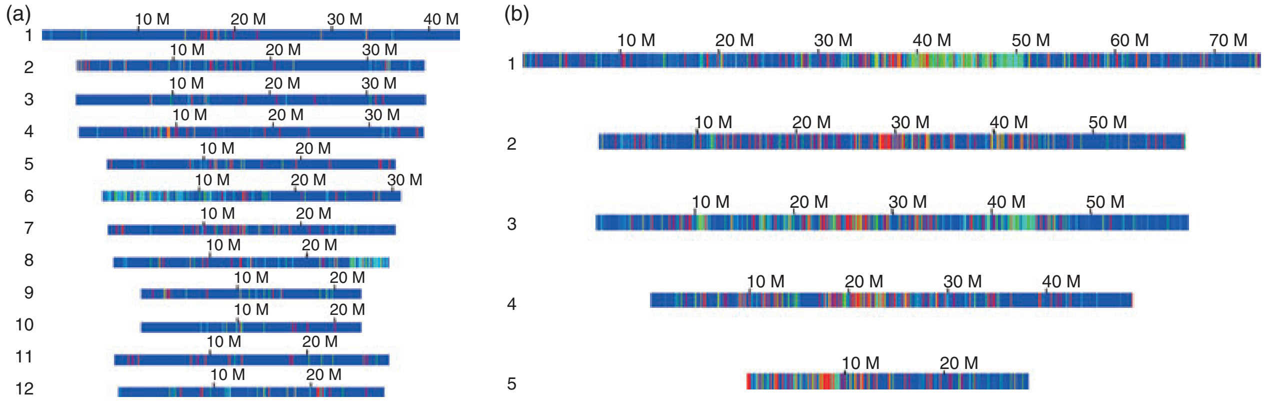

Figure 1 Wheat 7DS read pairs were mapped onto the genomes of *Oryza sativa* (a) and *Brachypodium distachyon* (b). The heat map depicts read density across each of the chromosomes, with a blue-red colour scale (blue = 0, red = 18). Regions on chromosomes one and three of *B. distachyon*, and six and eight of *O. sativa* showing the highest density of 7DS reads are the known syntenic regions for the 7DS wheat chromosome arm.

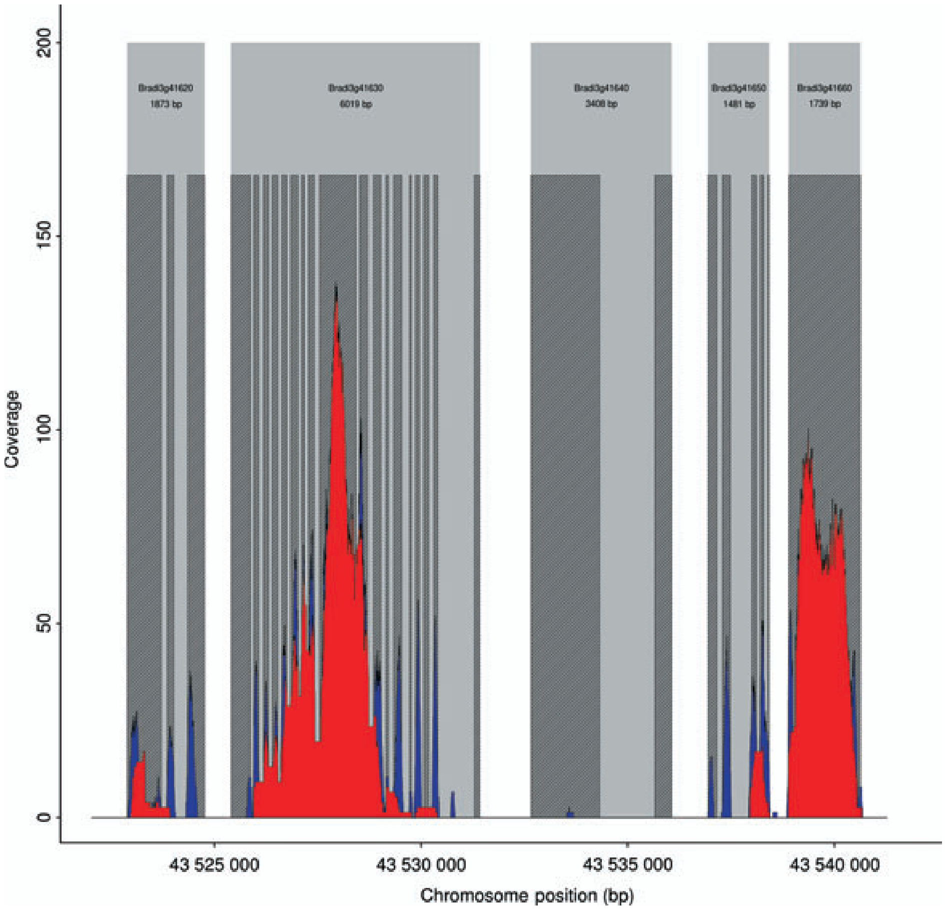

Figure 2 Wheat 7DS reads were mapped to a genomic region of *Brachypodium distachyon* containing five predicted genes (Bradi3g41620—Bradi3g41660). Predicted genes are in grey blocks, exons are hashed blocks. Paired (red) and single (blue) read density peaks map predominantly to conserved exons. Bradi3g41640 is not considered to have a homologue on wheat 7DS.

outside of the *B. distachyon* syntenic regions. The 7DS contigs were ordered and orientated with respect to the *B. distachyon* and *O. sativa* genomes to produce a syntenic build of 7 814 423 bp in length. Of the 60 contigs that matched an *Ae. tauschii* 7S genetic marker, 58 were located within the 7DS syntenic build, reflecting their predominantly genic origin and selection based on synteny with rice (Luo *et al.*, 2009). The majority of these markers were also collinear with the 7DS syntenic build.

