BMC Genomics

DATABASE Open Access

MASiVEdb: the Sirevirus Plant Retrotransposon Database

Alexandros Bousios[1*†], Evangelia Minga[1†], Nikoleta Kalitsou[1,2], Maria Pantermali[1,2], Aphrodite Tsaballa[1,2] and Nikos Darzentas[1,3]

Abstract

Background: Sireviruses are an ancient genus of the *Copia* superfamily of LTR retrotransposons, and the only one that has exclusively proliferated within plant genomes. Based on experimental data and phylogenetic analyses, Sireviruses have successfully infiltrated many branches of the plant kingdom, extensively colonizing the genomes of grass species. Notably, it was recently shown that they have been a major force in the make-up and evolution of the maize genome, where they currently occupy ~21% of the nuclear content and ~90% of the *Copia* population. It is highly likely, therefore, that their life dynamics have been fundamental in the genome composition and organization of a plethora of plant hosts. To assist studies into their impact on plant genome evolution and also facilitate accurate identification and annotation of transposable elements in sequencing projects, we developed MASiVEdb (Mapping and Analysis of SireVirus Elements Database), a collective and systematic resource of Sireviruses in plants.

Description: Taking advantage of the increasing availability of plant genomic sequences, and using an updated version of MASiVE, an algorithm specifically designed to identify Sireviruses based on their highly conserved genome structure, we populated MASiVEdb (http://bat.infspire.org/databases/masivedb/) with data on 16,243 intact Sireviruses (total length >158Mb) discovered in 11 fully-sequenced plant genomes. MASiVEdb is unlike any other transposable element database, providing a multitude of highly curated and detailed information on a specific genus across its hosts, such as complete set of coordinates, insertion age, and an analytical breakdown of the structure and gene complement of each element. All data are readily available through basic and advanced query interfaces, batch retrieval, and downloadable files. A purpose-built system is also offered for detecting and visualizing similarity between user sequences and Sireviruses, as well as for coding domain discovery and phylogenetic analysis.

Conclusion: MASiVEdb is currently the most comprehensive directory of Sireviruses, and as such complements other efforts in cataloguing plant transposable elements and elucidating their role in host genome evolution. Such insights will gradually deepen, as we plan to further improve MASiVEdb by phylogenetically mapping Sireviruses into families, by including data on fragments and solo LTRs, and by incorporating elements from newly-released genomes.

* Correspondence: alexandros.bousios@gmail.com
†Equal contributors
[1]Institute of Agrobiotechnology, Centre for Research and Technology Hellas, Thessaloniki 57001, Greece
Full list of author information is available at the end of the article

Background

The intense activity of long terminal repeat (LTR) retrotransposons has been among the major drivers (together with polyploidization events) for the often enormous size of plant genomes [1-3]. This phenomenon was initially suggested to lead plants to 'genomic obesity' [4], before it was shown that mechanisms of LTR retrotransposon removal counterbalance this propensity [5-7]. The relative success of these two opposing forces likely underlies the impressive variation in the LTR retrotransposon content of plant genomes. Just over 7% of the tiny genome (125 Mb) of *Arabidopsis thaliana* is occupied by this transposable element (TE) type [8,9], in contrast to approximately 25% of the rice genome (389 Mb) [10], ~75% of the maize genome (2,300 Mb) [11], and ~65% of the wheat genome (16,000 Mb) [12]. Although these vast genomic stretches were long dismissed as 'selfish' or 'junk' DNA [13,14], they have eventually emerged as a major evolutionary force with profound effects not only on the structure, organization and composition of the host epi/genome, but also on the evolution, function, and regulation of genes [15-18].

Sireviruses are an ancient LTR retrotransposon genus of the *Copia* superfamily, and the only one (of either *Copia* or *Gypsy*) that has exclusively proliferated within the plant kingdom [19]. Due to their host specificity they were originally termed Agroviruses [19], before being renamed to Sireviruses (derived from the SIRE1 element of soybean [20]) by the International Committee on the Taxonomy of Viruses (ICTV) [21]. In contrast to ICTV which has divided the *Copia* (or *Pseudoviridae*) superfamily into three genera (i.e. Sireviruses, Hemiviruses and Pseudoviruses), the unified classification system for eukaryotic TEs [22] is devoid of an analogous genus-level taxonomy. This has left as yet open questions on the position of Sireviruses within the LTR retrotransposon order, and whether they should actually be considered as viruses. Based on all published work on Sireviruses so far [23-28], it is safe to assume that the characteristics of their life cycle correspond to that of typical LTR retrotransposons and not of viruses.

Sireviruses have infiltrated many phylogenetic branches of flowering plants, as several elements from a plethora of monocot and eudicot species have been classified as Sireviruses [23,27] (Additional file 1: Figure S1). More specifically, and based on the Sirevirus origin [23] of abundant families in rice [29], barley and wheat [12,30], they have extensively colonized the genomes of grasses. Notably, Sireviruses currently take up ~21% of the maize genome and ~90% of the maize *Copia* population, with the majority of copies accumulating during the last 600,000 years [24]. Moreover, experimental evidence suggests that Sireviruses are present in high numbers in other species such as legumes [20,31,32], beets [33], bananas [34], and agaves [35], while approximately half of the *Copia* sequences deposited in GenBank belong to this genus [26]. Therefore, through their widespread, intense and complex colonization patterns in plant genomes, Sireviruses seem to have been critical in the evolution of their hosts.

Sireviruses are also unique among LTR retrotransposons in terms of their own genome structure [23,25]. It is the only *Copia* genus whose members often possess a putative envelope-like (*ENV*-like) gene [36], which however shares little sequence similarity among elements apart from the presence of transmembrane and coiled-coil domains [19,27]. The origin of this *ENV*-like gene in Sireviruses (and some *Gypsy* plant LTR retrotransposons) is currently unknown. Moreover, apart from elucidating the mechanisms by which it is expressed (i.e. stop codon suppression, internal promoter) [28,37,38], its function (if any) has not been experimentally proven. As a result, its role remains highly controversial [39], and it is likely that it may not even represent a true envelope gene. Another intriguing characteristic of Sireviruses is the presence of a variety of highly conserved sequence motifs within their extremely divergent genome (Additional file 1: Figure S1), regardless of the evolutionary distance between their hosts. The motifs are located in key non-coding domains known to decisively participate in the life cycle of LTR retrotransposons, and may be the underlying factors for the affinity of Sireviruses for plants [23].

Due to their abundance and complex insertion patterns, efficient annotation of TEs is among the most cumbersome and problematic tasks of genome sequencing projects. Despite the use of several structural [40-42] and homology-based methods [43], identification is hampered or misguided by the often recombined, degraded and nested genome structure of LTR retrotransposons, or by their low copy number or uniqueness that renders them invincible to detection by comparison with previously characterized elements of abundant families. To alleviate such issues in the analysis of Sireviruses, we recently developed an algorithm able to identify Sirevirus elements with high accuracy and sensitivity [44]. Initially implemented in maize [24], it yielded >2,700 previously unidentified intact Sireviruses and offered insights into their crucial role in the evolution of the maize genome. Herein, the algorithm was updated and applied on a curated collection of 14 fully-sequenced plant and algal genomes to create MASiVEdb (Mapping and Analysis of SireVirus Elements Database).

