Gene and repetitive sequence annotation in the Triticeae

Wicker, T ; Buell, C R

DOI: https://doi.org/10.1007/978-0-387-77489-3_15

Originally published at:
Wicker, T; Buell, C R (2009). Gene and repetitive sequence annotation in the Triticeae. In: Feuillet, C; Muehlbauer, G J. Genetics and Genomics of the Triticeae. New York, US: Springer, 407-426.
DOI: https://doi.org/10.1007/978-0-387-77489-3_15

Gene and Repetitive Sequence Annotation in the Triticeae

Thomas Wicker[1] and C. Robin Buell[2]
[1] Thomas Wicker, Institute of Plant Biology, University Zurich, Zollikerstrasse 107, CH-8008 Zurich; Email: wicker@botinst.uzh.ch
[2] C. Robin Buell, Department of Plant Biology, Michigan State University, East Lansing MI 48824 USA; Email: buell@msu.edu

Abstract. The Triticeae tribe contains some of the world's most important agricultural crops (wheat, barley and rye) and is perhaps, one of the most challenging for genome annotation because Trticeae genomes are primarily composed of repetitive sequences. Further complicating the challenge is the polyploidy found in wheat and particularly in the hexaploid bread wheat genome. Genomic sequence data are available for the Triticeae in the form of large collections (>1 million) of Expressed Sequence Tags and an increasing number of bacterial artificial chromosome clone sequences. Given that high repetitive sequence content in the Triticeae confounds annotation of protein-coding genes, repetitive sequences have been identified, annotated, and collated into public databases. Protein coding genes in the Triticeae are structurally annotated using a combination of *ab initio* gene finders and experimental evidence. Functional annotation of protein coding genes involves assessment of sequence similarity to known proteins, expression evidence, and the presence of domain and motifs. Annotation methods and tools for Triticeae genomic sequences have been adapted from existing plant genome annotation projects and were designed to allow for flexibility of single sequence annotation while allowing a whole community annotation effort to be developed. With the availability of an increasing number of annotated grass genomes, comparative genomics can be exploited to accelerate and enhance the

quality of Triticeae sequences annotation. This chapter provides a brief overview of the Triticeae genomes features that are challenging for genome annotation and describes the resources and methods available for sequence assembly and annotation with a particular emphasis on problems caused by the repetitive fraction of these genomes.

2.1 Triticeae Genomics

Although the Triticeae contains some of the world's most important agricultural crops, this group of plants have only begun to enter the genomics era. This is not due to a lack of interest or need for genomics of Triticeae species. It results from the technical challenges of obtaining the genomic sequence from large, repetitive and sometimes polyploid species. The genome of hexaploid, or bread wheat (T*riticum aestivum* L., 2N=6X=42), is reported to be 16 Gb (Arumuganathan and Earle 1991) and to contain more than 90 % repetitive sequences (Li et al. 2004) thereby presenting limitations primarily fiscal in nature, to current sequencing methodologies. Barley (*Hordeum vulgare* L., 2N=2X=14) is diploid and has a genome size comparable to that of diploid wheat (5.7 Gb (Bennett and Smith 1976)) with a similar content of repetitive DNA (Smith DB and RB 1975). In the past decade, however, the development of new genomic resources such as bacterial artificial chromsome (BAC) libraries, large collections of markers including Expressed Sequence Tags (ESTs; see chapter 2.1) have allowed the establishment of robust genomics programs in the Triticeae including a wheat and a barley genome sequencing initiative (see chapter 4.4). The International Wheat Genome Sequencing Consor-

tium (IWGSC; http://www.wheatgenome.org/)) initiative is focusing its effort on hexaploid wheat, specifically the cultivar Chinese Spring (Gill et al. 2004), and members of the initiative have already generated a number of resources (physical contigs) that allow targeted genome sequencing. The International Barley Sequencing Consortium (IBSC; http://barleygenome.org) was also launched to develop genomic resources for genome sequencing of cultivar Morex. For more details on the genome sequencing initiatives for the Triticeae see chapter 4.4.

The ideal outcome of a whole genome annotation effort would be a set of genes accurately identified with information about their location on linkage maps and their putative functions. Ancillary annotations such as expression patterns, promoter sequences, orthologous and paralogous sequences are also informative for biologists, breeders, and geneticists but they are not part of a "core" genome annotation. The foundation of an annotation project is the accurate identification of protein coding genes. This is obtained through a combination of computational predictions such as *ab initio* gene finders and through experimental evidence such as transcripts and protein alignments. Accurately weighting these data types and constructing accurate gene models for an entire genome is generally extremely challenging and becomes a major issue for genomes as complex and large as those of the Triticeae. Thus, successful genome annotation projects result in different gene subsets ranging from well annotated genes (i.e., genes with full length cDNA support) to reasonably annotated (i.e., genes with EST and/or protein support) and genes annotated with less confi-

dence (i.e., genes predicted solely by an *ab inito* gene finder). With respect to functional annotation of large genomes, putative function is primarily assigned through sequence similarity with other sequenced genomes which is highly prone to transitive annotation errors. This can be addressed by manual curation or through annotation of functional domains such as Pfam domains (Finn et al. 2006) rather relying on "best hits" to a large non-redundant amino acid database of primarily uncurated entries (UniProt_Consortium 2007). The high repetitive sequence content of the Triticeae genomes complicates the annotation process in two ways: First because of their abundance it contributes in a significant manner to the bulk of sequence that needs to be processed during the annotation phase and second because some of the repetitive elements are expressed and have features of protein coding genes that can confound gene annotation efforts. Efforts have already begun to address these challenges by developing adequate and efficient bioinformatics tools and resources for interpretation of the Triticeae genome sequences and to ensure a large accessibility to the scientific community..