The majority of 7DS contigs did not match any *B. distachyon* gene. Reviewing the annotation of these contigs suggests that they are predominantly made up of nested TE insertions (data not shown).

Discussion

Genome sequencing projects are gradually moving away from the traditional BAC by BAC approach towards the application of new, second-generation sequencing technologies and associated advanced bioinformatics tools. These approaches are now routinely used for relatively small bacterial and fungal genomes and have been implemented for larger genomes such as the giant panda, cucumber and *Brassica* species (Huang *et al.*, 2009; Li *et al.*, 2010; http://www.uq.edu.au/news/index.html?article=20010).

Sequencing the wheat genome remains a challenge owing to its repetitive nature; however, the approach of chromosome arm isolation overcomes many of the problems of genome complexity. Sequences that are repetitive across the whole genome are likely to have reduced representation or be represented uniquely on individual chromosome arms. This reduced repetitiveness decreases the complexity of assembling sequence data in these regions, regardless of the sequence read length used. Another key benefit of a chromosome arm sequencing approach is that, when applied to polyploid genomes such as wheat, homoeologous chromosomes are separated from the outset (Doležel *et al.*, 2007). The isolation of individual chromosome arms of wheat has previously been demonstrated (Kubaláková *et al.*, 2002; Janda *et al.*, 2006). Furthermore, an individual chromosome of barley has been isolated and sequenced to 1.3 × coverage by GSFLX pyrosequencing, which enabled the production of a virtual gene map of barley 1H by comparison of these sequence data with the genomes of rice and sorghum (Mayer *et al.*, 2009). Here, we demonstrate that it is possible to assemble chromosome arm-specific sequence data generated by Illumina short-read sequencing technology and produce a syntenic build representing the majority of genes from the chromosome arm.

Synteny comparison of 7DS paired sequence reads with *B. distachyon* and *O. sativa* showed sequence identity with chromosomes six and eight in *O. sativa*, as well as chromosomes one and three in *B. distachyon* (Figure 1), highlighting the purity of the isolated arms. This analysis also confirmed previous studies on the genomic relationship between these three species. These regions of chromosomes six and eight in rice were predicted to be syntenic with wheat 7DS based on EST deletion bin mapping (Sorrells *et al.*, 2003). The sequencing of *B. distachyon* in early 2010 similarly identified regions of chromosomes one and three as being syntenic with wheat 7DS (The International Brachypodium I, 2010). Our result confirmed that sequencing amplified wheat chromosome arm DNA can accurately target specific chromosomal regions.

Comparing 7DS reads with *B. distachyon* and *O. sativa* gene structures in their genomic context demonstrated conservation within annotated genes, but little conservation outside of genic regions (Figure 2), except for regions of conserved repetitive elements, and low complexity regions such as simple sequence repeats (data not shown). Mapping paired 7DS reads to the *B. distachyon* genome also permitted the precise delineation of the three syntenic regions to be between Bradi1g52510 and Bradi1g41880, Bradi3g10210 and Bradi3g20950, as well as Bradi3g36630 and Bradi3g43040.

One of the greatest challenges of complex genome assembly is the deconvolution of repetitive sequences. Our aim, however, was not to assemble the complete repetitive fraction of the genome but to focus on the genic and low copy regions which are likely to be responsible for the majority of phenotypic variation and provide a reference for the discovery of molecular genetic markers (Imelfort *et al.*, 2009b). The assembled sequence data amounted to a total of 154 Mbp or 40% of the predicted size of the chromosome arm. The wheat genome is predicted to be made up of 80% repetitive sequences (Bennetzen and Kellogg, 1997). Sequence assembly algorithms such as Velvet (Zerbino and Birney, 2008) perform poorly with repetitive regions, because repeats cause tangles in the de Bruijn graph which cannot be easily resolved. Low copy and unique regions have clear paths in the graph and assemble well. Thus, our assembly should be highly enriched for low copy regions and should include the majority, if not all, of the unique and low copy sequences as well as a substantial portion of the repetitive fraction.

Following assembly, we attempted to estimate the proportion of 7DS genes which were represented within contigs. A total of 88.5% of 7DS bin-mapped cDNAs could be identified within the assembled contigs. Several factors affect the assignment of a cDNA to a bin, and the error rate is estimated to be approximately 10% (M. Sorrells personal communication). The bin-mapped cDNAs which were absent from our assembly did not have homologues within the syntenic region of *B. distachyon*, further suggesting that they had been erroneously mapped to 7DS. In a second form of validation, we compared 65 sequence-based *Ae. tauschii* genetic markers with our assembly. Sixty of these sequences (92%) matched the 7DS assembly. All 60 of these markers were mapped as SNPs, which require significant sequence identity for marker assay design. In contrast, the five markers that were not found in our 7DS assembly were mapped by hybridization as RFLPs, which require substantially less sequence identity than SNP assays and may correspond to different but related genes. We conclude from the above analysis that we have assembled all, or nearly all, genes present on wheat 7DS.

