



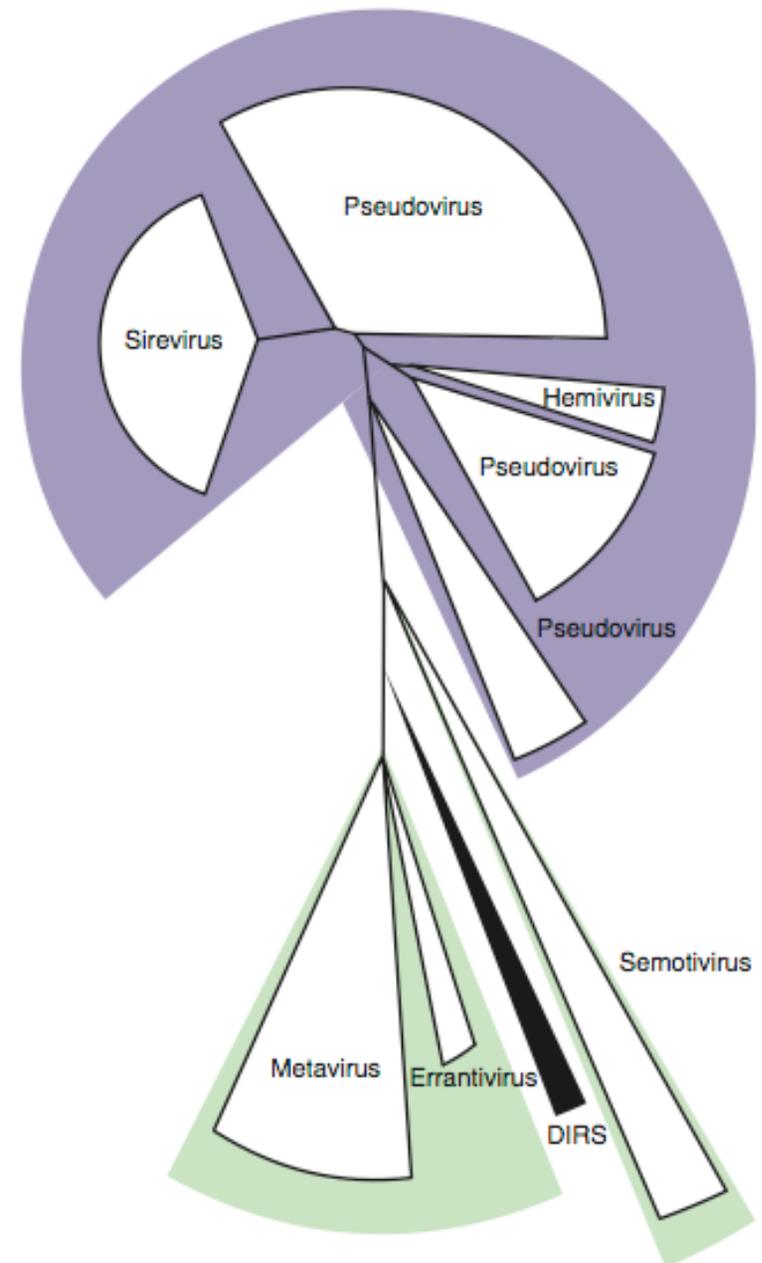
Unraveling the life dynamics of Sirevirus LTR retrotransposons: major players in the organization and evolution of maize and other angiosperm genomes

Alexandros Bousios

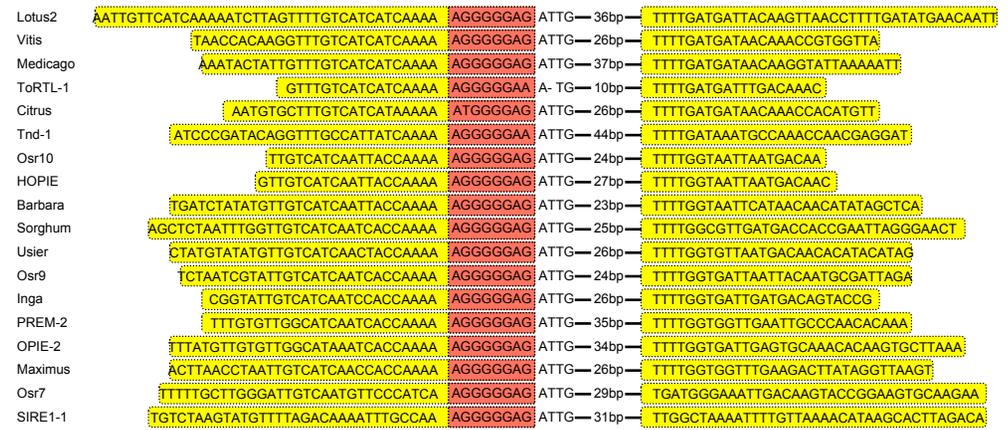
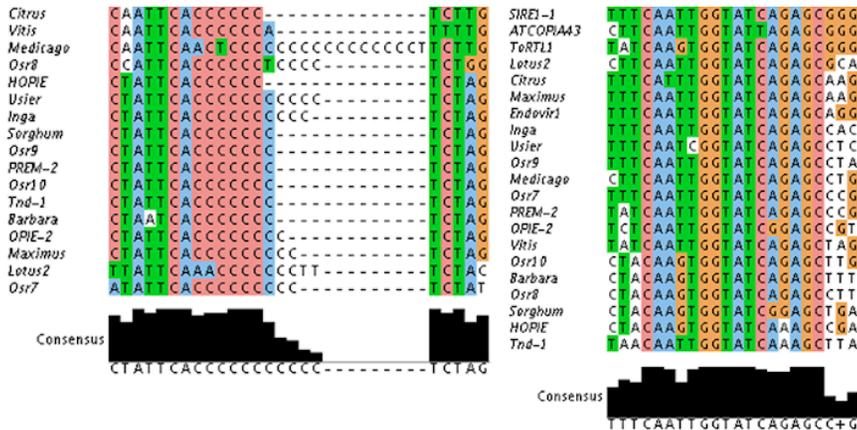
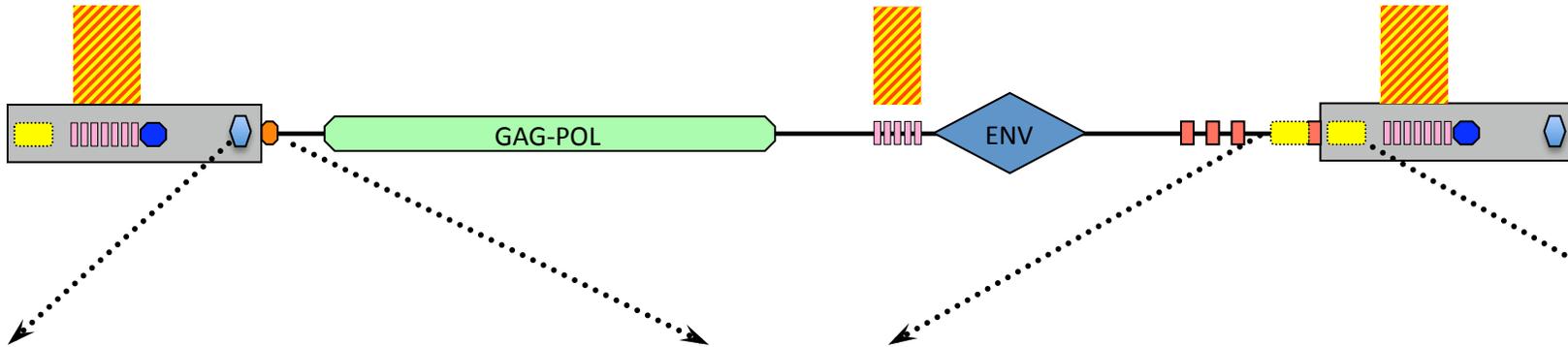


