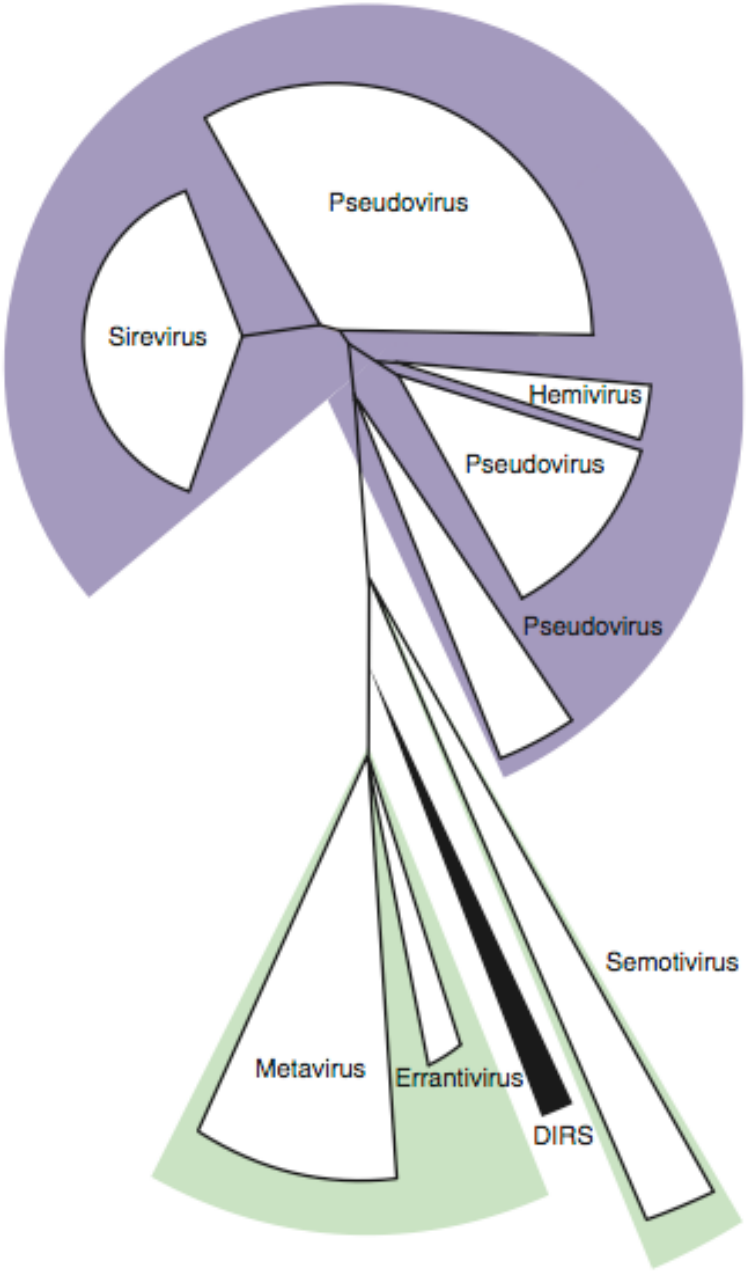
Unraveling the life dynamics of Sirevirus LTR retrotransposons: major players in the organization and evolution of maize and other angiosperm genomes

