Chromosome Research (2007) 15:21–31
DOI: 10.1007/s10577-006-1107-9

Triticeae genomics: advances in sequence analysis of large genome cereal crops

Nils Stein*
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstrasse 3, 06466 Gatersleben, Germany; Tel: +49-39482-5522; Fax: +49-39482-5595; E-mail: stein@ipk-gatersleben.de
**Correspondence*

Key words: genome analysis, genome sequence, genome structure, genomics, rice, Triticeae

Abstract

Whole genome sequencing provides direct access to all genes of an organism and represents an essential step towards a systematic understanding of (crop) plant biology. Wheat and barley, two of the most important crop species worldwide, have two- to five-fold larger genomes than human – too large to be completely sequenced at current costs. Nevertheless, significant progress has been made to unlock the gene contents of these species by sequencing expressed sequence tags (EST) for high-density mapping and as a basis for elucidating gene function on a large scale. Several megabases of genomic (BAC) sequences have been obtained providing a first insight into the complexity of these huge cereal genomes. However, to fully exploit the information of the wheat and barley genomes for crop improvement, sequence analysis of a significantly larger portion of the Triticeae genomes is needed. In this review an overview of the current status of Triticeae genome sequencing and a perspective concerning future developments in cereal structural genomics is provided.

Introduction

Access to the complete genome sequence of an organism provides a direct path to complete gene content information of a species. It paves the way for a comprehensive understanding of genome structure and evolution and it is one of the prerequisites towards systems biology of an organism. Over the past 15 years improvements in sequencing technology combined with drastically reduced costs triggered a development that led to hundreds of complete genome sequences, mostly from microbial genomes (see: http://www.genomenewsnetwork.org/resources/sequenced_genomes/genome_guide_index.shtml, http://www.jgi.doe.gov/sequencing/cspseqplans2007.html, http://www.genomesonline.org/; Bernal *et al*. 2001). Meanwhile the first sequences of higher eukaryotic genomes have been completed or reached an advanced stage by either a clone-by-clone-based strategy (i.e. human, Human Genome Sequencing Consortium 2004) or by whole genome shotgun sequencing (e.g. *C. elegans*, The *C. elegans* Sequencing Consortium 1998; *Drosophila*, Adams *et al*. 2000; bufferfish, Aparicio *et al*. 2002; rat, The Rat Genome Sequencing Project Consortium 2004; chimpanzee, The Chimpanzee Sequencing and Analysis Consortium 2005). Further low-pass shotgun sequencing and targeted resequencing of orthologous regions is under way for a number of additional higher eukaryote genomes (Thomas *et al*. 2003; Margulies *et al*. 2005a). This enables large-scale comparative sequence analysis between distantly (phylogenetic footprinting, Duret & Bucher 1997) and closely (phylogenetic shadowing, Boffelli *et al*. 2003) related species, allowing us to uncover relevant functional information of a genome that is not contained in exons.

In the plant kingdom less than a handful of genomes have been completely sequenced either following clone-by-clone (*Arabidopsis thaliana*, The *Arabidopsis* Genome Initiative 2000; rice, International Rice Genome Sequencing Project 2005) or whole genome shotgun sequencing strategies (rice, Goff *et al*. 2002; poplar, http://genome.jgi-psf.org/Poptr1/Poptr1.home.html). Similarly to the situation in animal genomics, rice and *Arabidopsis* genome annotation is becoming increasingly efficient with the availability of whole genome shotgun sequences of related (sub-)species such as *O. sativa* ssp. *indica* (Yu *et al*. 2001) and *Brassica oleracea* (Ayele *et al*. 2005).

Sequencing costs are directly related to the size of the target genome, and the funding volume for plant genome research has traditionally been orders of magnitude smaller than for biomedical research. Therefore, only relatively small plant genomes have been targeted so far by whole genome sequencing; i.e. *Arabidopsis* and rice feature 25–8 times smaller genomes than human. This situation might change in the next few years in prospect of decreasing sequencing costs and the development of new technologies (reviewed in Shendure *et al*. 2004, Metzker 2005, Service 2006).

The Triticeae tribe includes wheat (*Triticum aestivum*) and barley (*Hordeum vulgare*), that represented the largest and fourth-largest acreage worldwide in the year 2004 (http://faostat.fao.org/). They possess genome sizes of 5500 (barley) and 17 000 Mb (bread wheat) (Bennett & Smith 1976), that are composed of more than 80% of repetitive DNA (Smith & Flavell 1975). Thus, they have two to five times larger genomes than human with similar or higher complexity and content of repetitive DNA. The good level of genome colinearity to other grass species, including cereal crops such as wheat, maize, barley, sorghum, pearl millet, rye, and oats (Devos 2005), combined with its own agronomical importance, have promoted rice as the model for genome sequencing and gene isolation in cereals. Due to improved genomic resources for Triticeae species, map-based gene isolation in wheat and barley became feasible. The availability of the whole rice genome sequence repeatedly facilitated gene isolation in Triticeae (reviewed by Stein & Graner 2004). However, on a genome-wide scale, because of genomic rearrangements, rice can serve as a model for Triticeae gene isolation for about 50% of the target genes only (Gaut 2002, Stein *et al*. 2006). Based on the economic importance of these crop species a continuing and – in the prospect of global climatic change, increasing world population, and scarcity of arable land – an increasing need for gene isolation in Triticeae can be foreseen for accelerating improvement of these essential crops. Thus, an efficient strategy has to be established for sequencing the Triticeae genomes, or at least obtaining access to their entire gene content. In the following sections, the status of Triticeae genome sequencing is reviewed in the context of ongoing efforts in related crop species, and the paper concludes with a vision of future developments in cereal genome analysis.