Sireviruses – What was known

- ❖ One of three Copia genera
- ❖ Plant-specific and young genus
- ❖ Putative retroviral properties
- ❖ Few publications on Sireviruses
- ❖ Scarce reference on the Sirevirus origin of some elements
- ❖ Difficulty in assigning LTR-RTNs into genera, research at this classification level is missing



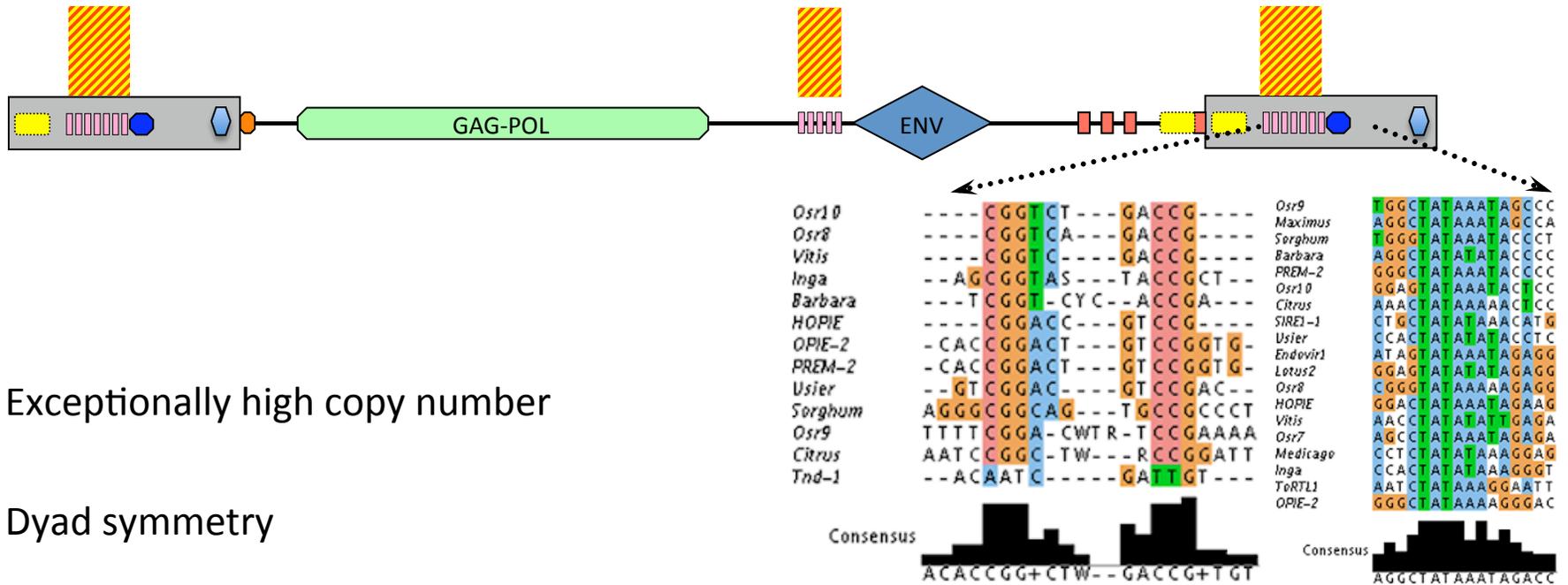
The unique genome structure of Sireviruses



primer binding site
 integrase signal

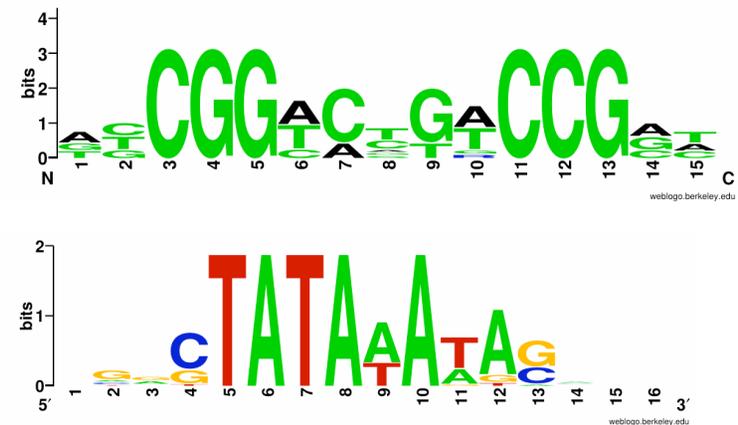
multiple PPT signature
 inverted repeat

Novel type of repeated motif with modular organization

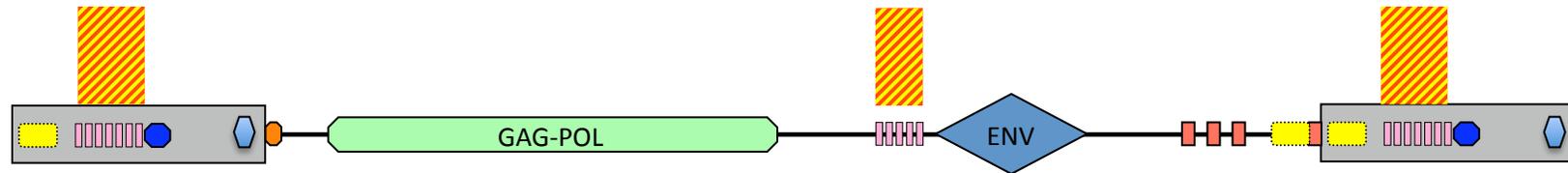


- ❖ Exceptionally high copy number
- ❖ Dyad symmetry
- ❖ Core CGG-CCG signature
- ❖ Form CpG islands
- ❖ Upstream of a conserved promoter in the typical regulatory locus of LTR-RTNs
- ❖ Surprisingly, also upstream of the ENV

■ repeated motif ● TATA box ▨ CpG island



at the sequence level...



