RepetDB: a unified resource for transposable element references

Joëlle Amselem, Guillaume Cornut, Nathalie Choisne, Michael Alaux, Françoise Alfama Depauw, Véronique Jamilloux, Florian Maumus, Thomas Letellier, Isabelle Luyten, Cyril Pommier, et al.

Mobile DNA

SOFTWARE Open Access

RepetDB: a unified resource for transposable element references

Joëlle Amselem*, Guillaume Cornut, Nathalie Choisne, Michael Alaux, Françoise Alfama-Depauw, Véronique Jamilloux, Florian Maumus, Thomas Letellier, Isabelle Luyten, Cyril Pommier, Anne-Françoise Adam-Blondon and Hadi Quesneville

Abstract

Background: Thanks to their ability to move around and replicate within genomes, transposable elements (TEs) are perhaps the most important contributors to genome plasticity and evolution. Their detection and annotation are considered essential in any genome sequencing project. The number of fully sequenced genomes is rapidly increasing with improvements in high-throughput sequencing technologies. A fully automated de novo annotation process for TEs is therefore required to cope with the deluge of sequence data.

However, all automated procedures are error-prone, and an automated procedure for TE identification and classification would be no exception. It is therefore crucial to provide not only the TE reference sequences, but also evidence justifying their classification, at the scale of the whole genome. A few TE databases already exist, but none provides evidence to justify TE classification. Moreover, biological information about the sequences remains globally poor.

Results: We present here the RepetDB database developed in the framework of GnpIS, a genetic and genomic information system. RepetDB is designed to store and retrieve detected, classified and annotated TEs in a standardized manner. RepetDB is an implementation with extensions of InterMine, an open-source data warehouse framework used here to store, search, browse, analyze and compare all the data recorded for each TE reference sequence. InterMine can display diverse information for each sequence and allows simple to very complex queries. Finally, TE data are displayed via a worldwide data discovery portal. RepetDB is accessible at urgi.versailles.inra.fr/repetdb.

Conclusions: RepetDB is designed to be a TE knowledge base populated with full de novo TE annotations of complete (or near-complete) genome sequences. Indeed, the description and classification of TEs facilitates the exploration of specific TE families, superfamilies or orders across a large range of species. It also makes possible cross-species searches and comparisons of TE family content between genomes.

Keywords: RepetDB, Transposable element, Database

Background

Transposable elements (TEs) are major players in the structure and evolution of eukaryote genomes. Thanks to their ability to move around and replicate within genomes, they are probably the most important contributors to genome plasticity [1]. Indeed, genome size is generally correlated with TE abundance: with up to 90% of the genome consisting of TE sequences in some species, such as wheat [2] and wheat powdery mildew fungus [3] [4]. The difference of the genome size of *Zea luxurians* and *Zea mays* is directly correlated to the abundance of TE in these 2 genomes [5]. The insertion of TEs close to genes can affect gene structure, expression and function, contributing to the genetic diversity underlying species adaptation [6]. Many studies have shown that TEs are generally silenced through epigenetic defense mechanisms, and that these elements play an important role in epigenetic genome regulation [7].

TE detection and annotation is, thus, now considered essential for any genome study. With the development of new high-throughput sequencing technologies, a large number of genomes have been sequenced, resulting in very large amounts of sequence data. Automated TE de novo detection software can provide up to thousands of

* Correspondence: joelle.amselem@inra.fr
URGI, INRA, Université Paris-Saclay, 78026 Versailles, France

TE reference sequences per genome, generally in the form of consensuses. TEs are then classified into two classes, with various subclasses, orders and superfamilies defined on the basis of mechanistic, enzymatic and sequence similarity criteria. The two classes of TEs are defined on the basis of their transposition mechanisms: with (Class I) or without (Class II) an RNA intermediate [8]. TE identification often makes use of popular pipelines, such as RepeatModeler [9] and TEdenovo pipeline from the REPET package [10], including tools for the identification of repeats, their grouping into TE families, and the generation of multiple alignments to build TE consensus sequences. TE classification is usually based on similarity to known TEs and protein domains and/or according sequence structure [11].