MASiVEdb offers, through multiple and often novel ways, a comprehensive, highly curated and detailed report on the full-length Sirevirus complement of each species, while it additionally includes an integrated system for analyzing user-provided sequences against MASiVEdb elements. In this way, MASiVEdb is unlike any other TE database, thus complementing the efforts of research groups to collect and organize repetitive sequences. Such widely used and useful databases include the TIGR plant repeat database [45], Repbase [46], TREP (the Triticeae Repeat Sequence Database) [47], the maize TE database [48,49], the species-specific RetrOryza [50] and SoyTEdb [51] of rice LTR retrotransposons and soybean TEs respectively, and the GyDB *Gypsy* database [52].

This unique directory of Sireviruses can aid the scientific community in a variety of ways. Firstly, it is a consistent and up-to-date source of full-length Sireviruses (at present totaling 16,243 elements), which can significantly improve TE annotation not only of species currently included in the database but also of other plant genomes. Among other, the above will enable comparative TE studies at whole genome levels, and analyses of interactions between Sireviruses and host genes. Finally from an evolutionary perspective, MASiVEdb provides the foundation for studying the depth and impact of infiltration of this intriguing TE genus across plants, and for discerning what underlies their success or failure in massively colonizing different phylogenetic branches of the plant kingdom.

Construction and content of MASiVEdb

Sequence data of twelve fully-sequenced plant genomes were downloaded from several websites (Table 1), alongside the genomes of two green algae representing the basal branch that gave rise to higher plants [53]. The algae were included as outgroups to investigate whether Sireviruses have been present during the time when land plants emerged.

All sequences were manually inspected to remove smaller contigs and scaffolds, properly formatted and split into chromosomes. For each chromosome, intact Sireviruses were identified and analysed with an updated version of the MASiVE algorithm [44]. MASiVE is based on the step-by-step identification of Sirevirus-specific and other critical sequence motifs of LTR retrotransposons, which have been shown to provide base-pair accuracy in outlining the element. The update mainly concerned the removal of a preliminary run of the LTRharvest algorithm [42] for detecting generic LTR retrotransposons. The exclusion of this step increased sensitivity without sacrificing accuracy, as was confirmed with large-scale manual inspection of the resulting data. Further improvements included optimized order of steps, element overlap detection, and data output. Specifically for the purposes of MASiVEdb we developed custom-built PERL scripts for the detection of the integrase (*INT*), monocot/eudicot *ENV*-like core domains, the multiple zf-CCHC motif of the *gag* gene [19], and the target site duplication of each element.

Intact Sireviruses were not detected in three out of the 14 species, the green algae *Chlamydomonas reinhardtii* and *Ostreococcus lucimarinus*, and the tree *Populus trichocarpa*, hence providing preliminary evidence that

Table 1 Properties of the host species and their Sirevirus populations included in MASiVEdb

host species[a]	common name	clade	genome size (Mb)[b]	number of chr.	intact SVs	avg age (my)	with *ENV*
Arabidopsis thaliana	thale cress	eudicot	116	5	4	0.95	2
Brachypodium distachyon	brome	monocot	262	5	22	1.97	14
Chlamydomonas reinhardtii	n/a	alga	99	17	0	n/a	0
Fragaria vesca	strawberry	eudicot	203	7	1	0.45	0
Glycine max	soybean	eudicot	915	20	1337	0.45	1294
Lotus japonicus	lotus	eudicot	291	7	282	0.35	270
Oryza sativa indica	rice	monocot	362	12	25	2.18	14
Oryza sativa japonica	rice	monocot	370	12	91	1.29	42
Ostreococcus lucimarinus	n/a	alga	13	21	0	n/a	0
Populus trichocarpa	poplar	eudicot	304	19	0	n/a	0
Sorghum bicolor	sorghum	monocot	633	10	522	0.80	227
Theobroma cocoa	cacao	eudicot	214	10	77	3.17	52
Vitis vinifera	grapevine	eudicot	414	19	49	1.73	45
Zea mays	maize	monocot	1969	10	13833	1.29	516
Total					16243	1.20	2476

[a]The chromosome sequence data for *Arabidopsis* were downloaded from http://www.arabidopsis.org/, for brachypodium, *Chlamydomonas*, soybean, poplar, sorghum and grapevine from http://www.phytozome.net/, for strawberry from http://www.strawberrygenome.org/, for lotus from http://www.kazusa.or.jp/lotus/, for rice (*indica*) from http://rice.genomics.org.cn/rice/, for rice (*japonica*) from http://rgp.dna.affrc.go.jp/E/IRGSP/Build5/build5, for *Ostreococcus* from http://genome.jgi-psf.org/, for cacao from http://cocoagendb.cirad.fr/, and for maize from http://www.maizesequence.org/; [b]The genome sizes do not represent the real estimates for each species, but correspond to the cumulative Mb found in sequence files for chromosomes after removing unanchored contigs and scaffolds; chr, chromosome; SVs, Sireviruses; my, million years; *ENV*, envelope-like gene.

Sireviruses have neither been present when land plants emerged from algae, nor have they successfully colonized all branches of the plant kingdom.

The remaining 11 species provided a total of 16,243 full-length elements (Table 1), with a highly variable abundance, ranging from just one Sirevirus in strawberry and four in *Arabidopsis*, to 1,337 in soybean and 13,833 in maize. The total length of the detected elements exceeds 158 Mb. Sireviruses appear to have been active in different time periods in the genomes of their hosts, with an average insertion age of 0.35 million years ago (mya) in lotus (median of 0.16) to 3.17 mya in cacao (median of 3.06). There are also stark differences in the distribution of elements containing the *ENV*-like gene. Nearly all Sireviruses identified within the eudicot genomes of soybean, lotus, cacao and grapevine carry the *ENV*-like gene, in contrast to approximately half of the populations present in grasses, including brome, rice and sorghum. Notably, and as shown in recent work [24], the vast majority of maize Sireviruses are devoid of it. Given the availability of a large number of such sequences in the Sireviruses of MASiVEdb, as well as in *Gypsy* LTR retrotransposons and other TEs available in aforementioned databases, it may now be possible to elucidate its evolutionary origin and test recently formed hypotheses [54], and also investigate whether its function (if any) can attribute retrovirus-like properties to the carrier-elements.

All data were organized per species into tab-delimited, GFF-formatted, FASTA-formatted, and database-ready files. The latter were loaded into a two-table schema in the postgreSQL software system, whilst the rest are available for download as described below. Data are divided in four categories: '*basic*', which includes the date of the run and version of the MASiVE algorithm used, the host species, Sirevirus identifier, chromosome, direction, coordinates, and distance to centromere (where available); '*advanced*', which includes phylogenetic information on Sireviruses (currently unavailable – see Future development), presence of the *ENV*-like gene, age or time of insertion (in million years, e.g. 0.1 equals to an age of 100,000 years), and length of the element and its LTRs; '*genes*', which includes the starting position (within the element) and length of the core domains of the reverse transcriptase (*RT*), *INT* and *ENV*-like genes; and '*motifs*', which include the target site duplication, the primer binding site (PBS) sequence and starting position, and detailed analysis of the zf-CCHC motifs and multiple polypurine tract (PPT) signature [23]. The Sirevirus identifier of the MASiVEdb is constructed with the host species four- or five-letter code (also available online in the home page), and the direction and start coordinate of the element, e.g. Ljap_chr_3-D-12553650 stands for a lotus Sirevirus that was identified on position 12,553,650 of the sense strand (D for direct, in contrast to P for palindromic) of chromosome 3 of the lotus genome.