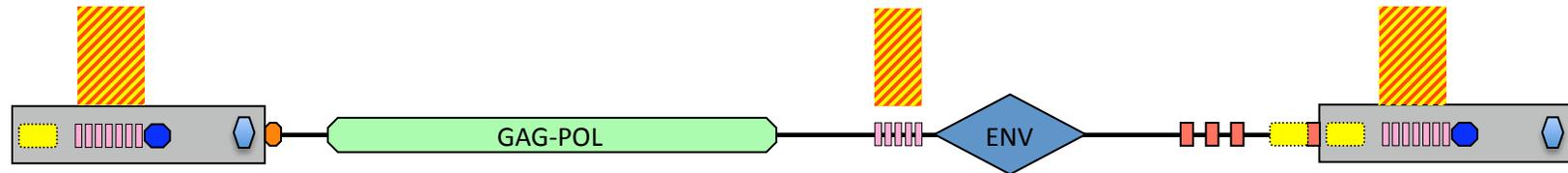
>Inga

```

AACTAGGCATGTTGCACCTCACTTCGCATTCCATTTTTGGTCTAGATGTAGGCATGGACATAGGGGGAGTGTGTTCTCTCAATGAACTG
TCCCTCCCCTCATTATGCATAAATCAATCTAGTCTTTCACTTTAGCCATTGTCAAATGGTACTTGTGCTTCAAAGACGAGCATTGGTCA
TGAACCCAAGGATAATTCTTCGGGTGTCATAACCATTGTCTCAAACATAGTGGCCCTCGGCCACCGCCCCTCCATCCTCTCTTGGCTAT
AAAGGAGAGGATATACAATGACAAGTCGGTACTTTTTCTTAGAAGTCATCCCTTTTTACTCTCTTGTGCATGTGCCCAAGTTTTCCCGCT
TTCTTAGCCGTGCTGAAACATGTACAGCGGTACTACCGCCAGGGTAGCGGTACTACCGCCAGGACCCCAGCGGTACTACCGCCATGGG
TAGCGGTACTACCGCCAGAGTTCAAATCTAACTCGGCCGGCCGGAATATTTCTAAGGTTTCGAGGGGGAGCGGTACTACCGTCAGGGG
TAGCGGTACTACCGCCAGGACCCCAGCGGTACTACCGCTGCACGCCAGCGGTACTACCGCTAGGTAGCGGTACTACCGCTCTAGTCCA
CGCGGTACTACCGCTGGACCCCAGCGGTACTACCGCTGCCCATACCCCTTGGGCTATATAAAGGGAGGGGGAGCCCGATTTTCAATCTGT
TTCTTCTTCTCGGGCTCTCCCTCCCCTACCCCGAAAACCCGCTGTTTGGATCTTTCTTGGAGGAGCTCCCTCTTCCCCTTGGTGTG
GAAGGGCTGATTTACTTCTTCTTCTCCTCAAAGCCATGAATCCCGGTAACAAAAGCTCCTCCCCTCTTCTATTTGGTTGTTTTTCGT
GTCTAGGGTTAGGAGCATTGGGGATTGCATTCATCTTGTGATCCAATGGCTTGTGCTAAGATTGCTGCCGTGGTTAGGGTTTCAT
CGTCTAGATCGGATATTTGAACCTTTTTCTGATTAGATCTAAAACGTGCTCTCTCTTGCCTATGCATGTTTGTACCTCGTGTACTAT
GTGTTCTTTTTGGCAATAGGGCTGATGTATACGTATTAATGATTATATTCAGTCCATGCCCTAGAATAAGAGTGTTTTATATGTCTAGATC
TGACAGTCAAACCCCCAGCGGTACTACCGCCAGGGGGTAGCGGTACTACCGCTATGACCCCCAGCGGTACTACCGCCAGAAGTGGCGGT
ACTACCGCCAGGACTCCCAGCGGTACTACCGCCAGGACTGGTGGTACTACCGCCAGGTGTGGCGGTACTACCGCCATGA-----
-----ENV GENE-----
CTACCTGTGACATTAGGTGTGCCCCCTTTTTGGTGTCTTGTGCCAAAGGGGGAGAGTCTAGGGATTTGATTTCAATCGAACTTTGCA
TTGCTTTGCTGATTTGCTTTGAACTTGGTTTAGGGGGAGCTTTTGTTTGCTATCGTGTGTGAGACATGATCTATCCTATGTGAGATTAG
