

Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation

Paul J. Berkman · Adam Skarshewski · Sahana Manoli · Michał T. Lorenc · Jiri Stiller · Lars Smits · Kaitao Lai · Emma Campbell · Marie Kubaláková · Hana Šimková · Jacqueline Batley · Jaroslav Doležel · Pilar Hernandez · David Edwards

Received: 19 April 2011 / Accepted: 27 September 2011 / Published online: 15 October 2011
© Springer-Verlag 2011

Abstract Complex Triticeae genomes pose a challenge to genome sequencing efforts due to their size and repetitive nature. Genome sequencing can reveal details of conservation and rearrangements between related genomes. We have applied Illumina second generation sequencing technology to sequence and assemble the low copy and unique regions of *Triticum aestivum* chromosome arm 7BS, followed by the construction of a syntenic build based on gene order in *Brachypodium*. We have delimited the position of a previously reported translocation between 7BS and 4AL with a resolution of one or a few genes and report approximately 13% genes from 7BS having been translocated to 4AL. An additional 13 genes are found on 7BS which appear to have originated from 4AL. The gene content of the 7DS and 7BS syntenic builds indicate a total

of ~77,000 genes in wheat. Within wheat syntenic regions, 7BS and 7DS share 740 genes and a common gene conservation rate of ~39% of the genes from the corresponding regions in *Brachypodium*, as well as a common rate of colinearity with *Brachypodium* of ~60%. Comparison of wheat homoeologues revealed ~84% of genes previously identified in 7DS have a homoeologue on 7BS or 4AL. The conservation rates we have identified among wheat homoeologues and with *Brachypodium* provide a benchmark of homoeologous gene conservation levels for future comparative genomic analysis. The syntenic build of 7BS is publicly available at <http://www.wheatgenome.info>.

Introduction

Wheat is probably one of the most important crops in the world, yet it has one of the most complex genomes, confounding many genomic applications for wheat crop improvement. Bread wheat is hexaploid, being derived from a combination of three diploid donor grass species each with seven pairs of chromosomes, resulting in a total of 21 pairs of chromosomes. The donor species are proposed to have diverged from an ancestral diploid species between 2.5 and 6 MYA (Huang et al. 2002; Chantret et al. 2005), and subsequently underwent two interspecies hybridisations which were each followed by a chromosome doubling event that produced a series of allopolyploid genomes. The first polyploidisation event between 0.5 and 3 MYA combined the genomes of *Triticum urartu* (A^uA^u) and an unidentified species (BB), which bears high similarity to *Aegilops speltoides*, to produce the allotetraploid genome of wild emmer wheat, *Triticum turgidum* (A^uA^uBB) (Chantret et al. 2005; Eckardt 2001; Huang et al. 2002). A second polyploidisation event occurred following

Communicated by T. Close.

P. J. Berkman · A. Skarshewski · S. Manoli · M. T. Lorenc · J. Stiller · L. Smits · K. Lai · D. Edwards (✉)
School of Agriculture and Food Sciences and Australian Centre for Plant Functional Genomics, University of Queensland, Brisbane, QLD 4072, Australia
e-mail: dave.edwards@uq.edu.au

E. Campbell · J. Batley
Centre for Integrative Legume Research,
University of Queensland, School of Agriculture and Food Sciences, Brisbane, QLD 4072, Australia

M. Kubaláková · H. Šimková · J. Doležel
Centre of the Region Haná for Biotechnological and Agricultural Research, Institute of Experimental Botany, Sokolovská 6,
77200 Olomouc, Czech Republic

P. Hernandez
Instituto de Agricultura Sostenible, Consejo Superior de Investigaciones Científicas (IAS, CSIC), Alameda del Obispo s/n, 14080 Cordoba, Spain

domestication, with the hybridisation of *T. turgidum* (A^uA^u BB) and *Ae. tauschii* (DD), to produce the allohexaploid genome of *T. aestivum* (A^uA^u BBDD). Each of these diploid progenitor genomes is between 4,000 and 6,000 million base pairs, almost twice the size of the human genome (Šafář et al. 2010). The bread wheat genome consists predominantly of repetitive elements which make up between 80 and 90% of the genome (Flavell et al. 1977; Wanjugi et al. 2009). The size, abundance of repeats and hexaploid nature of the wheat genome make it one of the most challenging crop genome sequences to assemble.

Reference genome sequences are valuable tools, allowing researchers to relate the heritable variation in agronomic traits with underlying variation in the genome (Edwards and Batley 2010; Duran et al. 2010; Buell and Last 2010). The sequence content of the genes or specifically the allelic variants of the sequenced genes are responsible for many of the heritable differences between crop varieties. An increasing number of crop genomes are becoming available, however, elucidating the larger genomes of some cereal crops such as wheat and barley has been hampered by their size and complexity. Recently, second generation sequencing (2GS) methods have been applied to characterise isolated fractions of these genomes, leading to a greater understanding of gene content and genome structure (Mayer et al. 2009; Berkman et al. 2011).

2GS platforms produce large amounts of short DNA sequence reads of length typically between 25 and 500 bp. Illumina produce some of the leading technologies for second generation sequencing. These systems use reversible terminator chemistry, and their current flagship system, the HiSeq 2000, can generate more than 200 thousand million bases of usable data per run. Illumina sequence reads are relatively short (100–150 bp), but can be produced as read pairs where the ends of a DNA fragment of known length are sequenced which provides additional information about the context of the sequence reads. The use of paired read information has greatly improved the applicability of short read sequence data for genome assembly, as reviewed by Imelfort and Edwards (2009).

While it is unlikely that a large eukaryote genome, including repeats, could be completely assembled using current second generation sequencing technology alone, the sequencing and assembly of the gene-rich non-repetitive regions is becoming relatively routine. In addition to focussing on assembling unique and low copy regions, it is possible to dissect the wheat genome into isolated chromosome arms using flow sorting (Doležel et al. 2004). The power of this approach, when applied to the complex hexaploid wheat genome, is that it provides a means to differentiate between homoeologous sequences, with chromosome arms separated prior to DNA sequencing (Šafář et al. 2010). We have previously tested the utility of

this strategy in wheat by assembling shotgun sequence data from isolated chromosome arm 7DS (Berkman et al. 2011). Sequences assembled into contigs representing the unique and low copy regions, including all known 7DS genes, the majority of which could be placed within a sequence scaffold build based on synteny with a close relative (syntenic build).

A number of translocations are known to have occurred in the bread wheat genome and its donor species. The best characterised of these occurred as a series of translocation and inversion events between chromosomes 4A, 5A and 7B (Naranjo et al. 1987). Consequently, chromosome arm 5AL contains a region which originated from 4AL, chromosome arm 4AL contains regions from 5AL and 7BS, and chromosome arm 7BS also contains a small region from 5AL (Devos et al. 1995). A translocation has been also identified by genetic mapping between chromosomes 2B and 6B (Conley et al. 2004; Devos et al. 1993) and while other minor rearrangements may have occurred on other chromosomes, it is understood that syntenic blocks in wheat are largely intact (Akhunov et al. 2003).