2.2 Triticeae Genome Sequence and Annotation Data

2.2.1 The Triticeae Transcriptome

ESTs provide a rapid form of gene discovery as they represent the genic portions of the genomes thereby bypassing the large tracts of genome sequence that does not encode for RNA or proteins. They provide a mechanism for gene discovery for species with large and

unsequenced genomes such as those of the Triticeae. In 1998, the Triticeae community established an international collaborative network, the International Triticeae EST Cooperative (ITEC, http://wheat.pw.usda.gov/genome/) to produce large collections of ESTs. To date about 1.6 million of ESTs for wheat, barley and rye are present in the EST database at NCBI (http://www.ncbi.nlm.nih.gov/dbEST/index.html). The May 2008 release of dbEST (050208) contained 1,051,465 ESTs from *Triticum aestivum* (bread wheat), 478,682 ESTs from *Hordeum vulgare*, 17,381 ESTs from *Triticum turgidum* subsp. *durum* (durum wheat), 10,139 ESTs from *Triticum monococcum*, 1,938 ESTs from *Triticum turgidum*, and 4,315 ESTs from *Aegilops speltoides*. As these ESTs are primarily derived from non-normalized cDNA libraries, redundancy is rampant making them difficult to work with on an individual basis. Thus, these ESTs, along with cloned mRNAs and cDNAs, are typically clustered and assembled into a smaller, representative set of transcripts (unigenes, transcript assemblies, tentative consensus sequences) prior to their use by biologists or bionformaticians. A number of laboratories provide these clustered assemblies as part of their resource efforts including PlantGDB (http://www.plantgdb.org/prj/ESTCluster/index.php), Dana Farber Gene Indices (http://compbio.dfci.harvard.edu/tgi/), the TIGR Plant Transcript Assemblies (http://plantta.tigr.org/), HarvEST (http://harvest.ucr.edu/) and GenoplanteDB (http://urgi.versailles.inra.fr/GnpSeq/). Gramene (http://gramene.org),

a major database for research on grasses, has also mapped EST assemblies to the rice and maize genomes.

The availability of the Triticeae EST collections has allowed for studies on the Triticeae transcriptome (Chao et al. 2006; Houde et al. 2006; Kawaura et al. 2005; Laudencia-Chingcuanco et al. 2006; Mochida et al. 2006; Ogihara et al. 2004), the development of bin-mapped markers for wheat genetic mapping (Conley et al. 2004; Hossain et al. 2004; Lazo et al. 2004; Linkiewicz et al. 2004; Miftahudin et al. 2004; Munkvold et al. 2004; Peng et al. 2004; Qi et al. 2004; Randhawa et al. 2004), EST maps in barley (Stein et al, 2007), and comparative studies of the syntenic relationships between wheat and rice (Conley et al. 2004; Francki et al. 2004; La Rota and Sorrells 2004; Linkiewicz et al. 2004; Peng et al. 2004; Salse et al. 2008; See et al. 2006; Sorrells et al. 2003).

ESTs, along with full-length cDNA clones, are valuable not only for gene discovery but also for empirical evidence that can be used in structural annotation of genomic sequences. The optimal transcript resource is a set of full length (FL) cDNA sequences that provide a complete representation of the full 5' and 3' untranslated regions and the precise location of intron/exon splice junctions for an unambiguous annotation of the gene structure. They have been instrumental in the annotation of genome sequences, including Arabidopsis and rice (Castelli et al. 2004; Haas et al. 2003; Ohyanagi et al. 2006; Ouyang et al. 2007; Tanaka et al. 2008). In addition to their use in structural an-

notation, the cDNA clone from which the EST or FLcDNA sequence is derived is highly desirable as a resource for functional genomics studies such as overexpression studies. Several thousands of full length cDNA sequences are available for wheat and barley. A query of Genbank (May 2, 2008) revealed 1,980 and 5,504 full length cDNA sequences for wheat and barley, respectively. A project is also in progress to generate full length cDNAs for Chinese Spring, the hexaploid wheat cultivar selected for genome sequencing by the International Wheat Genome Sequencing consortium (IWGSC) (see chapter 4.4). To date, ~4,200 full length cDNA sequences have been produced (http://wheat.pw.usda.gov/ITMI/ITMI2005_Proceedings/Abstracts/Ogihara.html). A similar project is in progress for barley (http://www.shigen.nig.ac.jp/barley/).

2.2.2 The Triticeae Genomes

Both the wheat and barley communities are pursuing BAC-based sequencing initiatives to obtain the genome sequence (see chapter 4.4) after physical maps have been established (see chapter 2.3). In the near future, the sequence of ~200 BACs randomly selected from Chinese Spring, will be made available as part of a survey of the wheat genome landscape (Devos et al. 2005). Targeted sequencing of wheat chromosome or chromosome arm specific BAC libraries are underway including chromosome 3B (Gill et al. 2004; Paux et al. 2006) (http://www.international.inra.fr/research/some_examples/sequencing_the_wheat_genome) and 3AS (http://wheat.tigr.org/tdb/e2k1/tae1/). A continually updated list of bread wheat sequencing activities can be

seen on The IWGSC web page (http://www.wheatgenome.org/). These random as well as targeted sequencing projects will provide ample sequence for optimization and improvement of genome annotation tools. Indeed, to date, genome sequence (excluding ESTs) is available for 67.8 Mb of *Triticum* species in Genbank including 30.3 Mb from the HTG division (draft sequences of BACs), 27.1 Mb from the GSS division (single pass sequences, typically end sequences of BACs), and 10.4 Mb from the PLN division (finished sequence). For *Hordeum* species, 21.8 Mb of genome sequence (excluding ESTs) is available in Genbank including 216.5 Kb from the HTG division, 1.3 Mb from the GSS division, and 20.3 Mb from the PLN division.

2.2.3 Genome Annotation: Structural and Functional Annotation

In the framework of the IWGSC, a working group of Triticeae biologists and bioinformaticians has been established to set up guidelines and develop a community effort for annotating the Triticeae genomic sequences. The guideline is focused on establishing a minimum set of annotations and processes that is provided to the research community for accurate, homogeneous and insightful interpretation of the sequence. The current focus of the guideline is structural annotation and to a smaller extent functional annotation with putative functions. The guidelines, summarized below, are available at the IWGSC web site (http://www.wheatgenome.org/tool.html).

The starting point for annotating Triticeae genomic sequences is the identification and annotation of repetitive sequences that compose

most of the Triticeae genomes (>80%). Their composition and identification are described in the section below. Following identification, the repetitive sequences are then "masked" to prevent them from confounding identification of protein coding genes. Genes are identified in the repeat-masked sequence using *ab initio* gene finders. Although multiple gene finders can be used, at a minimum, FGENESH (monocot matrix; (Salamov and Solovyev 2000)) must be run on the sequence. The sensitivity and specificity of various gene finders on wheat sequences have not been compared and documented although anecdotal evidence suggests that FGENESH is the most accurate *ab initio* gene finder currently available. Gene structure can be improved using transcript and protein evidence to construct an improved gene model. Nomenclature of the transcriptional unit and loci are outlined in the IWGSC annotation guideline. Standardization of the nomenclature, even at the early stages of a genome effort, is essential to minimizing population of databases with genes, gene models, and transcripts with divergent annotations.