Several databases for describing TE reference sequences already exist. Repbase Update [12, 13] contains (i) complete consensus sequences for TE families (70% of the database) built with homemade pipelines based principally on Recon [14] and LTR-FINDER [15], and (ii) TE copy sequences (complete or incomplete) extracted from individual genomic loci. DFAM [16], another database for repeats DNA families is based on profiles HMM constructed from multiple sequence alignment. The entries of Dfam are Repbase-derived library (humans, mouse, zebrafish, fly and nematodes). The Gypsy database (GyDB) [17] stores LTR-retroelements, including those of the Gypsy/Copia superfamilies in particular, and *Retroviridae*-like elements including sequences from the *Caulimoviridae* (plants Endoviruses). The SINE database (SINEBase) is dedicated to SINE elements prediction [18]. The P-MITE database hosts Miniature inverted-repeat transposable elements (MITEs) from 41 plant genomes [19]. The TIGR Plant Repeat Database [20] is a set of resources for the identification of repeats, including TEs, rDNAs, and telomere-associated sequences, populated with repeat sequences from 12 plant genomes from GenBank. This initial dataset is extended on the basis of sequence similarities between these sequences and GenBank sequences. The Transposable Elements Platform (TREP) is a curated TE database (http://botserv2.uzh.ch/kelldata/trep-db/index.html) mostly for monocotyledons and fungi. The MIPS Repeats database (PGSB-REdat) and Catalog (PGSC-REcat) are parts of the Plant Genome and Systems Biology platform (PGSB) including PlantsDB [21]. They contain TEs retrieved from TREP, TIGR repeats, Repbase and detected de novo in the genome sequences stored in PlantsDB. All these databases are accessed through web browsers and/or quick or advanced search forms. The curation of TE reference sequences is generally poor and based on automatic procedures.

However, intrinsic automated TE classification is an error-prone process. There was, therefore, an urgent need for annotation with TE consensus sequences together with the evidence justifying TE classification. None of the existing TE databases can provide the evidence on which the proposed classification is based. We filled this gap by developing the RepetDB database in the framework of GnpIS, a genetic and genomic Information System [22]. RepetDB is an instance of InterMine, a public open-source data warehouse that has been specially customized and enhanced [23, 24] with JBrowse [25] for searching, browsing, analyzing and comparing all the data provided for each TE consensus sequence in the many genomes analyzed. We use the powerful capacities of InterMine to create user-friendly interfaces allowing researchers to search and query data in multiple manners. Data can be exported in many commonly used formats, including fasta, GFF, BED or other tab-delimited formats. In addition, InterMine provides APIs in various languages, including Perl, Java, Python, Ruby, and JavaScript, for accessing its features via web services. The full capabilities of these web services are described at http://iodocs.apps.intermine.org/.

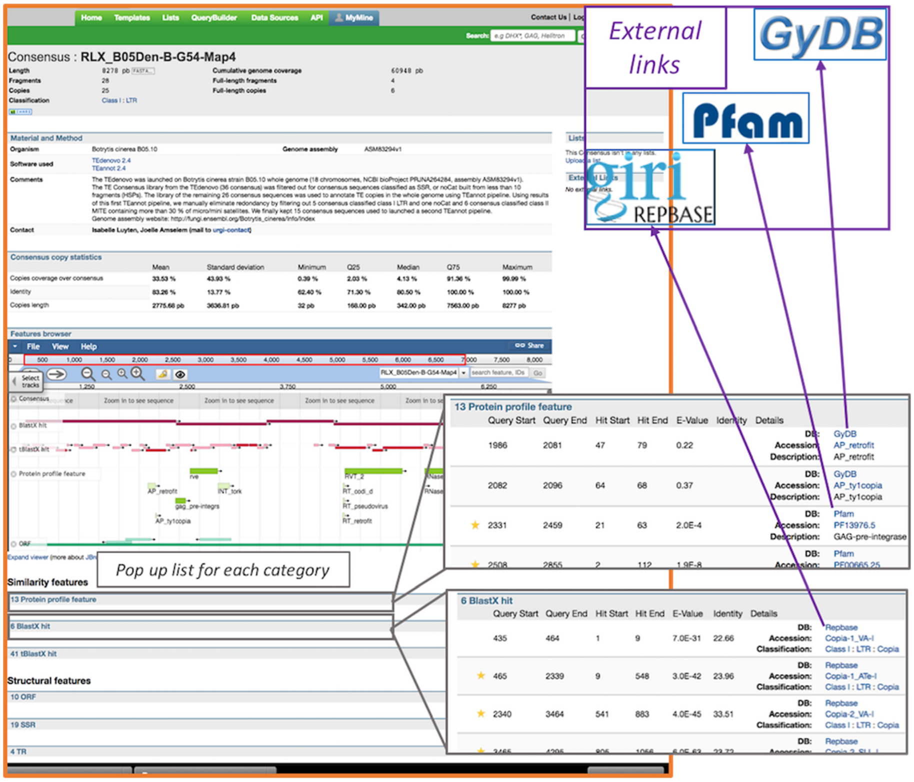
Implementation

InterMine instantiation and customization

RepetDB is an InterMine-based data warehouse. It was built with InterMine 1.4.2, with the InterMine web application used as the basis for data integration and presentation. Various comprehensive categories of reports through user-friendly interfaces for queries and visualization were developed with the templates provided by InterMine. RepetDB data integration is based on the InterMine data integration process built with Apache Ant 1.8.4, running in the Oracle Java 1.7.0_79 SE Runtime Environment and building a PostgreSQL 9.4.15 database. The RepetDB web application is deployed on Apache Tomcat 7.0.59 running in the same Java environment as used for data integration.