Identifying genome rearrangement events allows a detailed analysis of the syntenic relationships between crop species and assists our understanding of the timeline of wheat genome evolution associated with human cultivation over the last few 1,000 years. Understanding genome rearrangement in this important crop is of particular value in that it can assist our understanding of the genomic basis for phenotypic differences between closely related species and varieties.

We have applied the techniques developed for the assembly of 7DS to assemble Illumina paired read sequence data for isolated chromosome arm 7BS. Comparison of the 7DS and 7BS assemblies reveals the previously reported 7BS–4AL translocation to have relocated the region of 7BS between a *T. aestivum* orthologue of the Bradi1g49550 gene and the 7BS telomere to 4AL. The comparison of gene sequences from wheat homoeologous chromosome arms provides a basis for distinguishing between gene homoeologues, which will in turn help our understanding of genome evolution following polyploidisation and differential expression of gene homoeologues.

Materials and methods

Data generation and validation

Seeds of double ditelosomic lines 7B and 4A of *Triticum aestivum* cv. Chinese Spring were provided by Professor Bikram Gill (Kansas State University, Manhattan, 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 arms 7BS were flow-sorted as telocentric chromosomes in two batches of 28,000 chromosomes representing 20 ng DNA whereas 4AL arms were sorted in one batch of 50,000 chromosomes corresponding to 54 ng DNA. In order to estimate contamination with other chromosomes, 1,000 chromosomes were sorted onto a microscope slide in three replicates and used for fluorescence in situ hybridization (FISH) with probes for *Afa* family and telomeric repeats. The average purity in sorted fractions was 93.4 and 89% for 7BS and 4AL, respectively. Chromosomal DNA was purified and subsequently amplified using Illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare, Chalfont St. Giles, United Kingdom) as previously described (Šimková et al. 2008). Two and three independent amplifications were performed for 7BS and 4AL, respectively. A total of 200 ng of pooled, amplified DNA from 7BS was used to prepare an Illumina paired-end library which was sequenced on the Illumina GAIIx and HiSeq platforms using standard protocols. This single 7BS library, which had an expected insert size of ~350 bp, was sequenced across four lanes in three separate runs. Amplified DNA from 4AL was used to prepare a Roche shotgun library, which was sequenced on the GSFLX platform using the Titanium chemistry (three full sequencing runs) at the Lifesequencing S.L. facilities (Valencia, Spain). *B. distachyon* chromosome sequences were downloaded from the Bd21 8× assembly databank hosted at Brachypodium.org (Vogel et al. 2010).

A custom ‘double-barrelled BLAST’ script based on the TAGdb algorithm (Marshall et al. 2010) was used to compare query sequences with the 7BS 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 was converted into a blue–red colour scale (blue = 0, red = 18) and plotted as heat maps.

Syntenic build and assembly annotation

Several trimming, filtering and assembly parameters were assessed prior to the production of the final assembly. The 7BS sequence data was 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 30 bp. The resulting read-set was filtered to remove any reads containing k-mers of length 35 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 using used a k-mer of size 33 bp

and an expected coverage of 21.0, which represents the read depth after filtering. A comparative genomics approach was applied to order and orientate the wheat contigs into a draft syntenic build by identifying reciprocal best blast (RBB) hits of assembled contigs and *Brachypodium* genes using MEGABLAST (Zhang et al. 2000) with default e-value cut-off, as previously described (Mayer et al. 2009; Berkman et al. 2011).

The assembled contigs from chromosome arm 7BS were divided into two groups, the first containing contigs corresponding to genes that were identified to be present on 7BS via the syntenic build process and the second group containing contigs that did not. All ESTs from the organism *T. aestivum* available in the NCBI database as of 4 August 2011, were downloaded. The two groups of 7BS contigs were compared with all ESTs from the dataset from NCBI using TBLASTX (Altschul et al. 1990; Karlin and Altschul 1993) with an e-value cut-off of $1e^{-10}$, and the number of contigs from each group with a hit against the NCBI EST dataset was calculated. ESTs corresponding to 18,785 loci bin-mapped to specific chromosomal regions of *T. aestivum*, 345 of which had been predicted to be located on 7BS, were downloaded from GrainGenes (Carollo et al. 2005; Matthews et al. 2003; Qi et al. 2004). The ESTs for the 345 loci bin-mapped to 7BS were compared to the assembled 7BS contigs using BLAST and an e-value cut-off of $1e^{-10}$. All bin-mapped ESTs were compared separately to both the assembled 7BS and 7DS (Berkman et al. 2011) contigs using BLAST and an e-value cut-off of $1e^{-50}$.

454 data generated from chromosome arm 4AL was assembled with gsAssembler using default parameters (minimum overlap between reads 40 nt, minimum overlap identity 90%, alignment score identity +2 and alignment difference score -3.) (Margulies et al. 2005). All assembled contigs were compared to the predicted genes from *B. distachyon* using MEGABLAST (Zhang et al. 2000). RBB hits were identified from the results and RBB hits of *B. distachyon* predicted genes against 7BS, as well as 4AL and 7DS (Berkman et al. 2011), were plotted with respect to their position on *B. distachyon* chromosome 1 using in-house scripts. Code can be made available for academic purposes upon request.

Results

Data generation and validation

Two lanes of paired-end data were generated on the Illumina GAIIx with a read length of 100 bp, as well as an additional two lanes of paired-end data generated on the Illumina HiSeq with a read length of 35 bp. 173,915,402

reads were generated in total, representing 10.8 Gbp of sequence data. The mean insert size of the library was identified to be ~ 360 bp with a standard deviation of ~ 30 bp. Based on the estimated 360 Mbp size of this chromosome arm (Šafář et al. 2010), total coverage of 7BS was calculated to be $30.0\times$. All data has been submitted to the NCBI short read archive, reference SRA028115.1.

The short reads from 7BS were compared with the genome of *B. distachyon* to validate the sequence data and define the syntenic regions. Regions identified as syntenic to 7BS or 7DS are displayed as heat maps in Fig. 1. The 7BS regions are consistent with those identified as syntenic to 7DS (Berkman et al. 2011), with the exception of a short region of *B. distachyon* chromosome 1 which was absent in 7BS.

Data assembly and bin-mapped cDNA comparison

Following data pre-processing, the 7BS sequences were assembled using Velvet (Zerbino and Birney 2008) with a k-mer size of 33 and estimated coverage of $21.0\times$ representing the coverage level after filtering. The assembly contained 1,038,681 contigs, with an N50 of 472 bp and maximum contig length of 29,196 bp. The total assembly length was 176,154,889 bp, approximately 49% of the predicted size of this chromosome arm (Šafář et al. 2010).