Alexandros Bousios

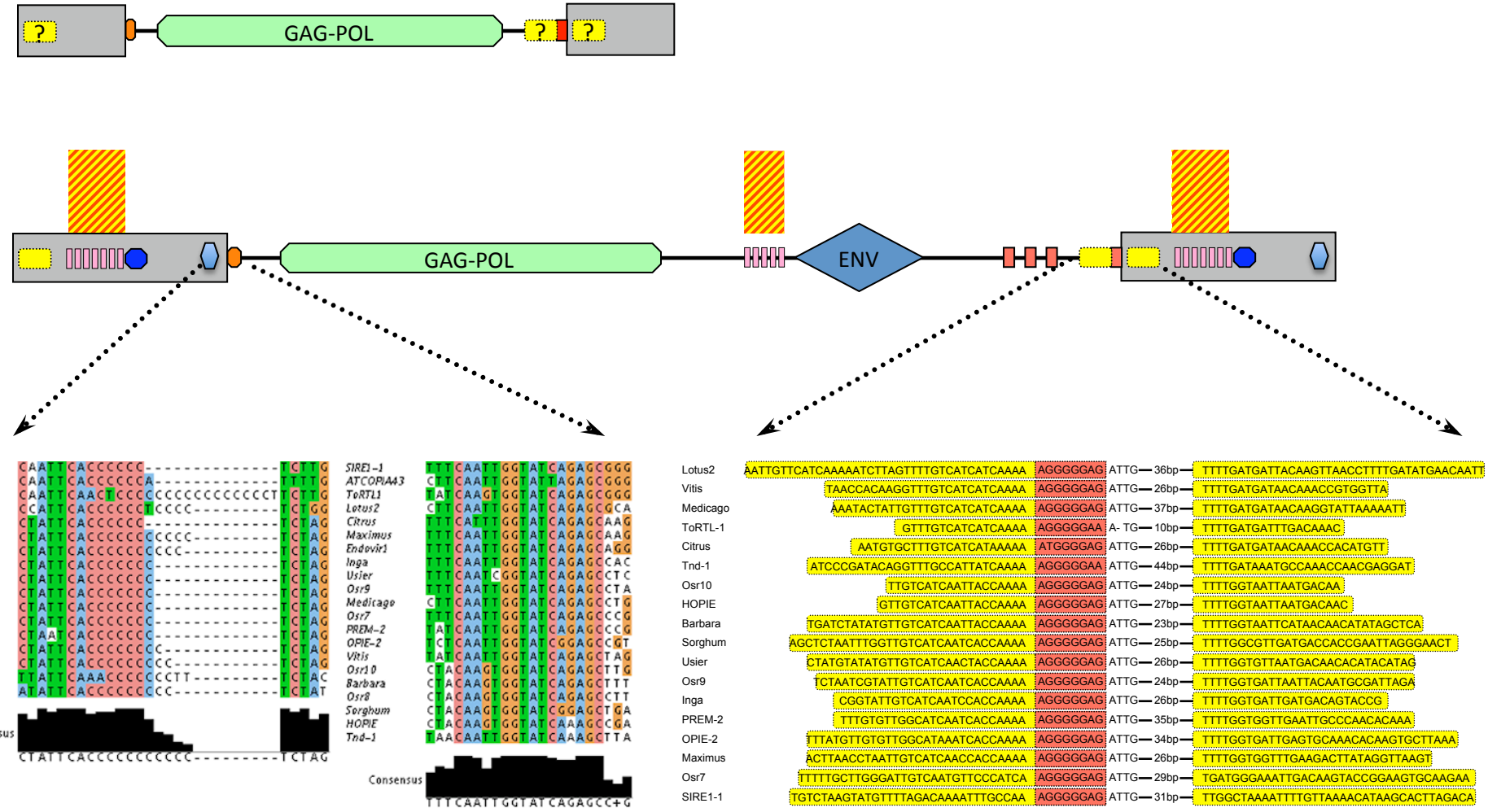
Sireviruses – What was known

- 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



Havecker et al. 2004 – Genome Biology

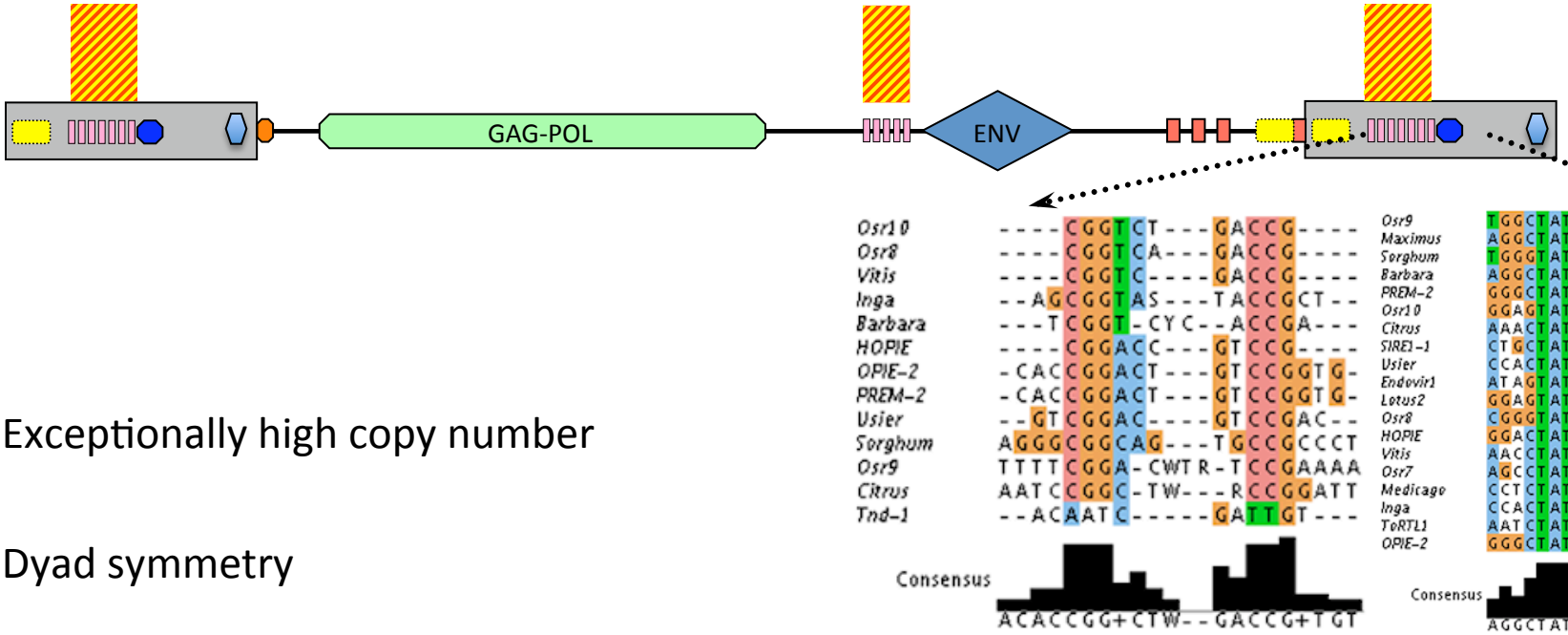
The unique genome structure of Sireviruses



- primer binding site
- integrase signal
- multiple PPT signature
- inverted repeat