Status of sequencing the Triticeae genomes

Accessing the genome sequence of an organism can be performed at different levels of complexity. The problem of genome size, the major obstacle for sequencing crop plant genomes, may be efficiently tackled by combining a bouquet of different approaches of deep sample sequencing anchored to an underlying BAC-based physical map of such genomes (recently reviewed by Paterson 2006, Rabinowicz & Bennetzen 2006). Gene sequence information, however, can be retrieved without large-scale genomic sequencing via random sequencing of cDNA clone libraries yielding expressed sequence tags (EST). Alternatively, sequencing of genomic shotgun clones produces a random picture of the overall genome structure and composition even if applied at low genome coverage (Genome Survey Sequencing GSS, definition at http://www.ncbi.nlm.nih.gov/dbGSS/). More detailed contiguous information concerning the structure and organisation of a genome can be obtained by selecting individual BAC (at random or from specific loci) for complete sequencing. This will eventually lead to an idea of the composition of DNA elements and the distribution of genes within genomes. Finally, whole genome sequencing delivers the completed picture of gene content and genome organization.

Triticeae genomes have so far been targeted by most of these approaches – except whole genome sequencing (clone-by-clone or shotgun), which remains a visionary task for future efforts to be tackled by international consortia (International Wheat Genome Sequencing Consortium, IWGSC, http://www.wheatgenome.org/; European Triticeae Genomics Initiative, ETGI, http://www.etgi.org, International Barley Sequencing Consortium, IBSC, http://barleygenome.org).

EST sequencing – assessing the Triticeae transcriptome

Generating large datasets of EST is a straightforward strategy to comprehensively access gene content and composition information of an organism; especially for large genome species, which are, under current costs, not accessible to whole genome sequencing (Rudd 2003). EST sequence information is furthermore of great value for supporting gene annotation in sequenced genomes. Thus, among the five largest publicly available plants EST datasets (Table 1), two originate from rice and *Arabidopsis* – species with whole genome information being available, and three from the major important cereal crop species maize, wheat and barley. For 19 grass species (including five Triticeae species) EST datasets larger than 5000 entries are available (Table 1). In barley and wheat, intermediate sets (>100 000) of EST were employed for estimating the gene content by sequence assembly of the redundant datasets leading to the prediction of around 30 000 unique genes (unigenes) (Ogihara *et al*. 2003, Lazo *et al*. 2004, Zhang *et al*. 2004). Independent results of EST sequence assemblies considering different proportions of publicly available EST can be accessed via various websites: CR-EST: http://pgrc.ipk-gatersleben.de/est/, HarvEST: http://harvest.ucr.edu/, NCBI UniGene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene. So-called 'gene indices' are provided for numerous organisms by 'The Institute of Genomics Research' (TIGR; http://www.tigr.org/tdb/tgi/plant.shtml). Their estimate indicates, the number of unique genes to be about 122 000 (~40 000 per homoeologous genome) and 50 000 for bread wheat and barley, respectively, based on assemblies of about 580 000 wheat and 370 000 barley EST. EST-based predictions of gene content can only be an approximation, since low or tissue/stage specifically expressed genes may not be represented in the datasets. These are also 'contaminated' with coding sequences from transposable elements and alternatively spliced genes that can be interpreted as multiple unigenes. Nonetheless, the number of unigenes deduced for wheat and barley is well within the range of the ~40 000 genes predicted from the complete rice genome sequence (International Rice

Table 1. Status of genome sequencing in different grass species

	EST[1]	GSS[2]	Genomic sequences (>70 000 bp)	
Oryza sativa	1 188 545	252 775	>400 Mb	Full genome[3]
Zea mays	879 619	2 014 160	295 Mb	2491 BAC[4]
Triticum aestivum[5]	855 066	7 999	9 Mb	67 BAC
Hordeum vulgare* ssp. *vulgare	437 321	2 033	2.4 Mb	16 BAC
Saccharum officinarum + spec.	255 937	–	–	–
Sorghum bicolor	204 208	573 724	4.9 Mb	36 BAC
Festuca arundinacea	41 834	–	–	–
Hordeum vulgare* ssp. *spontaneum	24 150	–	–	–
Sorghum propinquum	20 881	23 674	–	–
Brachypodium distachyon	20 449	–[6]	0.5 Mb	3 BAC
Panicum virgatum	11 990	–	–	–
Triticum monococcum	10 139	–	2.5 Mb	15 BAC
Secale cereale	9 195	–	–	–
Agrostis stolonifera	9 018	–	–	–
Triticum turgidum* ssp. *durum	8 924	–	2.9 Mb	18 BAC
Avena sativa	7 632	–	–	–
Agrostis capillaris	7 542	–	–	–
Lolium multiflorum	5 852	–	–	–
Lolium temulentum	5 738	–	–	–
Aegilops tauschii	–	5 055	0.62 Mb	5 BAC
Arabidopsis thaliana	622 972	440 517	>150 Mb	Full genome[7]

[1] Expressed sequence tags, dbEST release 081806 (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html).
[2] Genomic Survey Sequences, dbGSS release 081806 (http://www.ncbi.nlm.nih.gov/dbGSS/dbGSS_summary.html).
[3] International Rice Genome Sequencing Project 2005).
[4] Status 25 August, 2006.
[5] Members of the Triticeae tribe are indicated in bold letters.
[6] JGI/DOE agreed-upon shotgun sequencing of the *Brachypodium* genome http://www.jgi.doe.gov/sequencing/cspseqplans2007.html.
[7] The *Arabidopsis* Genome Initiative 2000), *A. th*. – first completely sequenced higher plant genome provided as reference.

Genome Sequencing Project 2005), thus it can be considered that a majority of all 'core' Triticeae genes has been matched either by public barley or wheat EST.

EST resources can directly be exploited to identify differentially expressed genes. Groups of co-regulated genes exhibiting similar tissue- or stage-specific expression have been found in barley and wheat (Ogihara *et al*. 2003, Zhang *et al*. 2004, Chao *et al*. 2006) using EST clustering data directly. Furthermore, commercial microarray products carrying 21 000, respectively 61 000, unique barley (Close *et al*. 2004) or wheat genes (Close *et al*. unpublished, http://harvest.ucr.edu/) have been developed from the public EST resources, and provide a standardized platform for Triticeae functional genomics.