ATTTATTTCCATCATATGCTGTGAGTCATATGTCATATATCTCTATCCTTTTTATTTCTCATATTATATCCATGCAAATCGGTATTGT
CATCAATCCACCAAAAAGGGGGAGATTGTTCAGGCATAATTTATGCCTAAGTAAATTTTGGTGATTGATGACAGTACCGTACAGGACTAA
TCGTGTGTCAAGGTTTCAGGCAACATTGTCACAGGCACAAGACGACAGACTCCTCTCTTCGGGAACGGAAACACGGCGCTATCCT
AGATTCTCTTCATTTGAGTCATAGGAAAGCCGTACTATTAAGAGGGGATCCGTAGTGGAAAGGTTTGGGTGGAATCTATCTTGCACAC
GCACACCTCTATTTCTCCCTTTTCTTTATCCTTGGAGCGGCCCTCGCCGGTTTGTCCCTTTCGCTCTCGGCAAAATGGTCCCAGCGGT
AGTACCGCTGGTTCCAGCGGTAGTACCGCTGGACTGCCAGCGGTAGTACCACTGCAGCCAGCGGTAGTACCGCTGCAGCCAGCGGTATT
ACCGCTGGCTGGCGGTAGTACCGCCCTTGGTCAGCGGTAGTACCGCTCCAGGTCCCTGTTCCACCTACCTAGCGGTAGTTCGTGGACG
GACCCCTTTTGCGAAGACTTTCTCGCGGTAGTGTGTTATGCACTACCAAGGGCCAGCGGTAGTACCGCTGGTGCCAGCGGTAGTACC
GCTGGAAGCCCCAGCGGTAGTACCGGTGCATGCAGCGGTAGTACCGCCAGGCAGCGGTAGTACCGCTCCTTCCCAGCGGTAGTACCGCT
GGATCTCGGGCAGAGAGTGGGAAAACGGTTTTGAATTTTCCCCCACTATATAAAGGGTCTTCTAGCTGAGGAACCCCTATCTTTACCT
CTCTAAGCTCCATTGTTGCTCCACAAGCTTAAAAGTGCCCGATCTCTCCTTAGCCAATCAAACCTGTTGATTCTTTCCGGGATTGGTT
GAGAAGGCCTAGATCCACACTTTCACCAAGAGAAAAGTTGATTCCCCCACCTATCCCTTGCGGATCTTGTACTCTTGGGTGTTTGGAGC
ATCCTAGACGGTTGAGGTCACCTCGAAGCCATATTCATTGTGGTGAAGCTTCGTGGTCTTGTGGGAGCCTCCAAGCTTTGTGTGGAG
TTGCCCAACCTTGTGTTAAAGTTTCGGTCGCCCGCTTCAAGGGCACCTATAGTGGAAATCACGGTACCTTGCATCGTGCAGGGCGTG
AGGAGAATACGGTGGCCTTAGTGGCTTTTTGGGGAGCATTGTGCCTCCACACCGCTCCAACGGAGACGTATCTTCTATCAAAGGGAAGG
AACTTCGGTAACACATCCTCGTCTCCATCGGTTCCACTTGTGGTATCTCTAACCTTTACTTTGTATTTGCTTTTCTTGTGACAAAAT
CTTACTTGTCTAATAGGTCTTGTGATAGCTTCTATAGGGGTCACTCTTCGTATATTTGTGAAATCGTATAGTGTACCTTA
ACTTGTAAAGATTAATAAAAAGTGGTCGTTCTCTATTCACCCCCCCCCCTAGCCAACCATATCGATCCTTTCA

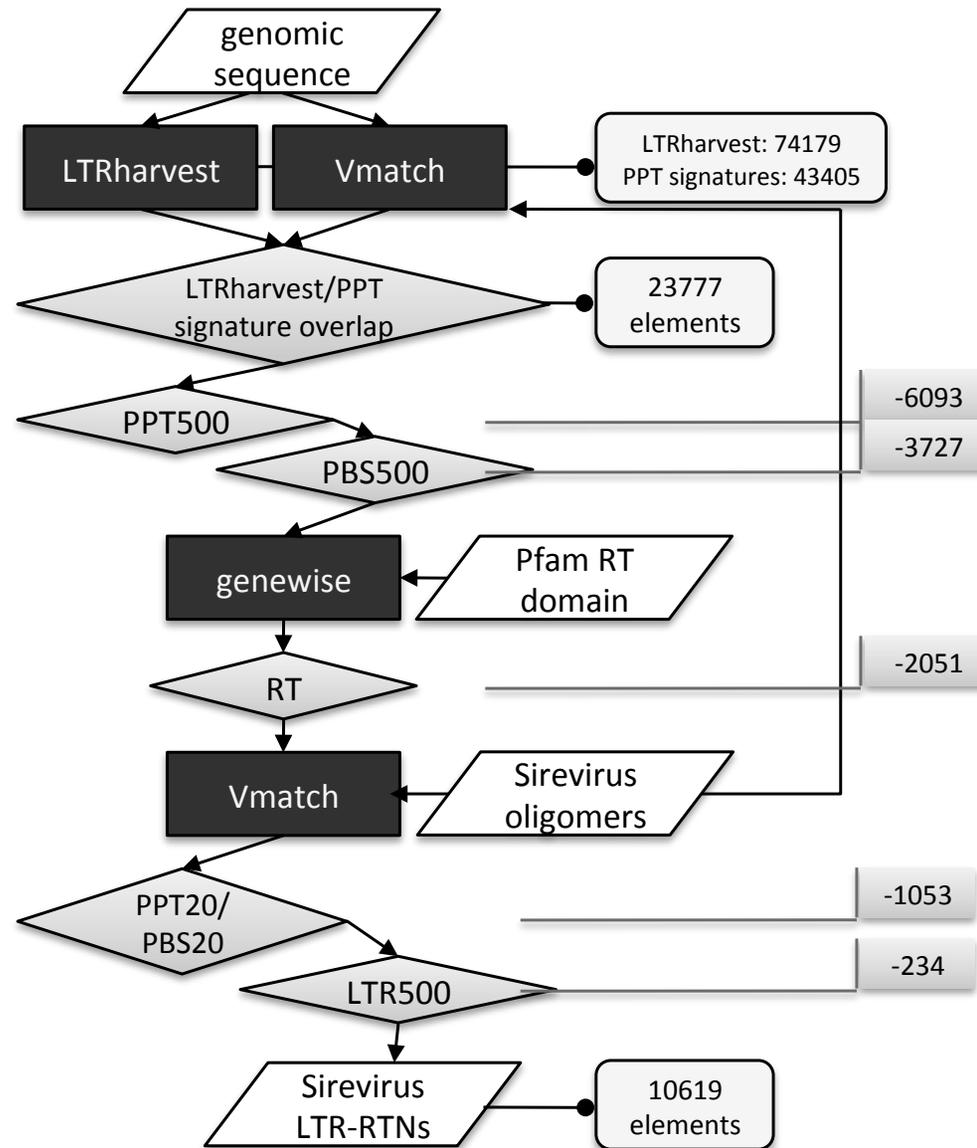
```

