Maize TE (transposable element) database users' guide

July 8, 2008
modified June 29, 2009

Overview: The maize TE (transposable element) database – here after referenced as TEDB – is designed to store information about TEs and make the information available in a variety of different formats.

Notes: The TEDB programs can change and so this guide may not reflect an exact step-by-step work flow. If you have questions or comments on this guide then please send email to Rick Westerman (westerman@purdue.edu) or Dave Jacoby (jacoby@purdue.edu) .

Organization of the guide: The first part of this guide deals with the retrieval programs. The second part deals with the submission of data. This guide is just for the web interface to the TEDB. There is also a command line interface to the TEDB.

Main web page

The entry web page is: http://**www.genomics.purdue.edu/~maize/**. From this page you can either put data into or get data from the TEDB. There are also other utility programs available; e.g., 'update account information & password'.

Retrieving information

There are three retrieval programs.

1) The "*Show TE classes and sub-classes*" program simply shows the relationship between all of the known classes, sub-classes, orders, super-families, families and sub-families within the TEDB. There is no user interaction with this program. It is just a listing of information.

2) The "*List the number of TEs associated with*" are also just listings of information, shown by annotator, organism, and TE classification. The latter works like the "Show TE classes and subclasses" program, with more information.

3) The "*Search the database*" program is the main way to retrieve targeted TEs from the database. By using successive search queries you can drill down through the database in order to narrow down the number of TEs that your retrieve or you can look at the whole database. The program is broken down into three sections (or parts.) as follows.

Results Section (search)

The top section shows how many TEs you have currently selected. See figures 1 and 2. The middle section contains 5 buttons which control the program. See figure 3. The bottom section shows parameters that can be used to narrow the search. See figures 8 through 14.

Fig. 1: Screen shot of the top section of the retrieval program. Shows a search that includes all TEs in the database.

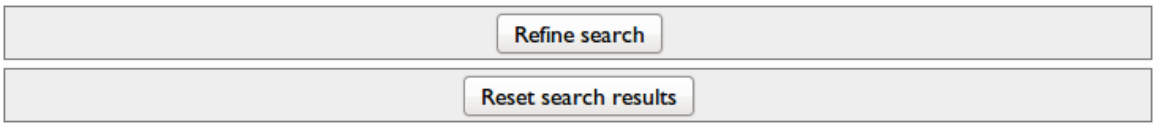
Fig. 2: Top section of the retrieval program after one or more refinements (searches) have been done.

Search Options (search)

The TE Classifications section will allow you to search by Classes, Subclasses, Orders, Superfamilies, Families and Subfamilies. More about this and the other search choices will be found below.

Output Options (search)

Fig. 3: Output Options in the retrieval program



For the Output Options:

The first button -- "*refine search*" -- allows you to select various parameters that refine your current selection. In other words all of your parameters will be merged together on an "AND" basis. Example: Classification AND Selected date AND Sequence.

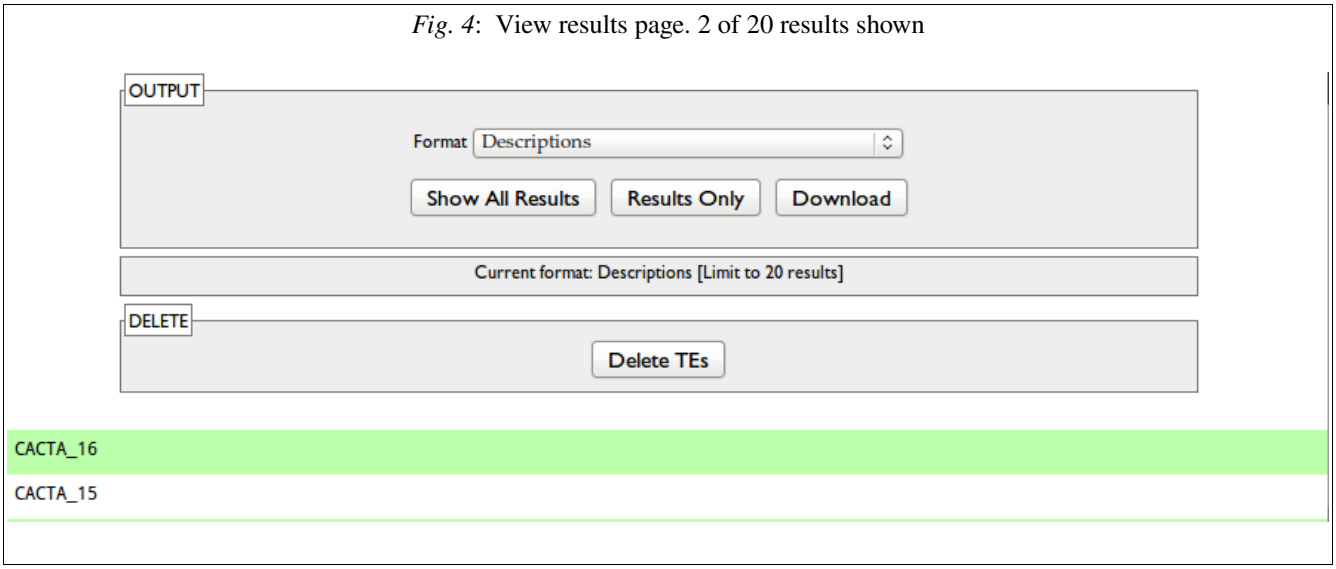
The second button -- "*reset search results*" -- will simply reset your search results back to selecting all TEs in the database. This button will not be further discussed.