In addition to traditional InterMine configurations, some more advanced customizations were developed in Java, JSP (Java server page) and JavaScript. All of the RepetDB programs are hosted on a CentOS 6.8 virtual machine.

The central object in RepetDB is the TE reference sequence, which is a consensus sequence. The main user interface visualization is a “TE reference page” referred to hereafter as the “Consensus card” (Fig. 1), on which information is gathered together and organized into five sections: (i) the “Header information” panel containing basic information, such as the consensus identifier, its length, a link to its sequence, its classification according to Wicker nomenclature [8] and basic occurrence metrics concerning annotation with this consensus in the genome (ii) a “Materials and Methods” section with a short description of the dataset (genome assembly version) and workflow used to provide the reference and other relevant information (from the point of view of the person submitting the



External links

Pop up list for each category

Consensus : **RLX_B05Den-B-G54-Map4**

Length	8278 pb	**Cumulative genome coverage**	60948 pb
Fragments	28	**Full-length fragments**	4
Copies	25	**Full-length copies**	6
Classification	Class I : LTR		

Material and Method

Organism	Botrytis cinerea B05.10	**Genome assembly**	ASM83294v1
Software used	TEdenovo 2.4 TEannot 2.4		
Comments	The TEdenovo was launched on Botrytis cinerea strain B05.10 whole genome (18 chromosomes, NCBI bioProject PRJNA264284, assembly ASM83294v1). The TE Consensus library from the TEdenovo (36 consensus) was filtered out for consensus sequences classified as SSR, or noCat built from less than 10 fragments (HSPs). The library of the remaining 26 consensus sequences was used to annotate TE copies in the whole genome using TEannot pipeline. Using results of this first TEannot pipeline, we manually eliminate redondancy by filtering out 5 consensus classified class I LTR and one noCat and 6 consensus classified class II MITE containing more than 30 % of micro/mini satellites. We finally kept 15 consensus sequences used to launched a second TEannot pipeline. Genome assembly website: http://fungi.ensembl.org/Botrytis_cinerea/Info/Index		
Contact	**Isabelle Luyten, Joelle Amselem (mail to urgi-contact)**		

Consensus copy statistics

	Mean	Standard deviation	Minimum	Q25	Median	Q75	Maximum
Copies coverage over consensus	**33.53 %**	**43.93 %**	**0.39 %**	**2.03 %**	**4.13 %**	**91.36 %**	**99.99 %**
Identity	**83.26 %**	**13.77 %**	**62.40 %**	**71.30 %**	**80.50 %**	**100.00 %**	**100.00 %**
Copies length	**2775.68 pb**	**3636.81 pb**	**32 pb**	**168.00 pb**	**342.00 pb**	**7563.00 pb**	**8277 pb**

Features browser

Similarity features

13 Protein profile feature

6 BlastX hit

41 tBlastX hit

Structural features

10 ORF

19 SSR

4 TR

13 Protein profile feature

	Query Start	Query End	Hit Start	Hit End	E-Value	Identity	Details
	1986	2081	47	79	0.22		**DB:** GyDB **Accession:** AP_retrofit **Description:** AP_retrofit
	2082	2096	64	68	0.37		**DB:** GyDB **Accession:** AP_ty1copia **Description:** AP_ty1copia
★	2331	2459	21	63	2.0E-4		**DB:** Pfam **Accession:** PF13976.5 **Description:** GAG-pre-integrase
★	2508	2855	2	112	1.9E-8		**DB:** Pfam **Accession:** PF00665.25

6 BlastX hit

	Query Start	Query End	Hit Start	Hit End	E-Value	Identity	Details
	435	464	1	9	7.0E-31	22.66	**DB:** Repbase **Accession:** Copia-1_VA-I **Classification:** Class I : LTR : Copia
★	465	2339	9	548	3.0E-42	23.96	**DB:** Repbase **Accession:** Copia-1_ATe-I **Classification:** Class I : LTR : Copia
★	2340	3464	541	883	4.0E-45	33.51	**DB:** Repbase **Accession:** Copia-2_VA-I **Classification:** Class I : LTR : Copia
★	3465	4295	805	1056	6.0E-63	23.72	**DB:** Repbase **Accession:** Copia-2_SU-I

Fig. 1 The RepetDB central object: the "Consensus card", with magnifications of "protein profile" and "Blast hit" result categories with external links

dataset) (iii) a "TE copies statistics" section with various metrics concerning the genome copies annotated with the current TE reference, (iv) an embedded browser (JBrowse 1.12.3) displaying the sequence features on the consensus reference sequence and allowing the user to browse the tracks corresponding to all the structural and similarity-based results used by the TEdenovo pipeline to justify the TE classification proposed and (v) a detailed report corresponding to each track displayed in JBrowse. For each similarity to a protein profile or TE sequence, the coordinates of the alignment are displayed in a table. External links to the appropriate web page (i.e. PFAM, GYDB, Repbase) have been set up for the relevant identifier of hits based on similarity matching.