Two approaches were applied for the assembly of the draft 7DS syntenic build. The first joined multiple contigs that represented single genes. This method has the potential risk of joining two related but different genes to produce a chimeric construct. In practice, the majority of split contigs were clearly identified as unique sequences, and the process of joining by comparison to syntenic genic sequence could be automated. Where the relationship between contigs and syntenic genes was unclear owing to the local gene duplication, manual joining was performed. The second approach applied the genome zipper method, which orders and orientates the gene-containing contigs based on synteny with related grass species (Mayer *et al.*, 2009). While it was accepted that the resulting syntenic build may not reflect the exact order of genes on 7DS owing to recent local chromosomal rearrangement, the approach provided an approximation of gene order suitable for further refinement by genetic and physical mapping.

A total of 443 7DS gene-containing contigs were non-syntenic to *B. distachyon*. These sequences are likely to represent genes that have moved since the divergence between wheat and *B. distachyon*. It is currently not possible to position these genes in the 7DS syntenic build; however, their location may be determined by subsequent physical or genetic mapping, or by improving the assembly process. A high proportion of non-syntenic genes have also been observed in the distal regions of chromosome 3B (Choulet *et al.*, 2010). Positional cloning of quantitative trait loci in wheat is frequently based on synteny with *O. sativa* or *B. distachyon*, which might fail if the gene controlling the trait is non-syntenic. Therefore, shotgun sequencing of chromosome arms provides unique data on non-syntenic genes for positional cloning.

The assembly of contigs and a draft syntenic build is of little value without suitable annotation. We compared the 7DS assembly with all predicted *B. distachyon* genes and the complete Uniref90 database and parsed the results into GFF3 format for display and searching using GBrowse2 (Arnaoudova *et al.*, 2009) (Figure 3). We believe that the resulting database is the most comprehensive catalogue of chromosome arm-specific

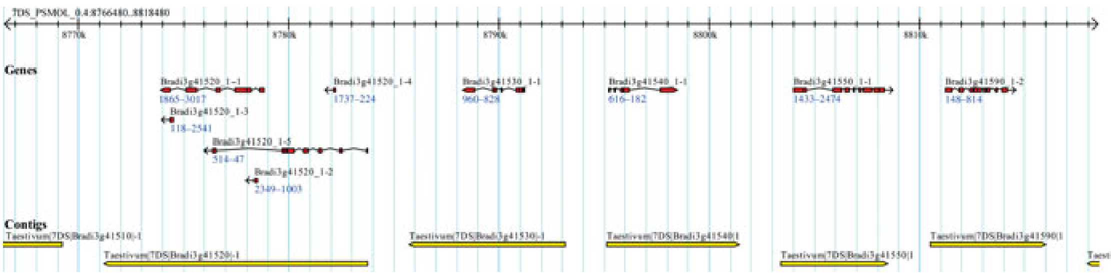

Figure 3 GBrowse2 view of the annotated wheat 7DS syntenic build, highlighting conserved *Brachypodium distachyon* and annotated Uniref90 matches.