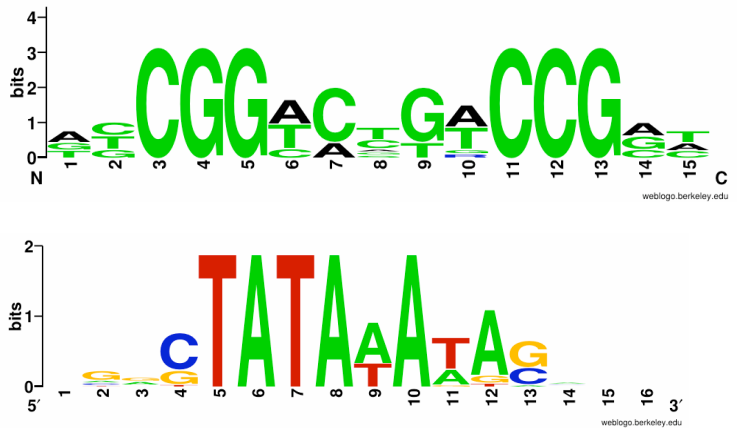
Bousios et al. 2010 – BMC Genomics

Novel type of repeated motif with modular organization



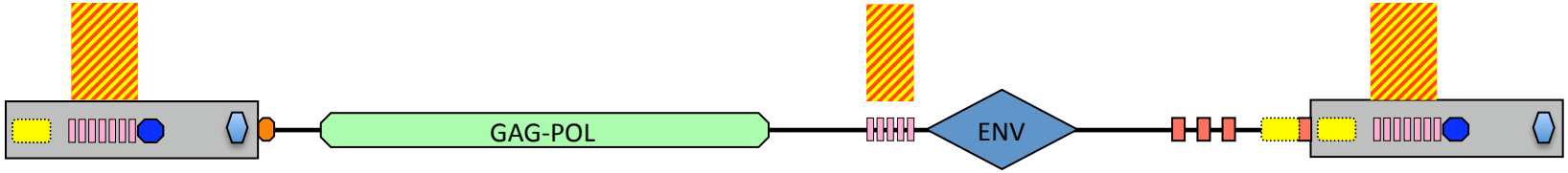
- Exceptionally high copy number
- Dyad symmetry
- Core CGG-CCG signature
- Form CpG islands
- Upstream of a conserved promoter in the typical regulatory locus of LTR-RTNs
- Surprisingly, also upstream of the ENV

