

Wheat genome structure and function: genome sequence data and the International Wheat Genome Sequencing Consortium

P. Moolhuijzen^A, D. S. Dunn^A, M. Bellgard^A, M. Carter^B, J. Jia^C, X. Kong^C,
B. S. Gill^D, C. Feuillet^E, J. Breen^A, and R. Appels^{A,F}

^ACentre for Comparative Genomics, Murdoch University, Murdoch, WA 6150, Australia.

^BState Agric Biotechnology Centre, Murdoch University, Murdoch, WA 6150, Australia.

^CInstitute of Crop Sciences, Chinese Academy of Agricultural Sciences, 100081, Beijing, China.

^DDepartment of Plant Pathology, Kansas State University, Manhattan, KS 66506, USA.

^EUMR ASP 1095, INRA, University Blaise Pascal, 63100 Clermont-Ferrand, France.

^FCorresponding author. Email: rappels@ccg.murdoch.edu.au

This paper is dedicated to Professor Bob McIntosh, in recognition of his tireless efforts to critically analyse the work carried out on the genetics of wheat.

Abstract. Genome sequencing and the associated bioinformatics is now a widely accepted research tool for accelerating genetic research and the analysis of genome structure and function of wheat because it leverages similar work from other crops and plants. The International Wheat Genome Sequencing Consortium addresses the challenge of wheat genome structure and function and builds on the research efforts of Professor Bob McIntosh in the genetics of wheat. Currently, expressed sequence tags (ESTs; ~500 000 to date) are the largest sequence resource for wheat genome analyses. It is estimated that the gene coverage of the wheat EST collection is ~60%, close to that of *Arabidopsis*, indicating that ~40% of wheat genes are not represented in EST collections. The physical map of the D-genome donor species *Aegilops tauschii* is under construction (<http://wheat.pw.usda.gov/PhysicalMapping>). The technologies developed in this analysis of the D genome provide a good model for the approach to the entire wheat genome, namely compiling BAC contigs, assigning these BAC contigs to addresses in a high resolution genetic map, filling in gaps to obtain the entire physical length of a chromosome, and then large-scale sequencing.

Introduction

Wheat is adapted to temperate regions and was among the first of the domesticated crops. The crop occupies approximately 17% of areas planted to cereals (210 Mha in 2002 v. 147 Mha for rice and 139 Mha for maize; Gill *et al.* 2004). It is predicted that to meet human needs by 2050, grain production must increase at an annual rate of 2% on an area of land similar to that currently utilised (Gill *et al.* 2004). In order to increase absolute yields and protect the crop from an estimated average annual loss of 25% caused by biotic stresses (pests) and abiotic stresses (heat, frost, drought, and salinity), it has been argued that a new level of understanding of the structure and function of the genome needs to be achieved (Gill *et al.* 2004). Genome sequencing and the associated bioinformatics is now a widely accepted research tool for accelerating the analysis of genome structure and function because it leverages similar work from other crops and plants. The International Wheat Genome Sequencing Consortium (<http://wheat.pw.usda.gov/PhysicalMapping>) addresses the challenge of wheat genome structure and function and builds on the research efforts of Professor Bob McIntosh in the genetics of wheat.

The genome of common wheat is distributed between 7 groups of chromosomes, each group containing a set of 3 homeologous chromosomes belonging to the A, B,

and D genomes. The diploid progenitors of the A, B, and D genomes have been identified (reviewed in Gill *et al.* 2004). Although common wheat functions much like a diploid organism (even though the genomes are closely related), the genome can tolerate aneuploidy because most gene functions are present in 3 doses. Sears (1954) demonstrated that mono-, tri-, and tetrasomic cytogenetic stocks were viable, as were nullisomics for 11 chromosomes. Since the loss of a pair of chromosomes can be compensated by 2 additional doses of a homeologous chromosome, Sears (1966) was able to compile a set of 42 compensating nulli-tetrasomic cytogenetic stocks, and subsequently Endo and Gill (1996) utilised the gametocidal chromosome introduced from *Aegilops cylindrical* Host, and developed 436 segmental deletion lines in Chinese Spring (CS). These cytogenetic stocks provide the foundation for studies on wheat genomics.