An implementation suitable for standardized TE detection, classification and annotation

It is now common practice to propose gene structures with all the multiple sources of evidence used by the predictors to combine them. We have designed RepetDB to function in a similar manner. In databases such as Repbase, useful evidences such as coding sequence, terminal repeats evidences are usually presented in "full text description fields" of the consensus. In addition to this "description field", DFAM, a database based on DNA profile prediction [16], also presents for each entry, the overlapping Repbase entries. The added value of RepetDB is that it displays the multiple sources of structural or similarity-based evidence used by PASTEC. Indeed, all the evidences used to automatically detect and classify TEs are presented by category in a text form, included coordinates and external links to Repbase, GyDB and PFAM and also displayed as annotation tracks in a Browser.

The TEdenovo pipeline [10] is used for the de novo detection of repeats. It uses Blaster [26], and groups repeats in a combined approach using Grouper [26], Recon [14] and Piler [27]. It then generates multiple alignments from the sequence clusters with MAP [28] and builds consensuses. Each consensus representing a structural variant of a TE family is classified with PASTEC [11], according to its structural and functional features (LTR, TIR, ORF, polyA tail), based on functional domain similarities to HMM profiles from the protein family databases PFAM [29] and GyDB [17]), and characterized TEs from the Repbase Update database [13]. Consensus sequences with no known structure or similarity are classified as noCat (i.e. "unknown"). Note that PASTEC may also be used on a TE library not provided by the TEdenovo pipeline.

RepetDB was also designed to integrate data relating to the annotation of TEs in a genome. The location of the TE copies (start, end) in the consensus and in the genome, are reported through the browser integrated into the RepetDB consensus card. A set of 26 optional metrics relating to the copies annotated are reported in a dedicated panel of the consensus card (e.g. cumulative coverage of the copies in the genome, number of copies, identity, copy lengths, copy coverage over the consensus (for the last three parameters: mean, standard deviation, median, 1st and 3rd quartiles, minimum, maximum)). The insertion of copy locations in RepetDB is based on standard GFF3 data formats (with the consensus used as the reference and the genome as the target) and a tabulated file (.tsv) containing the consensus copy statistics. When the TEannot pipeline is used, the PostAnalyse.py script (included in the REPET package) can be used to provide the statistics file in the appropriate format used by RepetDB (http://urgi.versailles.inra.fr/repetdb/docs/Data-submission.html#i.3.a.-consensus-copy-statistics). However, although any TE genome annotation pipeline can provide data in a suitable format for input into this system, the easiest way to insert data in RepetDB is to use outputs from the TEannot pipeline [30]. TEannot is based on three alignment methods (Blaster, Censor and RepeatMasker [31]). The HSPs obtained are filtered and combined. TE copies covered with SSRs over more than 75% (by default) of their length are then removed. Finally, a "long join procedure" [10] is used to address the problem of nested TEs. This procedure finds and connects fragments of TEs interrupted by other TEs inserted more recently, to build a TE copy. The nesting patterns of such insertions must respect three constraints: fragments must be collinear (both on the genome and the same TE consensus reference), of the same age (nucleotide identity with the consensus as a proxy) and separated by a more recent TE insertion.

Data insertion and integration

The dataset metadata are integrated in RepetDB in xml format. TEdenovo and TEannot output files can be used for direct insertion, for each dataset: consensus sequences (fasta format), consensus classification file (tabulation-separated format), consensus annotation files including structural and functional annotations (GFF3 format). The data are displayed in the Consensus card JBrowse panel and textual feature categories.

Data integration occurs in two major steps: "integrate" and "post process". The "integrate" step involves the loading of data from the provided files into the InterMine database. The "post process" step involves several operations, such as linking objects between data sources (e.g. linking consensuses to their wicker classification ontology terms), and data enrichment (e.g. fetching the NCBI taxonomy tree for all dataset organisms).

TE query forms

We present various methods for querying RepetDB data (Fig. 2). Like any InterMine-based information system, RepetDB benefits from standard features, such as a keyword search on the main home page, a query builder, and template queries. One of the features of the InterMine-based system providing added value is the list management system, which makes it possible to create personal lists of data (e.g. a list of consensuses in response to a query) and to perform actions on these lists, such as unions, intersections and differences. In addition to the functions intrinsic to InterMine, the RepetDB home page contains a customized form enabling the user to search for repeats by organism (taxon group selection), by classification (Wicker classification [8]), potentially chimeric or other elements, such as virus-like elements) or by similarity features (e.g. protein profile features, blast hits on transposable elements databases, such as Repbase [13]). "Potentially chimeric" elements are defined as TEs with an ambiguous classification. We also present a full-text search with the GnpIS [22] data discovery tool (https://urgi.versailles.inra.fr/gnpis), which uses autocompletion to assist the user typing keywords for the search. The results are displayed in a table with the source of the data. Results can be restricted to one or several species. Finally, we also provide a web blast tool for searching for similarities to one or more consensus sequences present in RepetDB. All the TE consensus libraries are available through the URGI BLAST [32] server (https://urgi.versailles.inra.fr/blast/).