Putative function for the protein encoded in a gene model is determined based on either the presence of a Pfam domain or through sequence similarity evidence. A gene model can be annotated as encoding a "known", "putative", "XX-domain containing", "expressed", "conserved hypothetical" or "hypothetical" protein depending on the extent of sequence similarly detected. For annotating a gene model as encoding a "known" protein, high sequence similarity (>90-100% identity and coverage) to a characterized protein within an amino acid

database such as UniProt (Suzek et al. 2007) must be detected. Expression evidence in the form of alignment to an EST, cDNA or mRNA is optional, additional annotation, for the gene model. When a lower level of similarity with an entry in an amino acid database (>45% identity, > 50% coverage), and thereby a lower confidence, is observed, the gene model is annotated as encoding a "putative XX" protein. Again, expression evidence is an optional yet informative layer of annotation. For gene models encoding proteins that lack similarity to an entry in an amino acid database but have a Pfam domain above the trusted cutoff, the gene model is annotated as encoding a "XX-domain containing protein. Here, although expression evidence is optional, its availability is highly informative for deducing the function of the gene model. Sometimes, gene models can have strong sequence similarity with proteins in the amino acid database without known function. In this case, they are referred to as expressed or hypothetical genes. Triticeae genes that match such an entry (>45% identity, >50% coverage) and lack sequence similarity with Triticeae ESTs, mRNAs, or cDNAs (<95% ID, <70% length), are annotated as encoding "conserved hypothetical proteins". For the gene models that lack substantial sequence similarity (>45% identity, >50% coverage) with a known or putative protein entry in an amino acid database as well as a Pfam domain over the trusted cutoff, but have sequence identity to an EST, cDNA, or mRNA (>95% ID, >70% length), the gene model is annotated as encoding an "expressed protein". When a gene model lacks any sequence similarity (>45% identity, >50% coverage) with an entry in an amino acid database or with an EST, cDNA or mRNA, the gene

model is annotated as encoding a “hypothetical protein”. The availability of transcript support is highly valuable as this is empirical evidence that the gene is transcribed thereby providing more confident annotation than that of hypothetical gene models.

According to the guidelines, additional annotations should be made for Triticeae genes. For example, the top match to the predicted rice and Arabidopsis proteomes should be provided. The rationale behind this is to provide links to well characterized plant genomes in which not only a complete genome sequence and genome annotation datasets are available, but functional resources and data are available to test hypothesis regarding the function of the wheat or barley homolog.

Annotation can be done manually, semi-automatically or automatically. A large factor in determining the approach is the available manpower and the level of quality of annotation desired. Certainly, manual annotation provides a high quality of interpretation as individual evidence can be weighted and new data from the literature or expert knowledge can be evaluated and incorporated on an ad hoc basis. However, manual annotation is it is very time consuming and cannot be envisaged for the Triticeae genomes. Thus, the majority of annotation for the wheat and barley genomes will be automated or semi-automated. This is similar to the trends in a number of plant genome projects in which the genome has been annotated using automated and semi-automated methods with targeted curation of genes and gene families (Jaillon et al. 2007; Ouyang et al. 2007; Tuskan et al. 2006).

For wheat, a semi-automated publicly available annotation pipeline, the TriAnnot pipeline (http://urgi.versailles.inra.fr/projects/TriAnnot/) has been established and proposed for the semi-automated annotation of Triticeae genomic sequences. It proposes an annotation of wheat and barley BAC sequences through gene and transposon prediction and modeling. Through a simple submission process, users can submit their BAC for annotation by the TriAnnot pipeline. Annotation output is provided in a number of formats for downstream analysis including editing in graphical viewers.

Clearly, annotation of Triticeae genomic sequences is in its infancy. As more genome sequence becomes available, training sets (genomic DNA and cognate full length cDNA sequences) will be available allowing for training and improvement of *ab initio* gene finders. Consequently, better characterization and cataloguing of Triticeae repetitive elements will allow for refinement of the gene space and reduce contamination of the gene complement with transposable elements (TEs). However, perhaps the greatest improvement in wheat genome annotation will be from comparative alignments with genome sequences from other Poaceae species. In addition, all annotation is iterative in nature and even for Arabidopsis, in which all of the genes were manually curated (Arabidopsis_Genome_Initiative 2000), the genome is continually re-annotated as new evidence types and computational methods become available ((Haas et al. 2005), http://arabidopsis.org/). Thus, with the large size of the wheat and barley genomes, it will be important that efficient automated/semi-automated annotation pipe-

lines are established which can handle the large sequences such as pseudomolecules as optimal annotation is performed on a genome-scale, not on a small representation (~120 kb BAC size) scale.

2.2.4 Comparative Genome Annotation

For genes, comparative genome annotation is a powerful tool because it can highlight conserved and diverged features among genomes. The availability of the complete rice genome sequence along with genomic sequences of several hundred kb from wheat and barley has allowed comparison between these genomes at the sequence level providing data on the degree of conservation between the grass genomes (Bossolini et al. 2007; Dubcovsky et al. 2001; Griffiths et al. 2006) (see chapter 2.7). In the next few years, the sequence of multiple Poaceae species will be available (Table 1) and this will provide important resources for improving genome annotation in this family. This has already been seen in annotation of the rice genome (Zhu and Buell 2007). The use of comparative alignments between rice, maize and sorghum provide information that can 1) improve the structural annotation due to sequence conservation of coding regions, 2) increase confidence of gene predictions in which no transcript support is available, and 3) provide new evidence for functional annotation as inferences can be drawn from experimental and literature reports between orthologous genes. In a comparison of Brachypodium with rice (Bossolini et al. 2007), the annotation of both rice and Brachypodium could be improved through such comparative analyses. For rice, seven

of the 47 annotated genes could be updated in their structure based on comparative alignments with the collinear Brachypodium sequence (Bossolini et al. 2007). Recent reports of comparative analyses between wheat and Brachypodium have confirmed the close relationship between genes, gene structure and gene order within the Pooideae (Bossolini et al. 2007; Griffiths et al. 2006). The availability of the Brachypodium sequence in the near future (see chapter 4.5), will greatly facilitate efforts in understanding the Triticeae genomes structure and composition

Table 1. List of Poaceae species with genome sequence and/or pending genome sequence (International_Rice_Genome_Sequencing_Project 2005).