Sireviruses – What we know now

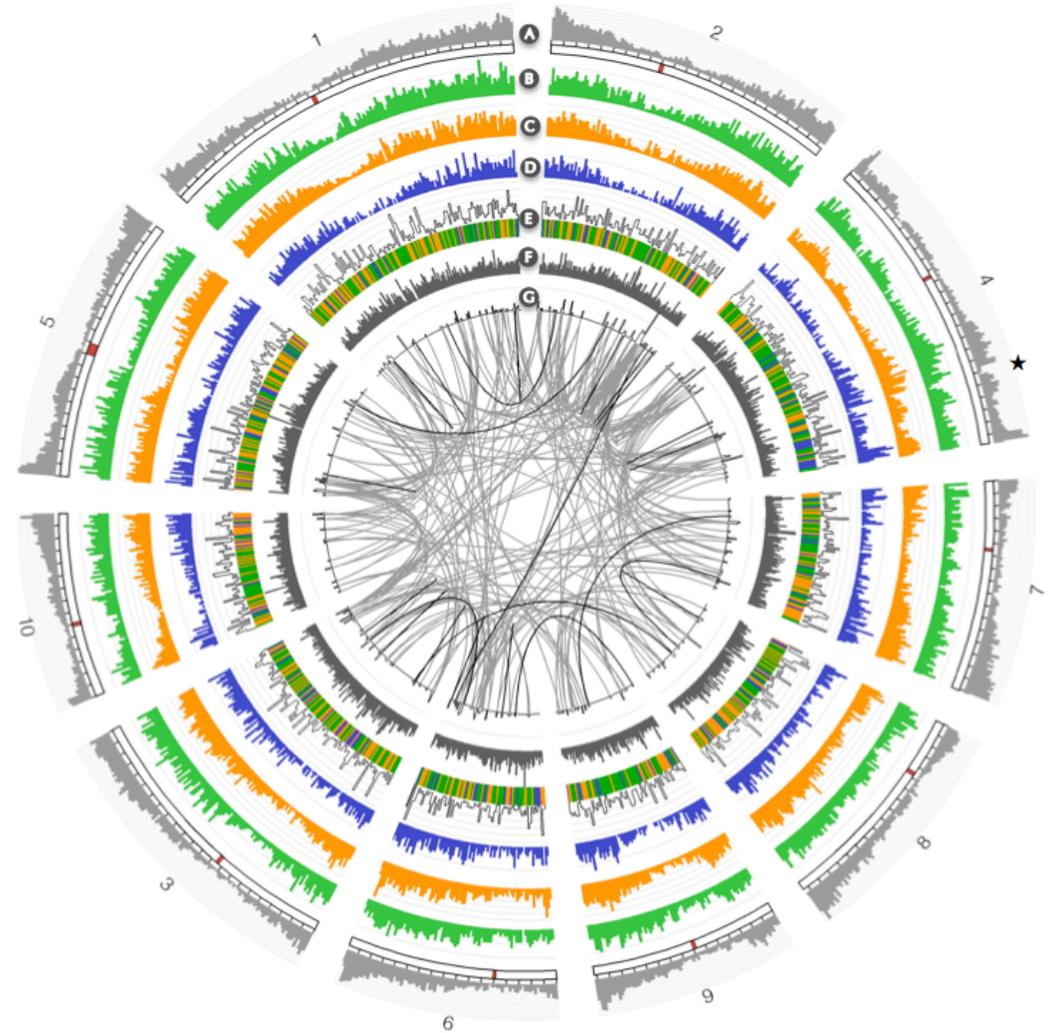
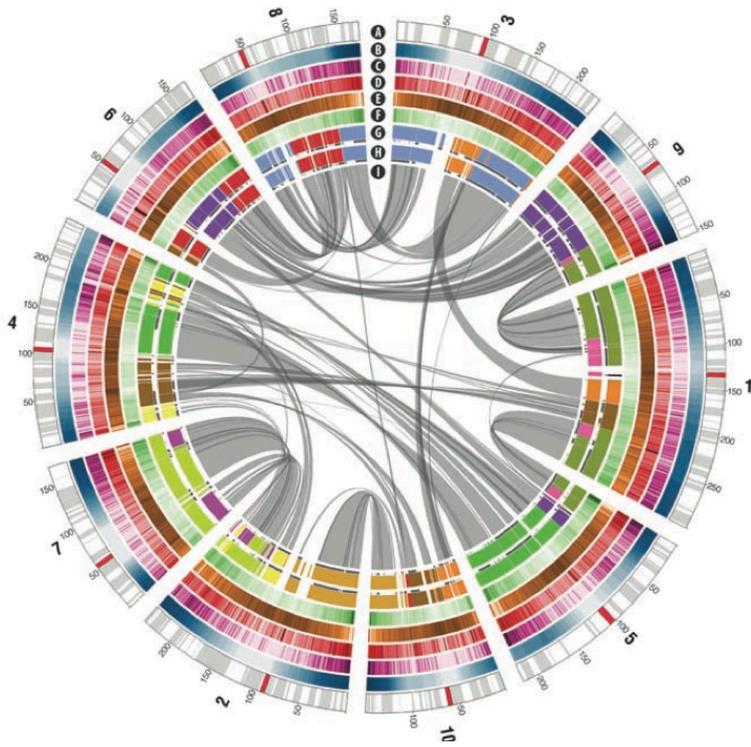


- ❖ Ancient genus with vastly divergent members
- ❖ Short highly conserved motifs in key non-coding domains that are critical for the life cycle of LTR-RTNs
- ❖ Astonishing conservation – hosts diverged 140-150 mya
- ❖ Only LTR-RTN genus with such an intriguing genome structure