Large sets of several thousands of EST-based unigenes have been mapped recently to genetic and physical maps in wheat and barley (Qi *et al*. 2004, Nasuda *et al*. 2005, Cho *et al*. 2006, Stein et al. 2006). These provided the opportunity to reveal rice/Triticeae genome colinearity at much higher density (Sorrells *et al*. 2003, Rota & Sorrells 2004) and resolution than previously available (Figure 1A). New patterns of genome colinearity and traces of ancient genome duplication in the Triticeae genomes have been observed (Stein *et al*. 2006, Figure 1B). This refined knowledge regarding cereal genome colinearity will help to improve any Triticeae genome physical maps, and thus facilitate current and future attempts of map-based gene isolation in Triticeae species.

Genomic survey sequencing (GSS)

In order to retrieve more general information about gene content and overall genome organization, GSS offers an alternative strategy that does not address genes exclusively. Here, the whole genome is targeted by random shotgun sequencing at about 1-fold coverage ('low-pass' or 'single-pass' random shotgun approach) and the resulting sequences are subsequently anchored to an available reference genome for assessing the level of genome conservation and colinearity.

Permutations of the GSS approach do not target the genome as a whole, but either physically defined subfractions (i.e. BAC end sequences, BES) or gene-enriched portions of the genome that can be selected by taking advantage of the physical properties of genomic DNA.

In general, very limited numbers of GSS sequences are available for the Triticeae species in public databases (Table 1). In maize, the next cereal species that has been selected for genome sequencing (NSF/USDA/DOE, http://www.nsf.gov/news/news_summ.jsp?cntn_id=104608), random large-scale assessment of ~470 000 BES (0.12-fold coverage of the maize genome) provided a representative snapshot of overall maize genome organization (Messing *et al*. 2004). Almost 56% of the genome belonged to class I mobile elements (retroelements) compared to about 1% being related to class II mobile elements (DNA transposons). After comparing with EST unigene assemblies roughly 9% of the maize BES (masked against repetitive elements) could be assigned to transcribed sequences (Messing *et al*. 2004). In a similar but smaller analysis, 11 Mb of wheat BES were derived from chromosome 3B-specific BAC (0.01-fold chromosome coverage): 86% of BES were derived from repetitive elements and only 3% originated from genic regions (1.2% from transcribed sequences) (Paux *et al*. 2006). A slightly higher content of genic sequences (~5%) was found in a limited source of 140 kb BES derived from BAC of *Triticum urartu, Aegilops speltoides* and *Ae. tauschii* (Akhunov *et al*. 2005).

Additionally, BES are helpful for establishing a BAC-based physical map of a genome, due to their physical assignment to library clone addresses. They are used as so-called sequence tagged connectors (STC) during contig establishment (Venter *et al*. 1996): overlapping BAC are determined by BES identity to low-pass sequencing assemblies of seeding BAC. These BAC are subsequently selected for fingerprinting and local contig construction, allowing for an iterative development of a physical map and sequence with the potential to particularly focus on regions of interest (Venter *et al*. 1996).

Gene-enrichment strategies can be performed either by exploiting differences in the methylation state or in the renaturation characteristics of low-copy DNA/the gene fraction compared to the repetitive fraction of the genome. Hypomethylated genic regions can be preferentially cloned either by utilizing bacterial strains that destroy cloned methylated DNA (methyl-filtration = MF) (Rabinowicz 2003) and thus eliminate methylated DNA inserts from the clone libraries; or hypomethylated DNA can be obtained and cloned by partial restriction with methylation-sensitive enzymes (hypomethylated partial restriction; HMPR) (Emberton *et al*. 2005). Alternatively, the bulk of repetitive DNA can be subtracted from less-frequent sequences or