The third button -- "*delete TEs*" -- deletes the selected TEs. It only appears if you have permissions to delete TEs. Obviously you should use this button with care.

Viewing Results

The results view is now part of the main search page.

Fig. 4: View results page. 2 of 20 results shown



There is one drop-down menu, labeled '*format*'. There are three buttons that control output.

The first button toggles between "Show All Results" and "Limit to 20 Results". The second opens a new page where you just see the results, in the format chosen from that drop-down. The third saves the results as a text file.

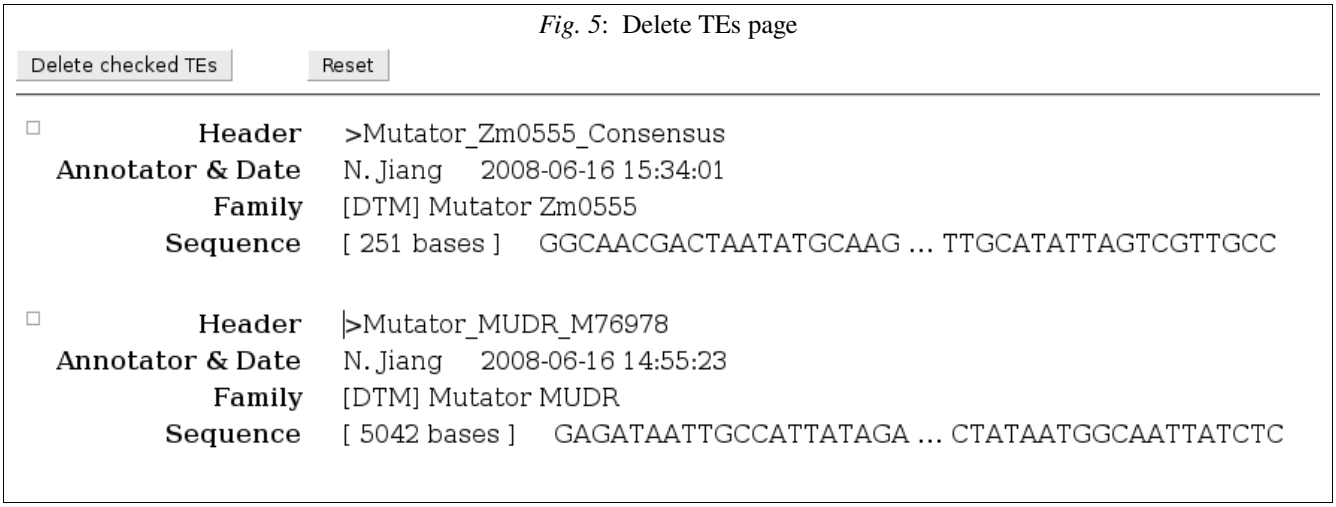
The 'format' menu contains many common output formats:

1. *Description* (the default). Just the FastA header line of each TE.
2. *FastA*. Full FastA format with header line and sequence.
3. *Quality file only*. FastA format with header line and quality values. Note that a TE may not have any quality values associated with it.
4. *GCG*.
5. *Genbank*.
6. *Tab delimited*. Useful for spreadsheets.
7. *Maize with references*. This is a custom format which can be used to re-import the data via the command line program. In general it will not be useful for anything beside that task.
8. *Maize w/o reference*s. As above, the custom format but without references.

Delete TEs

This button only appears if you have delete privileges. It will lead you to another page. On that page each of the selected TEs is displayed along with a checkbox. Check the TEs that you wish to actually delete and then press the '*delete checked TEs*' button. Note that any annotator can delete TEs so be careful about what you delete. See figure 5.

Fig. 5: Delete TEs page



Search Options (search)

The first parameter is the ability to search for TEs with a particular classification. See figures 7 and 8. The parameter is dynamic and requires that you have Javascript turned on in your web browser. In other words when you select a particular class then only the sub-classes available for that class are shown. Ditto with the order, super-family, and so on.

Fig. 7: Searching for all LINE order TEs

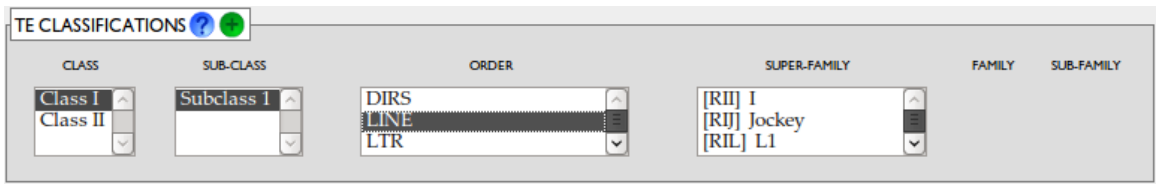
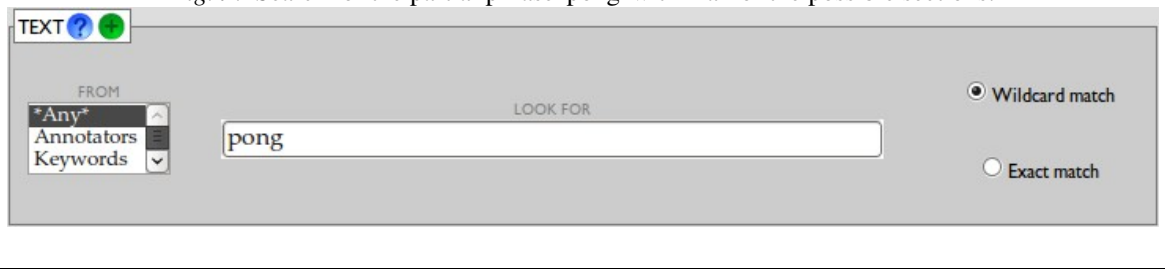