Get data from RepetDB

RepetDB data export makes use of the default exporters of InterMine. Common standard formats, such as tabulated/csv-separated, fasta, BED, and GFF (https://genome.ucsc.edu/FAQ/FAQformat.html) are proposed on the query results and list analysis pages. The user can customize the table (addition or deletion of columns, column sorting etc.). For each data set, we provide a link to download the genome annotation performed with the consensus library of the dataset.

Results

TE reference sequences with annotation

TE reference sequences are available for 23 genomes in RepetDB, which currently stores 39,039 TE consensus. These TE consensuses were detected, classified and annotated in the framework of whole-genome sequencing projects or comparative genome analyses of transposable elements in various groups of species or organisms, such

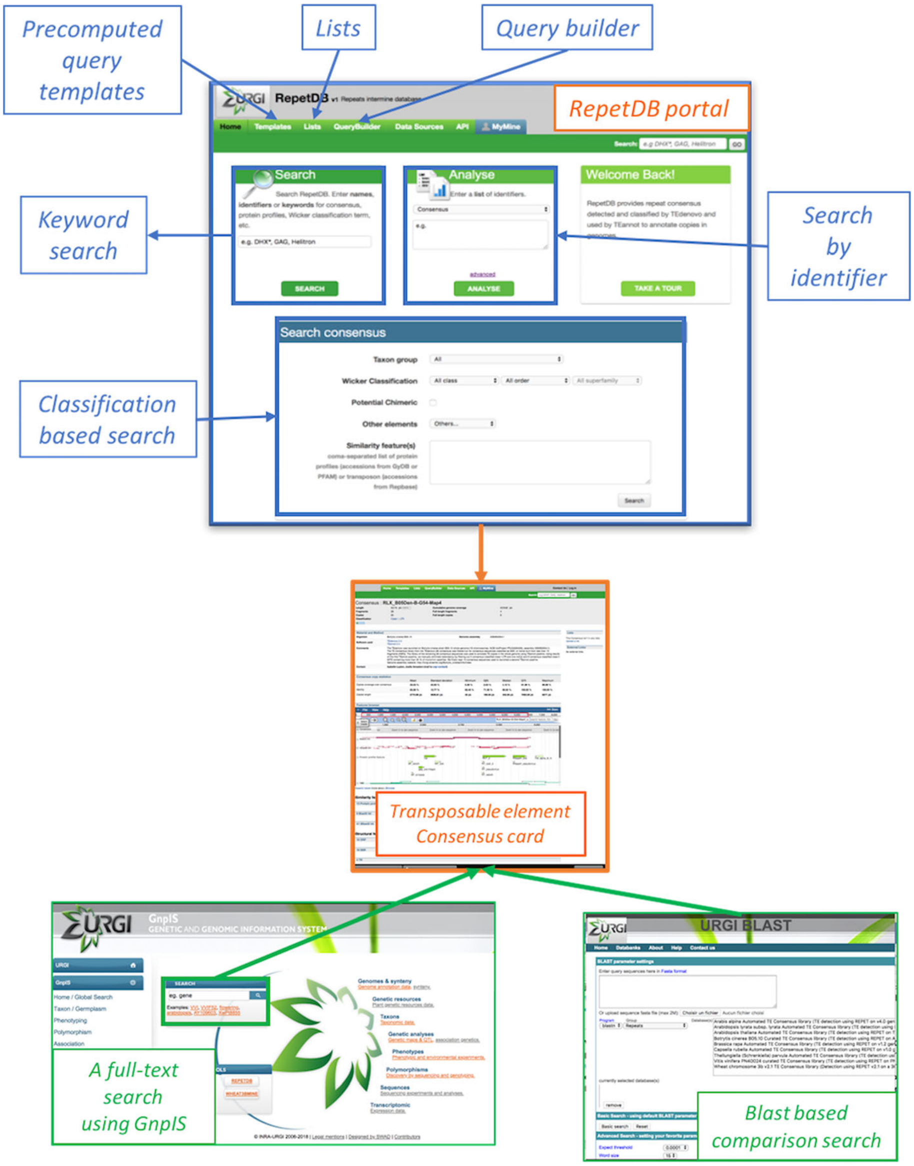


Fig. 2 The different ways to query RepetDB data. Blue boxes are RepetDB internal forms and green boxes are external queries with results linked to RepetDB

as the Brassicaceae family, Rosaceae, monocots and fungi (Table 1).