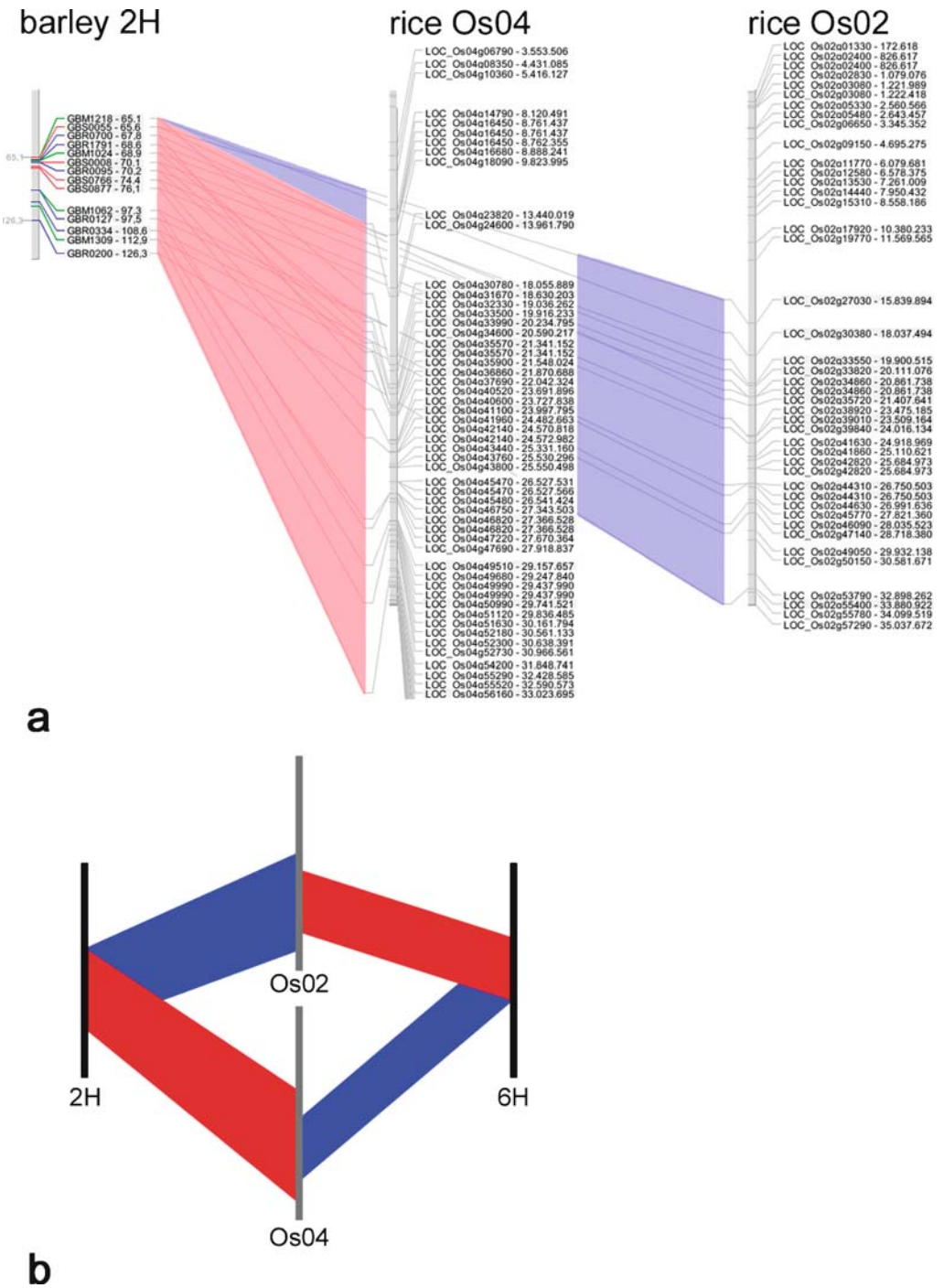

Figure 1. High-density transcript maps for comparative cereal genomics. Gene-based markers provide the possibility to compare genetic map position of genes in one species (barley) with the chromosomal position of the (putatively) orthologous gene in the genome sequence of another species (rice). **A**: A detailed view of the chromosome 2HL transcript map (taken from a 1000-loci transcript map of barley, Stein *et al*. 2006) and its relationship to rice chromosomes Os02 and Os04 is shown. Sequence comparison was based on BLASTN (Altschul *et al*. 1990, E-value ≤ E-10) to TIGR rice pseudomolecules v3.0 (http://www.tigr.org/tdb/e2k1/osa1/pseudomolecules/info.shtml). Best homologues (putative orthologues) of 2HL genes are located in a colinear part of rice Os04 (highlighted by red shading), whereas second-best homologues are present in colinear order on part of rice Os02 (highlighted by blue shading). **B**: A schematic drawing summarizes the relationship (red = orthologous region, blue = paralogous region) of 2H and Os04/Os02 as well as the reciprocal relationship of barley chromosome 6H to the same rice chromosomal regions, highlighting indirectly a conserved ancient genome duplication present in the rice (Guyot & Keller 2004) and barley genomes (Stein *et al*. 2006).

Table 2. Triticeae gene loci characterized by individual BAC sequencing

Species	Locus	Reference
Ae. tauschii	*Hardness* (*Ha*)	Chantret *et al*. 2005
	HMW Glutenin	Anderson *et al*. 2003
	Ph2	Whitford *et al*. 2006
	Lr21	Brooks *et al*. 2002
H. vulgare	D-hordein	Gu *et al*. 2003
	Ha homologous region	Caldwell *et al*. 2004
	Hv-eIF4E	Wicker *et al*. 2005
	Mla	Wei *et al*. 2002
	photoperiod response QTL	Szücs *et al*. 2006
	Ppd-H1	Turner *et al*. 2005
	Rph15	Falk, unpublished
	Rph7	Brunner *et al*. 2003
	Vrn2 homologous region	Yan *et al*. 2004
	Wx1 homologous region	Rostoks *et al*. 2002
	BCD135	Rostoks *et al*. 2002
	BCD1434	Rostoks *et al*. 2002
	RZ567	Rostoks *et al*. 2002
	WG644	Dubcovsky *et al*. 2001
T. monococcum	*Fr-Am2*	Miller *et al*. 2006
	Hardness (*Ha*)	Chantret *et al*. 2005
	LMW Glutenin	Wicker *et al*. 2003
	Lr10 orthologous region	Wicker *et al*. 2001
	Vrn1	Yan *et al*. 2003
	Vrn2	Yan *et al*. 2004
	Wx1 homologous region	Ma *et al*., unpublished
	WG644	Ramakrishna *et al*. 2002
	Q	Faris *et al*. 2003
T. turgidum	*Hardness* (*Ha*)	Chantret *et al*. 2005
	HMW Glutenin	Kong *et al*. 2004
	LMW Glutenin	Wicker *et al*. 2003
	Lr10 orthologous region	Isidore *et al*. 2005
	Snn1	Faris *et al*., unpublished
	Tsn1	Faris *et al*., unpublished
	unknown	Dvorak *et al*., unpublished
T. aestivum	*Glu-1*	Gu *et al*. 2006
	Hardness	Chantret *et al*. 2005
	Lr10	Isidore *et al*. 2005
	Ph1	Griffiths *et al*. 2006

low-copy DNA (enriched for gene sequences) by their higher re-association kinetics after DNA heat denaturation (= High-$C_o t$; cot based cloning and sequencing = CBCS, or cot filtration = CF) (Peterson *et al*. 2002). Repetitive DNA renatures more quickly after heat denaturation than single/low-copy DNA, forming faster double-stranded molecules, which can be purified from single-/low-copy (thus still single-stranded) and gene-enriched fractions of the genome based on single-stranded DNA affinity column (hydroxyapatite) chromatography.