Currently, expressed sequence tags (ESTs; ~500 000 to date) are the largest sequence resource for wheat genome analyses. ESTs are cDNA clones, and as such they do not contain promoters, introns, and other functional elements. It is estimated that the gene coverage of wheat EST collection is ~60%, close to that of *Arabidopsis* (W. Li and B. S. Gill, cited in Gill *et al.* 2004), indicating that ~40% of wheat genes are not represented in EST collections. The physical

map of the D-genome donor species *Aegilops tauschii* is under construction (<http://wheat.pw.usda.gov/PhysicalMapping>). Five BAC libraries have been constructed and fingerprinted using improved high-resolution methodologies (Luo *et al.* 2003). At the same time, wheat RFLP markers and ESTs have been placed onto the physical map to anchor the BAC contigs to genetic maps and deletion bins (Dvorak *et al.* cited in Gill *et al.* 2004). The technologies developed in this analysis of the D genome provide a good model for the approach to the entire wheat genome, i.e. compiling BAC contigs, assigning these BAC contigs to addresses in a high resolution genetic map, filling in gaps to obtain the entire physical length of a chromosome, and then large-scale sequencing.

Status of genome sequence organisation

The cereals rice, maize, barley, and wheat evolved from a common ancestor about 70–55 million years ago (Kellogg 2001); however, they differ greatly in genome size. Hexaploid wheat (*Triticum aestivum* L., $2n = 6x = 42$, AABBDD) has the largest genome at 16 000 Mb, about 8-fold larger than that of maize and 40-fold larger than that of rice (Arumuganathan and Earle 1991). Variation in the numbers of transposable and retrotransposable elements, and duplication of chromosome segments, are largely responsible for differences in sizes of the genomes. Schulman and Kalendar (2005) have reviewed the diversity of repetitive elements in cereal genomes, with a special reference to barley. The Class I elements comprise repetitive sequences that can be clearly assigned to sequences that could have undergone so-called retrotransposition, via an RNA intermediate, leading to cDNA copies that insert into new locations. Class II elements are repetitive sequences that can be assigned to transposable elements that move via a cut-and-paste mechanism and are not as prominent as the Class I elements.

Class I, retrotransposable, elements have 3 broad groups, LINEs, SINEs, and LTRs. The LTRs represent a group of retrotransposons that are recognisable across a wide range of organisms and are divided into copia-like elements (characterised by genes encoding a capsid protein, GAG, aspartic proteinase, AP, integrase, IN, reverse transcriptase, RT, and RNaseH) and gypsy-like elements (characterised by IN located in a different position). The LTRs appear to be the main components for changing the structure of genomes (Schulman and Kalendar 2005). LINEs contain an internal promoter and encode proteins that could function in RNA-binding and encapsidation, as an endonuclease and as a reverse transcriptase. SINEs do not encode proteins that could be assigned to a replication function. Underpinning the structural interpretation of these retrotransposable elements is that they originated from retroviruses. Although elements such as Tnt 1 in tobacco (Grandbastien *et al.* 1989) and Tos17 in rice (Miyao *et al.* 2003) can be shown to be active in retrotransposition, this has not been shown for the wheat and barley elements. Within the wheat genome, most of the DNA consists of repeated sequences, and 70% of the DNA consists of sequences referred to as retrotransposable elements (W. Li and B. S. Gill, cited in Gill *et al.* 2004; Schulman and Kalendar 2005). Low-copy retrotransposable elements and miniature inverted repeat retrotransposable elements (MITES) are most often associated

with active genes. High-copy retrotransposable elements tend to be located in the intergenic space (SanMiguel *et al.* 2002). Gene distribution along chromosomes is relatively homogeneous in the small genome of rice, but in wheat the gene clusters (gene-rich regions) tend to be separated by long stretches of retrotransposable elements (gene-poor or gene-free regions), as demonstrated by deletion mapping (Endo and Gill 1996; Faris *et al.* 2003) and BAC-based physical mapping (Dvorak *et al.* cited in Gill *et al.* 2004). Within some gene-rich regions of the wheat genome, gene density is similar to that of smaller genomes (Feuillet and Keller 2002).