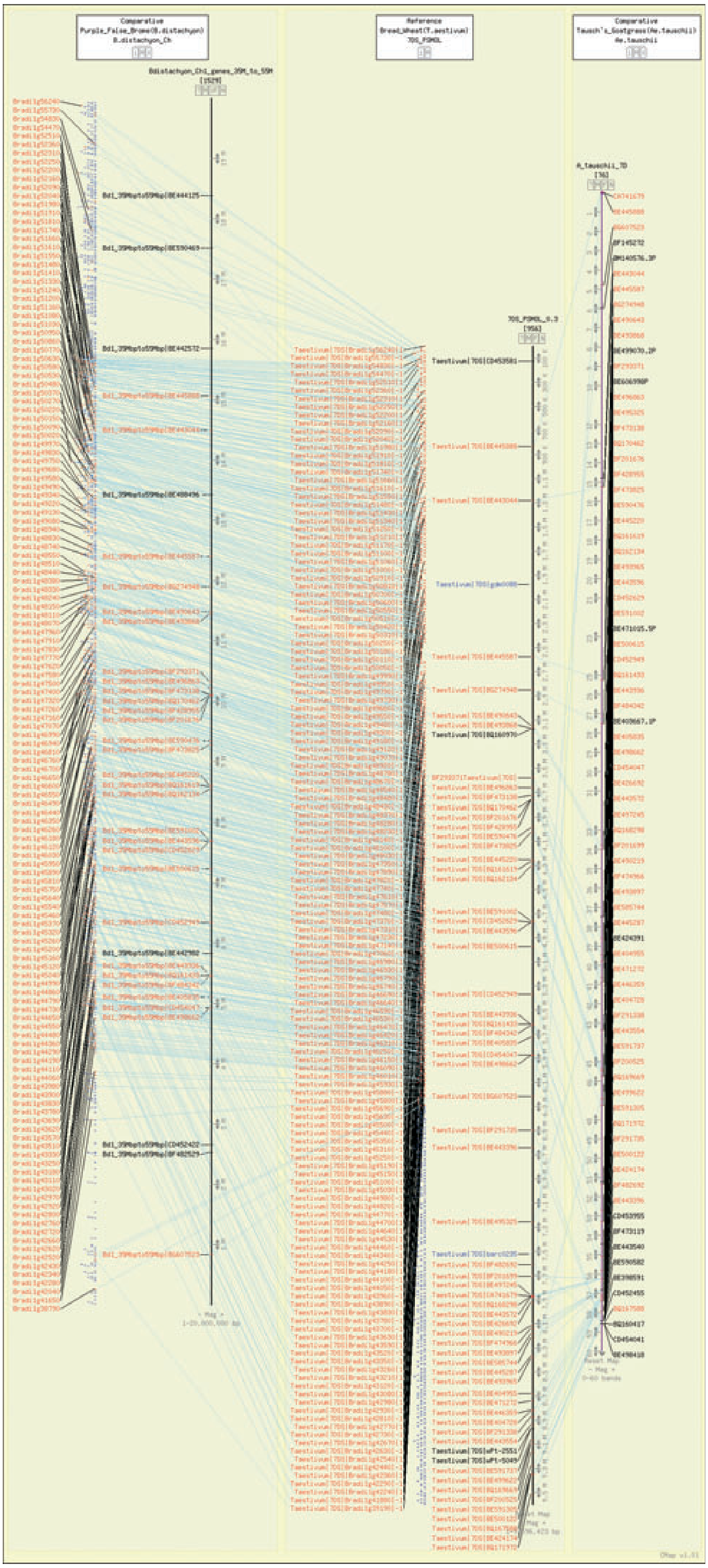

Figure 4 CMap view comparing part of the 7DS syntenic build with regions of *Brachypodium distachyon* and *Aegilops tauschii*.

wheat genes currently available and provides a public resource for gene and molecular genetic marker discovery.

Finally, to validate the general assembly structure of the syntenic build, the contigs were compared with sequence-based molecular markers that had been genetically mapped in *Ae. tauschii* (Luo *et al.*, 2009) (Figure 4). Two genetic markers matched contigs that could not be scaffolded into the syntenic build. As a result of the low resolution of genetic maps, it was not considered possible to accurately position these contigs.

By providing annotated wheat 7DS chromosome arm sequences as a syntenic build and additional contigs, we are allowing wheat researchers to enjoy the benefits of genome information much sooner than traditional sequencing methods would permit. This draft allows the identification of candidate genes underlying genetically mapped agronomic traits and may also be applied for high-density SNP discovery, either by aligning current EST sequences or by variety-specific whole-genome Illumina paired read data to the 7DS assembly, followed by the application of a redundancy-based SNP discovery assay (Edwards and Batley, 2009; Imelfort *et al.*, 2009b).

It may be possible to improve the current syntenic build by comparison with high-resolution maps of wheat or related cereals, such as barley. Improvements in sequence assembly methods or the addition of additional sequence data such as long mate pairs or BAC end sequences may also further refine this assembly.

There are several ongoing wheat sequencing projects, and it is likely that draft sequences for all chromosome arms will be produced by Illumina shotgun sequencing within the next 12 months, opening the way for genome-wide SNP discovery by variety re-sequencing and read mapping. These draft assemblies will also pave the way for complete genome sequencing through a combination of advanced sequencing and bioinformatics methods combined with current physical mapping approaches.

Experimental procedures

Data generation and validation

Seeds of double ditelosomic line 7D of *Triticum aestivum* cv. Chinese Spring were provided by Professor Bikram Gill (Kansas State University, Manhattan, KS, USA). The seeds were germinated, and root tips of young seedlings were used for the preparation of liquid suspensions of intact chromosomes as previously described (Vrána *et al.*, 2000). Chromosome arm 7DS was flow-sorted as telocentric chromosome in four batches of 39 000 chromosomes representing 30 ng DNA. To estimate contamination with other chromosomes, 1000 chromosomes were sorted onto a microscope slide in three replicates and used for fluorescence *in situ* hybridization with probes for *Afa* family and telomeric repeats. The average purity in sorted fractions was 84%. Chromosomal DNA was purified and subsequently amplified using Illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare, Chalfont St. Giles, UK) as previously described (Šimková *et al.*, 2008). A total of 200 ng of pooled, amplified DNA was used to prepare an Illumina paired-end library which was sequenced on the Illumina GAIIx platform using standard protocols.

O. sativa chromosome and predicted cDNA sequences were downloaded from the Phytozome Biomart (TIGR Release 6) (Ouyang *et al.*, 2007). *B. distachyon* chromosome sequences were downloaded from the Bd21 8× assembly databank hosted at Brachypodium.org (The International Brachypodium I, 2010).

A custom 'double-barrelled BLAST' script based on the TAGdb algorithm (Marshall *et al.*, 2010) was used to compare query sequences with the 7DS sequence data. Paired-end coverage was defined as the number of read-pair regions aligned at each nucleotide of the query sequence. This numeric coverage data were converted into a blue-red colour scale (blue = 0, red = 18) and plotted as heat maps. Local coverage plots were produced to compare read coverage with known *B. distachyon* gene structures. In addition to paired read coverage depth depicted as red peaks, coverage where only one read from a pair matched the *B. distachyon* sequence is displayed as blue peaks.

