



RepetDB: A TE database

Nicolas Francillonne, Mariène Wan, Nathalie Choisne, Françoise Alfama,
Raphaël Flores, Joelle Amselem, Johann Confais, Hadi Quesneville

► To cite this version:

Nicolas Francillonne, Mariène Wan, Nathalie Choisne, Françoise Alfama, Raphaël Flores, et al..
RepetDB: A TE database. International congress on transposable elements ICTE, Apr 2024, Saint
Malo, France. hal-04570060

HAL Id: hal-04570060

<https://hal.science/hal-04570060>

Submitted on 6 May 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Nicolas FRANCILLONNE ^{1,2}, Mariène WAN ^{1,2}, Nathalie CHOISNE ^{1,2}, Françoise ALFAMA ^{1,2}, Raphaël FLORES

^{1,2}, Joelle AMSELEM ^{1,2}, Johann CONFAIS ^{1,2} and Hadi QUESNEVILLE ^{1,2}

¹ Université Paris-Saclay, INRAE, URGI, 78026, Versailles, France

² Université Paris-Saclay, INRAE, BioinfOmics, Plant bioinformatics facility, 78026, Versailles, France

Transposable elements (TEs) are major players of structure and evolution of eukaryote genomes. Thanks to their ability to move around and to replicate within genomes, they are probably the most important contributors to genome plasticity. Their detection and annotation are considered essential and must be undertaken in any genome sequencing project.

- I wish to make an annotation of a Brassicaceae species close from *Arabidopsis thaliana*. - I go to repetdb website and check the related species of interest <https://urgi.versailles.inrae.fr/repetdb>

The screenshot shows the 'Search consensus' interface. It includes fields for Taxon group (set to 'Arabidopsis thaliana'), Wicker Classification (All classes, All orders, All superfamilies), Confused classification (checkboxes for 'Only confused', 'Only not confused', 'Unclassified or not TE', and 'Only TE'), Manual Validation (set to 'All'), and Similarity feature(s) (a text input field containing protein profile accessions). A 'Search' button is at the bottom right.

- I can also select multiple species, from example all *Brassicaceae* or just one close related species like *Arabidopsis thaliana*

The screenshot shows a tree view of selected species under 'Taxon group'. The root node is 'Arabidopsis thaliana', which branches into 'Rosa chinensis', 'Fragaria vesca', 'Populus trichocarpa', 'Brassicaceae', 'Arabidopsis halleri', 'Arabidopsis lyrata', and 'Arabidopsis thaliana' again. Other nodes like 'All orders' and 'All superfamilies' are visible on the left.

- Selection of specific class, order or superfamily is also available to refine a search

The screenshot shows a table of search results with a red box highlighting the first row. The columns include Consensus Identifier, Consensus Copies, Consensus Full-length copies, Consensus Length, TE classification code, Consensus Manual validation, Consensus Confused, and Consensus Comments. The first row is highlighted with a red box. A callout bubble points to the 'Gypsy' classification code in the 'TE classification code' column.

- I can obtain statistics and filter some values out or download my result and consensus fasta sequences

The screenshot shows a histogram of Consensus TE classification codes by count. A red box highlights the first bar. Below it, a download dialog box is open, asking for a file name ('results') and format ('FASTA sequence'). It also includes options for 'Comma separated values', 'XML', 'JSON', and 'BED locations'.

To annotate transposable elements in genome, consensus are needed and RepetDB aims to make available these sequence to make easier the annotation of closely related genomes. You can select consensus of interest in the database and download fasta sequence.

- I have an interesting sequence that I suspect is a transposable element and I wish to check if a match exist in the database. I go to PlantbioinfoPF blast services <https://urgi.versailles.inrae.fr/blast/> and select in group « Repeats » a close related species

The screenshot shows the 'Inspect BLAST output' interface. It displays a list of databases and a table of best hits. The table has columns for Query, Databases, Subject, Score, Identities, Percentage, Expect, Start, and End. The first hit is for 'Arabidopsis_thaliana.fasta consensus from RepetDB v2' with a score of 5844.

- If a match is found I get a direct link to the consensus card

The screenshot shows the 'Consensus card' for 'MCL105_AthaCol0_TEedenovo-B-R1018-Map20'. It includes sections for Material and Method, Genome assembly, Consensus copy statistics, Features browser, and Structural features. A red circle highlights the 'Consensus card' title at the top right.

By selecting one consensus I can access a consensus card describing its classification

Visit RepetDB :
<https://urgi.versailles.inrae.fr/repetdb/begin.do>
 Use URGIBlast :
<https://urgi.versailles.inrae.fr/blast/>
 Supplementary information and tutorials :
<https://urgi.versailles.inrae.fr/Data/Transposable-elements/REPETDB>