The potential of MF and CF techniques for gene enrichment has been extensively assessed for the maize genome. In MF reads, an enrichment for gene sequences by a factor of 6-fold (Palmer *et al*. 2003), ~8-fold (Springer *et al*. 2004), or 13-fold (Rabinowicz *et al*. 2005) was determined. This was comparable to the factor of gene enrichment obtained by CF (Springer *et al*. 2004). Both MF and CF have been applied to wheat, and MF to barley, at a limited number of a few hundred to thousand reads. For example, in wild wheat

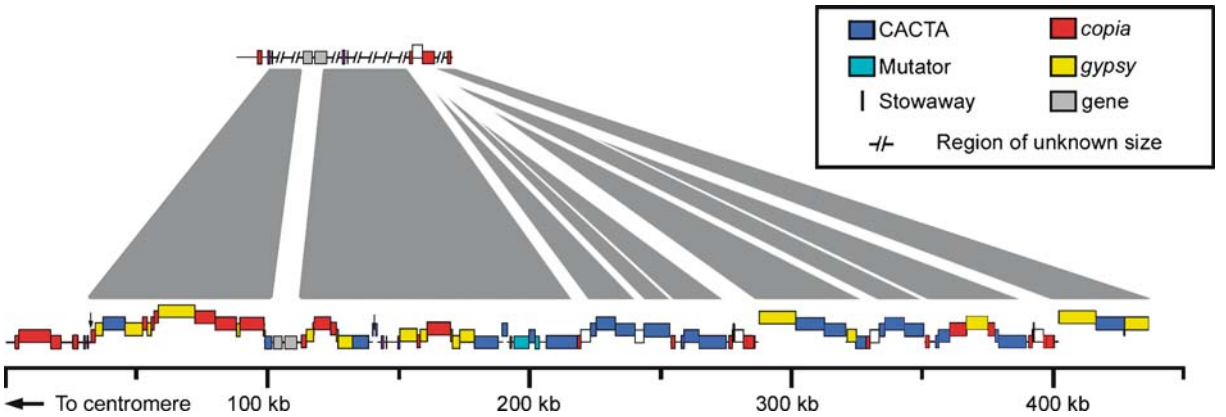

Figure 2. Genome expansion at the *Rym4* locus on barley chromosome 3HL. Complete LTR-retrotransposons were identified based on the presence of conserved target site duplications (TSD). Mutation rates in their respective two LTR indicated a 7 MY time scale for the expansion (indicated by grey shades) of the genomic region around *Hv-eIF4E* at the *Rym4* locus. It putatively expanded from originally 70 kb up to currently 440 kb (Wicker *et al*. 2005). Repeated cycles of transposon/retrotransposon insertion and DNA deletion formed the current structure of repetitive DNA (figure modified from Wicker *et al*. 2005).

(*Aegilops tauschii*) gene enrichment by MF reached only a factor of 2–4-fold (Li *et al*. 2004, Rabinowicz *et al*. 2005), whereas a significant enrichment of over 13-fold was obtained in hexaploid wheat using CF (Lamoureux *et al*. 2005). Methylfiltration works much more efficiently in barley, leading to over 18-fold gene enrichment in a single study (Rabinowicz *et al*. 2005). This difference in MF-based gene enrichment between wheat and barley possibly indicates some basic differences in genome structure or organization between these closely related Triticeae species.

Contiguous genomic sequencing

Sequencing larger and contiguous stretches of genomic DNA is necessary to reveal a detailed view of the organization and arrangement of genes and other DNA elements (i.e. repetitive DNA) within a genome. In Triticeae genomes such sequence information has so far been obtained basically by sequencing individual BAC or minimum tiling paths (MTP = clones covering a region or contig at minimum redundancy) of BAC contigs that had been developed for map-based cloning of agronomically important or just gene-containing loci (Table 2). These studies disclosed the expected presence of the enormous quantity of class I and II transposable elements making up the majority of the Triticeae genomes. Genes were frequently detected in small so-called 'gene islands', extending over 10–30 kb and containing a few genes, which were then separated from each other by large stretches (>100 kb) of repetitive DNA (Wicker *et al*. 2001, 2005, SanMiguel *et al*. 2002). Detailed characterization of the repetitive DNA allowed for the facilitated *de-novo* annotation of new large Triticeae sequences. To date over 1400 Triticeae repeats, representing more than 200 different repeat types, have been classified and collected in the Triticeae repeat database (TREP, Wicker *et al*. 2002). The pattern of intercalation between repetitive elements, as well as mutation rates, in long terminal repeats (LTR) of retrotransposons (which are identical at the time of insertion in the genome) can be used for dating genome rearrangements. Thus a model covering a period of 7 million years of events leading to genome expansion and contraction driven by repetitive element insertion and deletion (intra-/interelement recombination and or illegitimate recombination) was proposed for the barley *Rym4* locus (Figure 2, Wicker *et al*. 2005).

In a few cases the same locus was sequenced on BAC level either from two different cultivars in barley (Scherrer *et al*. 2005), between barley and diploid wheat (*T. monococcum*, Ramakrishna *et al*. 2002), between the orthologous A-genomes of diploid (2n), tetraploid (4n) and hexaploid (6n) wheat (Isidore *et al*. 2005) as well as between all homoeologous genomes of 2n–6n wheats (Chantret *et al*. 2005). In all cases, major changes in the composition of the intergenic repetitive DNA (lack of DNA-microcolinearity in repeat elements) were observed, indicating a rapid evolution of repetitive DNA within and between

Table 3. Estimating gene density in Triticeae genomes based on published BAC sequences

Species	No. of BAC[1]	No. of loci[2]	Accumulated fragment size (Mb)	Overall gene content[3]	Average gene density (1 gene/× kb)	Max. gene density (1 gene/× kb)	Min. gene density (1 gene/× kb)
Ae. tauschii	5	4	0.6	37	17	10	28
H. vulgare	16	14	2.4	85	28	12	220
T. monococcum	11	9	2.1	54	38	10	133
T. turgidum	6	7	1.3	29	44	23	154
T. aestivum	41	4	5.7	122	47	10	168
Triticeae core[4]	–	–	~5000	~30 000	160	–	–
genome (2n)	–	–	~5000	~40 000	125	–	–
	–	–	~5000	~50 000	100	–	–

[1] Only sequences >70 kb were retrieved from Genbank, status 25 August 2006; only annotated sequences were considered for evaluation.
[2] According to Table 2.
[3] According to published annotation of BAC sequences.
[4] A core 2n Triticeae genome was presumed to contain a 5000 Mb genome; gene content based on published EST assembly information.