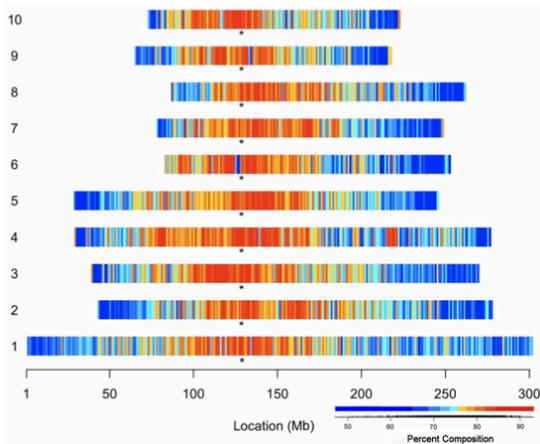
The MASiVE algorithm



Sirevirus infiltration patterns in maize



Schnable et al. 2009 – Science



Baucom et al. 2009 – Plos Genetics

Bousios et al. 2011 – The Plant Journal

Sirevirus phylogenetic diversity in maize

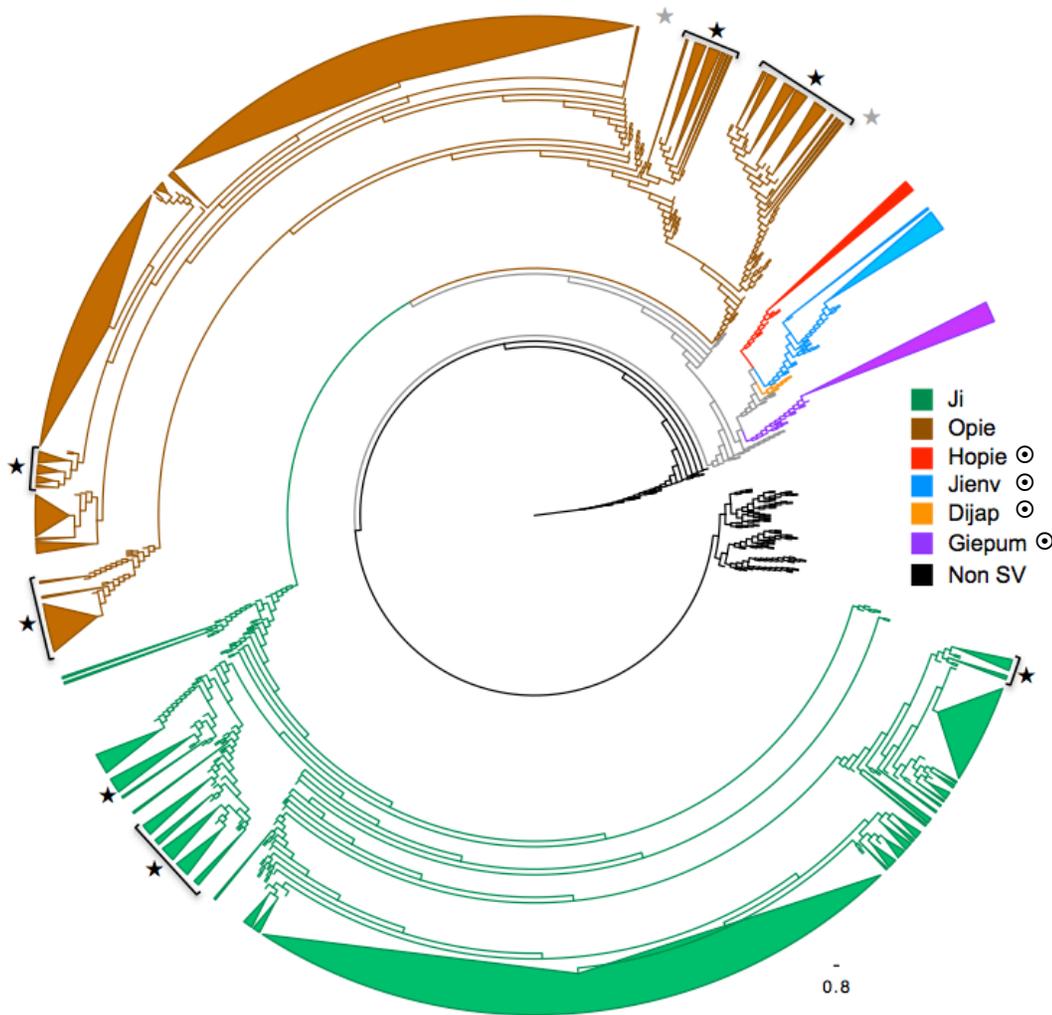
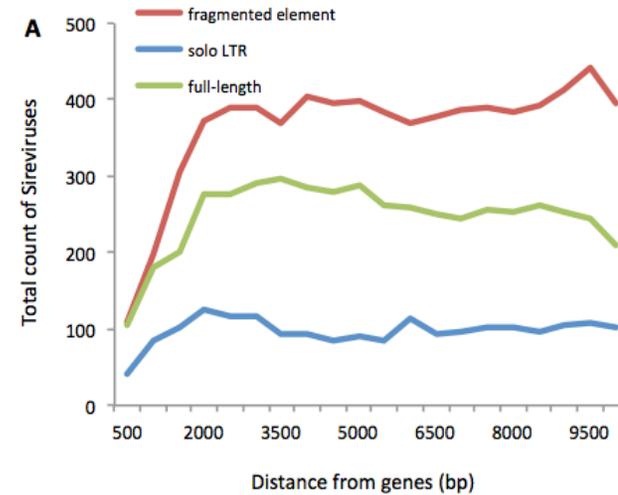
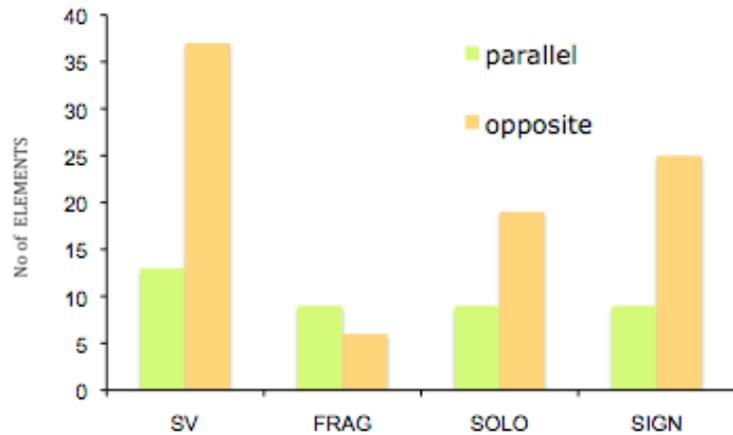


Table 1. Properties of the Sirevirus families identified in the maize genome

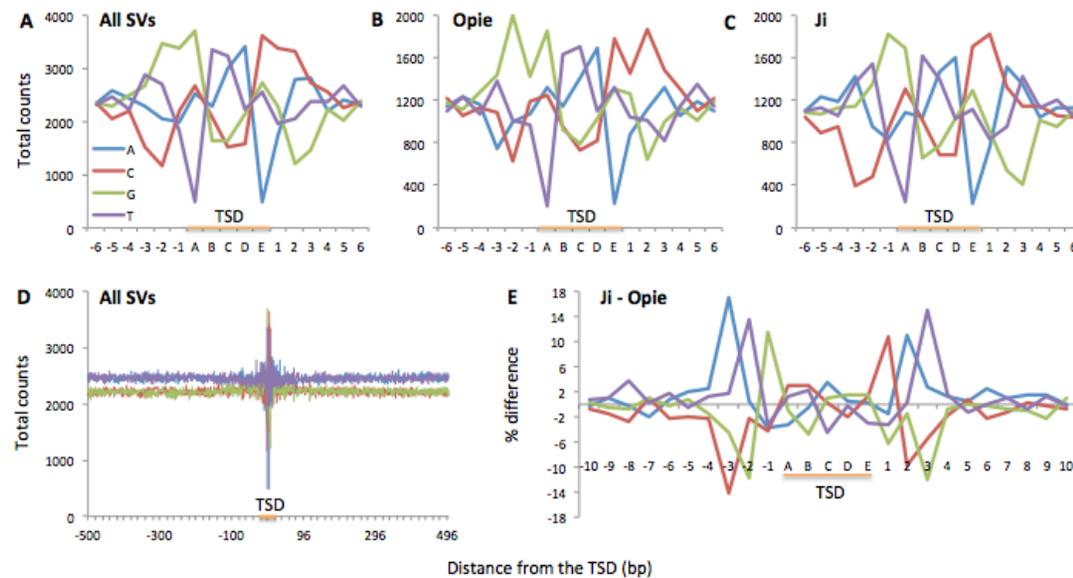
Family ^a	FL	solo	frag	FL:solo	FL:frag	solo:frag	avg age (my)	avg length	
								FL	LTR
<i>Opie</i>	5310 +1780 ^b	2028	9826	2.6	0.5	0.2	0.90	9117	1254
<i>Ji</i>	4865 +772 ^b	2421	11377	2.0	0.4	0.2	0.94	9519	1271
<i>Jienv</i> ^c	175 +175 ^b	103	469	1.7	0.4	0.2	0.71	12123	1534
<i>Giepum</i> ^c	143	180	698	0.8	0.2	0.3	0.76	12666	1469
<i>Hopie</i> ^c	74 +31 ^b	149	478	0.5	0.2	0.3	1.03	11696	1675
<i>Dijap</i> ^d	15 +6 ^b	38	28	0.4	0.5	1.4	2.0	10783	1525
Other	37	19	59	1.9	0.6	0.3	1.59	10449	1279
Total	10619	4938	22935	2.2	0.5	0.2	0.92	9424	1273

