Guidelines for Annotating Wheat Genomic Sequences: Release 1
Author: International Wheat Genome Sequencing Consortium Annotation Working Group
June 2006

I. Repetitive Sequences (Contact Person: Thomas Wicker)

A. Guidelines for repeat annotation

1. For the annotation of repeats, we recommend a search that includes several databases (TIGR rice repeat databases, RepBase and TREP). The TREP database only contains repeats which are well-characterised to avoid problems with classification. Mostly, we pay attention that TREP entries indeed only consist of one type of repeat and do not contain nested insertions of different repeat types that would result in confusing BLAST hits. Thus, TREP tends to be smaller than other repeat databases. The content of TREP will be integrated into the other repeat databases with each new release.
The first step of repeat identification should always be a BLASTN search against the above databases. Later steps include BLASTX and search for structural features (see below).

2. TREP will be updated twice a year (June 1st and December 1st) so that authors can refer to specific releases of TREP in their publications.

3. Researchers are invited to submit repetitive elements (especially new ones) to TREP.

B. Transposable element classification

Here, we present only the very basic guidelines for repeat annotation and classification. For more detailed descriptions and examples, please visit the TREP website (http://wheat.pw.usda.gov/ITMI/Repeats). These will also be updated and revised based on the discussions among the members of the TREP email group.

1- TE classes: We propose a relatively simple classification system that constitutes the following hierarchy: *Class, Subclass, Superfamily, Family, Subfamily, Name*. The ***Subclasses*** are so far only used with retrotransposons to divide the two main blueprints LTR and non-LTR retrotransposons.

The main classes and families are:

Class 1: retrotransposons

Subclass: LTR retrotransposons
Superfamilies : *copia*
gypsy
athila
TRIM

Subclass: non-LTR retrotransposons
Superfamilies: LINEs (long interspersed nuclear elements)

SINEs (short interspersed nuclear elements)

Class 2: DNA transposons
Superfamilies: CACTA
Mutator
Ac/Ds
Stowaway
Tourist
hAT

Class 3: Helitrons
Superfamilies: to be identified

2. Definition of Superfamilies

Superfamilies are in principle defined by specific structural criteria. For example *gypsy* and *copia* both contain the typical reverse transcriptase (RT) and integrase (INT) domains but their order is inverted. In gypsy, the domains are in the order (RT-INT) whereas in *copia* their order is (INT-RT). An extensive description of criteria defining superfamilies can be found at http://wheat.pw.usda.gov/ITMI/Repeats.

3. Definition of families

Each Superfamiliy contains multiple **families** of elements (e.g. *Sabrina* or *Isaac*). An element belongs to a family if it has >80% DNA sequence identity over at least 500 bp to other members of that family (Usually elements from the same family also give several blast hits more or less along the entire element). The 80% cutoff is for practical reasons because 80% sequence identity is approximately threshold for producing strong sequence alignments with **BLASTN** (at default settings) against nrTREP or totalTREP. Elements from different families show no or only very little sequence conservation at the DNA level. ***Subfamilies*** are used to further classify highly repetitive elements (e.g. the *BARE-1* family contains the *BARE-1*, *WIS* and *Angela* subfamilies, respectively).

4. The name of a transposable element

The **name** refers to one specific element of a family and includes either the GenBank accession number or the address of the BAC on which the element was found (e.g. *Sabrina_123A4-1 or Isaac_AF123456-2*). The hyphen after the BAC/Accession number indicates the specific copy of a family member on a sequence (e.g. if there are four *Sabrina* elements on BAC 123A4, they are named *Sabrina_123A4-1* through *Sabrina_123A4-4*).

5. Specific features

Since repetitive elements are often truncated by deletions or fragmented by nested insertions of other transposable elements, there are a few standardised attributes to further characterise an element. Examples can be found at the TREP website:

- “Complete”: Any element that has intact ends (i.e. a target site duplication is present). This only means that the exact borders of the transposed unit can be identified and does NOT mean that the element is intact or potentially functional.
- “Fragmented”: The element is cut into two pieces by a nested insertion.

- "Truncated": can be defined more precisely as 3' or 5' truncated. Often repeats are affected by deletions that delete one end of the element (this is independent of nested insertions of other elements).
- "Partial": TE identified at the 5' or 3' boundary of the BAC sequence and is only partially covered by the BAC. This also includes repeats that carry a nested insertion due to which one part of the repeat is shifted out of the window that is covered by the BAC clone.
- "Degenerated": Only a small fragment with similarity to known elements can be identified (e.g. only at the protein level). No clear boundaries can be identified.

For the annotation process, it is not important if a particular element is non-autonomous (i.e. does not have functional proteins) or potentially functional. Such characteristics can be the topic of specialised studies. The main goal of repeat annotation is to reliably classify sequences as repetitive to avoid them being wrongfully annotated as genes. Additionally, each identified repeat can be added to the existing databases, making future annotation easier.