TE consensuses can be inserted into RepetDB via the TEdenovo [10, 11] and TEannot [30] pipelines, from the REPET package, which provides high-quality, standardized TE classification and genome annotation. The added value of TEdenovo stems from its ability to detect and build structural variants of TE families, representing their sometimes complex evolutionary dynamics [10]. In RepetDB, we store and provide access to the consensus built for each structural variant. The TE consensus library is used as an input of the TEannot pipeline, for the annotation of copies in the genome. For the automatic curation of some poorly defined consensuses, we run a second TEannot pipeline, using a library of consensuses filtered for sequences without full-length fragments or full-length copies in the genome. A copy may be built with one or more fragments joined by the TEannot long-join procedure. Depending on the context of TE annotation and the additional analyses performed, some of the datasets are manually curated and filtered for unreliable TE consensuses. The version, specific parameters and level of manual curation performed on the datasets are indicated in the "Comments" section of the datasets.

Table 1 Transposable element annotation metrics

Species	Genome assembly annotated (without gap)	Cumulative coverage	Genome coverage	No. of consensus sequences	No. of genome copies	No. of full-length genome copies
Brassicaceae						
Arabidopsis lyrata	206667935	76899516	37.21	2408	112563	9527
Arabidopsis thaliana	119146348	22954742	19.27	641	37129	2513
Arabis alpina	309171870	152175264	49.22	3204	268936	11729
Brassica rapa	283841084	101457103	35.74	2660	239373	10881
Capsella rubella	134834574	27975436	20.75	873	54560	3326
Schrenkiella parvula	123600562	19838473	16.05	455	37597	1356
Rosaceae						
Fragaria vesca	211673467	58062323	27.43	1543	112822	8576
Malus domestica	624851326	365363669	58.47	2456	564270	25280
Prunus persica	227411381	99590159	43.79	1738	170681	9056
Pyrus communis	577335413	194166715	33.63	975	482345	11435
Vitaceae						
Vitis vinifera	486205130	290981308	59.85	2473	475119	10551
Monocots						
Triticum aestivum	986092508	894245831	90.69	6671	785986	15905
Zea mays	2059701728	1768705851	85.87	7319	1381303	41666
Fungi						
Blumeria graminis hordei	87976437	59069666	67.14	733	122756	8909
Botrytis cinerea B0510	42630066	1583714	3.72	15	1927	263
Botrytis cinerea T4	37887365	254124	0.67	24	611	62
Colletotrichum higginsianum	50819261	3505545	6.90	41	1482	440
Magnaporthe oryzae	40949321	4549294	11.11	37	4358	463
Melampsora larici populina	97682699	49975736	51.16	1779	88708	6942
Microtryum violaceum	25201507	4423374	17.55	286	9620	640
Puccinia graminis	81521292	37620112	46.15	1625	69167	6648
Sclerotinia sclerotiorum	38001451	3459261	9.10	178	13868	622
Tuber melanosporum	123533734	73821108	59.76	905	72212	3845
Total				39039	5107393	190635

Enhancing the discoverability of RepetDB data

This well-structured database should improve the visibility of the wealth of TE reference sequences already identified. In particular, its intrinsic functions, combined with its inclusion in the international data discovery network, make it possible to improve the FAIRness of these data [33]. Indeed, RepetDB ensures data accessibility, interoperability through the use of standard formats and ontologies, such as Wickers classification terminology, and reusability. Findability is increased through the GnpIS information system data discovery portal (https://urgi.versailles.inra.fr/gnpis), but the content of RepetDB is also indexed in various international discovery portals that are currently emerging in the field of plant biology. These portals can be used to search data with free keywords, across a set of databases displaying indices on several portals based on the same distributed full-text search technology and data model [34]. The content of RepetDB is currently available in the WheatIS data discovery tool for the international wheat research community (http://wheatis.org/Search.php) and from the IFB portal (https://urgi.versailles.inra.fr/ifb/), which aims to generalize the work of the wheat community to any plant.

Discussion

A database for the reuse of rich and often underexploited annotation data

The REPET package is one of the most widely used TE annotation tools in eukaryotic genome projects (mainly for plants and fungi). Its outputs are rich and often underexploited, not going beyond initial genome annotation in new genome annotation projects. The TE identification procedure provides access to large amounts of information for each reference sequence that are often lost after publication. We developed RepetDB for the storage and sharing of this information for other potential uses. The database is populated with the results of previous analyses already performed on species of agronomic interest in the framework of several unrelated projects. We group together results obtained in studies on plants (Brassicaceae, wheat, maize, grapevine, apple, oak, Rosaceae), symbiotic fungi and pathogens [35, 36]. However, there is no reason to limit the content of this system to the branches corresponding to these species. We will consider other eukaryotic species in the future. The system will be open to external submissions, making it possible to increase its diversity and the number of species represented.