The analysis of a random wheat BAC clone (contig 4, phase 1 assembly of AuE15 from Renan BAC library) is shown in Fig. 1 to illustrate the complexity of the interpretation of wheat genome sequence information. The 35 kb of DNA shown is largely composed of repetitive DNA sequence as judged from the similarity hits between the sequence (*x*-axis) and the named repetitive elements in the available databases (*y*-axis). Below the dot matrix analysis is a summary of the named repetitive elements identified using 2 databases (Rebase and TREP) and 3 programs (Blast, Censor, and RepeatMasker) commonly used for annotation. Although the annotation differs depending on the database selected for the annotation, there is some consistency in the 2 basic patterns of repetitive sequence structure within the sequence. One pattern is 'nested' where one or more repetitive elements appear to be inserted within a pre-existing element (summarised in Fig. 1; SanMiguel *et al.* 2002), and the other is where the repetitive element structure is simply that of a single element. A relatively simple structure is evident in the 7.5–20 kb region, in contrast to a complex nested structure in the 22–33 kb region (boxed in Fig. 1) where elements annotated as 'Cerebra' are nested within an element annotated as 'Fatima'. In the 2–6.5 kb region (also boxed in Fig. 1) a 'Barbara I' element is apparently inserted into a 'Romani I' element. As more genome sequence data become available and the databases become more refined (Schulman and Kalendar 2005), it is expected that the annotation issues will become more defined and genuine genes will become easier to recognise. The careful annotation of repetitive elements in the wheat genome provides new molecular markers, based on the junctions between different elements, for use in analysing crosses (Devos *et al.* 2005).

The wheat genomic DNA listed in sequence databases is indicated in Table 1, with the database expanding quickly as the commitment to wheat sequencing increases (Devos *et al.* 2005). BAC clones selected for sequencing include those hybridising with specific genes and they revealed that gene density varies greatly, ranging from 1 gene (Faris *et al.* 2003) to 16 genes (Brooks *et al.* 2002) per BAC, and that genes tend to be clustered into gene islands (Brooks *et al.* 2002; Wicker *et al.* 2001; SanMiguel *et al.* 2002). Feuillet and Keller (2002) reported a gene density of 1 per 4–5 kb within a small segment on chromosome 1A. Lu and Faris (2005) have sequenced over 600 kb at the *Tsn1* locus on chromosome 5B. The 600 kb contig contained 13 genes for an average gene density of 1 gene per 46 kb. However, 9 of these genes were located within a 90 kb segment, resulting in a gene density of 1 per 10 kb. In contrast, Faris *et al.* (2003) found only 3 known genes within a BAC contig of 300 kb spanning the *Q* locus on chromosome 5A, yielding an estimate of 1 gene per 100 kb. Large tracts of