Syntenic build assembly and annotation

Several trimming, filtering and assembly parameters were assessed prior to the production of the final assembly. The 7DS sequence data were filtered and trimmed using an in-house script, trimConverter.py, to produce reads with a quality score of at least 15 at each nucleotide position and a minimum length of 40 bp. The resulting read set was filtered to remove any reads containing kmers of length 39 bp that occurred only once. If a read was discarded, its respective read pair was passed into a single-read file for inclusion in the assembly. The trimmed and filtered read set was assembled using Velvet version 1.0.09 (Zerbino and Birney, 2008) on a DELL R905 server with 128-GB RAM. The final assembly used a kmer size of 39 bp and an expected coverage of 17.6x, which represents the read depth after filtering.

A genome zipper approach (Mayer *et al.*, 2009) was applied to order and orientate the wheat contigs into a draft syntenic build. The assembled Velvet contigs were compared with predicted *B. distachyon* cDNA sequences using MEGABLAST (Zhang *et al.*, 2000). RBB hits between assembled 7DS contigs and cDNAs were identified, and 7DS contigs were assembled into a syntenic build based on the order and orientation of genes in *B. distachyon*. Additional non-overlapping 7DS contigs representing wheat genes which were split into two or more contigs were included in the syntenic build. The above process was repeated using *O. sativa* cDNAs and additional syntenic genes shuffled into the syntenic build.

Sequence-based mapped genetic markers from the wheat D genome donor *Ae. tauschii* (Luo *et al.*, 2009) were compared to the 7DS contigs using BLAST (Altschul *et al.*, 1990) and an E value cut-off of 10^{-5}. Corresponding matches were imported into a CMap database (Youens-Clark *et al.*, 2009). This database is publicly available at http://www.wheatgenome.info. An additional 354 cDNAs, which had been predicted to be located on 7DS by bin mapping, were downloaded from GrainGenes (Matthews *et al.*, 2003; Carollo *et al.*, 2005) and compared to the assembled 7DS contigs using BLAST and an E value cut-off of 10^{-5}.

The contigs were annotated to predict the location and potential function of expressed genes by comparison with the Uniref90 database (Suzek *et al.*, 2007) and predicted *B. distachyon* genes, using BLAST and an E value cut-off of 10^{-5}. The resulting high-scoring pairs were converted to GFF3 format and parsed to a public GBrowse2 database hosted at http://www.wheatgenome.info.

Acknowledgements

We thank Dr. Jarmila Číhalíková, Romana Nováková, Bc. and Ms. Zdeňka Dubská for their assistance with chromosome sorting. The authors acknowledge funding support from the Australian Research Council (Projects LP0882095, LP0883462 and DP0985953), the Czech Science Foundation (grant No. 521/07/1573) and the European Union (grant No. ED0007/01/01 Centre of the Region Haná for Biotechnological and Agricultural Research). Support from the Australian Genome Research Facility (AGRF), the Queensland Cyber Infrastructure Foundation (QCIF), the Australian Partnership for Advanced Computing (APAC) and Queensland Facility for Advanced Bioinformatics (QFAB) is gratefully acknowledged.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Arabidopsis Genome I (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature (London)*, **408**, 796–815.

Arnaoudova, E.G., Bowens, P.J., Chui, R.G., Dinkins, R.D., Hesse, U., Jaromczyk, J.W., Martin, M., Maynard, P., Moore, N. and Schardl, C.L. (2009) Visualizing and sharing results in bioinformatics projects: GBrowse and GenBank exports. *BMC Bioinformatics*, **10** (Supp. 7), 1471–2105.

Bennett, M.D. and Leitch, I.J. (1995) Nuclear-DNA amounts in angiosperms. *Ann. Bot.*, **76**, 113–176.

Bennetzen, J.L. and Kellogg, E.A. (1997) Do plants have a one-way ticket to genomic obesity? *Plant Cell*, **9**, 1509–1514.

Carollo, V., Matthews, D.E., Lazo, G.R., Blake, T.K., Hummel, D.D., Lui, N., Hane, D.L. and Anderson, O.D. (2005) GrainGenes 2.0. An improved resource for the small-grains community. *Plant Physiol.*, **139**, 643–651.

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.X., Kong, X.Y., Jia, J.Z., Gut, M., Brunel, D., Anderson, J.A., Gill, B.S., Appels, R., Keller, B. and 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.

Devos, K.M., Dubcovsky, J., Dvorak, J., Chinoy, C.N. and Gale, M.D. (1995) Sructural evolution of wheat chromosomes 4A, 5A, AND 7B and its impact on recombination. *Theor. Appl. Genet.*, **91**, 282–288.

Doležel, J., Kubaláková, M., Bartoš, J. and Macas, J. (2004) Flow cytogenetics and plant genome mapping. *Chromosome Res.*, **12**, 77–91.

Doležel, J., Kubaláková, M., Paux, E., Bartoš, J. and Feuillet, C. (2007) Chromosome-based genomics in the cereals. *Chromosome Res.*, **15**, 51–66.

Edwards, D. and Batley, J. (2009) Plant genome sequencing: applications for crop improvement. *Plant Biotechnol. J.*, **7**, 1–8.