Species	Tribe	Reference
Oryza sativa (rice)	Ehrhartoideae	IRGSP, 2005
Zea mays (maize)	Panicoideae	http://www.maizesequence.org
Brachypodium distachyon	Pooideae	http://www.jgi.doe.gov/sequencing/why/CSP2007/brachypodium.html
Sorghum bicolor	Panicoideae	http://www.phytozome.net/sorghum
Setaria italica (foxtail millet)	Panicoideae	http://jgi.doe.gov/sequencing/why/CSP2008/foxtailmillet.html

2.3 Repetitive Sequences in the Triticeae

2.3.1 Methods for the identification of transposable elements

The easiest way to identify transposable elements (TEs) is by BLAST (Basic Local Alignment Search Tool) search of the sequence of inter-

est against a database containing known TE sequences. BLAST searches can be done at the DNA (BLASTN) or protein level (BLASTX). BLASTN helps identify closely related TEs which belong to the same family. Usually, the entire element (coding and non-coding parts) can be detected that way. If a TE is more divergent and does not belong to a family already present in the databases, a BLASTX search can help identify protein coding regions and thus allow determination to which superfamily the TE belongs. The non-coding portions of the TE cannot be characterized by BLASTX and other methods have to be used to determine the exact borders of the element (see below). The ability to identify TEs by BLAST entirely depends on the completeness of the TE database. Whenever a novel repeat is present on the sequence, it will remain undetected.

De novo detection of repeats is more labour intensive and requires a great expertise in the structure and characteristics of repeats. However, it is an important process because as soon as one member of a family is newly identified and characterized, it can be added to the existing databases and further copies of that family can then be identified by sequence comparisons. *De novo* detection is mainly done by searching for coding sequences that are similar to those of known TEs and identification of terminal repeat sequences. Coding sequences are again identified by BLASTX against a series of databases which can (and should) also include animal, fungal and bacterial sequences.

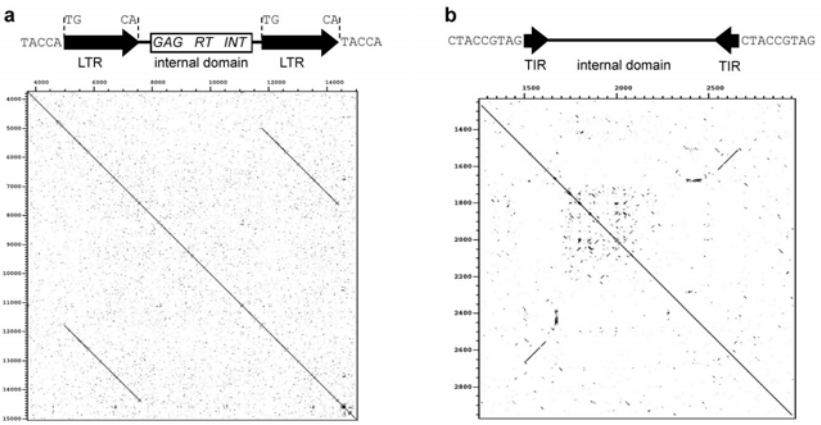

Fig. 1. Identification of TEs based on structural characteristics. The structure of the TE is displayed above a DotPlot in which the sequence containing the TE is aligned against itself to visualize repeat structures such as long terminal repeats (LTRs) or terminal inverted repeats (TIRs). A DotPlot is a visual alignment of two sequences, one horizontally and one vertically (the case illustrated here correspond to the alignment of a sequences against itself). The full diagonal line from the top left to the bottom right is the 100% match of the sequence on itself. Other diagonal lines represent repeat structures. Direct repeats (LTRs) are parallel to the main diagonal line while inverted repeats (TIRs) are perpendicular to it. Other diagnostic features such as canonical LTR termini and target site duplications (TSD) are also easy to detect on such representation (with zooming possibilities on specific

regions). a. DotPlot and characteristics of a *BARE1* LTR retrotransposon (5 bp TSD). b. DotPlot of a Mutator transposon with no coding capacity (9 bp TSD).

As mentioned in the chapter 2.8 on the genomics of TEs, many transposable elements are non-autonomous and do not contain any coding regions. It is also possible that a new TE family contains highly divergent coding regions that can not be identified based on homology to known elements. In such cases, the TE has to be identified based on structural characteristics such as their terminal repeat sequences (reviewed by (Wicker et al. 2007)). An efficient tool for this task is a so-called DotPlot (Fig. 1) which aligns two sequences graphically, one on the x-axis and one on the y-axis. Whenever there is a short stretch of homology (e.g. 5 bp), the programs produces a dot at this position, allowing to easily identify long regions of homology. If a sequence is aligned with itself, DotPlot can be used to identify repeat structures within that sequence (e.g. terminal repeats of TEs).

Almost all TE superfamilies create a so-called target site duplication (TSD) when they insert into the genome (Fig. 2). The TSD (also called a “genetic footprint”) is created because the Integrase or Transposase enzymes usually produce staggered ends with overhangs of 2-10 bp.
In the case of LTR retrotransposons, one would search for the presence of direct repeats that are separated from each other by about 3-5 kb (the usual size of an internal domain) and are flanked by a 5 bp TSD

(Figure 1a). Furthermore, LTRs almost always start with TG and end with a CA motif (Fig. 1a). DNA transposons can be identified based on their terminal inverted repeats (Fig 1b) as well as the characteristic length of TSDs which usually range form 2-10 bp, depending on the Superfamily (Fig. 2).

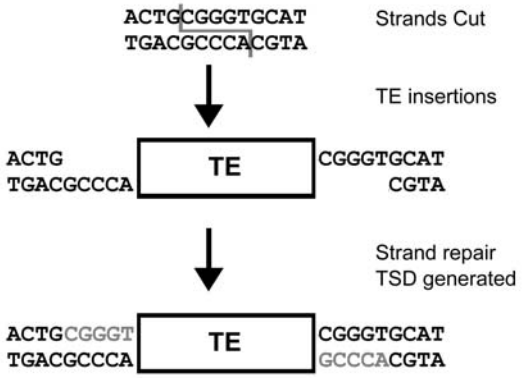

Fig. 2. Creation of a target site duplication (TSD) upon insertion of a TE into the genome.