Of the 2,807 contigs which were predicted to contain genes, 2,161 (76.99%) matched wheat ESTs, with an average of 45.5 ESTs per contig. A total of 94,214 (9.1%) of the contigs which were predicted not to contain genes matched wheat ESTs, with an average of 1.00 EST per contig.

We obtained 18,785 loci from the GrainGenes database (Carollo et al. 2005; Matthews et al. 2003), which had been mapped by hybridisation of ESTs with DNA from wheat deletion lines missing defined chromosomal regions (Hossain et al. 2004; Qi et al. 2004). Of the 345 loci which had been mapped to 7BS, 307 (89.0%) were identified within the 7BS assembly, 227 of which also had a match in the *B. distachyon* genome, with 148 of these matching within the syntenic regions. Of the 38 7BS bin-mapped loci without a match in the 7BS assembly, none matched the *B. distachyon* syntenic regions. These results are consistent with our previous results for 7DS and the expected error rate for bin mapping ESTs in wheat (Berkman et al. 2011) (M. Sorrells, personal communication) and suggest that the assembly represents all or nearly all the genes on 7BS.

Analysis of all 18,785 bin-mapped wheat loci (Fig. 2) showed that loci which had been bin-mapped to 7AS, 7BS or 7DS, matched contigs in both 7BS and 7DS assemblies. In addition, a greater number of 7BS bin-mapped loci had a match to a 7BS contig than 7DS mapped cDNAs. Almost twice as many 4AL mapped loci matched 7DS contigs than matched 7BS contigs. These results are consistent with a major translocation of genes from 7BS to 4AL.

Producing a syntenic build

Comparing the 7BS contigs with *B. distachyon* genes in the syntenic region, 780 genes were found to have a reciprocal best BLAST (RBB) hit, with an additional 187 genes found to be assembled into contigs hosting a neighbouring gene with a RBB hit, resulting in a predicted 967 wheat 7BS

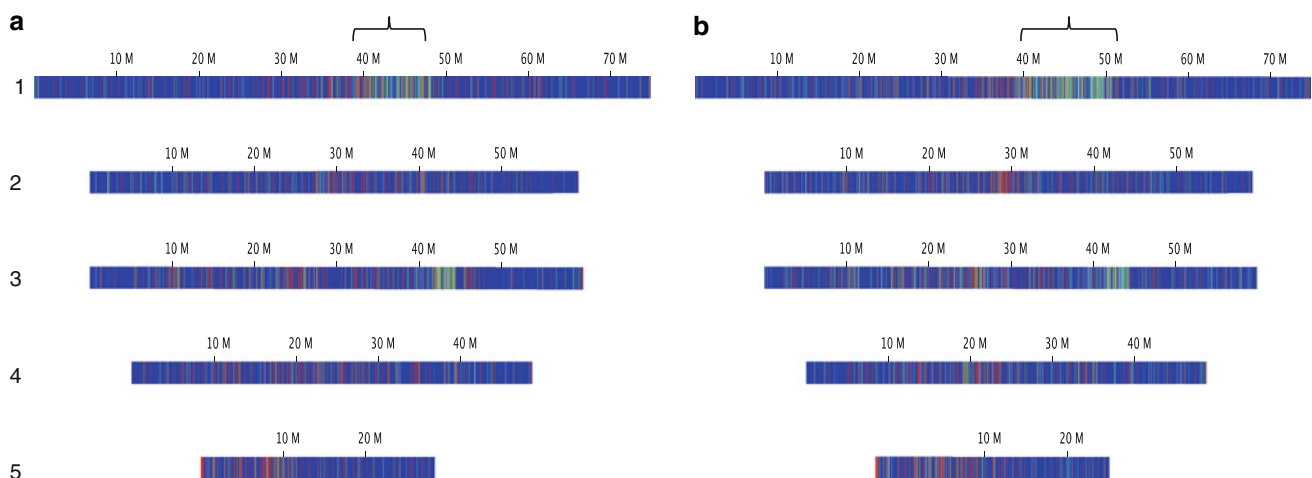


Fig. 1 Heatmaps of 7BS and 7DS read pair coverage against *B. distachyon* genome; Wheat 7BS read pairs (**a**) and 7DS read pairs (**b**) were mapped onto the genomes of *B. distachyon*. The heatmap depicts read density across each of the chromosomes, with a blue–red colour scale (blue = 0, red = 18). Regions on chromosomes one and three of *B. distachyon* showing the highest density of both 7BS and

7DS reads are the known syntenic regions for these wheat chromosome arms. Notably, the green syntenic region identified by 7BS read pairs mapped to *B. distachyon* chromosome 1 does not extend beyond the 50 Mbp measure, while 7DS read pairs mapped to this same region do extend beyond the 50 Mbp measure (highlighted by brackets)

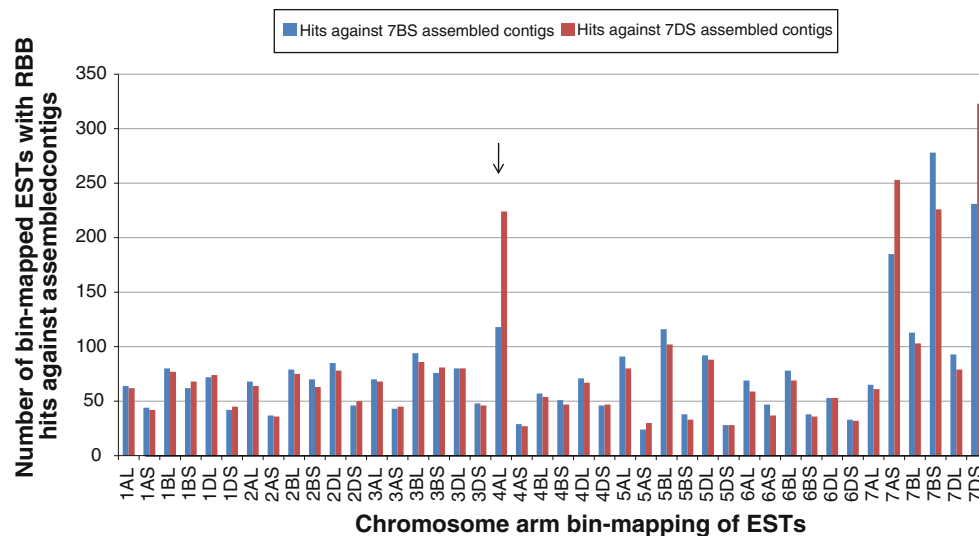


Fig. 2 Bin-mapped loci RBB hits with wheat 7BS and 7DS assembled contigs; The histogram depicts the number of RBB hits of loci which had previously been bin-mapped to individual

chromosome arms, with 7BS assembled contigs (*blue*) or 7DS assembled contigs (*red*). The *arrow* highlights a significant difference in 4AL bin-mapped loci found in these assemblies

genes which could be placed within a syntenic build. In total, 2,471 genes are predicted to be on the *Brachypodium* genome in these syntenic regions, of which only 967 (39.1%) have been retained on 7BS. The orthologue containing 7BS contigs were ordered and orientated with respect to the *B. distachyon* genome, as described previously (Berkman et al. 2011), to produce a syntenic build of 6.5 Mbp. The majority of 7BS contigs did not match any *B. distachyon* genes. Reviewing the annotation of these contigs suggests that they are predominantly made up of nested transposable element insertions (data not shown).