❖ Phylogenetic tree of maize Sireviruses based on the RT gene

Sireviruses integration biases in the maize genome

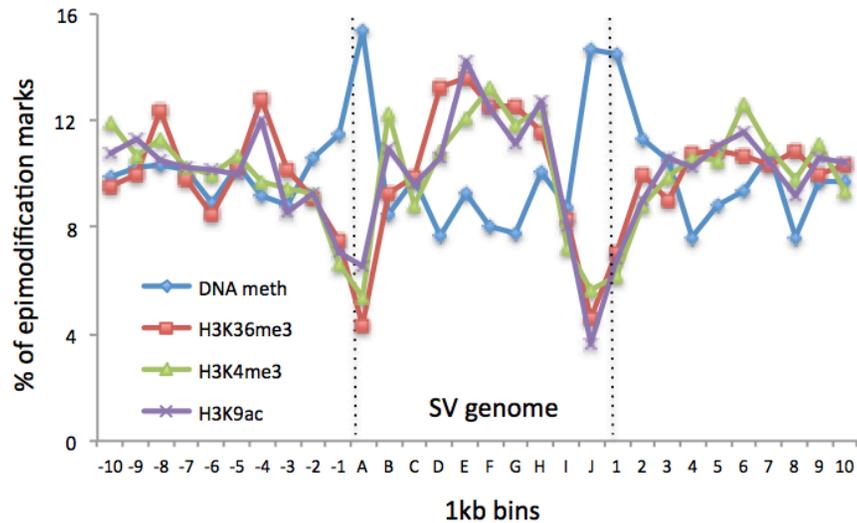
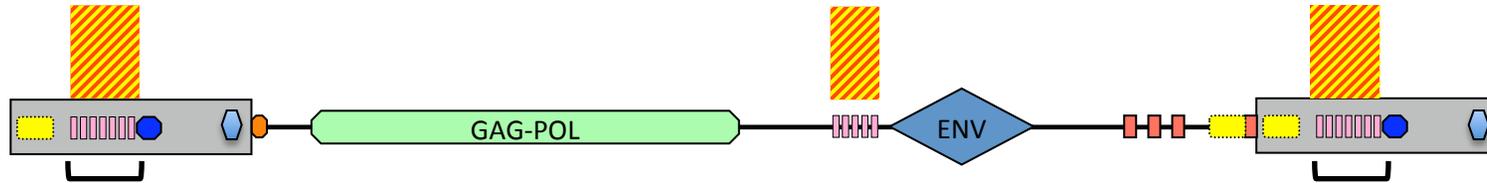


within and near genes...

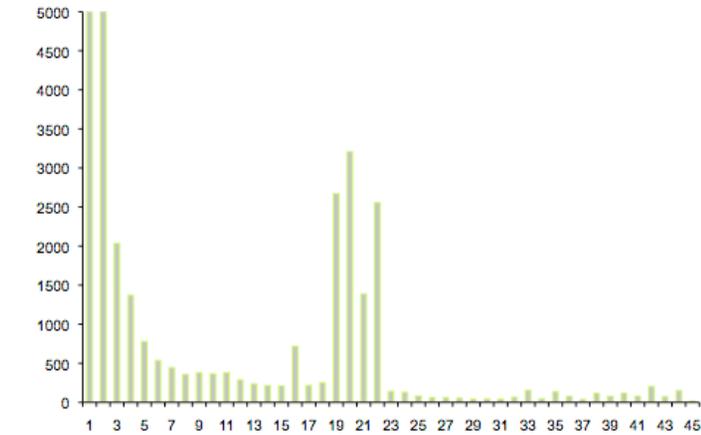


for a palindromic sequence motif...

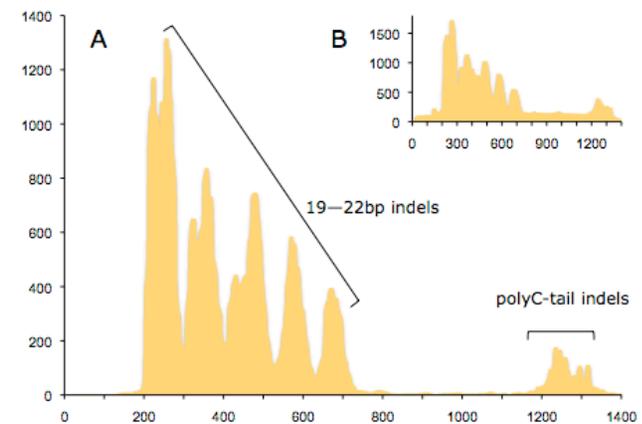
Sirevirus LTRs are methylation and recombination hotspots



◆ Epigenetic modification profiles of Sirevirus genomes



◆ LTR indels length (bp) distribution



◆ Indels topology on the Sirevirus LTRs

The turbulent life of Sireviruses in the maize genome

- ❖ 21% of the genome – 90% of the Copia complement
- ❖ Intense amplification period the past 600,000 years
- ❖ Plethora of families with different life and genome characteristics
- ❖ Colonize gene-rich areas, mediating gene diversification and the formation of gene islands
- ❖ Sirevirus LTRs may be the epicenter of interactions with the host control system

MASiVEDb: the Sirevirus Plant Retrotransposon Database

MASiVEDb

the Sirevirus Plant Retrotransposon Database

home | query | batch retrieval | sequence search | downloads | help | publications | acknowledgements | BAT cave | contact us | report a bug

plants LTR retrotransposon genus of the Copia superfamily, characterized by highly conserved motifs in key domains. Sireviruses have extensive diversity across plant species, and the accumulation of Copia elements, as is the case in maize (~90% of Copia elements), is thought to be the result of the fundamental role of Sireviruses on the evolutionary dynamics and impact of transposons in plant genomes. This database is a comprehensive resource of Sirevirus populations, aiming to assist in their further study and support research into the evolutionary dynamics and impact of transposons in plant genomes.

Deploying MASiVE, our tool for Sirevirus discovery and analysis, on all discovered Sireviruses, basic information, and primary results, can be found in this database. The tree *Populus trichocarpa* are missing from the database. MASiVE is designed to be straightforward yet powerful. It has four large buttons: query, batch retrieval, sequence search, and downloads. Help is provided at every page through the 'usage' link. Enjoy your MASiVEDb!