Hossain, K.G., Kalavacharla, V., Lazo, G.R., Hegstad, J., Wentz, M.J., Kianian, P.M.A., Simons, K., Gehlhar, S., Rust, J.L., Syamala, R.R., Obeori, K., Bhamidimarri, S., Karunadharma, P., Chao, S., Anderson, O.D., Qi, L.L., Echalier, B., Gill, B.S., Linkiewicz, A.M., Ratnasiri, A., Dubcovsky, J., Akhunov, E.D., Dvorak, J., Miftahudin, Ross, K., Gustafson, J.P., Radhawa, H.S., Dilbirligi, M., Gill, K.S., Peng, J.H., Lapitan, N.L.V., Greene, R.A. and Bermudez-Kandianis, C.E., Sorrells, M.E., Feril, O., Pathan, M.S., Nguyen, H.T., Gonzalez- Hernandez, J.L., Conley, E.J., Anderson, J.A., Choi, D.W., Fenton, D., Close, T.J., McGuire, P.E., Qualset, C.O. and Kianian, S.F. (2004) A chromosome bin map of 2148 expressed sequence tag loci of wheat homoeologous group 7. *Genetics*, **168**, 687–699.

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., Lucas, W.J., Wang, X., Xie, B., Ni, P., Ren, Y., Zhu, H., Li, J., Lin, K., Jin, W., Fei, Z., Li, G., Staub, J., Kilian, A., van der Vossen, E.A.G., Wu, Y., Guo, J., He, J., Jia, Z., Ren, Y., Tian, G., Lu, Y., Ruan, J., Qian, W., Wang, M., Huang, Q., Li, B., Xuan, Z., Cao, J., Asan, Wu, Z., Zhang, J., Cai, Q., Bai, Y., Zhao, B., Han, Y., Li, Y., Li, X., Wang, S., Shi, Q., Liu, S., Cho, W.K., Kim, J.-Y., Xu, Y., Heller-Uszynska, K., Miao, H., Cheng, Z., Zhang, S., Wu, J., Yang, Y., Kang, H., Li, M., Liang, H., Ren, X., Shi, Z., Wen, M., Jian, M., Yang, H., Zhang, G., Yang, Z., Chen, R., Liu, S., Li, J., Ma, L., Liu, H., Zhou, Y., Zhao, J., Fang, X., Li, G., Fang, L., Li, Y., Liu, D., Zheng, H., Zhang, Y., Qin, N., Li, Z., Yang, G., Yang, S., Bolund, L., Kristiansen, K., Zheng, H., Li, S., Zhang, X., Yang, H., Wang, J., Sun, R., Zhang, B., Jiang, S., Wang, J., Du, Y. and Li, S. (2009) The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.*, **41**, 1275–1281.

Imelfort, M. and Edwards, D. (2009) De novo sequencing of plant genomes using second-generation technologies. *Brief. Bioinform.*, **10**, 609–618.

Imelfort, M., Batley, J., Grimmond, S. and Edwards, D. (2009a) Genome sequencing approaches and successes. In *Plant Genomics* (Somers, D., Langridge, P. and Gustafson, J., eds), pp. 345–358, USA: Humana Press.

Imelfort, M., Duran, C., Batley, J. and Edwards, D. (2009b) Discovering genetic polymorphisms in next-generation sequencing data. *Plant Biotechnol. J.*, **7**, 312–317.

Janda, J., Safar, J., Kubalakova, M., Bartos, J., Kovarova, P., Suchankova, P., Pateyron, S., Cihalikova, J., Sourdille, P., Simkova, H., Faivre-Rampant, P., Hribova, E., Bernard, M., Lukaszewski, A., Dolezel, J. and Chalhoub, B. (2006) Advanced resources for plant genomics: a BAC library specific for the short arm of wheat chromosome 1B. *Plant J.*, **47**, 977–986.

Kubaláková, M., Vrána, J., Číhalíková, J., Šimková, H. and Doležel, J. (2002) Flow karyotyping and chromosome sorting in bread wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.*, **104**, 1362–1372.

Li, R.Q., Fan, W., Tian, G., Zhu, H.M., He, L., Cai, J., Huang, Q.F., Cai, Q.L., Li, B., Bai, Y.Q., Zhang, Z.H., Zhang, Y.P., Wang, W., Li, J., Wei, F.W., Li, H., Jian, M., Li, J.W., Zhang, Z.L., Nielsen, R., Li, D.W., Gu, W.J., Yang, Z.T., Xuan, Z.L., Ryder, O.A., Leung, F.C.C., Zhou, Y., Cao, J.J., Sun, X., Fu, Y.G., Fang, X.D., Guo, X.S., Wang, B., Hou, R., Shen, F.J., Mu, B., Ni, P.X., Lin, R.M., Qian, W.B.Wang, G.D., Yu, C., Nie, W.H., Wang, J.H., Wu, Z.G., Liang, H.Q., Min, J.M., Wu, Q., Cheng, S.F., Ruan, J., Wang, M.W., Shi, Z.B., Wen, M., Liu, B.H., Ren, X.L., Zheng, H.S., Dong, D., Cook, K., Shan, G., Zhang, H., Kosiol, C., Xie, X.Y., Lu, Z.H., Zheng, H.C., Li, Y.R., Steiner, C.C., Lam, T.T.Y., Lin, S.Y., Zhang, Q.H., Li, G.Q., Tian, J., Gong, T.M., Liu, H.D., Zhang, D.J., Fang, L., Ye, C., Zhang, J.B., Hu, W.B., Xu, A.L., Ren, Y.Y., Zhang, G.J., Bruford, M.W., Li, Q.B., Ma, L.J., Guo, Y.R., An, N., Hu, Y.J., Zheng, Y., Shi, Y.Y., Li, Z.Q., Liu, Q., Chen, Y.L., Zhao, J., Qu, N., Zhao, S.C., Tian, F., Wang, X.L., Wang, H.Y., Xu, L.Z., Liu, X., Vinar, T., Wang, Y.J., Lam, T.W., Yiu, S.M., Liu, S.P., Zhang, H.M., Li, D.S., Huang, Y., Wang, X., Yang, G.H., Jiang, Z., Wang, J.Y., Qin, N., Li, L., Li, J.X., Bolund, L., Kristiansen, K., Wong, G.K.S., Olson, M., Zhang, X.Q., Li, S.G., Yang, H.M. and Wang, J. (2010) The sequence and de novo assembly of the giant panda genome (vol 463, pg 311, 2010). *Nature*). **463**, 1106–1106.