2.3.2 Problems with transposable elements in Triticeae sequencing

The hundreds of thousands of TE sequences present in the Triticeae genomes have, so far, represented a major barrier to large scale se-

quencing. To date, most of the sequencing has been performed in the framework of map-based cloning projects in which a region of usually 2 to 4 BAC clones is established at the target locus and sequenced by the shotgun-sequencing method. The Sanger sequencing technology which was mostly used until recently only generates sequences of less than 1000 bp and therefore genomic regions have to be divided into smaller fragments for sequencing. Thus, during shotgun sequencing, the BAC DNA is sheared into small fragments of 3-10 kb which are then sequenced individually. Enough fragments are sequenced to reach a total of 8 to 10 times the size of the BAC (referred to as 8 to 10-fold sequencing coverage). The sequenced fragments are then collected together to find overlapping regions, in order to be able to reproduce the original BAC sequence. This process of reconstructing the original sequence is called "sequence assembly". The product of assembled overlapping sequences is called a "sequence contig".

The production of the primary (shotgun) sequence itself is not more labour-intensive in the Triticeae than for any other species. Difficulties arise during assembly of the shotgun reads when repetitive sequences are wrongly pooled into artificial contigs and when the sequence of the remaining gaps has to be determined. It is in this phase (called the "finishing" phase) that the TE sequences cause the problems that make Triticeae sequencing so costly and labour-intensive. As of June 2008, there were 377 Triticeae genomic sequences larger than 20 kb available in the NCBI public database (www.ncbi.nlm.nih.gov). Many of them corresponded to individual BAC sequences that were in an unfinished state mostly because of the difficulty to assemble TE regions . If

a BAC contains several copies of the same TE, they can cause confusion in the assembly as different copies are assembled into the same sequence contig thereby preventing the correct assembly of the whole sequence. The resolve such mis-assemblies, information from forward and reverse reads of the same shotgun clone can be used and detailed TE annotation of the unfinished sequence can provide hints as to the correct linear order of the sequence contigs. Often, the two LTRs of a LTR-retrotransposon cause the same effect as they are pooled into one single LTR consensus sequence while the internal domain is assembled into a separate sequence contig with no apparent connection to the rest of the BAC sequence.

Even if they are present in a single copy on the BAC i.e. they should behave like a normal low-copy sequence, TE can also cause gaps in the BAC sequence because of their sequence composition. For example, the highly repeated TEs of the *BARE1* group (*Angela*, *BARE1* and *WIS*) contain a G/C-rich region within their LTRs that almost in all cases causes sequencing problems. Interestingly, analysis of 16 Angela and *WIS* LTRs from several independent BACs showed that the gaps are all found in similar positions and that the region can be narrowed down to a few dozen base pairs that apparently contain the problematic motif (Fig. 3a). Similarly, many CACTA transposons contain regions that are very difficult to sequence. Most *Caspar* elements, for example, contain an extended region of low-complexity DNA, a GA-rich microsatellites, followed by its reverse complement, a T/C-rich motif (Fig. 3b). Additionally, many *CACTA* elements contain large arrays of

direct repeats with repeat units of dozens to hundreds of bp in size (Fig. 3c, (Wicker et al. 2003)). Because TEs of the *Caspar* of *BARE1* type are found on almost every Triticeae BAC, new bioinformatic tools will be needed for the Triticeae genomes sequencing projects.

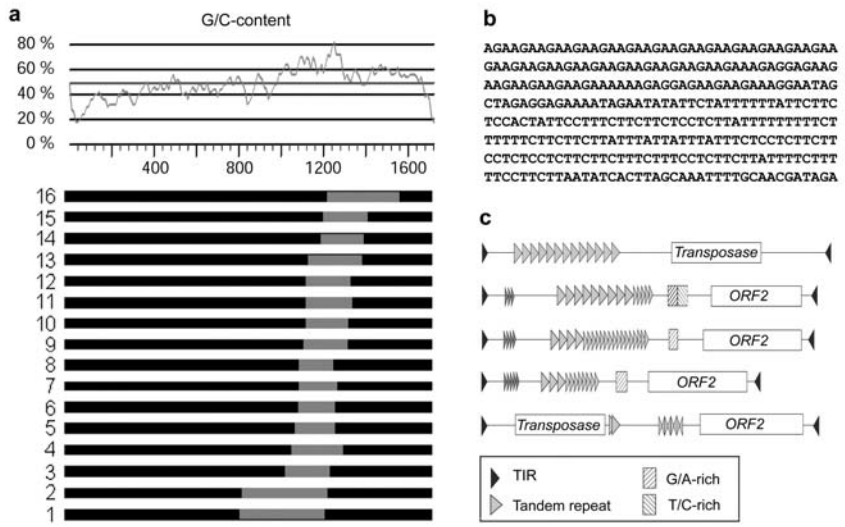

Fig. 3. Problematic sequence in TEs from Triticeae. **a.** Most LTRs of the BARE1-group TEs contain a G/C-rich region in the second half of their LTRs. A G/C-plot along the ~1800 bp of the LTR is shown at the top. Below are 16 LTR sequences obtained in a sequencing project with the positions of gaps in the sequence indicated as gray bars. Note

that the position of the gaps correlates well with the position of the G/C-rich region. **b.** Example of a large low-complexity region from a CACTA transposon of the Caspar family. A long region consisting almost exclusively of G/A is followed by its reverse complement, consisting mainly of T/C. **c.** repeat structures in CACTA transposons. Panel 3c is adapted from Wicker et al., 2003.

2.3.3 Software for repeat recognition and isolation

As the amount of genomic sequences from the Triticeae grows with an increasing speed, bioinformatics tools for efficient identification and annotation of TEs are urgently needed. Currently, a number of programs are available which assist the *de novo* identification of TEs and their annotation. The program LTR_STRUC (McCarthy and McDonald 2003), for example automatically searches a finished sequence (or even an entire genome) for the typical characteristics of LTR-retrotransposons (LTRs etc… as described above). It can be used for an efficient and quick identification of LTR retrotransposons without requiring a lot of specialized knowledge. The disadvantage of that program is that it does not really take into account the possibility of nested insertions i.e., TEs inserted into other TEs that are very frequent in the Triticeae genomes. Another example for automated annotation is the program TEnest (Kronmiller and Wise 2008) which identifies TEs based on a search against a TE database and also models their nesting patterns i.e., the order in which the TEs have inserted into one another. This allows a quick assessment of the genome evolution in a particular locus. However, the main disadvantage of these two programs is that

they work only on largely finished sequences. Further automated TE recognition pipelines have been developed (Bao and Eddy 2002; Quesneville et al. 2005). The former is a de-novo repeat identification software which defines the boundaries of repetitive sequences by multiple sequences alignments of regions that contain particular repeat. The latter employs a "combine evidence" strategy analogous that that used for gene prediction where results from homology based and de-novo TE identification methods are integrated.