simple

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

view: multiple values

when species is Sorghum bicolor and chromosome 1 and MASiVE ID select all plant species MASiVE ID chromosome direction from to % distance to centromere Envelope gene age full length 5' LTR length 3' LTR length retention position of RT

batch retrieval

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

load example

submit to MASiVEDb

downloads

date and version	species	Sireviruses	MASiVE matrix	full-length	5' LTR	3' LTR	signature	RT	INT	ENV	zf-OHCH
2011-06-28 21.06.11	Arabidopsis thaliana	4	14b	84b	14b	14b	14b	14b	14b	<14b	<14b
2011-06-28 21.06.11	Brachypodium distachyon	22	2b	24b	34b	34b	24b	24b	24b	24b	14b
2011-06-28 21.06.11	Fragaria vesca	1	<14b	34b	<14b	<14b	<14b	<14b	<14b	n/a	<14b
2011-06-28 21.06.11	Glycine max	1337	774b	15744b	1194b	1194b	334b	544b	334b	364b	224b
2011-06-28 21.06.11	Lotus japonicus	282	154b	4884b	294b	294b	134b	154b	84b	194b	44b
2011-06-28 21.06.11	Oryza sativa indica	25	2b	58b	74b	74b	24b	24b	24b	24b	14b
2011-06-28 21.06.11	Oryza sativa japonica	91	74b	2034b	194b	194b	64b	64b	44b	44b	24b
2011-06-28 21.06.11	Sorghum bicolor	522	334b	11574b	1094b	1094b	334b	334b	214b	54b	104b
2011-06-28 21.06.11	Theobroma cocoa	77	54b	1824b	154b	154b	84b	74b	54b	34b	24b
2011-06-28 21.06.11	Vitis vinifera	49	44b	634b	84b	84b	54b	54b	34b	34b	14b
2011-06-28 21.06.11	Zea mays	13833	8044b	233904b	25474b	25424b	11834b	8244b	5844b	214b	1624b

query

submit to MASiVEDb

batch retrieval

submit to MASiVEDb

sequence search

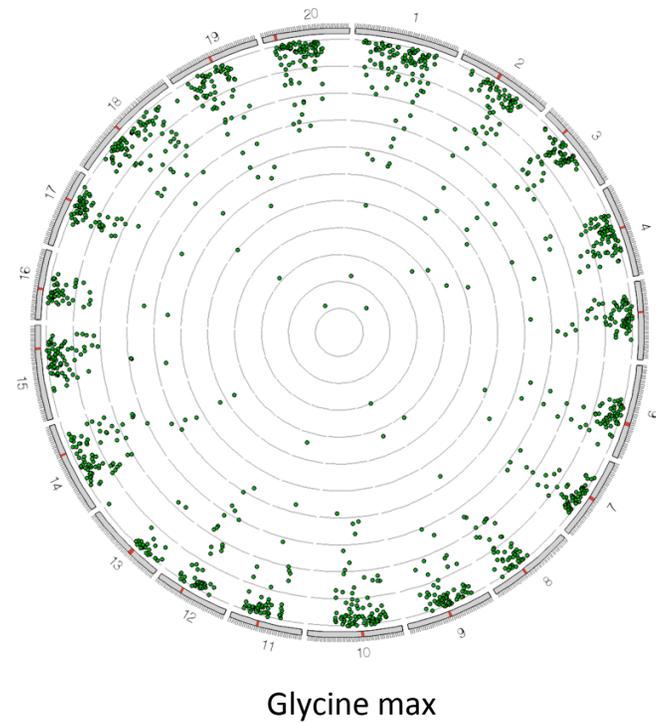
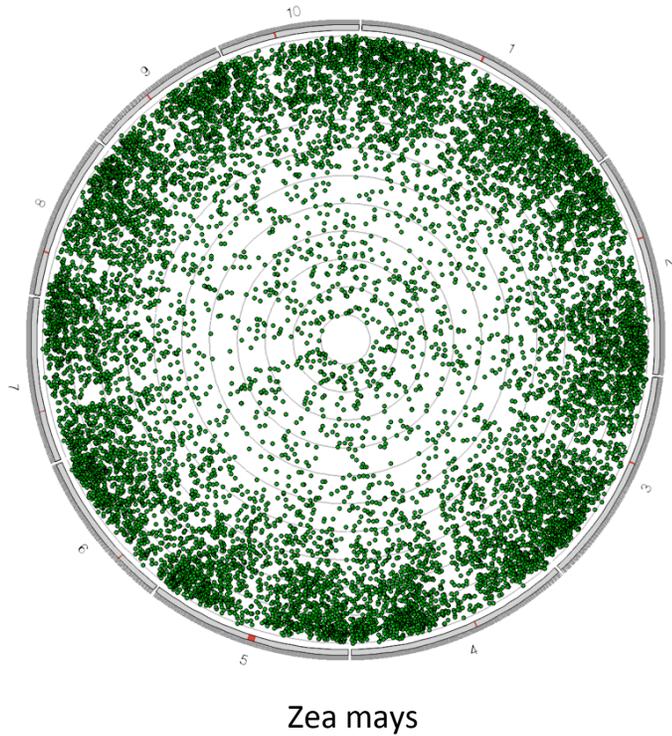
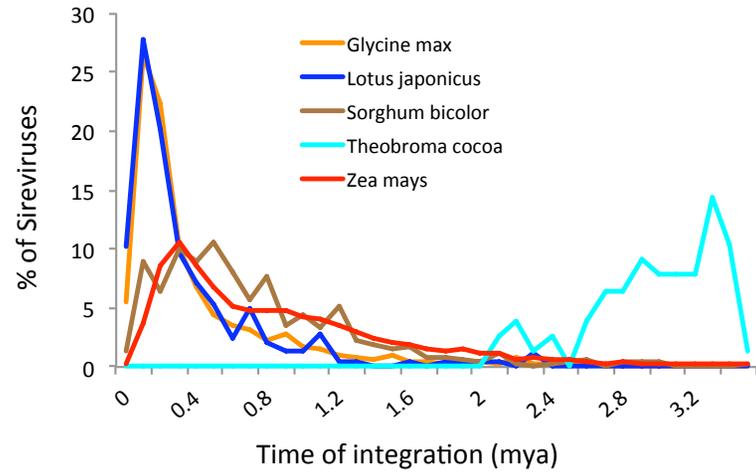
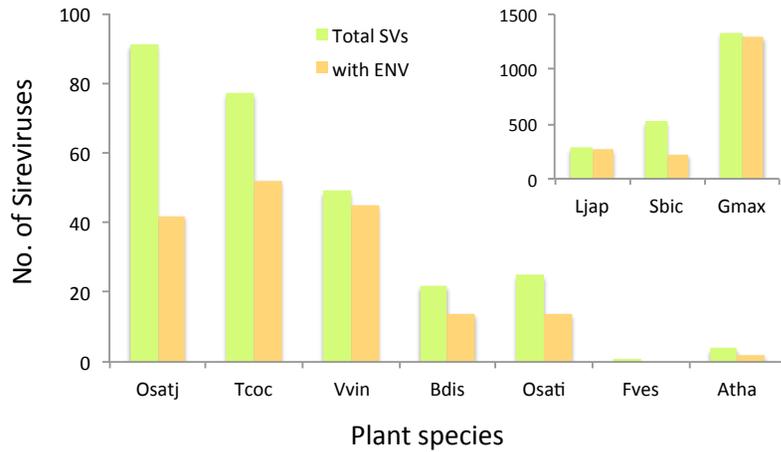
submit to LTRphlyer for MASiVEDb

submit to MASiVEDb

downloads

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	0	
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	Osati	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	

Sireviruses dynamics in other angiosperm genomes



Thank you for your attention...