The GFF-formatted files, viewable in appropriate browsers, provide the coordinates, direction, identifier and length of the element and each LTR, plus the age of the element; and the coordinates, sequences, and direction of the motifs of the multiple PPT signature of each element. Finally, the FASTA-formatted files provide the sequences of the full-length element and its LTRs, the multiple PPT signature, the zf-CCHC motifs, and the core domains of the *RT*, *INT*, and *ENV*-like genes.

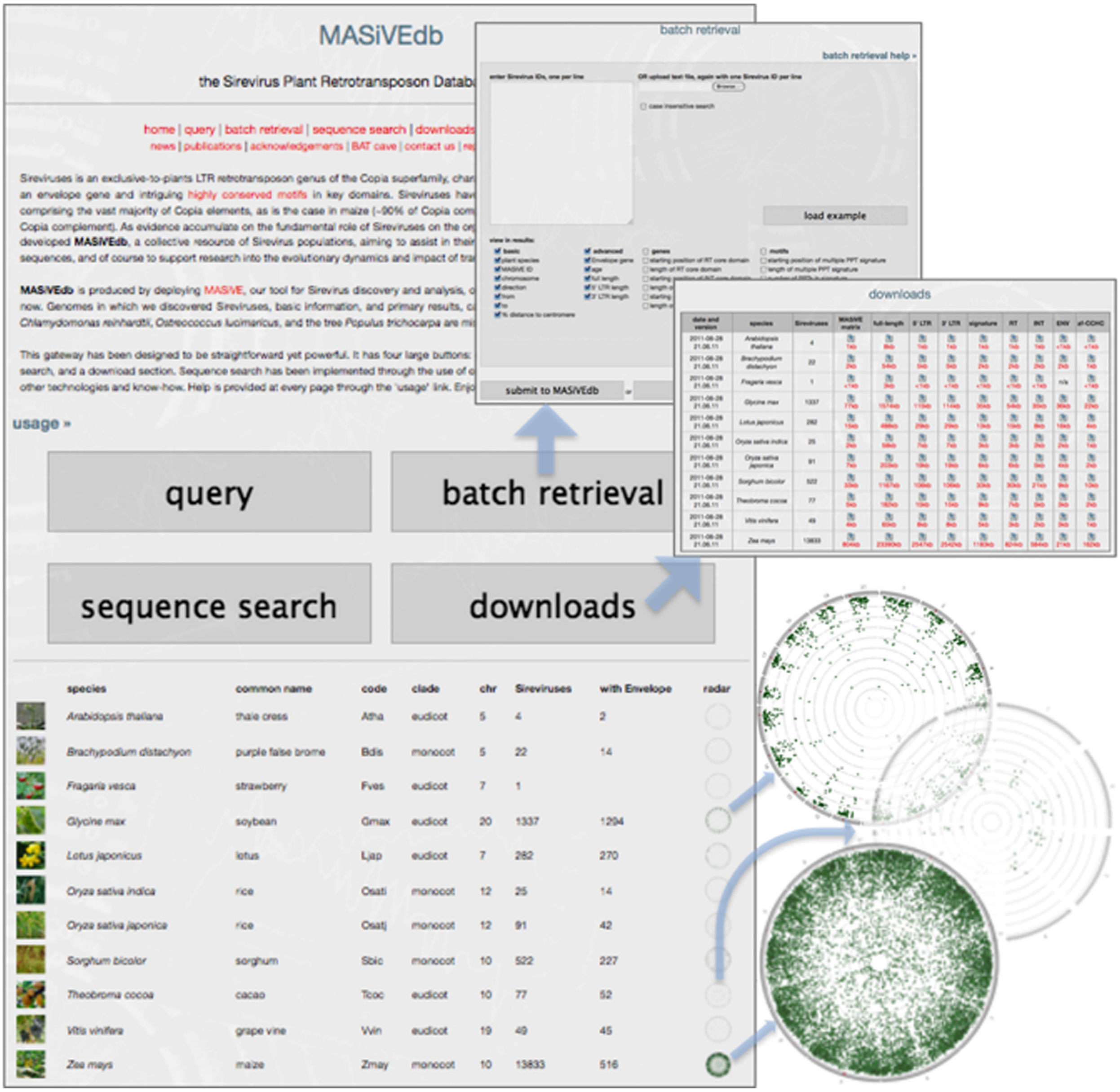
Utility

The top of the home page of MASiVEdb features a primary menu with the main sections, and a secondary menu with quick links to news, publications, acknowledgements (which include data sources, versions, dates and related information), and three pre-defined links for generic communication, error reporting, and genome inclusion requests (Figure 1). After a short introduction, four large buttons lead to the four main sections, presented below. Finally, a summary table provides information regarding the plant species and their Sirevirus content that are included in MASiVEdb, together with a Circos-based [55] circular representation of the Sirevirus abundance, localization and age across the chromosomes of each host.

Query

MASiVEdb offers both a simple and an advanced query interface, providing access to the vast majority of information available. We advise users to familiarize themselves with the simple form, its content and structure, before attempting an advanced query. The simple form begins with the choice of species, either one or all (Figure 2A). This opens up a multitude of fields, divided in the four categories analyzed previously (i.e. *basic*, *advanced*, *genes*, *motifs*) that are either queryable by text or drop-down lists, or just selectable for viewing in results. Alternatively, the user can submit without input, in which case all Sireviruses for the selected species will be returned. The advanced form provides a more streamlined and powerful way of accessing the data (Figure 2B). Fields allow the selection of which data to show in results, from which species, and the constraints the user wants to impose based on different relationship and logical operators. Importantly, for assisting research in specific chromosome segments, there is an option in both forms for retrieving all data within a user defined sequence window.

The interactive output matrix contains the information requested by the user (Figure 2C), where additional filters can be used to further process the data. The results



MASiVEdb

the Sirevirus Plant Retrotransposon Databa

home | query | batch retrieval | sequence search | downloads
news | publications | acknowledgements | BAT cave | contact us | re

Sireviruses is an exclusive-to-plants LTR retrotransposon genus of the Copia superfamily, char
an envelope gene and intriguing highly conserved motifs in key domains. Sireviruses hav
comprising the vast majority of Copia elements, as is the case in maize (~90% of Copia com
Copia complement). As evidence accumulate on the fundamental role of Sireviruses on the or
developed **MASiVEdb**, a collective resource of Sirevirus populations, aiming to assist in their
sequences, and of course to support research into the evolutionary dynamics and impact of tra

MASiVEdb is produced by deploying MASiVE, our tool for Sirevirus discovery and analysis, c
now. Genomes in which we discovered Sireviruses, basic information, and primary results, c
Chlamydomonas reinhardtii, *Ostreococcus lucimarinus*, and the tree *Populus trichocarpa* are mi

This gateway has been designed to be straightforward yet powerful. It has four large buttons:
search, and a download section. Sequence search has been implemented through the use of o
other technologies and know-how. Help is provided at every page through the 'usage' link. Enjo

usage »

species	common name	code	clade	chr	Sireviruses	with Envelope	radar
Arabidopsis thaliana	thale cress	Atha	eudicot	5	4	2	○
Brachypodium distachyon	purple false brome	Bdis	monocot	5	22	14	○
Fragaria vesca	strawberry	Fves	eudicot	7	1		○
Glycine max	soybean	Gmax	eudicot	20	1337	1294	○
Lotus japonicus	lotus	Ljap	eudicot	7	282	270	○
Oryza sativa indica	rice	Osati	monocot	12	25	14	○
Oryza sativa japonica	rice	Osatj	monocot	12	91	42	○
Sorghum bicolor	sorghum	Sbic	monocot	10	522	227	○
Theobroma cocoa	cacao	Tcoc	eudicot	10	77	52	○
Vitis vinifera	grape vine	Vvin	eudicot	19	49	45	○
Zea mays	maize	Zmay	monocot	10	13833	516	●

batch retrieval

batch retrieval help »

enter Sirevirus IDs, one per line

OR upload text file, again with one Sirevirus ID per line

Browse...

