



HAL
open science

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.**

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 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>

- 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

Query	Database	Subject	Score	Identifiers (Query length)	Percentage	Expect	Start	End
Query1	Arabidopsis thaliana fasta consensus from RepetDB v2	MCL105_AthaCol0_TEdenovo-B-R1018-Map20	5844	3240/3240 (100%)	100	0.0	1	3240
Query1	Arabidopsis thaliana fasta consensus from RepetDB v2	MCL106_AthaCol0_TEdenovo-B-R880-Map1_reversed	561	448/499 (90%)	90	0.0	932	434

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

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

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

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

Consensus card

Consensus : MCL105_AthaCol0_TEdenovo-B-R1018-Map20

Length: 3442 bp (3442) | Classification (Class/Order/Superfamily): Class I: LTR: Oypsy | Manual validation: Confused: no

Cumulative genome coverage: 58222 bp | Fragments: 56 | Full length fragments: 2 | Copies: 43 | Full length copies: 3

Material and Method

Organism: Arabidopsis thaliana | Genome assembly: Arabidopsis thaliana TAIR10 (JBrowse)

Software used: TEdenovo (REPET V2.5), TEannot (REPET V3.0), PASTEClassifier V2 (REPET V3.0)

Comments

Transposable Element (TE) Detection and Annotation Workflow

- TEdenovo on the genome produced a filtered consensus library without SSR rich consensus neither unclassified consensus built with less than 10 HSPs.
- This consensus library was clustered and the cluster number is in the header of each consensus.
- TEannot using this TE consensus library is performed to select consensus having at least one Full Length Copy (FLC) annotated in the genome.
- In this new release, classification of FLC consensus was updated with PASTEClassifier V2 (manually curated, a library of profiles from Plant2.0 and GYDB2.0 and RNA eukaryotes (Genbank) library. A post process was performed to curate some confounded classifications.
- After a step of manual curation to re-classify or remove some consensus, a TEannot using this curated consensus library was run to annotate TE copies in the whole genome and to obtain the final TE annotation.

Genome assembly fasta file, genome TE annotation (gff3) file and consensus library fasta file are available for download

Contact: Nathalie Choise, Johann Confais, Hadi Quesneville (mail tourgi-contact)

Consensus copy statistics

Similarity features

Query Start	Query End	HIT Start	HIT End	E-Value	Identity	Details
1	4965	4991	5668			DB: Repbase Accession: ATLANTY531 (Requires Repbase registration) Classification: Class I: LTR: Oypsy
4965	5460	1	494			DB: Repbase Accession: ATLANTY531LTR (Requires Repbase registration) Classification: Class I: LTR: Oypsy

5 Protein profile feature

Query Start	Query End	HIT Start	HIT End	E-Value	Identity	Details
775	969	64	102			DB: Pfam Accession: PF11690.3 Description: PF11690.3
1232	1463	26	111			DB: Pfam Accession: PF11690.3 Description: PF11690.3
2244	2348	189	210			DB: Pfam Accession: PF14362.1 Description: PF14362.1
2349	2720	10	128			DB: Pfam Accession: PF04111.7 Description: PF04111.7
2352	2618	49	136			DB: Pfam Accession: PF11559.3 Description: PF11559.3

Structural features

18 ORF

2 TR

Start	End	Details
41	3348	Type: non-termTR
2012	3384	Type: non-termTR

Consensus search result

Showing 1 to 25 of 406 rows

Consensus Identifier	Consensus Copies	Consensus Full-length copies	Consensus Length	Consensus TE classification code	Consensus Manual validation	Consensus Confused	Consensus Comments
MCL102_AthaCol0_TEdenovo-B-G1208-Map3	20	0	971	Class I: LINE: ?	classification	no	NO VALUE
MCL102_AthaCol0_TEdenovo-B-R178-Map3	122	8	2017	Class I: LINE: ?	classification	no	NO VALUE
MCL103_AthaCol0_TEdenovo-B-G1214-Map20	23	11	449	Class II: Helitron: ?	classification	no	NO VALUE
MCL103_AthaCol0_TEdenovo-B-P161.100-Map16	21	11	452	Class II: Helitron: ?	classification	no	NO VALUE

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



Visit RepetDB :

<https://urgi.versailles.inrae.fr/repetdb/begin.do>

Use URGI Blast :

<https://urgi.versailles.inrae.fr/blast/>

Supplementary information and tutorials :

<https://urgi.versailles.inrae.fr/Data/Transposable-elements/REPETDB>