Fig. 8: Searching for all 'Copia' super-family TEs

Most of the options are hidden, keeping the interface less complex. To open a particular hidden box, click on the corresponding green circle.

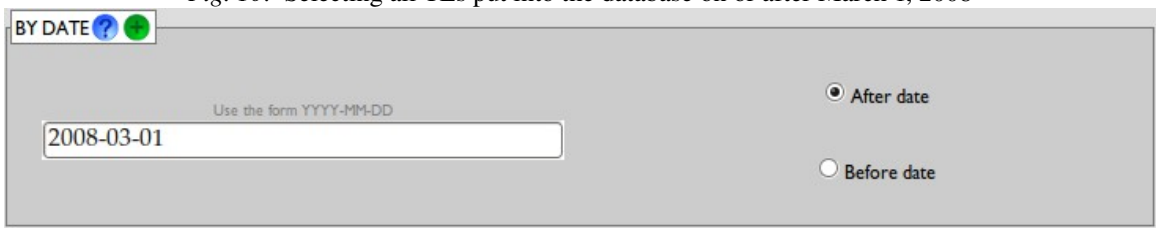
Text searching (figure 9) looks through the annotators, keywords, references, or TE names. You can specify a single word or phrase. Also there is an option for a wild card match (I.e., the search term is part of a longer word) or an exact match.

Fig. 9: Search for the partial phrase 'pong' within all of the possible sections.



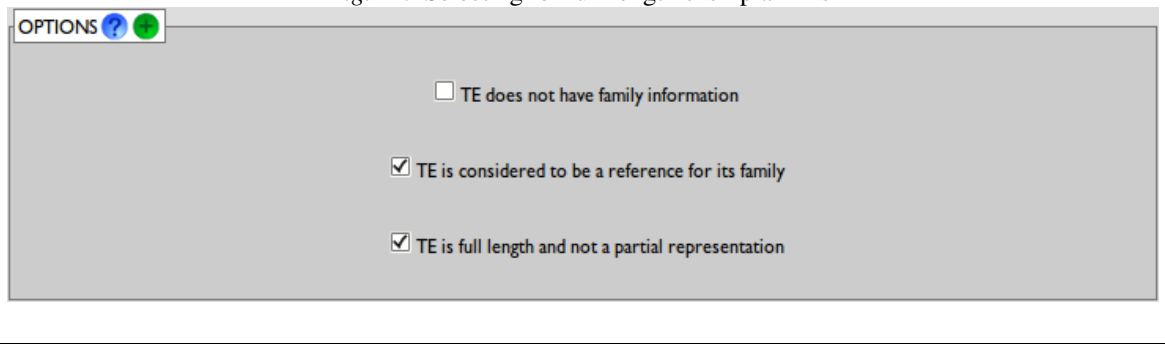
Date searching simply allows the selection of TEs that were entered into the database before or after (inclusive) a given date. See figure 10.

Fig. 10: Selecting all TEs put into the database on or after March 1, 2008



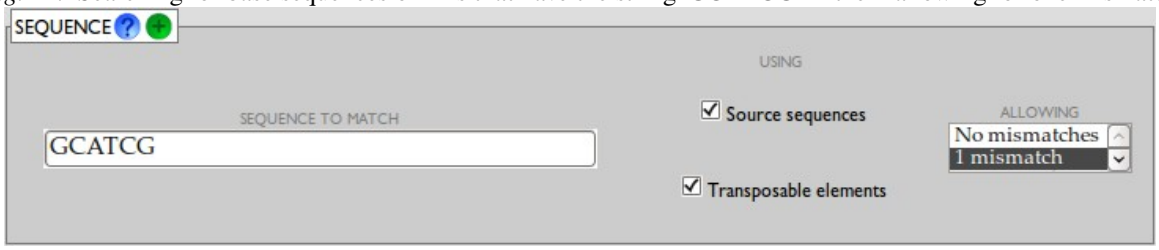
The are several characteristics (or options) of the TEs that can be selected for. See figure 11.