□ case insensitive search

load example

view in results:

☑ basic
☑ plant species
☑ MASiVE ID
☑ chromosome
☑ direction
☑ from
☑ to
☑ % distance to centromere

☑ advanced
☑ Envelope gene
☑ age
☑ full length
☑ 5' LTR length
☑ 3' LTR length

□ genes
□ starting position of RT core domain
□ length of RT core domain
□ starting position of INT core domain
□ length o
□ starting
□ length o

□ motifs
□ starting position of multiple PPT signature
□ length of multiple PPT signature
□ number of PPTs in signature

submit to MASiVEdb

downloads

date and version	species	Sireviruses	MASiVE matrix	full-length	5' LTR	3' LTR	signature	RT	INT	ENV	at-CCHC
2011-08-28 21.06.11	*Arabidopsis thaliana*	4	1kb	8kb	1kb	1kb	1kb	1kb	1kb	<1kb	<1kb
2011-08-28 21.06.11	*Brachypodium distachyon*	22	2kb	54kb	5kb	5kb	2kb	2kb	2kb	2kb	1kb
2011-08-28 21.06.11	*Fragaria vesca*	1	<1kb	3kb	<1kb	<1kb	<1kb	<1kb	<1kb	n/a	<1kb
2011-08-28 21.06.11	*Glycine max*	1337	77kb	1574kb	115kb	114kb	35kb	54kb	35kb	36kb	22kb
2011-08-28 21.06.11	*Lotus japonicus*	282	15kb	488kb	29kb	29kb	13kb	10kb	8kb	16kb	4kb
2011-08-28 21.06.11	*Oryza sativa indica*	25	2kb	58kb	7kb	7kb	3kb	3kb	2kb	2kb	1kb
2011-08-28 21.06.11	*Oryza sativa japonica*	91	7kb	203kb	19kb	19kb	6kb	6kb	5kb	4kb	2kb
2011-08-28 21.06.11	*Sorghum bicolor*	522	33kb	1167kb	106kb	106kb	33kb	30kb	21kb	9kb	13kb
2011-08-28 21.06.11	*Theobroma cocoa*	77	5kb	182kb	15kb	15kb	8kb	7kb	5kb	3kb	2kb
2011-08-28 21.06.11	*Vitis vinifera*	49	4kb	93kb	8kb	8kb	5kb	3kb	2kb	3kb	1kb
2011-08-28 21.06.11	*Zea mays*	13833	804kb	23390kb	2547kb	2542kb	1183kb	824kb	584kb	21kb	162kb

Figure 1 The web interface of MASiVEdb. The large buttons on the home page lead to the four main sections, including batch retrieval and downloads captured here. Clicking on the 'radar' icon of each species produces a Circos-based [55] image of the abundance, chromosomal localization and age distribution of its Sireviruses.

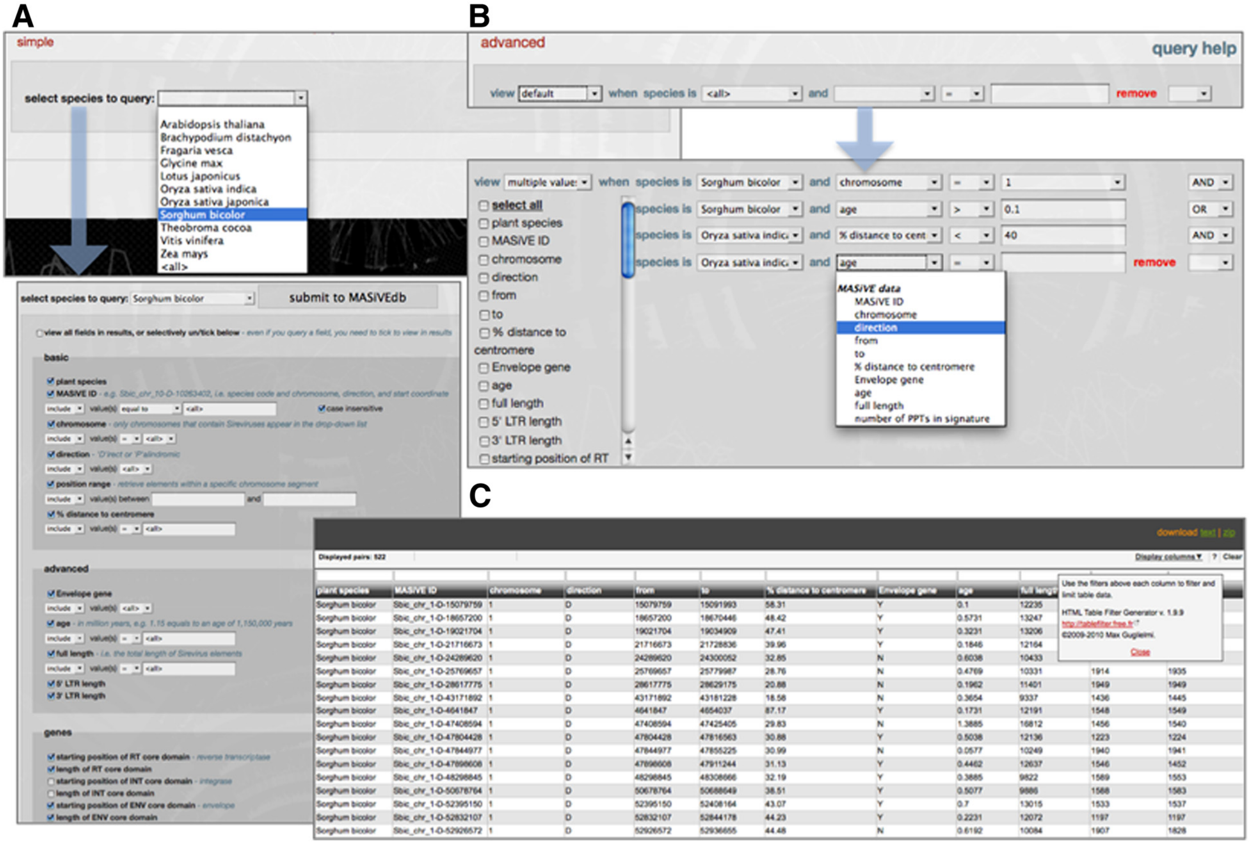
can be downloaded in tab-delimited text format, and the related sequences as FASTA-formatted files.

Batch retrieval

The batch retrieval function (Figure 1), as the name suggests, allows the simultaneous querying of MASiVEdb with one or more Sirevirus identifiers, which means that these have to be available to the user through e.g. a previous query. Again, data to be returned in the results can be selected.