Although such programs are very valuable and helpful tools for sequence analysis, one has to consider their outputs with caution. The automated annotation of TE is very complex and many exceptions and special cases are not handled by the programs because the programmer simply did not know about them at the time of development . A typical example is a deletion that eliminates part of a TE. The computer program might then find the first half of the TE and merge it with the second half of a similar TE further downstream. Such an artifact can cause inconsistencies when the evolution of a locus is being analyzed. Even worse, if a gene is located in between the two merged TEs as it can be interpreted as part of the TE if the results are not checked carefully. Thus, every automated annotation of automatically extracted TE dataset should be inspected carefully if one wants to ensure accurate information about TEs.

2.3.4 The challenge of the large number: Quality in quantity is needed

To ensure accurate repeat identification and characterization, it is essential that a high-quality repeat database is available. There are several criteria that define the quality of such a database. A few will be mentioned here:

- The size of the TEs and the structure of their terminal sequence needs to be well identified. This allows exact annotation of the borders of TEs on a given sequence and, thus, efficient making of a considerable fraction of the sequence for further gene identification.

- TEs in the database should not contain nested insertions of other TEs. This can lead to distorted estimates of copy numbers of TEs. If, for example, a low-copy transposon contains an insertion of a MITE which is present in 10,000 copy numbers in the genome, a BLAST searches against the TE database will often hit the high-copy element inside the low-copy one. If the BLAST output is not carefully read, one can gain the false impression of the abundance of the low-copy element.

-TEs in the databases should not contain genes or fragments thereof. Especially when the TE dataset is produced automatically, as described above, there is the danger that it contains artifact TEs which contain genes or gene fragments.

-TEs are often wrongly annotated as genes, since they may contain coding sequences which are not clearly homologous to typical TE pro-

teins such as Transposase of Reverse Transcriptase. Once a TE is wrongly labeled as a gene, the mistake will continue to be carried on, as future researchers, who come across that particular TE will annotate it again as a gene. This can result in potentially large artificial artifactual "gene families".

A number of TE databases have been created over the years, with RepBase being the pioneer (Jurka 2000). Several TE databases for plants have been generated as a result of the complete sequencing of the rice and Arabidopsis genomes. The only database dedicated to Triticeae is TREP (Triticeae repeat database, http://wheat.pw.usda.gov/ITMI/Repeats/). The most recent release contained over 1,400 TEs sequences representing 180 families. Considering the small set of sequences that is publicly available and the vast size of the Triticeae genome, one has to expect that there are thousands of different TE families yet to be discovered. Classification and annotation of such a large number of TEs can only be precise and reliable if a high quality of the repeat database is maintained even when the number of TEs reaches tens of thousands. So far, the TREP database was curated by a very small number of people, thus, providing a relatively consistency in quality. However, the challenges that lie ahead will require the definition of clear guidelines and quality control to provide a system for many dozens or even hundreds of researchers. First steps were taken by creation of the IWGSC annotation guideline and a proposal for a unified classification system for transposable elements [44].

2.4 Acknowledgments

Research in the Buell lab on wheat genomics is supported by the National Research Initiative (NRI) Plant Genome Program of the USDA Cooperative State Research, Education and Extension Service (CSREES).

2.5 Reference List

Arabidopsis_Genome_Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796-815

Arumuganathan K, Earle E (1991) Nuclear DNA content of some important plant species. Plant Molecular Biology Reporter 9:208-218

Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 12:1269-1276

Bennett MD, Smith JB (1976) Nuclear dna amounts in angiosperms. Philos Trans R Soc Lond B Biol Sci 274:227-274

Bossolini E, Wicker T, Knobel PA, Keller B (2007) Comparison of orthologous loci from small grass genomes Brachypodium and rice: implications for wheat genomics and grass genome annotation. Plant J 49:704-717

Castelli V, Aury JM, Jaillon O, Wincker P, Clepet C, Menard M, Cruaud C, Quetier F, Scarpelli C, Schachter V, Temple G, Caboche M, Weissenbach J, Salanoubat M (2004) Whole genome sequence com-

parisons and "full-length" cDNA sequences: a combined approach to evaluate and improve Arabidopsis genome annotation. Genome Res 14:406-413
Chao S, Lazo GR, You F, Crossman CC, Hummel DD, Lui N, Laudencia-Chingcuanco D, Anderson JA, Close TJ, Dubcovsky J, Gill BS, Gill KS, Gustafson JP, Kianian SF, Lapitan NL, Nguyen HT, Sorrells ME, McGuire PE, Qualset CO, Anderson OD (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
Conley EJ, Nduati V, Gonzalez-Hernandez JL, Mesfin A, Trudeau-Spanjers M, Chao S, Lazo GR, Hummel DD, Anderson OD, Qi LL, Gill BS, Echalier B, Linkiewicz AM, Dubcovsky J, Akhunov ED, Dvorak J, Peng JH, Lapitan NL, Pathan MS, Nguyen HT, Ma XF, 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 168:625-637
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 U S A 102:19243-19248
Dubcovsky J, Ramakrishna W, SanMiguel PJ, Busso CS, Yan L, Shiloff BA, Bennetzen JL (2001) Comparative sequence analysis of

colinear barley and rice bacterial artificial chromosomes. Plant Physiol 125:1342-1353
Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A (2006) Pfam: clans, web tools and services. Nucleic Acids Res 34:D247-251
Francki M, Carter M, Ryan K, Hunter A, Bellgard M, Appels R (2004) Comparative organization of wheat homoeologous group 3S and 7L using wheat-rice synteny and identification of potential markers for genes controlling xanthophyll content in wheat. Funct Integr Genomics 4:118-130
Gill BS, Appels R, Botha-Oberholster AM, Buell CR, Bennetzen JL, Chalhoub B, Chumley F, Dvorak J, Iwanaga M, Keller B, Li W, McCombie WR, Ogihara Y, Quetier F, Sasaki T (2004) A workshop report on wheat genome sequencing: International Genome Research on Wheat Consortium. Genetics 168:1087-1096
Griffiths S, Sharp R, Foote TN, Bertin I, Wanous M, Reader S, Colas I, Moore G (2006) Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat. Nature 439:749-752
Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Jr., Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31:5654-5666
Haas BJ, Wortman JR, Ronning CM, Hannick LI, Smith RK, Jr., Maiti R, Chan AP, Yu C, Farzad M, Wu D, White O, Town CD (2005)

Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release. BMC Biol 3:7
Hossain KG, Kalavacharla V, Lazo GR, Hegstad J, Wentz MJ, Kianian PM, Simons K, Gehlhar S, Rust JL, Syamala RR, Obeori K, Bhamidimarri S, Karunadharma P, Chao S, Anderson OD, Qi LL, Echalier 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 168:687-699
Houde M, Belcaid M, Ouellet F, Danyluk J, Monroy AF, Dryanova A, Gulick P, Bergeron A, Laroche A, Links MG, MacCarthy L, Crosby WL, Sarhan F (2006) Wheat EST resources for functional genomics of abiotic stress. BMC Genomics 7:149
International_Rice_Genome_Sequencing_Project (2005) The map-based sequence of the rice genome. Nature 436:793-800
Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyere C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chate-

let P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pe ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quetier F, Wincker P (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463-467
Jurka J (2000) Repbase update: a database and an electronic journal of repetitive elements. Trends Genet 16:418-420
Kawaura K, Mochida K, Ogihara Y (2005) Expression profile of two storage-protein gene families in hexaploid wheat revealed by large-scale analysis of expressed sequence tags. Plant Physiol 139:1870-1880
Kronmiller BA, Wise RP (2008) TEnest: automated chronological annotation and visualization of nested plant transposable elements. Plant Physiol 146:45-59
La Rota M, Sorrells ME (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
Laudencia-Chingcuanco DL, Stamova BS, Lazo GR, Cui X, Anderson OD (2006) Analysis of the wheat endosperm transcriptome. J Appl Genet 47:287-302
Lazo GR, Chao S, Hummel DD, Edwards H, Crossman CC, Lui N, Matthews DE, Carollo VL, Hane DL, You FM, Butler GE, Miller RE, Close TJ, Peng JH, Lapitan NL, Gustafson JP, Qi LL, Echalier B, Gill BS, Dilbirligi M, Randhawa HS, Gill KS, Greene RA, Sorrells ME, Akhunov ED, Dvorak J, Linkiewicz AM, Dubcovsky J, Hossain KG, Kalavacharla V, Kianian SF, Mahmoud AA, Miftahudin, Ma XF,

Conley EJ, Anderson JA, Pathan MS, Nguyen HT, McGuire PE, Qualset CO, Anderson OD (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
Linkiewicz AM, Qi LL, Gill BS, Ratnasiri A, Echalier B, Chao S, Lazo GR, Hummel DD, Anderson OD, Akhunov ED, Dvorak J, Pathan MS, Nguyen HT, Peng JH, Lapitan NL, Miftahudin, Gustafson JP, La Rota CM, Sorrells ME, Hossain KG, Kalavacharla V, Kianian SF, Sandhu D, Bondareva SN, Gill KS, Conley EJ, Anderson JA, Fenton RD, Close TJ, McGuire PE, Qualset CO, Dubcovsky J (2004) A 2500-locus bin map of wheat homoeologous group 5 provides insights on gene distribution and colinearity with rice. Genetics 168:665-676
McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19:362-367
Miftahudin, Ross K, Ma XF, Mahmoud AA, Layton J, Milla MA, Chikmawati T, Ramalingam J, Feril O, Pathan MS, Momirovic GS, Kim S, Chema K, Fang P, Haule L, Struxness H, Birkes J, Yaghoubian C, Skinner R, McAllister J, Nguyen V, Qi LL, Echalier B, Gill BS, Linkiewicz AM, Dubcovsky J, Akhunov ED, Dvorak J, Dilbirligi M, Gill KS, Peng JH, Lapitan NL, Bermudez-Kandianis CE, Sorrells ME, Hossain KG, Kalavacharla V, Kianian SF, Lazo GR, Chao S, Ander-

son OD, Gonzalez-Hernandez J, Conley EJ, Anderson JA, Choi DW, Fenton RD, Close TJ, McGuire PE, Qualset CO, Nguyen HT, Gustafson JP (2004) Analysis of expressed sequence tag loci on wheat chromosome group 4. Genetics 168:651-663
Mochida K, Kawaura K, Shimosaka E, Kawakami N, Shin IT, Kohara Y, Yamazaki Y, Ogihara Y (2006) Tissue expression map of a large number of expressed sequence tags and its application to in silico screening of stress response genes in common wheat. Mol Genet Genomics 276:304-312
Munkvold JD, Greene RA, Bermudez-Kandianis CE, La Rota CM, Edwards H, Sorrells SF, Dake T, Benscher D, Kantety R, Linkiewicz AM, Dubcovsky J, Akhunov ED, Dvorak J, Miftahudin, Gustafson JP, Pathan MS, Nguyen HT, Matthews DE, Chao S, Lazo GR, Hummel DD, Anderson OD, Anderson JA, Gonzalez-Hernandez JL, Peng JH, Lapitan N, Qi LL, Echalier B, Gill BS, Hossain KG, Kalavacharla V, Kianian SF, Sandhu D, Erayman M, Gill KS, McGuire PE, Qualset CO, Sorrells ME (2004) Group 3 chromosome bin maps of wheat and their relationship to rice chromosome 1. Genetics 168:639-650
Ogihara Y, Mochida K, Kawaura K, Murai K, Seki M, Kamiya A, Shinozaki K, Carninci P, Hayashizaki Y, Shin IT, Kohara Y, Yamazaki Y (2004) Construction of a full-length cDNA library from young spikelets of hexaploid wheat and its characterization by large-scale sequencing of expressed sequence tags. Genes Genet Syst 79:227-232
Ohyanagi H, Tanaka T, Sakai H, Shigemoto Y, Yamaguchi K, Habara T, Fujii Y, Antonio BA, Nagamura Y, Imanishi T, Ikeo K, Itoh T, Gojobori T, Sasaki T (2006) The Rice Annotation Project Database