Following the generation of the 7BS syntenic build, alignment to the 7DS syntenic build (Berkman et al. 2011) revealed 69.03% of the 1,072 genes from 7DS were conserved on 7BS, with 227 genes (23.5%) included in the syntenic build of 7BS not found in the 7DS assembly (Fig. 3). We calculate the number of genes predicted on 7BS and 7DS (Berkman et al. 2011) to be 967 and 1,072, respectively, within the syntenic builds and 665 and 663, respectively, outside the syntenic builds. By dividing these gene counts by the total size of chromosome arms 7BS (360 Mbp) and 7DS (381 Mbp) (Šafář et al. 2010) and then multiplying by the overall size of the wheat genome (17 Gbp), we estimate that there are around 77,000 genes in the whole wheat genome, with between 45,000 and 50,000 of these genes likely to be found in blocks syntenic to regions of the *Brachypodium* genome.

Delimiting the translocation between 4AL and 7BS

The similarity between the syntenic build of 7BS and 7DS dropped significantly between the genes orthologous to

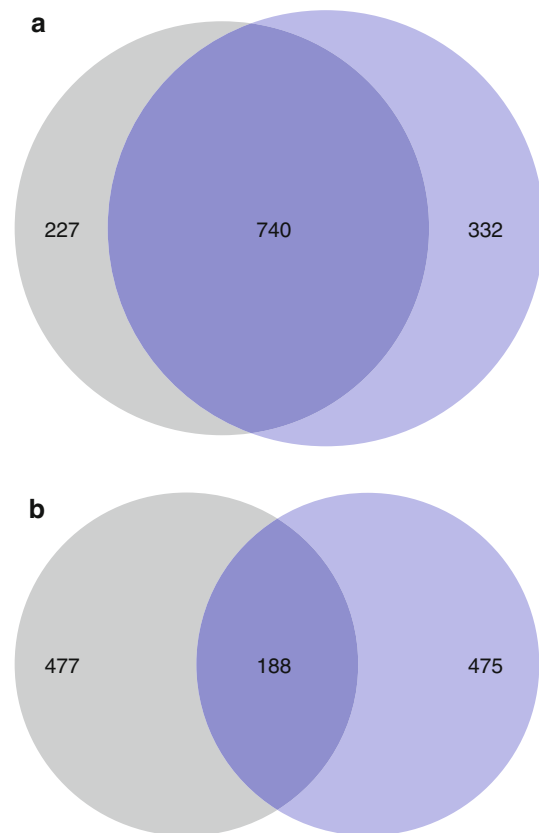


Fig. 3 Venn diagram displaying the common genes between 7BS (*grey*) and 7DS (*blue*) within the syntenic build (**a**) and genes outside of the syntenic build (**b**)

Bradi1g49510 and Bradi1g52510. Therefore two distinct regions can be described on 7DS, with the region from the centromere to the *T. aestivum* orthologue of the

Bradi1g49550 displaying a high degree of similarity with 7BS, and the region from this orthologue to the telomere displaying very little similarity with 7BS. This 7DS region which is missing from 7BS likely represents the translocation from 7BS to 4AL. In the region of high similarity between 7DS and 7BS, 740 7DS genes (84.6%) were found to be conserved in 7BS.

Recently, Hernandez et al. (2011) obtained a total of 2,987,532 reads of Roche 454 sequence data for chromosome arm 4AL, representing 900,594,638 bp and approximately 1.7× coverage (NCBI short read archive reference SRA034928.1). We assembled this data using Newbler (Margulies et al. 2005) with a minimum read length of 20, overlap seed length of 16, overlap seed step of 12 and overlap minimum match identity of 90%. This assembly obtained an N50 of 451 bp, a longest contig of 7,350 bp and a total assembly size of 70,326,673 bp, approximately 13% of the predicted the size of the chromosome arm. Comparison of the 7DS region corresponding to the predicted translocated region of 7BS revealed 166 7DS gene scaffolds (83.7%) having reciprocal best blast hits with the 4AL assembly. Comparison of the 4AL assembly with the genome of *B. distachyon* also revealed similarity between *B. distachyon* genes Bradi1g49470 and Bradi1g52330, consistent with the predicted 7BS translocation (Fig. 4).

In addition to the region identified to have translocated from 7BS to 4AL, a small region syntenic to the telomeric region of *B. distachyon* 1S, containing 13 genes and likely to have originated from 5AL, appears to have transferred from 4AL to 7BS (Fig. 4). Given the predicted position of the 7BS–4AL translocation, these genes were able to be positioned at the telomeric region of the 7BS syntenic build, increasing the number of genes in the 7BS syntenic build to 980 with final size of 6,508,016 bp.