Triticeae species since their divergence from a common progenitor. In general the overall gene order was well conserved between species in this limited number of analysed cases. However, other studies described a lack of colinearity or gene order even between closely related species such as diploid wheats and their homoeologous genomes of polyploid wheats (Chantret *et al*. 2005, Isidore *et al*. 2005).

Gene density in Triticeae genomes

Sequencing data of individual BAC has provided some insight into gene density in local regions of the Triticeae genomes. Based on the size of a basic (2n) Triticeae genome (i.e. 5000 Mb for barley) and a proposed number of 30 000–50 000 genes (based on EST data), physical distances between randomly distributed genes would be around 160–100 kb. A rough calculation of average gene densities can be performed based on the available annotated sequences of Triticeae BAC (length >70 kb each), providing values of between one gene in 17 kb for *A. tauschii* and one gene in 47 kb for *T. aestivum* (Table 3). These densities are consistently smaller than expected for a random distribution, but due to the small number of included and partially redundant loci (Table 2), these values are far from being representative for the Triticeae genomes. In order to address this issue of non-representative sampling of the Triticeae genome, a study aiming at sequencing 220 randomly chosen BAC of the *T. aestivum* genome has been initiated (Devos *et al*., oral presentation W47, Plant and Animal Genome Conference XIV, San Diego, 2006). An initial glimpse of only four of these clones provided an average gene density of one gene in 83 kb (Devos *et al*. 2005), which is higher than expected for a random distribution but at least 2-fold smaller than found in non-randomly selected published clones. However, the limited size of the sample does not allow reliable statistical statements on overall gene density in Triticeae genomes. It is known from other studies that gene distribution is not at random in Triticeae genomes (Barakat *et al*. 1997, Erayman *et al*. 2004, Varshney *et al*. 2006). Thus, selecting BAC clones based on the known presence of genes possibly increases the chance to select clones from a region with favourable gene densities, which may allow targeting specifically the gene space of a Triticeae genome for larger-scale sequencing.

Outlook and conclusions

Finishing the sequences of the first higher eukaryote genomes boosted similar attempts in a number of other eukaryotic organisms including crop plants. For more than 5 years a systematic and accelerated increase in the availability of EST sequence information has provided very comprehensive knowledge concerning the gene content in the very important crop species wheat and barley. Furthermore, detailed information on genome structure and organization has been accumulated for a number of important agronomic loci. A more comprehensive understanding of general Triticeae genome organization, and the forces and mechanisms involved in shaping them during evolution, will become available only after unbiased sequencing of larger parts of their homoeologous genomes. As a prerequisite the construction of physical maps that are needed as a backbone for larger-scale and genetically anchored genome sequencing have been initiated for individual wheat chromosomes (Paux *et al*. 2006), the D-genome of *A. tauschii* (J. Dvorak & M. Luo unpublished data, http://wheatdb.ucdavis.edu:8080/wheatdb/) as well as the barley genome as a whole (Stein *et al*. 2006). In addition sequencing a further ~400 BAC from wheat has been initiated to cover either larger contiguous regions for trait isolation or for sampling independent sites on chromosome 3B of hexaploid wheat (C. Feuillet, personal communication). Another 100 BAC originating from 13 loci of the A, B, and D genomes of diploid, tetraploid, and hexploid wheat will be sequenced to study mechanisms of polyploidization (http://www.genoscope.cns.fr/externe/English/Projets/Projet_LE/LE.html). Other projects are under way to perform comparative sequencing at orthologous loci on the homoeologous chromosomes of group 3 (see projects at http://www.wheatgenome.org). Still costs for sequencing larger parts of the Triticeae genomes, where individual chromosomes can be twice the size of the rice genome, are enormous. Only a dramatic advancement in sequencing technology, combined with a quantum leap decrease in sequencing costs, will make the sequencing of large plant genomes a feasible task. Developing new sequencing concepts and technology has become a very mobile field, as the race to sequence a genome for $1000 has been initiated (recently reviewed in: Service 2006, Shendure *et al*. 2004, Metzker 2005). Some of these innovations may also become relevant for Triticeae research, since high-throughput pyrosequencing (454 sequencing, Margulies *et al*. 2005b) also proved to be applicable to determine genomic sequences of the complex barley genome (Wicker *et al*. 2006). Thus the vision of international consortia such as IWGSC, ETGI and IBSC may become true, and the genomes of wheat and barley will become sequenced in large part within the next decade.

Acknowledgements

I deeply acknowledge Drs C. Feuillet and T. Wicker for critical and helpful comments and suggestions on the manuscript.

References

Adams MD, Celniker SE, Holt RA *et al*. (2000) The genome sequence of *Drosophila melanogaster*. *Science* **287**: 2185–2195.

Akhunov E, Akhunova A, Dvorak J (2005) BAC libraries of *Triticum urartu*, *Aegilops speltoides* and *Ae. tauschii*, the diploid ancestors of polyploid wheat. *Theor Appl Genet* **111**: 1617–1622.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Anderson OD, Rausch C, Moullet O, Lagudah ES (2003) The wheat D-genome HMW-glutenin locus: BAC sequencing gene distribution, and retrotransposon clusters. *Funct Integr Genomics* **3**: 56–68.

Aparicio S, Chapman J, Stupka E *et al*. (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**: 1301–1310.

Ayele M, Haas BJ, Kumar N *et al*.(2005) Whole genome shotgun sequencing of *Brassica oleracea* and its application to gene discovery and annotation in *Arabidopsis*. *Genome Res* **15**: 487–495.

Barakat A, Carels N, Bernardi G (1997) The distribution of genes in the genomes of Gramineae. *Proc Natl Acad Sci USA* **94**: 6857–6861.

Bennett MD, Smith JB (1976) Nuclear DNA amounts in angiosperms. *Phil Trans R Soc Lond B* **274**: 227–274.

Bernal A, Ear U, Kyrpides N (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res* **29**: 126–127.

Boffelli D, McAuliffe J, Ovcharenko D *et al*. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**: 1391–1394.

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.