Fig. 11: Selecting for full length exemplar TEs



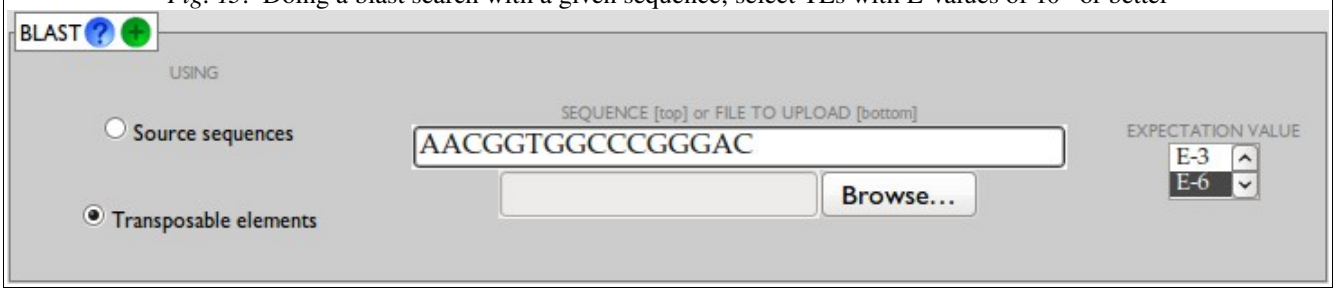
A search can be done through the actual sequence letters of a TE or of the base sequence that the TE is derived from. See figure 12. Note that not all TEs may be associated with a larger base sequence. Or the base sequence may be indirectly specified (e.g., a Genbank ID) and thus the actual sequence letters will be unknown and not able to be searched.

Fig. 12: Searching for base sequences or TEs that have the string 'GCATCG' in them allowing for one mismatch



A Blast search can be done through the TEs or base sequences. See figure 13. Also read the note in the above section in regards to base sequences.

Fig. 13: Doing a blast search with a given sequence; select TEs with E-values of 10^{-6} or better



Note: Figures 14-19 are reserved for later updates to the guide.

Submitting information

The web-based submission page has multiple sections. Each section has a help button (a question mark – see the figure to the right) next to it. A help screen will show up when your cursor is placed over the question mark. The help is also located at the bottom of the page.

Any of the 'submit TE(s)' buttons will work the same. There are several on the page merely for your convenience.

There is a lot of flexibility and options within the submission page. However for simple submissions all you will need to do is to enter the nucleotides within the first section of the web page. See figures 20, 21, 22 and 23. The TEs must be in FastA format with headers and nucleotides. The exception is if you have a single sequence in which case you can just enter the sequence but only if you enter the other required information later in the submission form.

Nucleotide section (submission)

The submission can contain multiple sequences. The header line for the sequences must be all in the same format. There are three types of header lines. The general format is SSS_FFF_QQQ-RRR where 'SSS' is the super-family name (or code) and 'FFF' is the family name. 'QQQ' is optional and is the Genbank name (or other reference) of the base sequence from which the TE is derived. 'QQQ' is the running number (or position) of the TE within the base sequence; this assumes multiple TEs have been derived from the same base sequence. See figures 20 through 23.

The different header lines are:

1. *Super-family with family*. Each TE header line must have both the super-family and family.
2. *Super-family without family.* Only the super-family is specified. Use this form if you do not know what family the TE belongs to. See note below.
3. *Family and no super-family*. Sometimes all you will want to do is to give the family name. See note below.
4. No super-family and no family. If you use this format then you will need to specify the family in the 'family' section as described below.

Important note; if you use format #2 or #3 (super-family w/o family or family w/o superfamily) then you will need to put an underscore after the super-family or family name. Otherwise the program will not be to differentiate this format and the last one.

If you use a family name that is not known then the submission program will report an error. You can override the behavior in the the 'change how the program runs' section.

It is also possible to replace the FastA header line with a different one. This option is not likely to be often used.

Fig. 20: Three TEs with super-family and family information. The first sequence is the 'ji' family which is part of the 'copia' super-family. The second sequence is the same only it uses the 'copia' three letter code of 'RLC'. The third TE is a member of the 'opie' family and was derived from the reference sequence 'AE4325' and is the 6th TE in that sequence. The other information on the line ('My best sequence') is ignored.

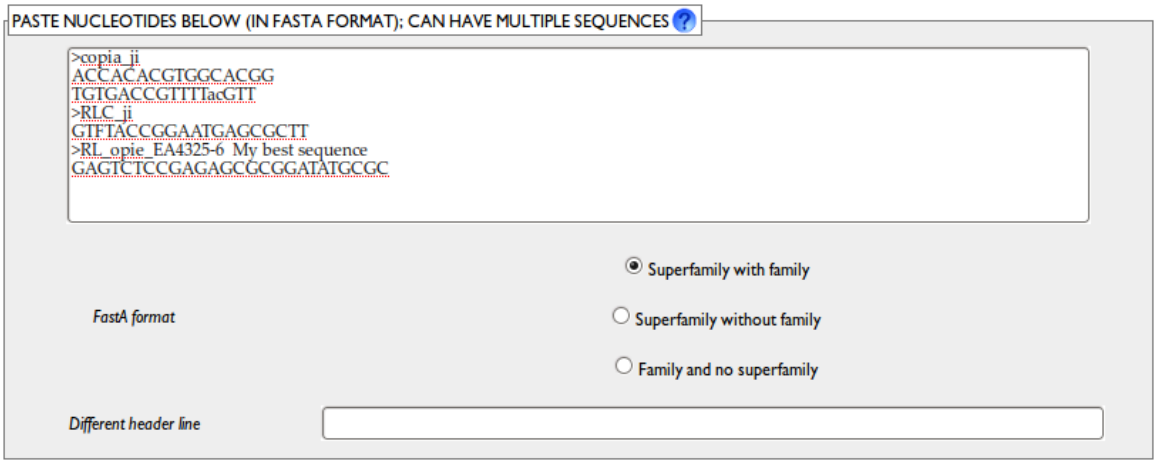


Fig. 21: Since the family names in figure 20 are unique within the database, the same information could be inserted via using the family names only. However note that the family names must be followed by an underscore so that they can be differentiated from sequences with arbitrary information.