RepetDB to assist the pan-genome revolution initiated by third-generation sequencers

RepetDB is designed to be a TE knowledge base populated with full de novo TE annotations of complete (or near-complete) genome sequences. Continual improvements in sequencing technologies (which are now entering their third generation) over the last few years have made it possible to obtained genomic sequences at a cost, speed and quality close to those required for use as a routine procedure [37]. This should open up access to full-genome sequencing, which should no longer be restricted to a few representative individuals from a small number of species of interest.

Short-read resequencing, which generally provides no more than an inventory of SNPs and small indels, will probably be replaced by long-read sequencing to generate high-quality full-length genomic sequences opening up new possibilities for more detailed investigations of sequence polymorphism and diversity, to which TEs make an important contribution. We are entering the pan-genomic era, in which the search for dispensable genes in population-specific genomic sequences will be accompanied by the identification of new TE insertions. In this context, full gene and TE annotations of each individual genome will be required for subsequent analyses. The knowledge gained with one individual must be capitalized upon to guide the annotation of closely related genomes, and RepetDB is designed to make this possible.

Conclusion

RepetDB, a database of consistent TE datasets

RepetDB aims to provide a highly homogenized TE classification, with supporting evidence. We therefore integrated into the data submission process the systematic use of PASTEC [11], a tool from the REPET TEdenovo pipeline. This tool searches for evidence of TEs through dedicated similarity and structural searches, and uses this evidence to classify TEs automatically. We would therefore expect to obtain similar levels of sequence description and classification for different TEs in different species, thereby facilitating comparative genomics.

This consistency of description, classification and annotation makes it possible to perform consistent cross-species or cross-variety searches to compare TE family content between genomes. This tool can be used to explore specific TE families, superfamilies or orders across a large range of species.

Like any automatic procedure, full de novo TE annotation is likely to generate false positives, and ways of assessing correctness of TE classification relative to the corresponding reference sequence are therefore required. RepetDB is designed to display the evidence supporting each consensus classification, so as to facilitate the curation of TE reference libraries. Links to external databases are also available when similarities are detected to sequences from other databases (RepBase, GyDB [13, 17]).

Availability and requirements

Project name: RepetDB

Project home page: http://urgi.versailles.inra.fr/repetdb/

The details of data submission are available from http://urgi.versailles.inra.fr/repetdb/docs/Data-submission.html

The RepetDB user guide is available from the RepetDB portal in the "Take a tour" panel or directly from http://urgi.versailles.inra.fr/repetdb/docs/User-guide.html.

Operating system(s): Any with a web browser

Programming language: Not applicable.

Other requirements: No.

Any restrictions to use by non-academics: No, RepetDB is publicly accessible

Acknowledgments

We thank the URGI platform support, particularly C. Michotey, R. Flores, C. Guerche and M. Loaec.

Funding

The main source of funding was the TransPlant project funded by the European Commission within its 7th Framework Program, under the thematic area "Infrastructures", contract number 283496.

Availability of data and materials

The datasets generated and/or analyzed during this study, are available in the database from the "Data Sources" tab accessible from any page of RepetDB. This datasets page (http://urgi.versailles.inra.fr/repetdb/dataCategories.do) contains the links for the direct downloading (at least the GFF3 file of consensus annotations and the reference genome assembly), browsing and export of all the data.

Authors' contributions

JA: designed and managed the project, wrote the manuscript and prepared several datasets. GC: set up and developed RepetDB and inserted the data. FA-D was responsible for RepetDB databank management. NC prepared several datasets and was involved in specification of the tool. TL contributed to InterMine development. FA-D, MA, CP and A-FA-B contributed to data management and repository design. FM was involved in specification of the tool. IL, VJ prepared several datasets. HQ supervised the project, wrote the paper and prepared several datasets. All authors contributed to the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 26 October 2018 Accepted: 24 January 2019
Published online: 29 January 2019