Brunner S, Keller B, Feuillet C (2003) A large rearrangement involving genes and low-copy DNA interrupts the microcollinearity between rice and barley at the *Rph7* locus. *Genetics* **164**: 673–683.

Caldwell KS, Langridge P, Powell W (2004) Comparative sequence analysis of the region harboring the hardness locus in barley and its colinear region in rice. *Plant Physiol* **136**: 3177–3190.

Chantret N, Salse J, Sabot F *et al*. (2005) Molecular basis of evolutionary events that shaped the *Hardness* locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). *Plant Cell* **17**: 1033–1045.

Chao S, Lazo GR, You F *et al*. (2006) Use of a large-scale Triticeae expressed sequence tag resource to reveal gene expression profiles in hexaploid wheat (*Triticum aestivum* L.). *Genome* **49**: 531–544.

Cho S, Garvin DF, Muehlbauer GJ (2006) Transcriptome analysis and physical mapping of barley genes in wheat–barley chromosome addition lines. *Genetics* **172**: 1277–1285.

Close TJ, Wanamaker SI, Caldo RA *et al*. (2004) A new resource for cereal genomics: 22K barley GeneChip comes of age. *Plant Physiol* **134**: 960–968.

Devos KM (2005) Updating the 'crop circle'. *Curr Opin Plant Biol* **8**: 155–162.

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. *Proc Natl Acad Sci USA* **102**: 19243–19248.

Dubcovsky J, Ramakrishna W, SanMiguel PJ *et al*. (2001) Comparative sequence analysis of colinear barley and rice bacterial artificial chromosomes. *Plant Physiol* **125**: 1342–1353.

Duret L, Bucher P (1997) Searching for regulatory elements in human noncoding sequences. *Curr Opin Struct Biol* **7**: 399–406.

Emberton J, Ma J, Yuan Y, SanMiguel P, Bennetzen JL (2005) Gene enrichment in maize with hypomethylated partial restriction (HMPR) libraries. *Genome Res* **15**: 1441–1446.

Erayman M, Sandhu D, Sidhu D, Dilbirligi M, Baenziger PS, Gill KS (2004) Demarcating the gene-rich regions of the wheat genome. *Nucleic Acids Res* **32**: 3546–3565.

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.

Gaut BS (2002) Evolutionary dynamics of grass genomes. *New Phytol* **154**: 15–28.

Goff SA, Ricke D, Lan TH *et al*. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). *Science* **296**: 92–100.

Griffiths S, Sharp R, Foote TN *et al*. (2006) Molecular characterization of *Ph1* as a major chromosome pairing locus in polyploid wheat. *Nature* **439**: 749–752.

Gu YQ, Anderson OD, Londeore CF, Kong XY, Chibbar RN, Lazo GR (2003) Structural organization of the barley *D-hordein* locus in comparison with its orthologous regions of wheat genomes. *Genome* **46**: 1084–1097.

Gu YQ, Salse J, Coleman-Derr D *et al*. (2006) Types and rates of sequence evolution at the high-molecular-weight glutenin locus in hexaploid wheat and its ancestral genomes. *Genetics* **174**: 1493–1504.

Guyot R, Keller B (2004) Ancestral genome duplication in rice. *Genome* **47**: 610–614.

Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431**: 931.

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

Isidore E, Scherrer B, Chalhoub B, Feuillet C, Keller B (2005) Ancient haplotypes resulting from extensive molecular rearrangements in the wheat A genome have been maintained in species of three different ploidy levels. *Genome Res* **15**: 526–536.

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 Mol Biol* **54**: 55–69.

Lamoureux D, Peterson DG, Li W, Fellers JP, Gill BS (2005) The efficacy of Cot-based gene enrichment in wheat (*Triticum aestivum* L.). *Genome* **48**: 1120–1126.

Lazo GR, Chao S, Hummel DD *et al*. (2004) Development of an expressed sequence tag (EST) resource for wheat (*Triticum aestivum* L.): EST generation, unigene analysis, probe selection and bioinformatics for a 16,000-locus bin-delineated map. *Genetics* **168**: 585–593.

Li W, Zhang P, Fellers JP, Friebe B, Gill BS (2004) Sequence composition, organization, and evolution of the core Triticeae genome. *Plant J* **40**: 500–511.

Margulies EH, Vinson JP, Program NCS *et al*. (2005a) An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc Natl Acad Sci USA* **102**: 4795–4800.

Margulies M, Egholm M, Altman WE *et al*. (2005b) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–377.

Messing J, Bharti AK, Karlowski WM *et al*. (2004) Sequence composition and genome organization of maize. *Proc Natl Acad Sci USA* **101**: 14349–14354.

Metzker ML (2005) Emerging technologies in DNA sequencing. *Genome Res* **15**: 1767–1776.

Miller A, Galiba G, Dubcovsky J (2006) A cluster of 11 *CBF* transcription factors is located at the frost tolerance locus *Fr-A^{m}2* in *Triticum monococcum*. *Mol Genet Genomics* **275**: 193–203.

Nasuda S, Kikkawa Y, Ashida T, Islam AKMR, Sato K, Endo TR (2005) Chromosomal assignment and deletion mapping of barley EST markers. *Gen Genet Syst* **80**: 357–366.

Ogihara Y, Mochida K, Nemoto Y *et al*. (2003) Correlated clustering and virtual display of gene expression patterns in the wheat life cycle by large-scale statistical analyses of expressed sequence tags. *Plant J* **33**: 1001–1011.

Palmer LE, Rabinowicz PD, O'Shaughnessy AL *et al*. (2003) Maize genome sequencing by methylation filtration. *Science* **302**: 2115–2117.

Paterson AH (2006) Leafing through the genomes of our major crop plants: strategies for capturing unique information. *Nat Rev Genet* **7**: 174.

Paux E, Roger D, Badaeva E *et al*. (2006) Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. *Plant J* (In press).

Peterson DG, Schulze SR, Sciara EB *et al*. (2002) Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. *Genome Res* **12**: 795–807.

Qi LL, Echalier B, Chao S *et al*. (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**: 701–712.