Sequence search

We consider the sequence similarity-based access to the Sirevirus data among the most important aspects of MASiVEdb, for which we developed an integrated system collectively termed 'LTRphyler', and linked it to the database sequences. Through LTRphyler users can examine whether their sequences contain Sirevirus-related fragments, visualize the sequence similarity in a highly informative way, and infer the phylogenetic position of their query (if successful) within the *Copia* tree. More



select species to query:

Arabidopsis thaliana
Brachypodium distachyon
Fragaria vesca
Glycine max
Lotus japonicus
Oryza sativa indica
Oryza sativa japonica
Sorghum bicolor
Theobroma cocoa
Vitis vinifera
Zea mays
<all>

select species to query: Sorghum bicolor

□ view all fields in results, or selectively un/tick below - *even if you query a field, you need to tick to view in results*

basic

☑ plant species
☑ MASiVE ID - *e.g. Sbic_chr_10-D-10063402, i.e. species code and chromosome, direction, and start coordinate*
include value(s) equal to <all> ☑ case insensitive
☑ chromosome - *only chromosomes that contain Sireviruses appear in the drop-down list*
include value(s) = <all>
☑ direction - *'D'irect or 'P'alindromic*
include value(s) <all>
☑ position range - *retrieve elements within a specific chromosome segment*
include value(s) between and
☑ % distance to centromere
include value(s) = <all>

advanced

☑ Envelope gene
include value(s) <all>
☑ age - *in million years, e.g. 1.15 equals to an age of 1,150,000 years*
include value(s) = <all>
☑ full length - *i.e. the total length of Sirevirus elements*
include value(s) = <all>
☑ 5' LTR length
☑ 3' LTR length

genes

☑ starting position of RT core domain - *reverse transcriptase*
☑ length of RT core domain
□ starting position of INT core domain - *integrase*
□ length of INT core domain
☑ starting position of ENV core domain - *envelope*
☑ length of ENV core domain

view default when species is <all> and = remove

view multiple values when species is Sorghum bicolor and chromosome = 1 AND
species is Sorghum bicolor and age > 0.1 OR
species is Oryza sativa indica and % distance to cent < 40 AND
species is Oryza sativa indica and age = remove

□ select all
□ plant species
□ MASiVE ID
□ chromosome
□ direction
□ from
□ to
□ % distance to centromere
□ Envelope gene
□ age
□ full length
□ 5' LTR length
□ 3' LTR length
□ starting position of RT

MASiVE data
MASiVE ID
chromosome
direction
from
to
% distance to centromere
Envelope gene
age
full length
number of PPTs in signature

Displayed pairs: 522

Use the filters above each column to filter and limit table data.

HTML Table Filter Generator v. 1.9.9
http://tablefilter.free.fr
©2009-2010 Max Guglielmi.

plant species	MASiVE ID	chromosome	direction	from	to	% distance to centromere	Envelope gene	age	full lengt		
Sorghum bicolor	Sbic_chr_1-D-15079759	1	D	15079759	15091993	58.31	Y	0.1	12235		
Sorghum bicolor	Sbic_chr_1-D-18657200	1	D	18657200	18670446	48.42	Y	0.5731	13247		
Sorghum bicolor	Sbic_chr_1-D-19021704	1	D	19021704	19034909	47.41	Y	0.3231	13206		
Sorghum bicolor	Sbic_chr_1-D-21716673	1	D	21716673	21728836	39.96	Y	0.1846	12164		
Sorghum bicolor	Sbic_chr_1-D-24289620	1	D	24289620	24300052	32.85	N	0.6038	10433		
Sorghum bicolor	Sbic_chr_1-D-25769657	1	D	25769657	25779987	28.76	N	0.4769	10331	1914	1935
Sorghum bicolor	Sbic_chr_1-D-28617775	1	D	28617775	28629175	20.88	N	0.1962	11401	1949	1949
Sorghum bicolor	Sbic_chr_1-D-43171892	1	D	43171892	43181228	18.58	N	0.3654	9337	1436	1445
Sorghum bicolor	Sbic_chr_1-D-4641847	1	D	4641847	4654037	87.17	Y	0.1731	12191	1548	1549
Sorghum bicolor	Sbic_chr_1-D-47408594	1	D	47408594	47425405	29.83	N	1.3885	16812	1456	1540
Sorghum bicolor	Sbic_chr_1-D-47804428	1	D	47804428	47816563	30.88	Y	0.5038	12136	1223	1224
Sorghum bicolor	Sbic_chr_1-D-47844977	1	D	47844977	47855225	30.99	N	0.0577	10249	1940	1941
Sorghum bicolor	Sbic_chr_1-D-47898608	1	D	47898608	47911244	31.13	Y	0.4462	12637	1546	1452
Sorghum bicolor	Sbic_chr_1-D-48298845	1	D	48298845	48308666	32.19	Y	0.3885	9822	1589	1553
Sorghum bicolor	Sbic_chr_1-D-50678764	1	D	50678764	50688649	38.51	Y	0.5077	9886	1588	1583
Sorghum bicolor	Sbic_chr_1-D-52395150	1	D	52395150	52408164	43.07	Y	0.7	13015	1533	1537
Sorghum bicolor	Sbic_chr_1-D-52832107	1	D	52832107	52844178	44.23	Y	0.2231	12072	1197	1197
Sorghum bicolor	Sbic_chr_1-D-52926572	1	D	52926572	52936655	44.48	N	0.6192	10084	1907	1828

Figure 2 The query interface and output of MASiVEdb. (A) Simple form: the user has to select a host species or all species from the drop-down menu, which then opens an adapted list of choices and filters. (B) Advanced form: here users can retrieve multiple information from multiple species simultaneously. (C) The output matrix is common for both forms and permits further interactive processing (i.e. sorting and filtering).

specifically, LTRphyler is a BLAST-based system that combines the use of Circoletto [56] for visualization, of the Wise2 package [57] for the detection of the *RT* and *INT* genes, and of MAFFT [58] for the construction of the *RT*- and *INT*-derived draft phylogenetic trees. The Circoletto visualization has been complemented with the highlighting of LTRs, the zf-CCHC motifs, and the *RT*, *INT* and *ENV*-like core domains.