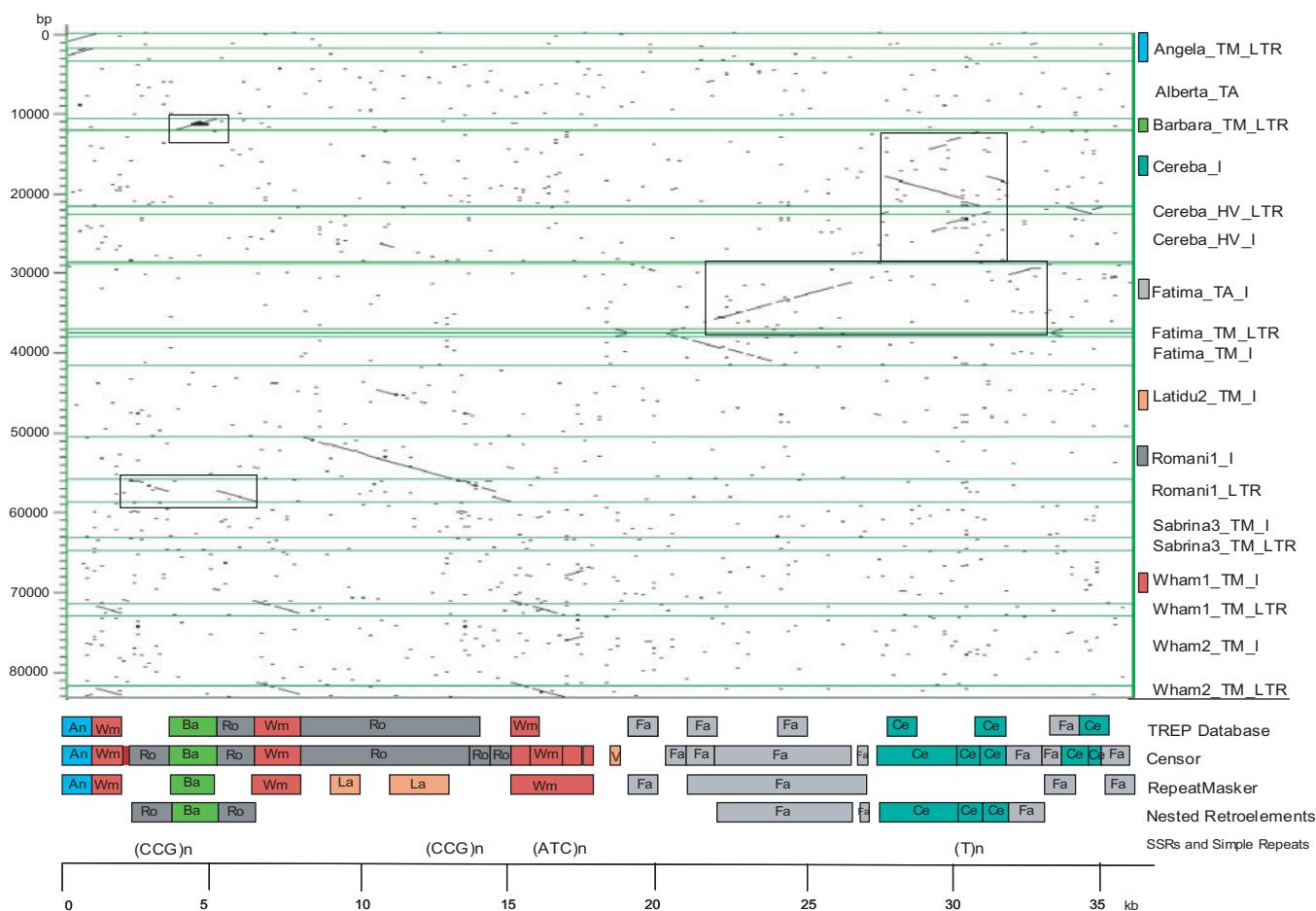


Fig. 1. Genomic structure of contig 4 in the wheat BAC AuE15 clone. Transposable elements constitute >99% of the contig. The 36.5 kb genomic sequence of contig 4 is on the x-axis and database sequences of some complete and LTR fragments of transposable elements related to this contig are on the y-axis of the dotplot. The first 3 rows below the dotplot are the identified elements as determined from the TREP database, Censor and RepeatMasker, respectively. The nested elements are shown on the next row and 4 simple repeats on the last row. The boxed areas show the alignment evidence that indicates a nested structure for named retrotransposable element DNA sequences.

Table 1. Wheat genomic DNA which has been sequenced (August 2005)

Species	No. of genomic sequences > 10 kb	Length sequenced (bp)	No. of BAC clones	Chrom. locations	Gene annotation
<i>A. tauschii</i>	13	243 759	13	5D, 1D	VRN-d1, Glo-2, Glu-D1 (HMW)
<i>T. monoccocum</i>	17	2 108 830	17	1A, 4A, 5A, 7A, 5AS, 5AL	5K14.1-9, GSP-Am1, Pina-Am1, Pinb-Am1, AGLG1, phytochelatin synthetase, Cyb5, Ap1, Glu-A3-3, Glu-A3-2, Balduin-transposase, Caspar-transposase, Caspar_ORF-2, RGL-1, STF-1
<i>T. aestivum</i>	10	674 813	10	Group 5	MnSOD, ACT-1, CCF, VRN-B1, VRN-A1, putative gag-pol polyprotein

repetitive elements with very few intervening low-copy non-coding sequences separated the 3 genes. This is also the case for the sequences determined at the *Glu-B1* and *Glu-D1* loci (Kong et al. 2004).

The International Wheat Genome Sequencing Consortium (IWGSC)

The IWGSC concept was first discussed among a broad group of scientists at the International Triticeae Mapping Initiative

(ITMI) held in Winnipeg, Canada, in June 2003, and the need for sequencing the genome of this important polyploid was more formally presented at a workshop in Washington DC, November 2003 (Gill et al. 2004). In 2004 the IWGSC was established by a group of plant scientists, breeders, and growers who were dedicated to defining the detailed structure of the wheat genome (www.wheatgenome.org/organisation.html). The IWGSC has maintained its close relationship with the broader ITMI alliance of scientists because it is clear that