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. and 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. USA*, **106**, 15780–15785.

Marshall, D., Hayward, A., Eales, D., Imelfort, M., Stiller, J., Berkman, P., Clark, T., McKenzie, M., Lai, K., Duran, C., Batley, J. and Edwards, D. (2010) Targeted identification of genomic regions using TAGdb. *Plant Methods*, **6**, 19.

Matsumoto, T., Wu, J.Z., Kanamori, H., Katayose, Y., Fujisawa, M., Namiki, N., Mizuno, H., Yamamoto, K., Antonio, B.A., Baba, T., Sakata, K., Nagamura, Y., Aoki, H., Arikawa, K., Arita, K., Bito, T., Chiden, Y., Fujitsuka, N., Fukunaka, R., Hamada, M., Harada, C., Hayashi, A., Hijishita, S., Honda, M., Hosokawa, S., Ichikawa, Y., Idonuma, A., Iijima, M., Ikeda, M., Ikeno, M., Ito, K., Ito, S., Ito, T., Ito, Y., Iwabuchi, A., Kamiya, K., Karasawa, W., Kurita, K., Katagiri, S., Kikuta, A., Kobayashi, H., Kobayashi, N., Machita, K., Maehara, T., Masukawa, M., Mizubayashi, T., Mukai, Y., Nagasaki, H., Nagata, Y., Naito, S., Nakashima, M., Nakama, Y., Nakamichi, Y., Nakamura, M., Meguro, A., Negishi, M., Ohta, I., Ohta, T., Okamoto, M., Ono, N., Saji, S., Sakaguchi, M., Sakai, K., Shibata, M., Shimokawa, T., Song, J.Y., Takazaki, Y., Terasawa, K., Tsugane, M., Tsuji, K., Ueda, S., Waki, K., Yamagata, H., Yamamoto, M., Yamamoto, S., Yamane, H., Yoshiki, S., Yoshihara, R., Yukawa, K., Zhong, H.S., Yano, M., Sasaki, T., Yuan, Q.P., Shu, O.T., Liu, J., Jones, K.M., Gansberger, K., Moffat, K., Hill, J., Bera, J., Fadrosh, D., Jin, S.H., Johri, S., Kim, M., Overton, L., Reardon, M., Tsitrin, T., Vuong, H., Weaver, B., Ciecko, A., Tallon, L., Jackson, J., Pai, G., Van Aken, S., Utterback, T., Reidmuller, S., Feldblyum, T., Hsiao, J., Zismann, V., Iobst, S., de Vazeille, A.R., Buell, C.R., Ying, K., Li, Y., Lu, T.T., Huang, Y.C., Zhao, Q., Feng, Q., Zhang, L., Zhu, J.J., Weng, Q.J., Mu, J., Lu, Y.Q., Fan, D.L., Liu, Y.L., Guan, J.P., Zhang, Y.J., Yu, S.L., Liu, X.H., Zhang, Y., Hong, G.F., Han, B., Choisne, N., Demange, N., Orjeda, G., Samain, S., Cattolico, L., Pelletier, E., Couloux, A., Segurens, B., Wincker, P., D'Hont, A., Scarpelli, C., Weissenbach, J., Salanoubat, M., Quetier, F., Yu, Y., Kim, H.R., Rambo, T., Currie, J., Collura, K., Luo, M.Z., Yang, T.J., Ammiraju, J.S.S., Engler, F., Soderlund, C., Wing, R.A., Palmer, L.E., de la Bastide, M., Spiegel, L., Nascimento, L., Zutavern, T., O'Shaughnessy, A., Dike, S., Dedhia, N., Preston, R.Balija, V., McCombie, W.R., Chow, T.Y., Chen, H.H., Chung, M.C., Chen, C.S., Shaw, J.F., Wu, H.P., Hsiao, K.J., Chao, Y.T., Chu, M.K., Cheng, C.H., Hour, A.L., Lee, P.F., Lin, S.J., Lin, Y.C., Liou, J.Y., Liu, S.M., Hsing, Y.I., Raghuvanshi, S., Mohanty, A., Bharti, A.K., Gaur, A., Gupta, V., Kumar, D., Ravi, V., Vij, S., Kapur, A., Khurana, P., Khurana, J.P., Tyagi, A.K., Gaikwad, K., Singh, A., Dalal, V., Srivastava, S., Dixit, A., Pal, A.K., Ghazi, I.A., Yadav, M., Pandit, A., Bhargava, A., Sureshbabu, K., Batra, K., Sharma, T.R., Mohapatra, T., Singh, N.K., Messing, J., Nelson, A.B., Fuks, G., Kavchok, S., Keizer, G., Llaca, E.L.V., Song, R.T., Tanyolac, B., Young, S., Il, K.H., Hahn, J.H., Sangsakoo, G., Vanavichit, A., de Mattos, L.A.T., Zimmer, P.D., Malone, G., Dellagostin, O., de Oliveira, A.C., Bevan, M., Bancroft, I., Minx, P., Cordum, H., Wilson, R., Cheng, Z.K., Jin, W.W., Jiang, J.M., Leong, S.A., Iwama, H., Gojobori, T., Itoh, T., Niimura, Y., Fujii, Y., Habara, T., Sakai, H., Sato, Y., Wilson, G., Kumar, K., McCouch, S., Juretic, N., Hoen, D., Wright, S., Bruskiewich, R., Bureau, T., Miyao, A., Hirochika, H., Nishikawa, T., Kadowaki, K., Sugiura, M. and Int Rice Genome Sequencing P (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.