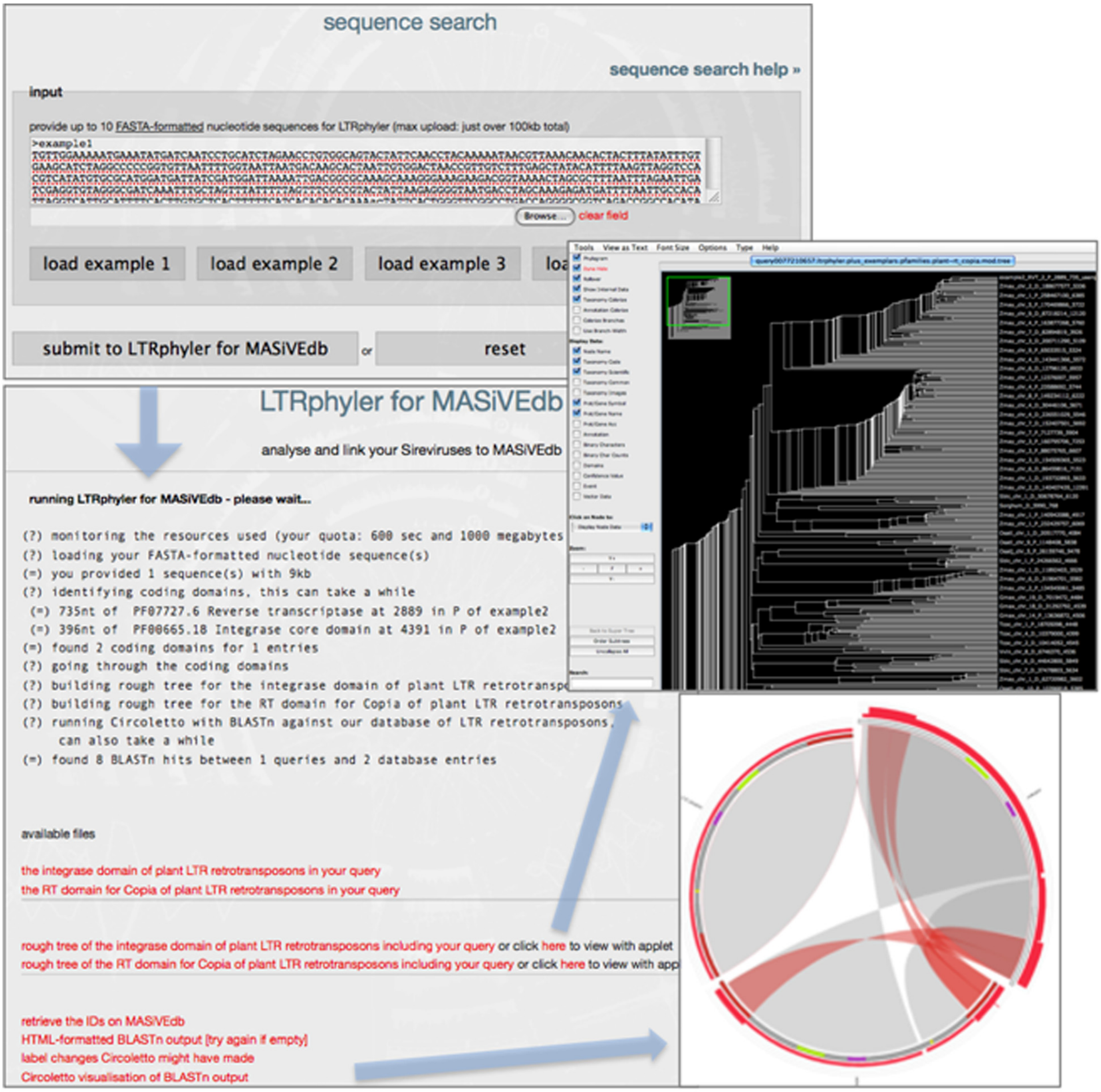
Input are FASTA-formatted nucleotide sequences (Figure 3) assumed to contain intact or fragmented Sirevirus elements. The report includes information on i) the 'health' of the user sequences (i.e. if the input is in the wrong format, not a nucleotide sequence, or above the maximum length of 100 kb, then the program will report and exit), ii) the detection of the *RT* and *INT* core domains, with details of alignment length, direction, and coordinates, iii) the tree building for the *RT* and *INT* core domains detected in the query, together with >400 MASiVEdb exemplars, and a precompiled dataset of known Sireviruses and other *Copia* elements, iv) the run of Circoletto against MASiVEdb. Output might include: i) the FASTA-formatted sequences of the *RT* and *INT* genes in the query, ii) downloadable or applet viewable trees, iii) BLASTn output in HTML format, iv) a link to load the Sirevirus identities in 'batch retrieval,' v) Circoletto visualization of the BLASTn results.

Downloads

Finally, compressed data files (with file size) are available for download. They are divided per species and per content type, with information provided for the date of the run and the version of the MASiVE algorithm used (Figure 1).

Discussion

Due to the difficulty in correctly assigning LTR retrotransposon families into the genera of the *Copia*/*Gypsy* superfamilies (which possibly contributed to the omission of a taxonomic step below 'superfamily' in the proposed TE classification system [22]), and to the scarce reference on the Sirevirus origin of elements, research on Sireviruses has been very limited so far. Hence, despite their abundance



provide up to 10 FASTA-formatted nucleotide sequences for LTRphyler (max upload: just over 100kb total)

```
>example1
TGTTGGAAAAATGAAATATGATCAATCCTGCATCTAGAACCTGTGGCAGTACTATTCAACCTACAAAAATAACGTTAAACAACACTACTTTATATTTGT
GAAGCATCTAGGCCCCCGGTGTTAATTTTGGTAATTAATGACAAGCACTAATTGTGGACTAACCGTTGTCTTTGAGCTATACATTTTAAGTTAGGTCCA
CGTCATATGTGCGCATGGATGATTATCGATGGATTAAAATTGACGGCGCAAAGCAAAGGGAAAGAAGACGGTAAAACTAGCGCTTTAATTTAGAATTGA
TCGAGGTGTAGGGCGATCAAATTTGCTAGTTTATTTTTAGTTTCGCCGTACTATTAAGAGGGGTAATGACCTAGCAAAGAGATGATTTTAATTGCCACA
TTAGGTCATTGCATTTTCACTTGTGCTCACTTTTTCATCACACACACAAAACTATTCACTGGGTTCGGCCTGACCAGGGCGGTCAGACCGGCCACATA
```

running LTRphyler for MASiVEdb - please wait...

```
(?) monitoring the resources used (your quota: 600 sec and 1000 megabytes
(?) loading your FASTA-formatted nucleotide sequence(s)
(=) you provided 1 sequence(s) with 9kb
(?) identifying coding domains, this can take a while
 (=) 735nt of  PF07727.6 Reverse transcriptase at 2889 in P of example2
 (=) 396nt of  PF00665.18 Integrase core domain at 4391 in P of example2
(=) found 2 coding domains for 1 entries
(?) going through the coding domains
(?) building rough tree for the integrase domain of plant LTR retrotransp
(?) building rough tree for the RT domain for Copia of plant LTR retrotransposons
(?) running Circoletto with BLASTn against our database of LTR retrotransposons,
     can also take a while
(=) found 8 BLASTn hits between 1 queries and 2 database entries
```

available files

the integrase domain of plant LTR retrotransposons in your query
the RT domain for Copia of plant LTR retrotransposons in your query

rough tree of the integrase domain of plant LTR retrotransposons including your query or click here to view with applet
rough tree of the RT domain for Copia of plant LTR retrotransposons including your query or click here to view with app

retrieve the IDs on MASiVEdb
HTML-formatted BLASTn output [try again if empty]
label changes Circoletto might have made
Circoletto visualisation of BLASTn output

Figure 3 The sequence search interface and output of MASiVEdb. Besides reporting on the results of the analysis, the output page also provides links to the *RT*- and *INT*-based *Copia* phylogenetic trees and to the visualization of the sequence similarity through Circoletto.

and wide distribution in plant genomes as suggested in a small number of earlier publications [26,27,32,33], the implications of their colonization dynamics are currently unknown. The recent discovery of their highly conserved genome structure [23,25], however, enabled their efficient and collective identification in one step [44], which has already proven crucial in elucidating their role in the structure and evolution of the maize genome [24]. Large-scale studies on other TE superfamilies like Helitrons [59-61] or subclasses like Pack-MULEs [62-64], which are distinguished by their structural characteristics or their amplification intricacies (i.e. carrying gene fragments), has shown that research at these higher classification levels can provide valuable insights into the mechanisms of plant genome evolution.

We argue that MASiVEdb is a step towards this direction. It represents the resource and methodological platform that can support research for uncovering the integrative impact of a specific TE genus on plant genomes - the first such attempt for LTR retrotransposons, excluding research on

Gypsy chromoviruses [65,66]. In this respect, but also based on the functionality and technologies it incorporates, MASiVEdb is unique, and hence, complementary to the compendium of related databases [45-52]. Consequently, our aim for MASiVEdb is to assist other similar resources in untangling the complex genomic landscape of plants, by means of accurate annotation of TEs and genes, by assisting studies on their interactions, and by enabling whole genome comparative analysis of their TE complement. Such insights will gradually deepen as MASiVEdb will be continuously expanding its phylogenetic coverage (see below).