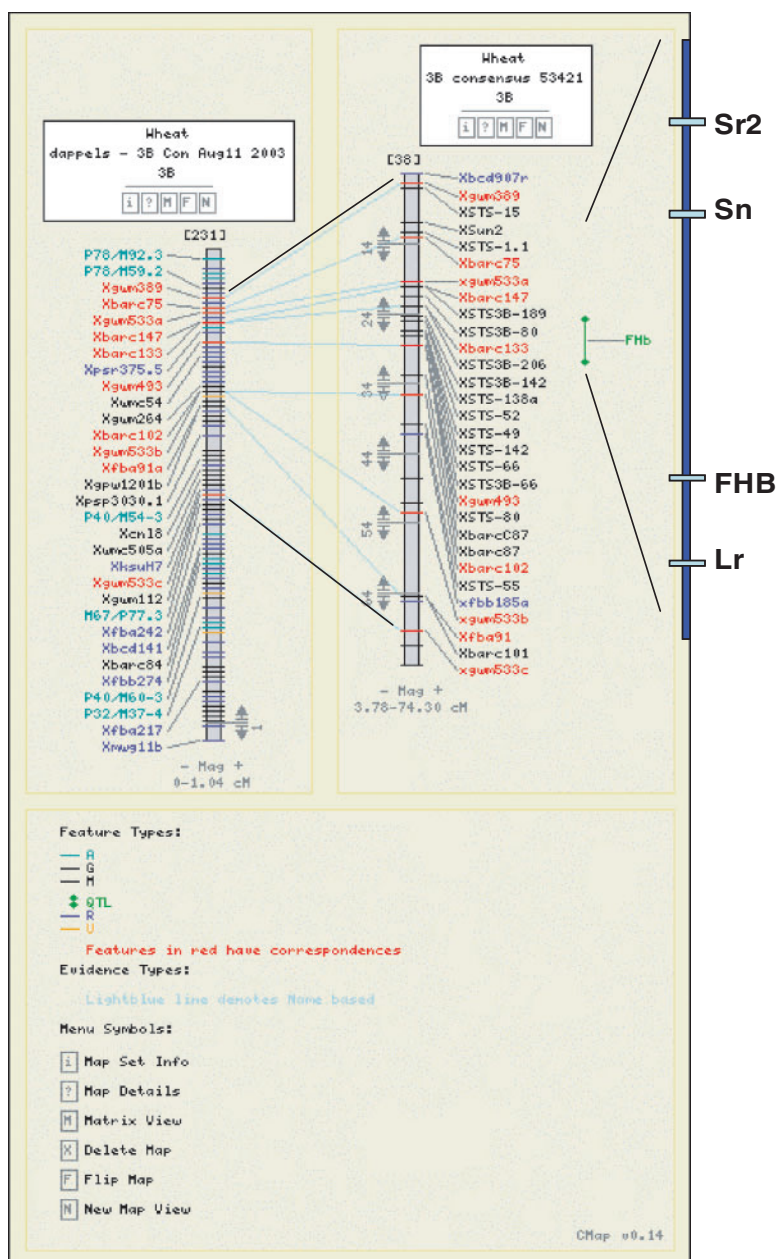


Fig. 2. Molecular genetic map of chromosome 3B indicating the region that is currently targeted for sequencing. The left hand map of chromosome 3B is from CMap (GrainGenes) and the right hand map is a high resolution map compiled by Liu *et al.* (2005) and is shown as an example of the resolution required for accurately defining the ordering of BAC clones before genome sequencing – all the XSTS loci in the composite map shown were added by Liu *et al.* (2005). The simple map on the far right indicates the current target region for sequencing and indicates the fusarium head blight (FHB) locus, a putative rust resistance locus (*Lr*), a *Septoria nodorum* blotch resistance locus (*Sn*), and the stem rust resistance locus (*Sr2*). The relative order of *Sr2* and *Sn* are unclear at present.

developments in defining the structure and function of the barley and Brachypodium genomes have significant impacts on work focused on the wheat genome.

The IWGSC has a world-wide representation through its 6 co-chairs, a coordinating committee, and an executive director.

A strategy of integrating data from random sequencing, gene-enriched DNA sequencing and flow-sorted chromosome-BAC-based sequencing is now well established, and defines areas of coordination especially for facilitating funding applications. The long-term goal is a sequenced wheat genome, and the strategy

of building physical 'sequence ready' maps (based on molecular genetic maps) and analysing flow-sorted chromosomes and chromosome arms of wheat helps to define short- to medium-term goals.