repeated motif TATA box CpG island



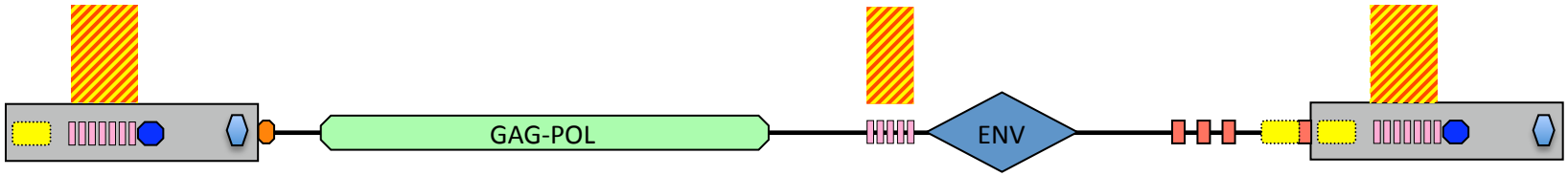
Bousios et al. 2010 – BMC Genomics

at the sequence level…



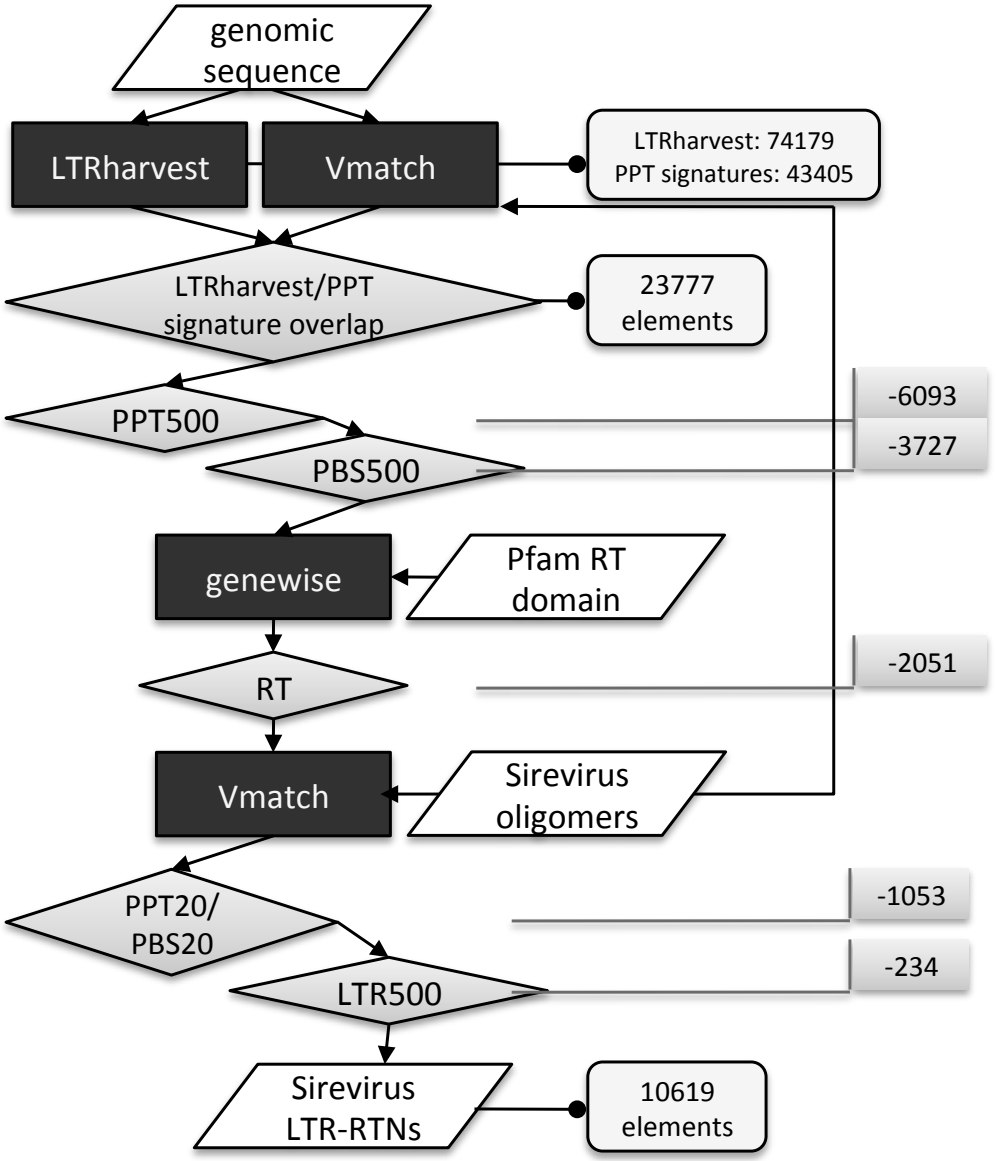

```
>Inga
AACTAGGCATGTTGCACCTCACTTCGCATTCCATTTTGGTCTAGATGTAGGCATGGACATAGGGGGAGTGTTGTTCTCTCAATGAACTG
TCCCTCCCCTCATTATGCATAAATCAATCTAGTCTTTCACTTTAGCCATTGTCAAATGGTACTTGTGCTTCAAAGACGAGCATTGGTCA
TGAACCCAAGGATAATTCTTCGCGGTGTCATACCATTGTCTCAAACATAGGTGGCCTCGGCCACCGCCCCTCCATCCTCTCTTGCGTAT

AAAGGAGAGGATATACAATGACAAGTCGGTACTTTTTCTTAGAAGTCATCCCTTTTTACTCTCTTGTCATGTGCCCAAGTTTTCCCGCT
TTCCTTAGCCGTGCTGAAACATGTACAGCGGTACTACCGCCAGGGTAGCGGTACTACCGCCAGGACCCCAGCGGTACTACCGCCATGGG
TAGCGGTACTACCGCCAGAGTTCAAAATCTAACTCGGCCGGCCGGAATATTTCTAAGGTTTCGAGGGGGAGCGGTACTACCGTCAGGGG