Future development

We plan to periodically update MASiVEdb with Sireviruses from other plant genomes as they become available, so as to delve deeper into their distribution across plants, and possibly uncover more branches (like maize) where Sireviruses have aggressively amplified to achieve massive numbers, or others in which they have spectacularly failed to establish. We also intend to enrich the database with entries from various species where only limited sequence information is available.

The next major update of MASiVEdb will include a comprehensive phylogenetic analysis of its elements, and their categorization into families within and across species. Although such an analysis could have been performed relatively easily by either sequence clustering with a number of pre-annotated elements, or by construction of an e.g. *RT*-based tree, we strongly believe that more intricate sequence and genome characteristics of Sireviruses should be taken into consideration, a considerable undertaking out of the context of this first version of MASiVEdb. Finally, we expect in the near future to be able to incorporate sequence data of fragmented Sireviruses and solo LTRs.

Conclusion

MASiVEdb is so far the most comprehensive directory for Sireviruses, an abundant and distinctive genus of plant LTR retrotransposons, in currently available fully-sequenced plant genomes. Although there are a number of databases (and methods behind them) dealing with the repetitive fraction of genomes, the methodology of MASiVEdb provides unprecedented accuracy in delineating and analyzing Sireviruses, in turn enabling robust and meaningful research into their own life and their impact on their hosts.

Availability and requirements

MASiVEdb is freely accessible without any restriction to its use by non-academics at http://bat.infspire.org/databases/masivedb.

Additional file

Additional file 1: Phylogenetic and genome structure analyses within the *Copia* superfamily. This file contains the composite Additional file 1: Figure S1 that shows i) the phylogenetic relationships (based on the *RT* core domain) of exemplars from all three *Copia* genera, and ii) the highly conserved genome organization of Sireviruses and its comparison with the genome of other non-Sirevirus *Copia* elements (Figure adapted from [24]).

Acknowledgements and Funding
We thank Prof. Athanasios Tsaftaris and Dr. Kostas Stamatopoulos for reading and improving the manuscript. This work was partially supported by the Hellenic General Secretariat for Research and Technology (GSRT). ND is currently supported by CEITEC MU (CZ.1.05/1.1.00/02.0068) and project SuPReMMe (CZ.1.07/2.3.00/20.0045). AT acknowledges a PhD scholarship from the “Alexander S. Onassis” Public Benefit Foundation.

Author details
[1]Institute of Agrobiotechnology, Centre for Research and Technology Hellas, Thessaloniki 57001, Greece. [2]Department of Genetics and Plant Breeding, Aristotle University of Thessaloniki, Thessaloniki 54006, Greece. [3]Central European Institute of Technology, Masaryk University, Brno Czech Republic.

Author's contributions

AB conceived the study, collected and curated data, drafted the manuscript. EM designed and implemented the database and the web-accessible content, querying system, interfaces, and results presentation, and helped in drafting the manuscript. NK and MP assisted in data collection and algorithm development. AT curated data and assisted in algorithm development. ND coordinated the study, developed algorithms, and drafted the manuscript. All authors read and approved the final manuscript.

Received: 19 October 2011 Accepted: 30 April 2012
Published: 30 April 2012

References

1. Bennetzen JL: **Mechanisms and rates of genome expansion and contraction in flowering plants.** *Genetica* 2002, **115**(1):29–36.
2. Bennetzen JL, Ma JX, Devos K: **Mechanisms of recent genome size variation in flowering plants.** *Ann Bot* 2005, **95**(1):127–132.
3. Devos KM: **Grass genome organization and evolution.** *Curr Opin Plant Biol* 2010, **13**(2):139–145.
4. Bennetzen JL, Kellogg EA: **Do plants have a one-way ticket to genomic obesity?** *Plant Cell* 1997, **9**(9):1509–1514.
5. Hawkins JS, Proulx SR, Rapp RA, Wendel JF: **Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants.** *Proc Natl Acad Sci U S A* 2009, **106**(42):17811–17816.
6. Vitte C, Panaud O: **Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice Oryza sativa L.** *Mol Biol Evol* 2003, **20**(4):528–540.
7. Vitte C, Panaud O: **LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model.** *Cytogenet Genome Res* 2005, **110**(1–4):91–107.
8. The Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.** *Nature* 2000, **408**(6814):796–815.
9. **The Arabidopsis Information Resource (TAIR v10.0).** [http://wwwarabidopsisorg/]
10. Int Rice Genome Sequencing Project: **The map-based sequence of the rice genome.** *Nature* 2005, **436**(7052):793–800.
11. Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon JM, Westerman RP, SanMiguel PJ, Bennetzen JL: **Exceptional Diversity, Non-Random Distribution, and Rapid Evolution of Retroelements in the B73 Maize Genome.** *Plos Genetics* 2009, **5**(11):e1000732.
12. Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, Schlub S, Le Paslier MC, Magdelenat G, Gonthier C, *et al*: **Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces.** *Plant Cell* 2010, **22**(6):1686–1701.
13. Ohno S: **SO MUCH JUNK DNA IN OUR GENOME.** *Brookhaven Symposia Biol* 1972, **23**:366–70.