6. Identification of novel elements

After annotating all known elements, the leftover sequence should be searched by BLASTX against the TREP protein collection (PTREP) to identify possible coding sequences of divergent transposable elements that have no (or only little) sequence conservation at the DNA level. For the classification of coding sequences, the same criteria can be used as for gene classification. Novel elements can also often be identified by structural features such as LTRs or CACTA signatures. Examples as well as information on naming conventions can be found at http://wheat.pw.usda.gov/ITMI/Repeats.

7. The coding sequences of all Triticeae genes identified according to the guidelines for gene annotation will also be deposited at TREP in a separate BLAST database. This will help identify repetitive elements which were wrongly annotated as genes (e.g. if one "gene" is found over and over again in several sequences, it is probably a repeat).

II. Identification and annotation of genes (Contact Person: Robin Buell). Gene finding and repeat annotation will be done in parallel to maximize identification of true genes and minimize mis-annotation of transposable elements as genes.

A. Orientation

All sequences will be oriented from the SP6 (base 1) to the T7 end of the vector.

B. Repetitive Sequences

Repeats will be identified as described in Section I.

C. Ab initio gene finders

Any number of *ab initio* gene finders can be run on the sequences (GeneMarkHMM (multiple matrices), GeneID, GeneScan, EUGENE (rice matrix)) but at a minimum, FGENESH (Monocot matrix) must be run.

D. Loci and gene model nomenclature

The genes (also know as loci or transcriptional units (TU)) will be annotated using the BAC name and a gene number that is oriented relative to the sequence. For example, BAC clone 27H32, the first gene located at base 10 to 1247 will be 27H32.t00001, the second gene located at base 1568 to 2700 will be 27H32.t00002, etc. Models should be named

with a “m” to distinguish models from TUs/loci. To provide a stable identifier for future updates of the annotation, a reduced gene/locus/TU can be used (27H32.1, 27H32.2, etc).

Example:
Stable Identifier : 27H32.1
Locus or TU: 27H32.t00001
Gene model: 27H32.m00001

E. Functional assignment

Putative function for the genes will be assigned via combination of BLASTP matches to a non-redundant amino acid database and Pfam trusted cutoff scores as well as searches of transcript evidence (ESTs and full length cDNAs). A table summarizing the putative function assignment guidelines is provided below. In addition, a comment field describing how the putative function was determined should be provided to allow others to ascertain the evidence used in assignment of putative function.

Putative Function	Match in Non-redundant amino acid (nraa) db	Pfam database Trusted Cutoff Score	Wheat ESTs/FL-cDNA alignment	Sample of annotation
Known	>90-100% ID, >90-100% length	May be above trusted cutoff, not essential	Optional for annotation	Aquaporin
Putative	>45% ID, >50% length	May be above trusted cutoff, not essential	Optional for annotation	chitinase, putative
XX-domain containing protein	N/A	Above trusted cutoff	Optional for annotation	WD-domain containing protein
Expressed	No similarity detected in nraa, or similarity to protein in nraa is < 45% ID and/or <50% coverage, or similarity is to 1) an expressed protein, or 2) a protein with no known	Below trusted cutoff	>95% ID, >70% length of EST	Expressed protein
Conserved Hypothetical Protein	>45% ID, >50% length to a protein annotated as hypothetical protein	Below trusted cutoff	<95% ID, <70% length of EST	Conserved hypothetical protein
Hypothetical Protein	No match to any db entry >45% ID, >50% length	Below trusted cutoff	<95% ID, <70% length of EST	Hypothetical protein

N/A : not applicable

Examples:
Annotation: 27H32.t00001: aquaporin
Comment Field: “based on 96.3% identity, 97.9% coverage to known wheat aquaporin, Genbank accession ##”

Annotation: 27H32.t00002: chitinase, putative
Comment Field: “based on 53.1% identity, 66% coverage to *Oryza sativa* chitinase, Genbank accession ##”

Annotation: 27H32.t00003: conserved hypothetical protein
Comment Field: “based on 53.1% identity, 66% coverage to At1g05670, hypothetical protein

Annotation: 27H32.t00005: hypothetical protein
Comment Field: “predicted from FGENESH”

F. Pseudogenes
Pseudogenes will be defined based on evidence of transcription yet have no clear ORF.

G. Rice Homolog
A top match (Rice Locus Name) to the predicted rice proteome should be provided

H. Arabidopsis Homolog
A top match (Arabidopsis Locus Name) to the predicted Arabidopsis proteome should be provided

I.Optional
All regions that do not contain a gene or repeat should be searched against the rice genome using tBLASTx.

III. Data distribution

A. Data to be made available via ftp to the TREP website

1. Sequence files
    a. Raw BAC sequence (.con)
    b. Locus sequence (.seq)
    c. CDS sequence (.cds)
    d. Peptide sequence (.pep)
2. Annotation data in GFF (sample will be generated once annotation is finalized by the working group)