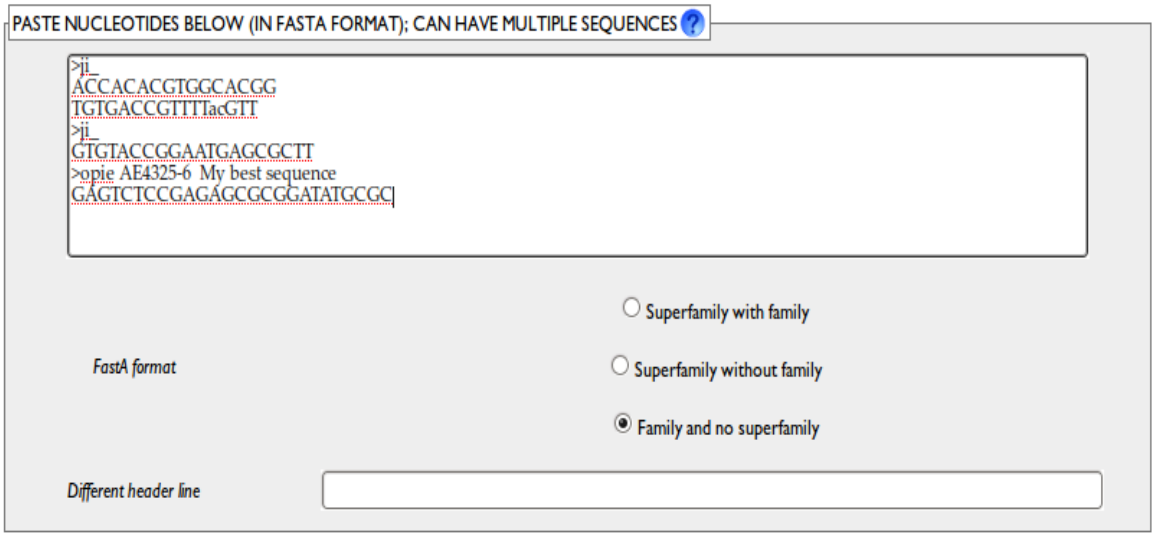


Fig. 22: Three TEs with super-family information only. The first two are of the 'jockey' (code: RIJ) super-family while the last is a member of the Viper super-family (note the required underscore after the super-family.) The first TE is the 5th TE derived from the reference sequence 'AE4325'. The third TE is derived from the reference 'AE12345'; its position in the reference is not defined. Note that, by default, unknown bases (Ns) are allowed and that case does not matter.

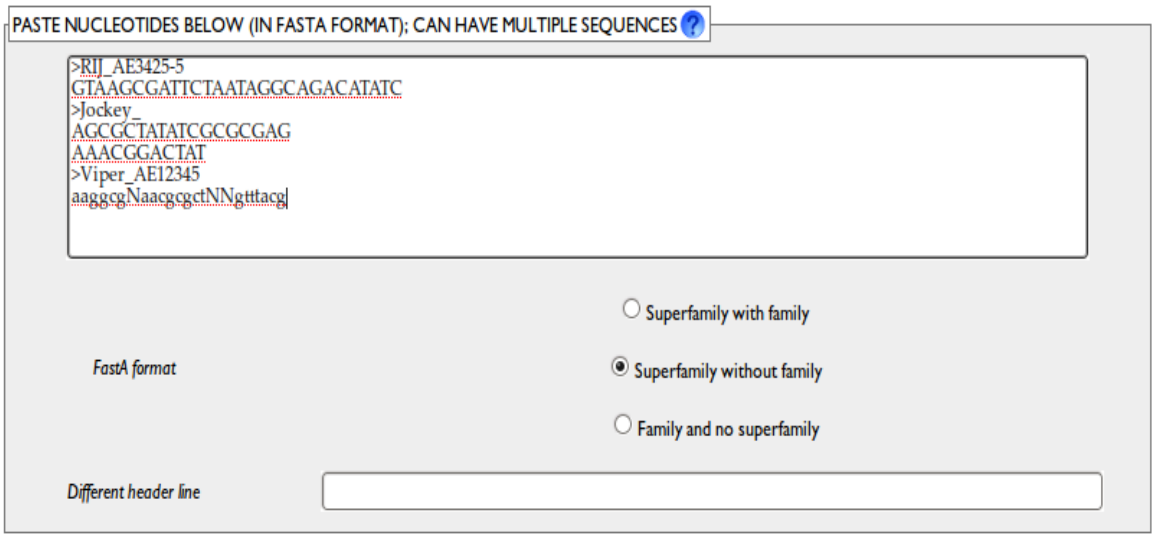
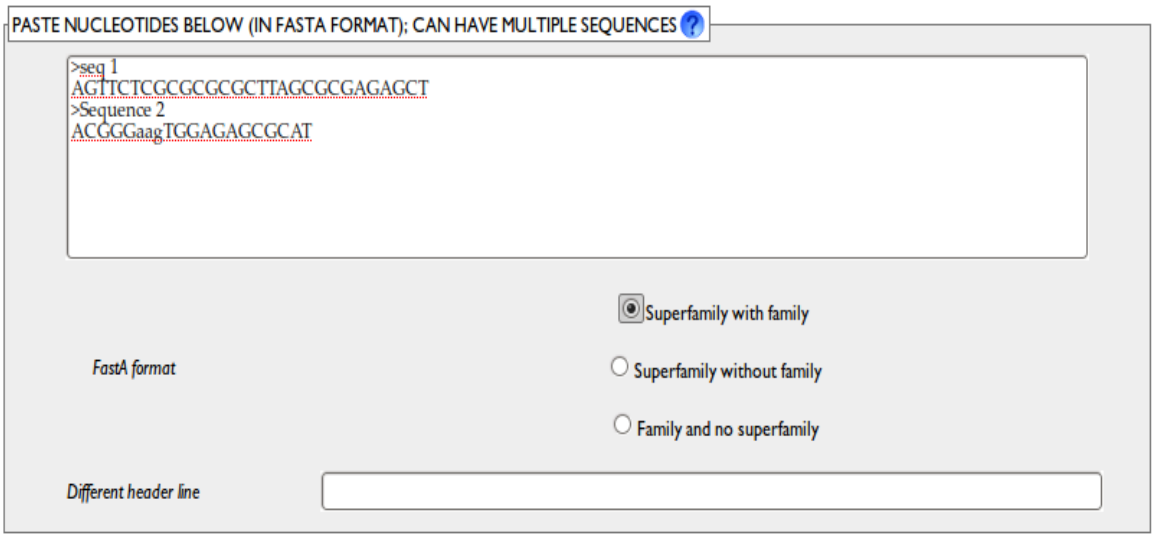