Discussion

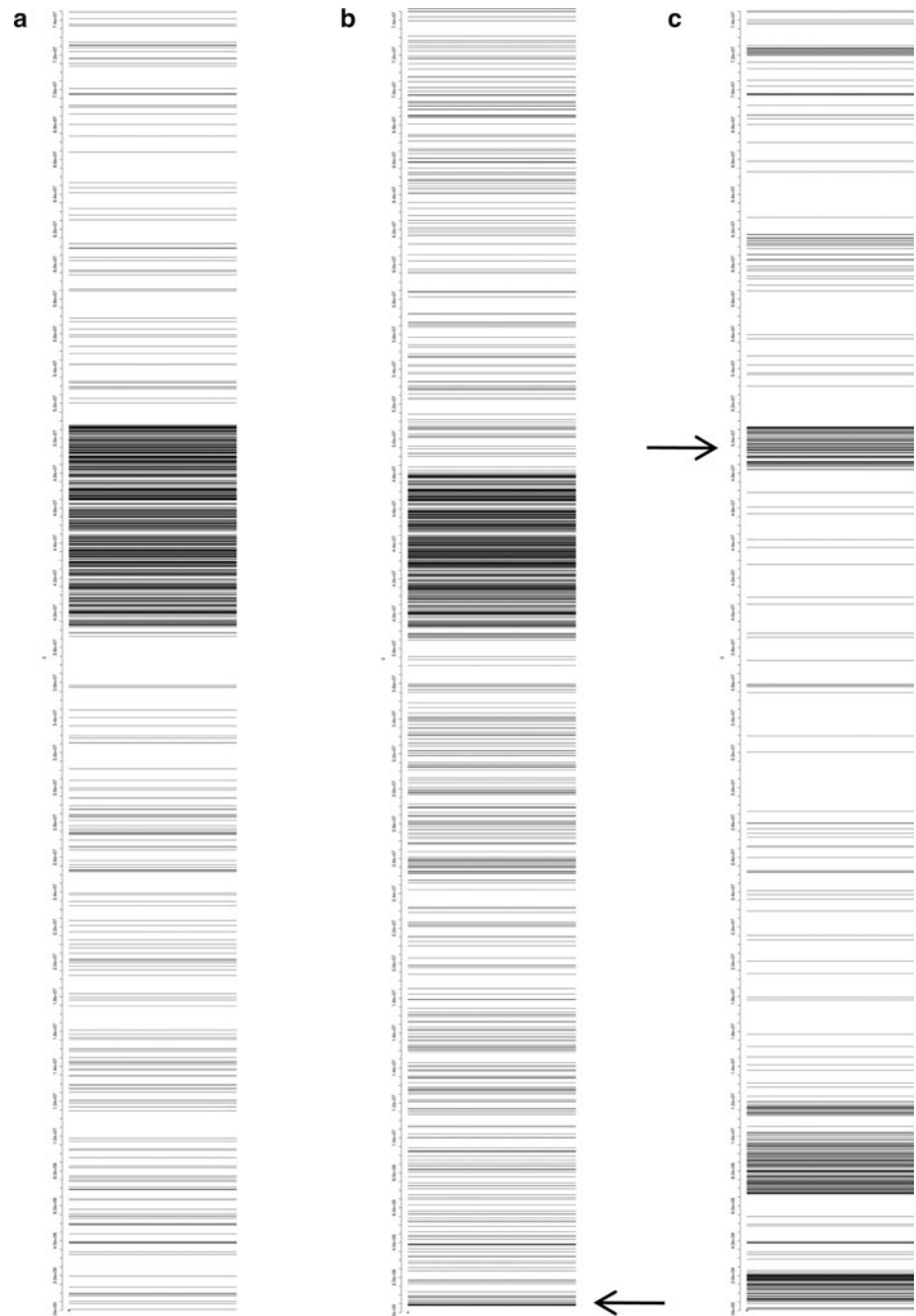
Based on the Lander–Waterman coverage probability (Lander and Waterman 1988) the probability that a nucleotide is missing is 9×10^{-14} , which suggests that every nucleotide of the chromosome arm has been included in the sequence data. The observation that the total length of the assembled sequence corresponds to only 49% of the chromosome arm can be explained by the collapsing of repetitive regions into single assemblies, while unique regions can be reasonably expected to assembly into unique, longer contigs. This preferential assembly of unique/genic regions provides the basis for the syntenic build methodology (Berkman et al. 2011). This conclusion is supported by the identification of all known 7DS and 7BS genes within the respective assemblies (Berkman et al. 2011).

Wheat 7DS contains a large number of genes which are not found on 7BS, however, the majority of these genes can be found in the 4AL assembly. Alignment of the 7BS and 7DS syntenic builds clearly places the position of the translocation between Taestivum|7BS|Bradi1g49500 and Taestivum|7BS|Bradi1g49550, with the four intervening genes missing from both 7DS and 7BS assemblies. Given the absence of these genes from 7DS as well as 7BS, it is likely that they are missing due to gene movement/loss events rather than the translocation. Sequencing of the remaining chromosome arms of wheat will identify if they are present elsewhere in the genome.

This study clearly demonstrates the value of 2GS technology in understanding the structure of complex genomes. While the translocation between 7BS and 4AL was originally identified using cytological and marker-based techniques, we have been able to pinpoint the position of this translocation at a far higher resolution than has previously been possible. By generating a syntenic build for individual chromosome arms in wheat, a sequence-based template can be produced which not only provides a detailed view of these genomic regions to assist in understanding past genomic rearrangement events, but also provides a sequence-based foundation for further analysis of genetically mapped traits. While the sequence of syntenic builds does not give the complete picture of the wheat genome, the assembly of genic and low copy regions from 2GS data provides sufficient high-resolution detail for comparative genomics, evolutionary studies, gene and genetic marker discovery, assisting researchers in applied crop improvement programs. There are likely to be several rearrangements which cannot be resolved at the resolution of a syntenic build, and these will be identified following further advances in wheat genome sequencing.

Previous studies have demonstrated the loss of gene colinearity between genomes of the grass family with increasing evolutionary distance (Wicker et al. 2010). The assembled syntenic builds enable a comparison of the level of this “gene movement” between the translocated and non-translocated regions. In this case, the level of gene conservation between 7BS/7DS and 4AL/7DS appears to be consistent (84.6% compared with 83.7%). In contrast, the level of gene conservation between the wheat chromosome arms and syntenic regions of *Brachypodium* is far lower than is found between the homoeologues (39.1%). This result is consistent with the analysis of wheat 7DS, where 1,072 genes (38.6%) were conserved in the *Brachypodium* region syntenic to 7DS (Berkman et al. 2011). While the level of gene conservation from *Brachypodium* to wheat is ~39% in both samples, the number of genes placed into the syntenic build compared with the number of genes in total identified in each chromosome arm (Fig. 3) indicate colinearity with *Brachypodium* of 59.3% for 7BS

Fig. 4 Comparison of *B. distachyon* chromosome 1 with **a** 7DS, **b** 7BS, and **c** 4AL assemblies; Reciprocal best blast hits of genes between the wheat arms and *B. distachyon* chromosome 1 are presented as *black lines*. The plots in this figure are based on chromosomal position from the short arm telomere on the y-axis and therefore display the short arm at the bottom and long arm at the top (i.e. reversed orientation from standard chromosomal display). *Arrows* indicate the translocated regions on 4AL and 7BS



(967/1632 genes) and 61.8% for 7DS (1072/1735 genes). The observation that gene colinearity between wheat homoeologues (~84%) is greater than gene colinearity between wheat and *Brachypodium* (~60%) reflects the divergence of these grasses 35–40 MYA, compared with only 2.5–6 million years since the divergence within the Triticeae (Huang et al. 2002). While levels of colinearity between wheat and *Brachypodium* appear lower than the reported 69% conservation of colinearity between rice, sorghum, and *Brachypodium* (Wicker et al. 2010), our rate of colinearity between wheat and *Brachypodium* is consistent with results recently reported by Massa et al. (2011).