14. Orgel LE, Crick FHC: **Selfish DNA - the ultimate parasite.** *Nature* 1980, **284** (5757):604–607.
15. Biemont C, Vieira C: **Junk DNA as an evolutionary force.** *Nature* 2006, **443** (7111):521–524.
16. Morgante M: **Plant genome organisation and diversity: the year of the junk!** *Curr Opin Biotechnol* 2006, **17**(2):168–173.
17. Schmidt T, Heslop-Harrison JS: **Genomes, genes and junk: the large-scale organization of plant chromosomes.** *Trends Plant Sci* 1998, **3**(5):195–199.
18. Volff JN: **Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes.** *Bioessays* 2006, **28**(9):913–922.
19. Peterson-Burch BD, Voytas DF: **Genes of the Pseudoviridae (Ty1/copia retrotransposons).** *Mol Biol Evol* 2002, **19**(11):1832–1845.
20. Laten HM, Havecker ER, Farmer LM, Voytas DF: **SIRE1, an endogenous retrovirus family from Glycine max, is highly homogenous and evolutionarily young.** *Mol Biol Evol* 2003, **20**(8):1222–1230.
21. Boeke J, Eickbush T, Sandmeyer S, Voytas D: **Index of Viruses - Pseudoviridae (2006).** *ICTVdB - The Universal Virus Database, version 4, Columbia University, New York, USABuchen-Osmond C* 2006.
22. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, *et al*: **A unified classification system for eukaryotic transposable elements.** *Nat Rev Genet* 2007, **8**(12):973–982.
23. Bousios A, Darzentas N, Tsaftaris A, Pearce SR: **Highly conserved motifs in non-coding regions of Sirevirus retrotransposons: the key for their pattern of distribution within and across plants?** *Bmc Genomics* 2010, **11:**89.
24. Bousios A, Kourmpetis YAI, Pavlidis P, Minga E, Tsaftaris A, Darzentas N: **The turbulent life of Sirevirus retrotransposons and the evolution of the maize genome: more than ten thousand elements tell the story.** *Plant J* 2012, **69**(3):475–488.
25. Gao X, Havecker ER, Baranov PV, Atkins JF, Voytas DF: **Translational recoding signals between gag and pol in diverse LTR retrotransposons.** *Rna-a Publication of the Rna Society* 2003, **9**(12):1422–1430.
26. Havecker ER, Gao X, Voytas DF: **The diversity of LTR retrotransposons.** *Genome Biology* 2004, **5**(6):225.
27. Havecker ER, Gao X, Voytas DF: **The sireviruses, a plant-specific lineage of the Ty1/copia retrotransposons, interact with a family of proteins related to dynein light chain.** *Plant Physiol* 2005, **139**(2):857–868.
28. Havecker ER, Voytas DE: **The soybean retroelement SIRE1 uses stop codon suppression to express its envelope-like protein.** *EMBO Rep* 2003, **4** (3):274–277.
29. McCarthy EM, Liu JD, Lizhi G, McDonald JF: **Long terminal repeat retrotransposons of Oryza sativa.** *Genome Biology* 2002, **3**(10):0053.1.
30. Wicker T, Keller B: **Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families.** *Genome Res* 2007, **17**(7):1072–1081.
31. Chesnay C, Kumar A, Pearce SR: **Genetic diversity of SIRE-1 retroelements in annual and perennial glycine species revealed using SSAP.** *Cell Mol Biol Lett* 2007, **12**(1):103–110.
32. Holligan D, Zhang XY, Jiang N, Pritham EJ, Wessler SR: **The transposable element landscape of the model legume Lotus japonicus.** *Genetics* 2006, **174**(4):2215–2228.
33. Weber B, Wenke T, Frommel U, Schmidt T, Heitkam T: **The Ty1-copia families SALIRE and Cotzilla populating the Beta vulgaris genome show remarkable differences in abundance, chromosomal distribution, and age.** *Chromosome Res* 2010, **18**(2):247–263.
34. Hribova E, Neumann P, Matsumoto T, Roux N, Macas J, Dolezel J: **Repetitive part of the banana (Musa acuminata) genome investigated by low-depth 454 sequencing.** *Bmc Plant Biology* 2010, **10:**204.
35. Bousios A: **Characterization of the retrotransposon population in Agave tequilana and their use as molecular genetic markers**. *Thesis (PhD), University of Sussex, The British Library Main Catalogue 015282218,* 2009, [http://www.bl.uk/]
36. Laten HM, Majumdar A, Gaucher EA: **SIRE-1, a copia/Ty1-like retroelement from soybean, encodes a retroviral envelope-like protein.** *Proc Natl Acad Sci U S A* 1998, **95**(12):6897–6902.
37. Vicient CM, Kalendar R, Schulman AH: **Envelope-class retrovirus-like elements are widespread, transcribed and spliced, and insertionally polymorphic in plants.** *Genome Res* 2001, **11**(12):2041–2049.
38. Wright DA, Voytas DF: **Athila4 of Arabidopsis and Calypso of soybean define a lineage of endogenous plant retroviruses.** *Genome Res* 2002, **12**(1):122–131.
39. Peterson-Burch BD, Wright DA, Laten HM, Voytas DF: **Retroviruses in plants?** *Trends Genet* 2000, **16**(4):151–152.
40. Xu Z, Wang H: **LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons.** *Nucleic Acids Res* 2007, **35:**W265–W268.
41. McCarthy EM, McDonald JF: **LTR_STRUC: a novel search and identification program for LTR retrotransposons.** *Bioinformatics* 2003, **19**(3):362–367.
42. Ellinghaus D, Kurtz S, Willhoeft U: **LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons.** *Bmc Bioinformatics* 2008, **9:**18.
43. **RepeatMasker.**[*http://wwwrepeatmaskerorg/*]
44. Darzentas N, Bousios A, Apostolidou V, Tsaftaris AS: **MASiVE: mapping and analysis of sirevirus elements in plant genome sequences.** *Bioinformatics* 2010, **26**(19):2452–2454.
45. 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:**D360–D363.
46. 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**(1–4):462–467.
47. Wicker T, Matthews DE, Keller B: **TREP: a database for Triticeae repetitive elements.** *Trends Plant Sci* 2002, **7**(12):561–562.
48. **Maize transposable element database**. [http://maizetedborg/]
49. Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, Liang CZ, Zhang JW, Fulton L, Graves TA, *et al*: **The B73 Maize genome: complexity, diversity, and dynamics.** *Science* 2009, **326**(5956):1112–1115.
50. Chaparro C, Guyot R, Zuccolo A, Piegu B, Panaud O: **RetrOryza: a database of the rice LTR-retrotransposons.** *Nucleic Acids Res* 2007, **35:**D66–D70.
51. Du JC, Grant D, Tian ZX, Nelson RT, Zhu LC, Shoemaker RC, Ma JX: **SoyTEdb: a comprehensive database of transposable elements in the soybean genome.** *Bmc Genomics* 2010, **11:**113.
52. Llorens C, Futami R, Covelli L, Dominguez-Escriba L, Viu JM, Tamarit D, Aguilar-Rodriguez J, Vicente-Ripolles M, Fuster G, Bernet GP, *et al*: **The Gypsy Database (GyDB) of mobile genetic elements: release 2.0.** *Nucleic Acids Res* 2011, **39:**D70–D74.
53. Lewis LA, McCourt RM: **Green algae and the origin of land plants.** *Am J Bot* 2004, **91**(10):1535–1556.
54. Du JC, Tian ZX, Hans CS, Laten HM, Cannon SB, Jackson SA, Shoemaker RC, Ma JX: **Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison.** *Plant J* 2010, **63**(4):584–598.
55. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: an information aesthetic for comparative genomics.** *Genome Res* 2009, **19**(9):1639–1645.
56. Darzentas N: **Circoletto: visualizing sequence similarity with Circos.** *Bioinformatics* 2010, **26**(20):2620–2621.
57. Birney E, Clamp M, Durbin R: **GeneWise and genomewise.** *Genome Res* 2004, **14**(5):988–995.
58. Katoh K, Toh H: **Recent developments in the MAFFT multiple sequence alignment program.** *Brief Bioinform* 2008, **9**(4):286–298.
59. Kapitonov VV, Jurka J: **Helitrons on a roll: eukaryotic rolling-circle transposons.** *Trends Genet* 2007, **23**(10):521–529.
60. Yang LX, Bennetzen JL: **Distribution, diversity, evolution, and survival of Helitrons in the maize genome.** *Proc Natl Acad Sci U S A* 2009, **106**(47):19922–19927.
61. Yang LX, Bennetzen JL: **Structure-based discovery and description of plant and animal Helitrons.** *Proc Natl Acad Sci U S A* 2009, **106**(31):12832–12837.
62. Bennetzen JL: **Transposable elements, gene creation and genome rearrangement in flowering plants.** *Curr Opin Genet Dev* 2005, **15**(6):621–627.
63. Jiang N, Bao ZR, Zhang XY, Eddy SR, Wessler SR: **Pack-MULE transposable elements mediate gene evolution in plants.** *Nature* 2004, **431**(7008):569–573.
64. Jiang N, Ferguson AA, Slotkin RK, Lisch D: **Pack-Mutator-like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition.** *Proc Natl Acad Sci U S A* 2011, **108**(4):1537–1542.

65. Gao X, Hou Y, Ebina H, Levin HL, Voytas DF: Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Res* 2008, **18**(3):359–369.

66. Gorinsek B, Gubensek F, Kordis D: Evolutionary genomics of chromoviruses in eukaryotes. *Mol Biol Evol* 2004, **21**(5):781–798.

doi:10.1186/1471-2164-13-158
Cite this article as: Bousios *et al.*: MASiVEdb: the Sirevirus Plant Retrotransposon Database. *BMC Genomics* 2012 **13**:158.