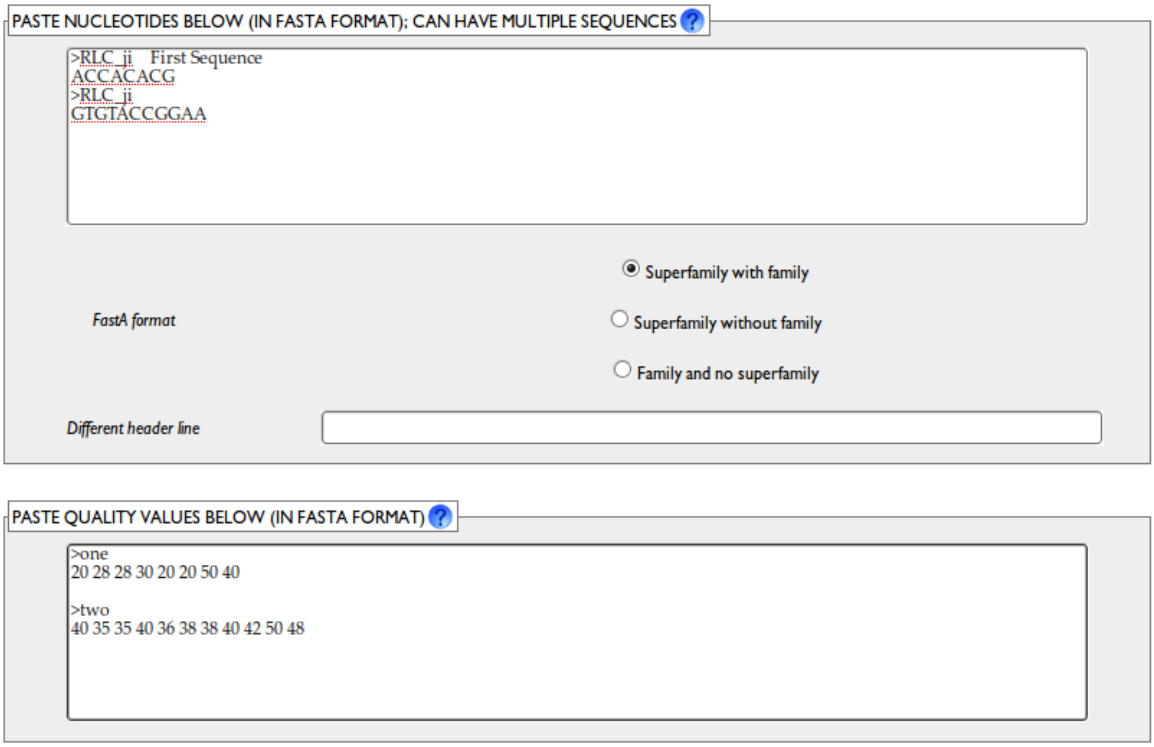
Fig. 23: Two sequences without any super family nor family information. This information will have to be defined later on the submission page.



Quality values section (submission)

Quality scores are optional but can useful to other people retrieving your TE. If you have quality scores then put them in FastA format and – important -- in the same order as the TE nucleotide order. In other words TE #1 will be associated with quality score #1. See figure 24.

Fig. 24: Two sequences with quality scores. Order is important and the header lines do not need to be the same. E.g., the first 'RLC_ji' (ACCACACG) is associated with quality 'one' while second 'is associated with the quality values labeled 'two'.