This could be explained by a high rate of transposable element activity in hexaploid wheat accelerating the erosion of colinearity, which is consistent with the hypothesis of TE-driven gene movement (Vogel et al. 2010; Wicker et al. 2010) and has recently been further supported by evidence of gene erosion being lineage-dependant with an accelerated erosion of colinearity in wheat (Massa et al. 2011). While it has been suggested that wheat genes are not affected by major structural rearrangements (Choulet et al. 2010; Massa et al. 2011), the full impact of polyploidy on colinearity is not yet known. Further analysis of the gene content in wheat may provide additional insight into this.

Based on the number of genes we have identified to be present and conserved on the 7BS and 7DS chromosome arms (Berkman et al. 2011) in both syntenic and non-syntenic regions of *B. distachyon*, the estimated size of these regions (Šafář et al. 2010) and the size of the entire wheat genome, we estimate that there are approximately 77,000 genes in the wheat genome. Based on the number of genes included in the 7BS and 7DS syntenic builds, we estimate that between 45,000 and 50,000 wheat genes remain in syntenic blocks relative to *B. distachyon*. This prediction of overall gene content in wheat is substantially lower than previous estimates, with recent estimates suggesting that there may be between 100,000 and 350,000 genes in wheat (Paux et al. 2006; Rabinowicz et al. 2005; Devos et al. 2008). One previous estimate of 108,000 genes in hexaploid bread wheat was based on an 11 Mbp sample of BAC-end sequence data from chromosome 3B (Paux et al. 2006). Another estimate of 295,000 genes was based on sequence analysis of less than 1 Mbp of sequence data (Rabinowicz et al. 2005), though the authors suggest that the majority of the predicted genes are likely to represent pseudogenes. Choulet et al. (2010) provide several estimates of gene content using different methods. Using BAC sequences, they predict a weighted total of 50,000 genes per diploid genome, and around 40,000 genes on the B genome using low coverage Illumina sequence data. They conclude with an estimate of around 36,000–50,000 genes in the B genome of wheat, though gene annotation was by comparison with ESTs and cDNA sequences which may lead to a larger number compared with our more conservative estimate. All of these methods extrapolate gene presence based on relatively small samples. Extrapolating results from small samples which may not reflect the overall genome structure may bias the results, a risk that is substantially reduced by increasing the sample size (Devos et al. 2008). By sequencing 360 Mbp of 7BS and 381 Mbp of 7DS, we base our estimate on a significantly larger sample (Šafář et al. 2010; Berkman et al. 2011). In a recent paper by Massa et al. (2011) the authors suggest a total of around 36,000 genes in *Ae. tauschii*, the D genome donor of hexaploid wheat. This is higher than our very conservative estimate of gene content in hexaploid wheat. In comparing our assembled contigs with wheat ESTs, we identified many additional contigs which may be predicted to contain genes. While some of these contigs are likely to contain expressed genes, they may also represent pseudogenes or transposon related expressed genes. De Bruijn graph based assemblers such as velvet are often confounded by repetitive regions and so produce longer contigs for low copy and unique regions compared to repeat regions (Pop 2009). This is reflected in our results by an average size of 3,521 bp for predicted gene containing contigs compared to an overall average of 161 bp. Genes

which contain repetitive DNA sequence may be assembled as fragments using our approach and contribute to an overall reduction in estimated gene number. Our predicted gene count in wheat may also be an under-estimate due to the possible exclusion of novel genes, though there is a limited likelihood that the inclusion of novel genes would substantially increase the total gene count in wheat.

Our estimate of gene number does not take into account the emergence of unique, wheat-specific genes that may be annotated using alternate methods; however, it has also been noted that early estimates of gene content can inflate the actual number of genes due to mis-annotations (Benetzen et al. 2004). The method of applying syntenic builds to estimate gene content in wheat may overcome the issue of mis-annotation by basing estimates on previously identified genes from related species, and our estimate is likely to suggest conservative minimum gene content in the genome of *T. aestivum*.

Conclusion

The application of 2GS data to chromosome arm 7BS of wheat has yielded results consistent with those we have previously described for 7DS, indicating that we have assembled all or nearly all genes on this chromosome arm, providing a strong basis for comparison of the homoeologues (Berkman et al. 2011).

Comparison of assemblies of two homoeologous arms 7DS and 7BS, together with a third assembly of 4AL using Roche 454 sequence data, has enabled the delimitation of the translocation between 7BS and 4AL. The previous identification of this translocation was based on genetic mapping and therefore provided limited resolution of the translocation position (Devos et al. 1995). Our method has allowed us to delimit the position of the translocation to the gene level. This high-resolution depiction of genomic rearrangements in *T. aestivum* provides the foundation required to undertake finer genomic analysis in wheat, particularly in deconvoluting relationships between homoeologous chromosome arms, by identifying the presence or absence of genes from specific chromosomal locations. In turn, this provides a detailed gene-rich reference enabling wheat crop improvement researchers to more effectively conduct their research.

We have provided an accurate measure of gene colinearity between homoeologues of 84% and between the homoeologues and *B. distachyon* of ~60%. Our estimate of wheat's gene content overcomes some of the sequencing bias, small sample size, and annotation errors, inherent in earlier estimates of gene content in the wheat genome (Paux et al. 2006; Rabinowicz et al. 2005). By applying a similar approach to the remaining chromosome arms of

wheat, a much more accurate estimate could be provided on the gene content of wheat.

Acknowledgments The authors would like to acknowledge funding support from the Australian Research Council (Projects LP0882095, LP0883462 and DP0985953), the Czech Republic Ministry of Education, Youth and Sports (grant no. LC06004), the European Regional Development Fund (Operational Programme Research and Development for Innovations No. CZ.1.05/2.1.00/01.0007) and the Spanish Ministry of Science and Innovation (MICINN grants BIO2009-07443 and AGL2010-17316). We thank Dr. Jarmila Čihalíková, Romana Nováková, Bc. and Ms. Zdeňka Dubská for their assistance with chromosome sorting. 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