References

1. Bennett PM. Genome plasticity. In: Woodford N, Johnson AP, editors. Genomics, proteomics, and clinical bacteriology: methods and reviews. Totowa: Humana Press; 2004. p. 71–113. https://doi.org/10.1385/1-59259-763-7:071.
2. Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, et al. Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and Transposable element spaces. Plant Cell. 2010;22:1686–701.
3. Parlange F, Oberhaensli S, Breen J, Platzer M, Taudien S, Simkova H, et al. A major invasion of transposable elements accounts for the large size of the Blumeria graminis f.Sp. tritici genome. Funct Integr Genomics. 2011;11:671–7.
4. Wicker T, Oberhaensli S, Parlange F, Buchmann J, Shatalina M, Roffler S, et al. The wheat powdery mildew genome shows the unique evolution of an obligate biotroph. Nat Genet. 2013;45:1092–6.
5. Tenaillon MI, Hufford MB, Gaut BS, Ross-Ibarra J. Genome size and transposable element content as determined by high-throughput sequencing in maize and Zea luxurians. Genome Biol Evol. 2011;3:219–29.
6. Lisch D. How important are transposons for plant evolution? Nat Rev Genet. 2013;14:49–61.
7. Slotkin R, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet. 2007;8:272–85.
8. Wicker T, Sabot F, Hua-Van A, Bennetzen J, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973–82.
9. Smith A, Hubley R. RepeatModeler Open-1.0. 2008. http://www.repeatmasker.org.
10. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering transposable element diversification in de novo annotation approaches. PLoS One. 2011;6:e16526.
11. Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, et al. PASTEC: an automatic transposable element classification tool. PLoS One. 2014;9:e91929.
12. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110:462–7.
13. Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6. https://doi.org/10.1186/s13100-015-0041-9.
14. Bao Z, Eddy S. Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. 2002;12:1269–76.
15. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;35(Web Server issue):W265–8.
16. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam database of repetitive DNA families. Nucleic Acids Res. 2016;44(Database issue):D81–9.
17. Llorens C, Futami R, Covelli L, Domínguez-Escribá L, Viu JM, Tamarit D, et al. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res. 2011;39(Database issue):D70–4.
18. Vassetzky NS, Kramerov DA. SINEBase: a database and tool for SINE analysis. Nucleic Acids Res. 2013;41(Database issue):D83–9.
19. Chen J, Hu Q, Zhang Y, Lu C, Kuang H. P-MITE: a database for plant miniature inverted-repeat transposable elements. Nucleic Acids Res. 2014;42(Database issue):D1176–81.
20. Ouyang S, Buell CR. The TIGR plant repeat databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res. 2004;32(Database issue):D360–3.
21. Nussbaumer T, Martis MM, Roessner SK, Pfeifer M, Bader KC, Sharma S, et al. MIPS PlantsDB: a database framework for comparative plant genome research. Nucleic Acids Res. 2013;41(Database issue):D1144–51.
22. Steinbach D, Alaux M, Amselem J, Choisne N, Durand S, Flores R, et al. GnpIS: an information system to integrate genetic and genomic data from plants and fungi. Database Oxf. 2013;2013:bat058.
23. Smith RN, Aleksic J, Butano D, Carr A, Contrino S, Hu F, et al. InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics. 2012;28:3163–5.
24. Kalderimis A, Lyne R, Butano D, Contrino S, Lyne M, Heimbach J, et al. InterMine: extensive web services for modern biology. Nucleic Acids Res. 2014;42(Web Server issue):W468–72.
25. Buels R, Yao E, Diesh CM, Hayes RD, Monica M-T, Helt G, et al. JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol. 2016;17:66.
26. Quesneville H, Nouaud D, Anxolabéhère D. Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae genomes. J Mol Evol. 2003;57:S50–9.
27. Edgar R, Myers E. PILER: identification and classification of genomic repeats. Bioinformatics. 2005;21(suppl 1):i152–8.
28. Huang X. On global sequence alignment. Comput Appl Biosci CABIOS. 1994;10:227–35.
29. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42:D222–30.
30. Quesneville H, Bergman C, Andrieu O, Autard D, Nouaud D, Ashburner M, et al. Combined evidence annotation of transposable elements in genome sequences. PLoS Comput Biol. 2005;1:166–75.
31. Huda A, Jordan K. Analysis of Transposable Element Sequences Using CENSOR and RepeatMasker. Methods in molecular biology (Clifton, NJ). 2009;537:323–36.
32. Deng W, Nickle DC, Learn GH, Maust B, Mullins JI. ViroBLAST: a stand-alone BLAST web server for flexible queries of multiple databases and user's datasets. Bioinformatics. 2007;23:2334–6.
33. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016. https://doi.org/10.1038/sdata.2016.18.
34. Spannagl M, Alaux M, Lange M, Bolser DM, Bader KC, Letellier T, et al. transPLANT resources for Triticeae genomic data. Plant. Genome. 2016;9. https://doi.org/10.3835/plantgenome2015.06.0038.
35. Amselem J, Lebrun M, Quesneville H. Whole genome comparative analysis of transposable elements provides new insight into mechanisms of their inactivation in fungal genomes. BMC Genomics. 2015;16:141.
36. Dallery J-F, Lapalu N, Zampounis A, Pigné S, Luyten I, Amselem J, et al. Gapless genome assembly of Colletotrichum higginsianum reveals chromosome structure and association of transposable elements with secondary metabolite gene clusters. BMC Genomics. 2017;18. https://doi.org/10.1186/s12864-017-4083-x.
37. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333–51.