Family sections (submission)

See figures 25 and 26. If you do not specify a family or super-family in the header lines of your submitted nucleotides then you will need to defined the family in this section. Also if you wish to override the header lines you can put information in this section – but see the 'change how the program runs' section for a required flag if you are going to do the override. In order to select a super-family or family you will need Javascript enabled in your web browser. Simply drill down through the list available classes, sub-classes and orders until you get to your selection.

Additionally ff your family is not defined then it would be useful to put some information in the 'not defined' text box so that a human annotator can classify the family.

Fig. 25: All of the TEs submitted in the 'nucleotide' section will be classified as being in the 'barbara' family.

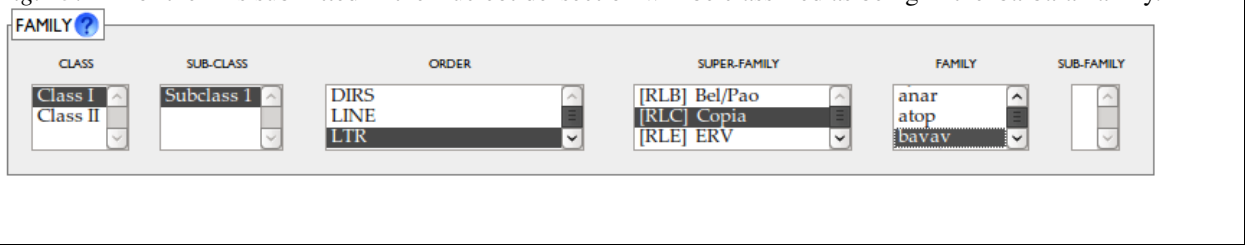
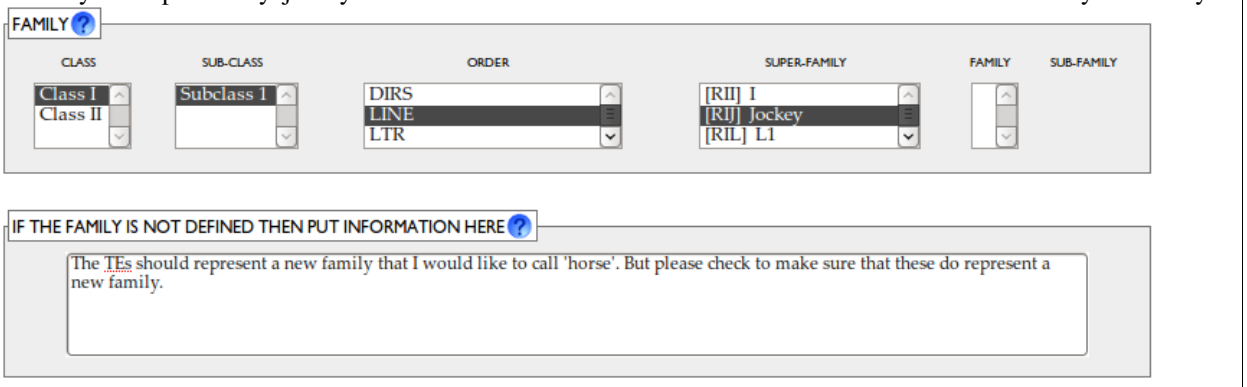


Fig. 26: Only the super-family 'jockey' was chosen. Information was entered so that a human could classify the family.



Attribute section (submission)

There are several attributes that can be assigned to the TE. Most of the time you will be working with only one TE when you use this section since multiple TEs will not often have the same attribute information. See figure 27.

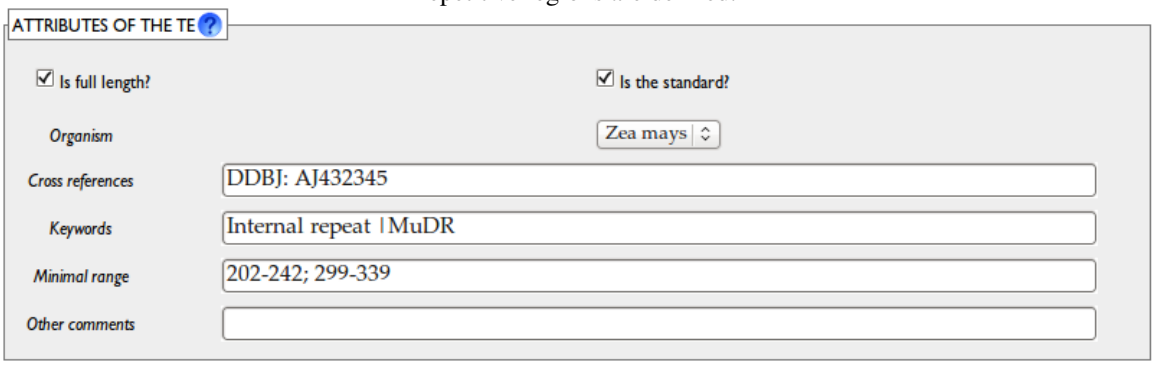
The most important attributes are

1. Is full length? Check this box if the TE is known to be complete and full length.
2. Is the standard? Check this box if the TE is the exemplar or standard for its family.

Less important attributes are:

3. Organism. Obviously it is always Maize for the TEDB but could change later.
4. Cross references. A free-form method to enter references to other databases.
5. Keywords. Keywords to help people find the TE. Separate keywords with vertical bars.
6. Minimal range. Any highly repetitive regions of the TE.
7. Other comments. Any other comments you wish to make about the TE.