(RAP-DB): hub for Oryza sativa ssp. japonica genome information. Nucleic Acids Res 34:D741-744
Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, Orvis J, Haas B, Wortman J, Buell CR (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res 35:D883-887
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 48:463-474
Peng JH, Zadeh H, Lazo GR, Gustafson JP, Chao S, Anderson OD, Qi LL, Echalier B, Gill BS, Dilbirligi M, Sandhu D, Gill KS, Greene RA, Sorrells ME, Akhunov ED, Dvorak J, Linkiewicz AM, Dubcovsky J, Hossain KG, Kalavacharla V, Kianian SF, Mahmoud AA, Miftahudin, Conley EJ, Anderson JA, Pathan MS, Nguyen HT, McGuire PE, Qualset CO, Lapitan NL (2004) Chromosome bin map of expressed sequence tags in homoeologous group 1 of hexaploid wheat and homoeology with rice and Arabidopsis. Genetics 168:609-623
Qi LL, Echalier 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 XF, Miftahudin, Gustafson JP, Conley EJ, Nduati V, Gonzalez-Hernandez JL, Anderson JA, Peng JH, Lapitan NL, Hossain KG, Kalavacharla V, Kianian SF, Pathan MS, Zhang DS, Nguyen HT, Choi

DW, 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 168:701-712
Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, Anxolabehere D (2005) Combined evidence annotation of transposable elements in genome sequences. PLoS Comput Biol 1:166-175
Randhawa HS, Dilbirligi M, Sidhu D, Erayman M, Sandhu D, Bondareva S, Chao S, Lazo GR, Anderson OD, Miftahudin, Gustafson JP, Echalier B, Qi LL, Gill BS, Akhunov ED, Dvorak J, Linkiewicz AM, Ratnasiri A, Dubcovsky J, Bermudez-Kandianis CE, Greene RA, Sorrells ME, Conley EJ, Anderson JA, Peng JH, Lapitan NL, Hossain KG, Kalavacharla V, Kianian SF, Pathan MS, Nguyen HT, Endo TR, Close TJ, McGuire PE, Qualset CO, Gill KS (2004) Deletion mapping of homoeologous group 6-specific wheat expressed sequence tags. Genetics 168:677-686
Salamov AA, Solovyev VV (2000) Ab initio gene finding in Drosophila genomic DNA. Genome Res 10:516-522
Salse J, Bolot S, Throude M, Jouffe V, Piegu B, Quraishi UM, Calcagno T, Cooke R, Delseny M, Feuillet C (2008) Identification and Characterization of Shared Duplications between Rice and Wheat Provide New Insight into Grass Genome Evolution. Plant Cell 20:11-24
See DR, Brooks S, Nelson JC, Brown-Guedira G, Friebe B, Gill BS (2006) Gene evolution at the ends of wheat chromosomes. Proc Natl Acad Sci U S A 103:4162-4167

Smith DB, RB F (1975) Characterisation of the wheat genome by renaturation kinetics. Chromosoma 50:223-242
Sorrells ME, La Rota M, Bermudez-Kandianis CE, Greene RA, Kantety R, Munkvold JD, Miftahudin, Mahmoud A, Ma X, Gustafson PJ, Qi LL, Echalier B, Gill BS, Matthews DE, Lazo GR, Chao S, Anderson OD, Edwards H, Linkiewicz AM, Dubcovsky J, Akhunov ED, Dvorak J, Zhang D, Nguyen HT, Peng J, Lapitan NL, Gonzalez-Hernandez JL, Anderson JA, Hossain K, Kalavacharla V, Kianian SF, Choi DW, Close TJ, Dilbirligi M, Gill KS, Steber C, Walker-Simmons MK, McGuire PE, Qualset CO (2003) Comparative DNA sequence analysis of wheat and rice genomes. Genome Res 13:1818-1827
Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23:1282-1288
Tanaka T, Antonio BA, Kikuchi S, Matsumoto T, Nagamura Y, Numa H, Sakai H, Wu J, Itoh T, Sasaki T, Aono R, Fujii Y, Habara T, Harada E, Kanno M, Kawahara Y, Kawashima H, Kubooka H, Matsuya A, Nakaoka H, Saichi N, Sanbonmatsu R, Sato Y, Shinso Y, Suzuki M, Takeda J, Tanino M, Todokoro F, Yamaguchi K, Yamamoto N, Yamasaki C, Imanishi T, Okido T, Tada M, Ikeo K, Tateno Y, Gojobori T, Lin YC, Wei FJ, Hsing YI, Zhao Q, Han B, Kramer MR, McCombie RW, Lonsdale D, O'Donovan CC, Whitfield EJ, Apweiler R, Koyanagi KO, Khurana JP, Raghuvanshi S, Singh NK, Tyagi AK, Haberer G, Fujisawa M, Hosokawa S, Ito Y, Ikawa H, Shibata M, Yamamoto M, Bruskiewich RM, Hoen DR, Bureau TE, Namiki N, Ohyanagi H, Sakai Y, Nobushima S, Sakata K, Barrero RA, Souvorov

A, Smith-White B, Tatusova T, An S, An G, S OO, Fuks G, Messing J, Christie KR, Lieberherr D, Kim H, Zuccolo A, Wing RA, Nobuta K, Green PJ, Lu C, Meyers BC, Chaparro C, Piegu B, Panaud O, Echeverria M (2008) The Rice Annotation Project Database (RAP-DB): 2008 update. Nucleic Acids Res 36:D1028-1033

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Dejardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjarvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leple JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouze P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y, Rokhsar D (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596-1604

UniProt_Consortium (2007) The Universal Protein Resource (UniProt). Nucleic Acids Res 35:D193-197
Wicker T, Guyot R, Yahiaoui N, Keller B (2003) CACTA transposons in Triticeae. A diverse family of high-copy repetitive elements. Plant Physiol 132:52-63
Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973-982
Zhu W, Buell CR (2007) Improvement of whole-genome annotation of cereals through comparative analyses. Genome Res

Stein, N., Prasad, M., Scholz, U., Thiel, T., Zhang, H., Wolf, M., Kota, R., Varshney, R.K., Perovic, D., Grosse, I. and Graner, A. (2007) A 1,000-loci transcript map of the barley genome: new anchoring points for integrative grass genomics. Theor. Appl. Genet. 114(5), 823-39.