- Akhunov ED, Goodyear AW, Geng S, Qi LL, Echalié B, Gill BS, Miftahudin, Gustafson JP, Lazo G, Chao S, Anderson OD, Linkiewicz AM, Dubcovsky J, La Rota M, Sorrells ME, Zhang D, Nguyen HT, Kalavacharla V, Hossain K, Kianian SF, Peng J, Lapitan NL, Gonzalez-Hernandez JL, Anderson JA, Choi DW, Close TJ, Dilbirligi M, Gill KS, Walker-Simmons MK, Steber C, McGuire PE, Qualset CO, Dvorak J (2003) The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. *Genome Res* 13(5):753–763. doi:10.1101/gr.808603GR-8086R
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410. doi:10.1006/jmbi.1990.9999S0022283680799990
- Bennetzen JL, Coleman C, Liu R, Ma J, Ramakrishna W (2004) Consistent over-estimation of gene number in complex plant genomes. *Curr Opin Plant Biol* 7(6):732–736. doi:10.1016/j.pbi.2004.09.003
- Berkman PJ, Skarshewski A, Lorenc M, Lai K, Duran C, Ling EYS, Stiller J, Smits L, Imelfort M, Manoli S, McKenzie M, Kubaláková M, Šimková H, Batley J, Fleury D, Doležel J, Edwards D (2011) Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. *Plant Biotechnol J* 9(7):768–775. doi:10.1111/j.1467-7652.2010.00587.x
- Buell CR, Last RL (2010) Twenty-First century plant biology: impacts of the Arabidopsis genome on plant biology and agriculture. *Plant Physiol* 154(2):497–500. doi:10.1104/pp.110.159541
- Carollo V, Matthews DE, Lazo GR, Blake TK, Hummel DD, Lui N, Hane DL, Anderson OD (2005) GrainGenes 2.0: an improved resource for the small-grains community. *Plant Physiol* 139(2):643–651. doi:10.1104/pp.105.064485
- Chantret N, Salse J, Sabot F, Rahman S, Bellec A, Laubin B, Dubois I, Dossat C, Sourdille P, Joudrier P, Gautier MF, Cattolico L, Beckert M, Aubourg S, Weissenbach J, Caboche M, Bernard M, Leroy P, Chalhou B (2005) Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). *Plant Cell* 17(4):1033–1045. doi:10.1105/tpc.104.029181
- Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, Schlub S, Le Paslier MC, Magdelenat G, Gonthier C, Couloux A, Budak H, Breen J, Humphrey M, Liu S, Kong X, Jia J, Gut M, Brunel D, Anderson JA, Gill BS, 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(6):1686–1701
- Conley EJ, Nduati V, Gonzalez-Hernandez JL, Mesfin A, Trudeau-Spanjers M, Chao S, Lazo GR, Hummel DD, Anderson OD, Qi LL, Gill BS, Echalié B, Linkiewicz AM, Dubcovsky J, Akhunov ED, Dvorak J, Peng JH, Lapitan NLV, Pathan MS, Nguyen HT, Ma X-F, Miftahudin, Gustafson JP, Greene RA, Sorrells ME, Hossain KG, Kalavacharla V, Kianian SF, Sidhu D, Dilbirligi M, Gill KS, Choi DW, Fenton RD, Close TJ, McGuire PE, Qualset CO, Anderson JA (2004) A 2600-Locus Chromosome Bin Map of Wheat Homoeologous Group 2 reveals interstitial gene-rich islands and colinearity with rice. *Genetics* 168(2):625–637. doi:10.1534/genetics.104.034801
- Devos KM, Millan T, Gale MD (1993) Comparative RFLP maps of the homoeologous group-2 chromosomes of wheat, rye and barley. *Theor and Appl Genet* 85(6):784–792. doi:10.1007/bf00225020
- Devos KM, Dubcovsky J, Dvorak J, Chinoy CN, Gale MD (1995) Structural Evolution of wheat chromosomes 4a, 5a, and 7b and its impact on recombination. *Theor Appl Genet* 91(2):282–288
- Devos KM, Costa de Oliveira A, Xu X, Estill JC, Estep M, Jogi A, Morales M, Pinheiro J, San Miguel P, Bennetzen JL (2008) Structure and organization of the wheat genome—the number of genes in the hexaploid wheat genome. Paper presented at the 11th International Wheat Genetics Symposium, Brisbane, Australia, 24–29
- Doležel J, Kubaláková M, Bartoš J, Macas J (2004) Flow cytogenetics and plant genome mapping. *Chromosome Res* 12(1):77–91
- Duran C, Eales D, Marshall D, Imelfort M, Stiller J, Berkman PJ, Clark T, McKenzie M, Appleby N, Batley J, Basford K, Edwards D (2010) Future tools for association mapping in crop plants. *Genome* 53(11):1017–1023. doi:10.1139/g10-057
- Eckardt NA (2001) A sense of self: the role of DNA sequence elimination in allopolyploidization. *Plant Cell* 13(8):1699–1704
- Edwards D, Batley J (2010) Plant genome sequencing: applications for crop improvement. *Plant Biotechnol J* 8(1):2–9. doi:10.1111/j.1467-7652.2009.00459.x
- Flavell RB, Rimpau J, Smith DB (1977) Repeated sequence DNA relationships in 4 cereal genomes. *Chromosoma* 63(3):205–222
- Hernandez P, Martis M, Dorado G, Pfeifer M, Gálvez S, Schaaf S, Jouve N, Simková H, Valárik M, Doležel J, Mayer KF (2011) Next generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *Plant J*. doi:10.1111/j.1365-313X.2011.04808.x (accepted)
- Hossain KG, Kalavacharla V, Lazo GR, Hegstad J, Wentz MJ, Kianian PM, Simons K, Gehlhar S, Rust JL, Syamala RR, Obeori K, Bhamidimarri S, Karunadharm P, Chao S, Anderson OD, Qi LL, Echalié B, Gill BS, Linkiewicz AM, Ratnasiri A, Dubcovsky J, Akhunov ED, Dvorak J, Miftahudin, Ross K, Gustafson JP, Radhawa HS, Dilbirligi M, Gill KS, Peng JH, Lapitan NL, Greene RA, Bermudez-Kandianis CE, Sorrells ME, Feril O, Pathan MS, Nguyen HT, Gonzalez-Hernandez JL, Conley EJ, Anderson JA, Choi DW, Fenton D, Close TJ, McGuire PE, Qualset CO, Kianian SF (2004) A chromosome bin map of 2148 expressed sequence tag loci of wheat homoeologous group 7. *Genetics* 168(2):687–699. doi:10.1534/genetics.104.034850
- Huang S, Sirikhachornkit A, Su X, Faris J, Gill B, Haselkorn R, Gornicki P (2002) Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the *Triticum/Aegilops* complex and the evolutionary history of polyploid wheat. *Proc Natl Acad Sci USA* 99(12):8133–8138. doi:10.1073/pnas.072223799
- Imelfort M, Edwards D (2009) De novo sequencing of plant genomes using second-generation technologies. *Brief Bioinform* 10(6):609–618. doi:10.1093/bib/bbp039