Matthews, D.E., Carollo, V.L., Lazo, G.R. and Anderson, O.D. (2003) GrainGenes, the genome database for small-grain crops. *Nucleic Acids Res.*, **31**, 183–186.

Mayer, K.F.X., Taudien, S., Martis, M., Simkova, H., Suchankova, P., Gundlach, H., Wicker, T., Petzold, A., Felder, M., Steuernagel, B., Scholz, U., Graner, A., Platzer, M., Dolezel, J. and Stein, N. (2009) Gene content and virtual gene order of barley chromosome 1H. *Plant Physiol.*, **151**, 496–505.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R.L., Lee, Y., Zheng, L., Orvis, J., Haas, B., Wortman, J. and Buell, C.R. (2007) The TIGR Rice Genome Annotation Resource: Improvements and new features. *Nucleic Acids Res.*, **35**, D883–D887.

Šafář, J., Šimková, H., Kubalákoá, M., Číhalíková, J., Suchánková, P., Bartoš, J. and Doležel, J. (2010) Development of chromosome-specific BAC resources for genomics of bread wheat. *Cytogenet. Genome. Res.*, **129**, 211–223.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F.S., Pasternak, S., Liang, C.Z., Zhang, J.W., Fulton, L., Graves, T.A., Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S.M., Belter, E., Du, F.Y., Kim, K., Abbott, R.M., Cotton, M. and Levy, A. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science (Washington)*, **326**, 1112–1115.

Šimková, H., Svensson, J.T., Condamine, P., Hřiboá, E., Suchánková, P., Bhat, P.R., Bartoš, J., Šafář, J., Close, T.J. and Doležel, J. (2008) Coupling amplified DNA from flow-sorted chromosomes to high-density SNP mapping in barley. *BMC Genomics*, **9**, 294.

Sorrells, M.E., La Rota, M., Bermudez- Kandianis, C.E., Greene, R.A., Kantety, R., Munkvold, J.D., Miftahudin, Mahmoud, A., Ma, X.F., Gustafson, P.J., Qi, L.L.L., Echalier, B., Gill, B.S., Matthews, D.E., Lazo, G.R., Chao, S.M., Anderson, O.D., Edwards, H., Linkiewicz, A.M., Dubcovsky, J., Akhunov, E.D., Dvorak, J., Zhang, D.S., Nguyen, H.T., Peng, J.H., Lapitan, N.L.V., Gonzalez- Hernandez, J.L., Anderson, J.A., Hossain, K., Kalavacharla, V., Kianian, S.F., Choi, D.W., Close, T.J., Dilbirligi, M., Gill, K.S., Steber, C., Walker- Simmons, M.K., McGuire, P.E. and Qualset, C.O. (2003) Comparative DNA sequence analysis of wheat and rice genomes. *Genome Res.*, **13**, 1818–1827.

Suzek, B.E., Huang, H.Z., McGarvey, P., Mazumder, R. and Wu, C.H. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, **23**, 1282–1288.

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

Vrána, J., Kubaláková, M., Šimková, H., Číhalíková, J., Lysák, M.A. and Doležel, J. (2000) Flow sorting of mitotic chromosomes in common wheat (Triticum aestivum L.). *Genetics*, **156**, 2033–2041.

Youens-Clark, K., Faga, B., Yap, I.V., Stein, L. and Ware, D. (2009) CMap 1.01: a comparative mapping application for the Internet. *Bioinformatics*, **25**, 3040–3042.

Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.

Zhang, Z., Schwartz, S., Wagner, L. and Miller, W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.

Supporting information

Additional Supporting information may be found in the online version of this article:

Data S1 Comparison of all bin-mapped cDNAs with the assembly of the 7DS data revealed primary alignment to the group 7 chromosomes. Interestingly, 187 4AL bin-mapped ESTs also had a match. 4AL has undergone a translocation with 7BS and is therefore partially syntenic to the 7DS chromosome arm (Devos et al., 1995). The low frequency of cDNAs mapped to other chromosomes and matching our 7DS assembly may be interpreted in several ways, such as noise in the analysis caused by related gene families or erroneous bin mapping. Due to the coverage required to assemble contigs using velvet, we would not expect to see assembly of contaminating non-7DS DNA sequence, however this possibility cannot be excluded.