The 21 wheat chromosomes can be readily identified by heterochromatic banding (Gill *et al.* 1991) or *in situ* hybridisation patterns using repetitive DNA probes (May and Appels 1980). A specific chromosome or chromosome arm can be flow-sorted at high purity using specific wheat genetic stocks (Vra'na *et al.* 2000), and the sorted chromosomes have been used for construction of chromosome-specific BAC libraries (Janda *et al.* 2004; Safar *et al.* 2004). The chromosome 3B BAC library, in particular, currently forms focus for a pilot project within the IWGSC. The Chinese Spring 3B BAC library consists of 67 968 clones (Safar *et al.* 2004), and these clones have been fingerprinted using an improved SNAPshot protocol for BAC fingerprinting and high-throughput facilities at the French National Sequencing Center (C. Feuillet, unpublished data; fingerprinting based on Luo *et al.* 2003). The fingerprints allowed 57 329 BACs to be assembled into contigs, and among these, 140 contigs have been anchored to wheat BINs (the detailed molecular genetic characterisation of the wheat BINs is in Sourdille *et al.* 2004). It has been estimated that anchored contigs cover approximately 80% of the genome in chromosome 3B (C. Feuillet, unpublished data). A particularly challenging issue is the detailed anchoring of BAC contigs to a molecular genetic map, and a 12 cM region encompassing the *Sr2* stem rust resistance and the fusarium head blight resistance loci is the region within chromosome 3B currently being tackled (C. Feuillet, unpublished data). Detailed molecular genetic maps are being developed (Fig. 2; Liu *et al.* 2005; see also Mester *et al.* 2003) and these are crucial for anchoring BAC contigs for sequencing. Currently a 1820 kb contig, anchored to the genetic map around the *Sr2* stem rust resistance locus is being sequenced (C. Feuillet, unpublished data).

The chromosome 3B pilot project is providing a template for the detailed analysis of the entire wheat genome by demonstrating the feasibility of large-scale contig assembly for the wheat genome, establishing international teams to contribute to the detailed analysis of the wheat genome and establishing a sufficiently large genetic population for the anchoring of BAC clones. The physical mapping of BAC contigs for the entire wheat genome is now widely discussed and will make use of the observation from the 3B chromosome analysis that BAC end-sequences and sequences defining junctions between repetitive elements are often good probes for analysing populations of lines used for the molecular genetic mapping of wheat (Devos *et al.* 2005; C. Feuillet, unpublished data).