Fig. 27: The full length exemplar for its family. Has a cross reference to the DDBJ and a couple of keywords. Two repetitive regions are defined.



Association section (submission)

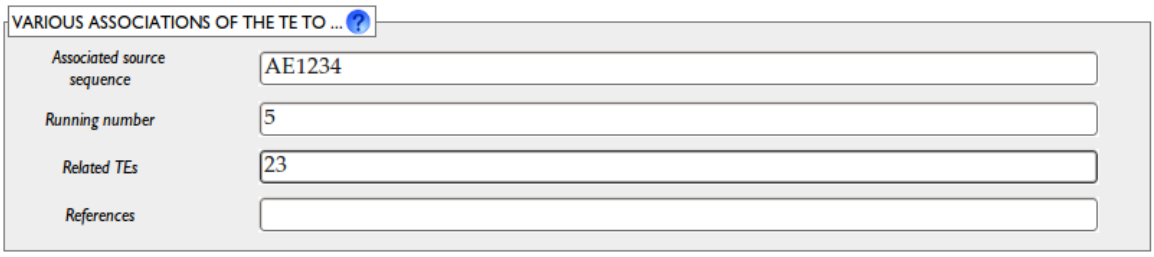
This is the section where you can defined associations of the TE to other parts of the TEDB. See figure 28. The first two entries in this section – 'associated source sequence' and 'running number' – are the same parameters that can be found in the FastA header lines of the nucleotide section. They are in this section in case you wish to override the header line and/or just manually enter the information – but see the 'change how the program runs' section for a required flag if you are going to do the override.

Note that the source sequence must already be in the TEDB. Or you can check the 'ignore source sequence errors' in the 'change how the program runs' section.

The 'related' TEs part allows you to associate the current TE with other TEs in the database. However – and this could be difficult – you must know the TEDB id of the other TEs. You can find this via the search page.

References are the TEDB journal reference ids to associate with the TE. Often these will not be used.

Fig. 28: Instead of putting the source sequence (AE1234) and position within the source (5) in the FastA header, this information is entered below. Also the TE is associated to the TE that has the TEID of '23'.



GFF Section (submission)

The GFF, or General Feature Format) lines are based on the GFF standard file format. They must be tab-delimited.

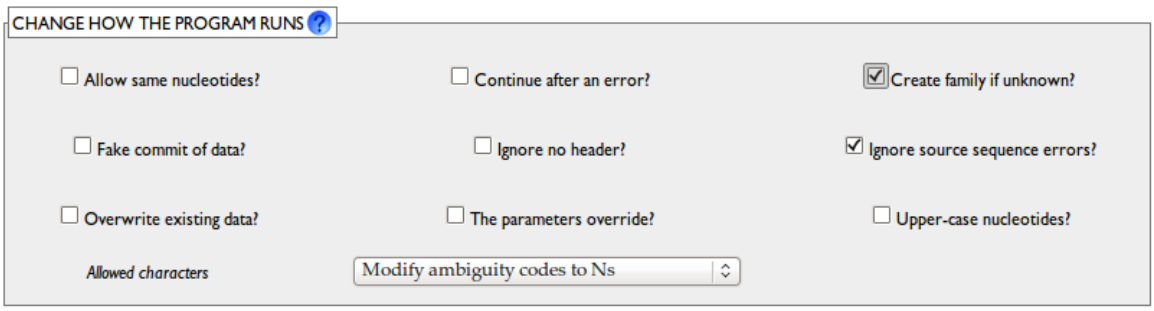
Change how the program runs section (submission)

How the submission program runs can be changed in a variety of ways. In general you should not change how the program runs since the program is trying to protect you from dumb mistakes; e.g., entering incorrect family names. But sometimes you will want to override the defaults. See figure 29.

The major options are:

- *Ignore source sequence errors* – used in case the source sequence is not in the TEDB.
- *Create family if unknown* – used when you want to create a new family.
- *The parameters override?* -- makes the parameters in the above sections override the information found via the FastA header line in the 'nucleotide' section.
- *Allowed characters* – change which characters are allowed. The default is to convert all non-ACGT into 'N' and then allow only ACGTN characters.

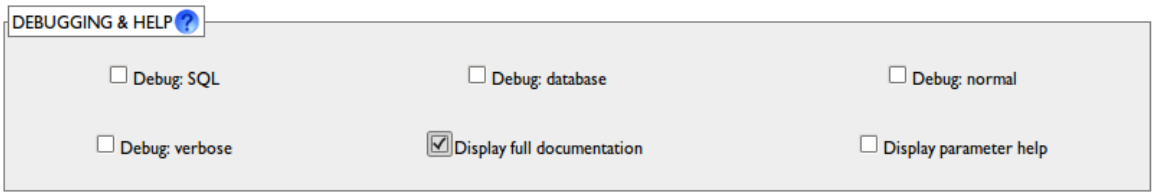
Fig. 29: Ignore source sequence errors and create new families if needed.



Debugging and help section (submission)

You probably will not need this section. However if you want to see the full documentation of the command line program (which the web-based program uses) then you can check that box. Doing so might help answer some of your questions. Or if you want to see debugging statements then you can look at those. See figure 30.

Fig. 30: Display the full command-line based documentation.



END OF DOCUMENT