Rabinowicz PD (2003) Constructing gene-enriched plant genomic libraries using methylation filtration technology. *Meth Mol Biol* **236**: 21–36.

Rabinowicz PD, Bennetzen JL (2006) The maize genome as a model for efficient sequence analysis of large plant genomes. *Curr Opin Plant Biol* **9**: 149–156.

Rabinowicz PD, Citek R, Budiman MA *et al*. (2005) Differential methylation of genes and repeats in land plants. *Genome Res* **15**: 1431–1440.

Ramakrishna W, Dubcovsky J, Park YJ *et al*. (2002) Different types and rates of genome evolution detected by comparative sequence analysis of orthologous segments from four cereal genomes. *Genetics* **162**: 1389–1400.

Rostoks N, Park Y, Ramakrishna W *et al*. (2002) Genomic sequencing reveals gene content, genomic organization, and recombination relationships in barley. *Funct Integr Genomics* **2**: 51–59.

Rota M, Sorrells M (2004) Comparative DNA sequence analysis of mapped wheat ESTs reveals the complexity of genome relationships between rice and wheat. *Funct Integr Genomics* **4**: 34–46.

Rudd S (2003) Expressed sequence tags: alternative or complement to whole genome sequences? *Trends Plant Sci* **8**: 321.

SanMiguel PJ, Ramakrishna W, Bennetzen JL, Busso C, Dubcovsky J (2002) Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome $5A^m$. *Funct Integr Genomics* **2**: 70–80.

Scherrer B, Isidore E, Klein P *et al*. (2005) Large intraspecific haplotype variability at the *Rph7* locus results from rapid and recent divergence in the barley genome. *Plant Cell* **17**: 361–374.

Service RF (2006) Gene sequencing: the race for the $1000 genome. *Science* **311**: 1544–1546.

Shendure J, Mitra RD, Varma C, Church GM (2004) Advanced sequencing technologies: methods and goals. *Nature Rev Genet* **5**: 335–343.

Smith DB, Flavell RB (1975) Characterisation of the wheat genome by renaturation kinetics. *Chromosoma* **50**: 223–242.

Sorrells ME, La Rota M, Bermudez-Kandianis CE *et al*.(2003) Comparative DNA sequence analysis of wheat and rice genomes. *Genome Res* **13**: 1818–1827.

Springer NM, Xu X, Barbazuk WB (2004) Utility of different gene enrichment approaches toward identifying and sequencing the maize gene space. *Plant Physiol* **136**: 3023–3033.

Stein N, Graner A (2004) Map-based gene isolation in cereal genomes. In Gupta P, Varshney R, eds., *Cereal Genomics*. Dordrecht: Kluwer, pp. 331–360.

Stein N, Prasad M, Scholtz U *et al*. (2006) A 1000 loci transcript map of the barley genome - new anchoring points for integrative grass genomics. *Theor Appl Genet* (in press).

Szücs P, Karsai I, Zitzewitz Jv *et al*. (2006) Positional relationships between photoperiod response QTL and photoreceptor and vernalization genes in barley. *Theor Appl Genet* **112**: 1277–1285.

The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815.

The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**: 2012–2018.

The Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**: 69–87.

The Rat Genome Sequencing Project Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**: 493–521.

Thomas JW, Touchman JW, Blakesley RW *et al*. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**: 788–793.

Turner A, Beales J, Faure S, Dunford RP, Laurie DA (2005) The pseudo-response regulator *Ppd-H1* provides adaptation to photoperiod in barley. *Science* **310**: 1031–1034.

Varshney RK, Grosse I, Haehnel U *et al*. (2006) Genetic mapping and BAC assignment of EST-derived SSR markers shows non-uniform distribution of genes in the barley genome. *Theor Appl Genet* **113**: 239–250.

Venter JC, Smith HO, Hood L (1996) A new strategy for genome sequencing. *Nature* **381**: 364.

Wei FS, Wong RA, Wise RP (2002) Genome dynamics and evolution of the *Mla* (powdery mildew) resistance locus in barley. *Plant Cell* **14**: 1903–1917.

Whitford R, Baumann U, Sutton T *et al*. (2007) Identification of transposons, retroelements, and a gene family predominantly expressed in floral tissues in chromosome 3DS of the hexaploid wheat progenitor *Aegilops tauschii*. *Funct Integr Genomics* **7**: 37–52.

Wicker T, Matthews D, Keller B (2002) TREP, a database for Triticeae repetitive elements. *Trends Plant Sci* **7**: 561–562.

Wicker T, Schlagenhauf E, Graner A *et al*. (2006) 454 sequencing put to the test using the complex genome of barley. *BMC Genomics* **7**: 275.

Wicker T, Stein N, Albar L, Feuillet C, Schlagenhauf E, Keller B (2001) Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. *Plant J* **26**: 307–316.

Wicker T, Yahiaoui N, Guyot R *et al*. (2003) Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and A(m) genomes of wheat. *Plant Cell* **15**: 1186–1197.

Wicker T, Zimmermann W, Perovic D *et al*. (2005) A detailed look at 7 million years of genome evolution in a 439 kb contiguous sequence at the barley *Hv-eIF4E* locus: recombination, re-arrangements and repeats. *Plant J* **41**: 184–194.

Yan L, Loukoianov A, Blechl A *et al*. (2004) The wheat *VRN2* gene is a flowering repressor down-regulated by vernalization. *Science* **303**: 1640–1644.

Yan L, Loukoianov A, Tranquilli G, Helguera M, Fahima T, Dubcovsky J (2003) Positional cloning of the wheat vernalization gene *VRN1*. *Proc Natl Acad Sci USA* **100**: 6263–6268.

Yu J, Hu SN, Wang J *et al*. (2001) A draft sequence of the rice (*Oryza sativa* ssp. *indica*) genome. *Chin Sci Bull* **46**: 1937–1942.

Zhang H, Sreenivasulu N, Weschke W *et al*. (2004) Large-scale analysis of the barley transcriptome based on expressed sequence tags. *Plant J* **40**: 276–290.