References

- Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. *Plant Molecular Biology Reporter* **9**, 208–218.
- Brooks SA, Huang L, Gill BS, Fellers JP (2002) Analysis of 106 kb of contiguous DNA sequence from the D genome of wheat reveals high gene density and a complex arrangement of genes related to disease resistance. *Genome* **45**, 963–972. doi: 10.1139/g02-049
- Devos KM, Ma J, Pontaroli AC, Pratt LH, Bennetzen JL (2005) Analysis and mapping of randomly chosen bacterial artificial chromosome clones from hexaploid bread wheat. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 19243–19248. doi: 10.1073/pnas.0509473102
- Endo TR, Gill BS (1996) The deletion stocks of common wheat. *Journal of Heredity* **87**, 295–307.
- Faris JD, Fellers JP, Brooks SA, Gill BS (2003) A bacterial artificial chromosome contig spanning the major domestication locus *Q* in wheat and identification of a candidate gene. *Genetics* **164**, 311–321.
- Feuillet C, Keller B (2002) Comparative genomics in the grass family: molecular characterization of grass genome structure and evolution. *Annals of Botany* **89**, 3–10. doi: 10.1093/aob/mcf008
- Gill BS, Appels R, Botha-Oberholster AM, Buell CR, Bennetzen JL *et al.* (2004) A workshop report on wheat genome sequencing: International Genome Research on Wheat Consortium. *Genetics* **168**, 1087–1096. doi: 10.1534/genetics.104.034769
- Gill BS, Friebe B, Endo TR (1991) Standard karyotype and nomenclature system for description of chromosome bands and structural aberrations in wheat (*Triticum aestivum*). *Genome* **34**, 830–839.
- Grandbastien M-A, Spielmann A, Caboche M (1989) *Tnt1* a mobile retroviral-like transposable element of tomato isolated by plant cell genetics. *Nature* **337**, 376–380. doi: 10.1038/337376a0
- Janda J, Bartoš J, Šafář J, Kubaláková M, Valárik M *et al.* (2004) Construction of a subgenomic BAC library specific for chromosomes 1D, 4D and 6D of hexaploid wheat. *Theoretical and Applied Genetics* **109**, 1337–1345. doi: 10.1007/s00122-004-1768-8
- Kellogg EA (2001) Evolutionary history of the grasses. *Plant Physiology* **125**, 1198–1205. doi: 10.1104/pp.125.3.1198
- Kong XY, Gu YQ, You FM, Dubcovsky J, Anderson OD (2004) Dynamics of the evolution of orthologous and paralogous portions of a complex locus region in two genomes of allopolyploid wheat. *Plant Molecular Biology* **54**, 55–69. doi: 10.1023/B:PLAN.0000028768.21587.dc
- Liu S, Zhang X, Pumphrey MO, Stack RW, Gill BS, Anderson JA (2005) Complex microcolinearity among wheat, rice, and barley revealed by fine mapping of the genomic region harboring a major QTL for resistance to Fusarium head blight in wheat. *Functional & Integrative Genomics*. doi: 10.1007/s10142-005-0007-y
- Lu H, Faris JD (2005) Macro- and microcolinearity between the genomic region of wheat chromosome 5B containing the *Tsn1* gene and the rice genome. *Functional & Integrative Genomics*. doi: 10.1007/s10142-005-0020-1
- Luo MC, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics* **82**, 378–389. doi: 10.1016/S0888-7543(03)00128-9
- May CE, Appels R (1980) Rye chromosome translocations in hexaploid wheat: a re-evaluation of the loss of heterochromatin from rye chromosomes. *Theoretical and Applied Genetics* **56**, 17–23. doi: 10.1007/BF00264422
- Mester DI, Ronin YI, Hu Y, Peng J, Nevo E, Korol AB (2003) Efficient multipoint mapping: making use of dominant repulsion-phase markers. *Theoretical and Applied Genetics* **107**, 1102–1112. doi: 10.1007/s00122-003-1305-1
- Miyao A, Tanaka K, Murata K, Sawaki H, Takeda S, Abe K, Shinozuka Y, Onosato K, Hirochika H (2003) Target site specificity of the *Tos17* retrotransposon shows a preference for insertion within genes and against insertion in retrotransposon-rich regions of the genome. *The Plant Cell* **15**, 1771–1780. doi: 10.1105/tpc.012559
- Safar J, Bartos J, Janda J, Bellec A, Kubaláková M *et al.* (2004) Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. *The Plant Journal* **39**, 960–968. doi: 10.1111/j.1365-3113X.2004.02179.x

- SanMiguel PJ, Ramakrishna W, Bennetzen JL, Busso CS, Dubcovsky J (2002) Transposable elements, genes and recombination in a 215 kb contig from wheat chromosome 5A. *Functional & Integrative Genomics* **2**, 70–80. doi: 10.1007/s10142-002-0056-4
- Schulman AH, Kalendar R (2005) A movable feast: diverse retrotransposons and their contribution to barley genome dynamics. *Cytogenetic and Genome Research* **110**, 598–605. doi: 10.1159/000084993
- Sears ER (1954) The aneuploids of common wheat. *Mo. Agric. Experiment Station Research Bulletin* **572**, 1–59.
- Sears ER (1966) Nullisomic-tetrasomic combinations in hexaploid wheat. In 'Chromosome manipulation and plant genetics'. (Eds R Riley, KR Lewis) pp. 29–45. (Oliver & Boyd: Edinburgh)
- Sourdille P, Singh S, Cadalen T, Brown-Guedira GL, Gay G, Qi L, Gill BS, Dufour P, Murigneux A, Bernard M (2004) Microsatellite-based deletion bin system for the establishment of genetic-physical map relationships in wheat (*Triticum aestivum* L.). *Functional & Integrative Genomics* **4**, 12–25. doi: 10.1007/s10142-004-0106-1
- Vra'na J, Kubala'kova' M, Simkova' H, Cý'halý'kova' J, Lysa'k MA *et al.* (2000) Flow sorting of mitotic chromosomes in common wheat (*Triticum aestivum* L.). *Genetics* **156**, 2033–2041.
- Wicker T, Stein N, Albar L, Feuillet C, Schlagenhauf E *et al.* (2001) Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. *The Plant Journal* **26**, 307–316. doi: 10.1046/j.1365-313X.2001.01028.x

Manuscript received 11 May 2007, accepted 1 June 2007