- Karlin S, Altschul SF (1993) Applications and statistics for multiple high-scoring segments in molecular sequences. *Proceedings of the National Academy of Sci* 90(12):5873–5877
- Lander ES, Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2(3):231–239. doi:10.1016/0888-7543(88)90007-9. <http://www.sciencedirect.com/science/article/pii/0888754388900079>
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057):376–380. doi:10.1038/nature03959
- Marshall DJ, Hayward A, Eales D, Imelfort M, Stiller J, Berkman PJ, Clark T, McKenzie M, Lai K, Duran C, Batley J, Edwards D (2010) Targeted identification of genomic regions using TAGdb. *Plant Methods* 6:19. doi:10.1186/1746-4811-6-19
- Massa AN, Wanjugi H, Deal KR, O'Brien K, You FM, Maiti R, Chan AP, Gu YQ, Luo MC, Anderson OD, Rabinowicz PD, Dvorak J, Devos KM (2011) Gene space dynamics during the evolution of *Aegilops tauschii*, *Brachypodium distachyon*, *Oryza sativa*, and *Sorghum bicolor* Genomes. *Mol Biol Evol* 28(9):2537–2547
- Matthews DE, Carollo VL, Lazo GR, Anderson OD (2003) Grain Genes, the genome database for small-grain crops. *Nucleic Acids Res* 31(1):183–186
- Mayer KF, Taudien S, Martis M, Šimková H, Suchánková P, Gundlach H, Wicker T, Petzold A, Felder M, Steuernagel B, Scholz U, Graner A, Platzer M, Doležel J, Stein N (2009) Gene content and virtual gene order of barley chromosome 1H. *Plant Physiol*. doi:10.1104/pp.109.142612
- Naranjo T, Roca A, Goicoechea PG, Giraldez R (1987) Arm homoeology of wheat and rye chromosomes. *Genome* 29(6):873–882. doi:10.1139/g87-149
- Paux E, Roger D, Badaeva E, Gay G, Bernard M, Sourdille P, Feuillet C (2006) Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. *Plant J* 48(3):463–474. doi:10.1111/j.1365-313X.2006.02891.x
- Pop M (2009) Genome assembly reborn: recent computational challenges. *Brief Bioinform* 10(4):354–366. doi:10.1093/bib/bbp026
- Qi LL, Echalié B, Chao S, Lazo GR, Butler GE, Anderson OD, Akhunov ED, Dvorak J, Linkiewicz AM, Ratnasiri A, Dubcovsky J, Bermudez-Kandianis CE, Greene RA, Kantety R, La Rota CM, Munkvold JD, Sorrells SF, Sorrells ME, Dilbirligi M, Sidhu D, Erayman M, Randhawa HS, Sandhu D, Bondareva SN, Gill KS, Mahmoud AA, Ma X-F, Miftahudin, Gustafson JP, Conley EJ, Nduati V, Gonzalez-Hernandez JL, Anderson JA, Peng JH, Lapitan NLV, Hossain KG, Kalavacharla V, Kianian SF, Pathan MS, Zhang DS, Nguyen HT, Choi D-W, Fenton RD, Close TJ, McGuire PE, Qualset CO, Gill BS (2004) A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics* 168(2):701–712. doi:10.1534/genetics.104.034868
- Rabinowicz PD, Citek R, Budiman MA, Nunberg A, Bedell JA, Lakey N, O'Shaughnessy AL, Nascimento LU, McCombie WR, Martienssen RA (2005) Differential methylation of genes and repeats in land plants. *Genome Res* 15(10):1431–1440. doi:10.1101/gr.4100405
- Šafář J, Šimková H, Kubaláková M, Číhalíková J, Suchánková P, Bartoš J, Doležel J (2010) Development of chromosome-specific BAC resources for genomics of bread wheat. *Cytogenet Genome Res* 129(1–3):211–223. doi:10.1159/000313072
- Šimková H, Svensson JT, Condamine P, Hřibová E, Suchánková P, Bhat PR, Bartoš J, Šafář J, Close TJ, Doležel J (2008) Coupling amplified DNA from flow-sorted chromosomes to high-density SNP mapping in barley. *BMC Genomics* 9:294. doi:10.1186/1471-2164-9-294
- Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D, Bevan MW, Barry K, Lucas S, Harmon-Smith M, Lail K, Tice H, Schmutz Leader J, Grimwood J, McKenzie N, Huo N, Gu YQ, Lazo GR, Anderson OD, Vogel Leader JP, You FM, Luo MC, Dvorak J, Wright J, Febrer M, Idziak D, Hasterok R, Lindquist E, Wang M, Fox SE, Priest HD, Filichkin SA, Givan SA, Bryant DW, Chang JH, Mockler Leader TC, Wu H, Wu W, Hsia AP, Schnable PS, Kalyanaraman A, Barbazuk B, Michael TP, Hazen SP, Bragg JN, Laudencia-Chingcuanco D, Weng Y, Haberer G, Spannagl M, Mayer Leader K, Rattei T, Mitros T, Lee SJ, Rose JK, Mueller LA, York TL, Wicker Leader T, Buchmann JP, Tanskanen J, Schulman Leader AH, Gundlach H, Bevan M, Costa de Oliveira A, da CML, Belknap W, Jiang N, Lai J, Zhu L, Ma J, Sun C, Pritham E, Salse Leader J, Murat F, Abrouk M, Mayer K, Bruggmann R, Messing J, Fahlgren N, Sullivan CM, Carrington JC, Chapman EJ, May GD, Zhai J, Ganssmann M, Guna Ranjan Gurazada S, German M, Meyers BC, Green Leader PJ, Tyler L, Wu J, Thomson J, Chen S, Scheller HV, Harholt J, Ulvskov P, Kimbrel JA, Bartley LE, Cao P, Jung KH, Sharma MK, Vega-Sanchez M, Ronald P, Dardick CD, De Bodt S, Verelst W, Inze D, Heese M, Schnittger A, Yang X, Kalluri UC, Tuskan GA, Hua Z, Vierstra RD, Cui Y, Ouyang S, Sun Q, Liu Z, Yilmaz A, Grotewold E, Sibout R, Hematy K, Mouille G, Hofte H, Michael T, Pelloux J, O'Connor D, Schnable J, Rowe S, Harmon F, Cass CL, Sedbrook JC, Byrne ME, Walsh S, Higgins J, Li P, Brutnell T, Unver T, Budak H, Belcram H, Charles M, Chalhoub B, Baxter I (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* 463(7282):763–768. doi:10.1038/nature08747
- Vrána J, Kubaláková M, Šimková H, Číhalíková J, Lysák MA, Doležel J (2000) Flow sorting of mitotic chromosomes in common wheat (*Triticum aestivum* L.). *Genetics* 156(4):2033–2041
- Wanjugi H, Coleman-Derr D, Huo N, Kianian SF, Luo M-C, Wu J, Anderson O, Gu YQ (2009) Rapid development of PCR-based genome-specific repetitive DNA junction markers in wheat. *Genome* 52(6):576–587
- Wicker T, Buchmann JP, Keller B (2010) Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Res*. doi:10.1101/gr.107284.110
- Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18(5):821–829. doi:10.1101/gr.074492.107
- Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. *J Comput Biol* 7(1–2):203–214. doi:10.1089/10665270050081478