TAGCGGTACTACCGCCAGGACCCCCAGCGGTACTACCGCTGCACGCCAGCGGTACTACCGCTAGGTAGCGGTACTACCGCTCTAGTCCA
GCGGTACTACCGCTGGACCCCAGCGGTACTACCGCTGCCCATACCCCTTGGGCTATATAAAGGGAGGGGGAGCCCGATTTTCAATCTGT

TTCTTCTTCCTCGCGGCTCCTCCCTCCCCTACCCCGAAAACCCGCTGTTTGGATCTTTCTTGGAGGAGCTCCCTCTTCCCGTTGGTGTG
GAAGGGCTGATTTACTTCTTCTTTCTCCTCCAAAGCCATGAATCCCGGTAACAAAAGCTCCTCCCCCTCTTCTATTTGGTTGTTTTTCGT
GTTCTAGGGTTAGGAGCATTGGGGATTGCATTCATCTTGTTGATCCAATGGCTTGTTGCTAAGATTGCTGCCGTGGTTTAGGGTTTCAT
CGTCTTAGATCGGATATTTGAACCTTTTTCTGATTAGATCTAAAACGTGCTCTCTCTTGCCTATGCATGTTTGTACCTCGTGTTACTAT
GTGTTCCTTTTGGCAATAGGGCTGATGTATACGTATTAATGATTATATTCAGTCCATGCCCTAGAATAGGAGTTTTATATGTCTAGATC
TGACAGTCAAACCCCCAGCGGTACTACCGCCAGGGGGTAGCGGTACTACCGCTATGACCCCCAGCGGTACTACCGCCAGAACTGGCGGT
ACTACCGCCAGGACTCCCAGCGGTACTACCGCCAGGACTGGTGGTACTACCGCCAGGTGTGGCGGTACTACCGCCATGA----------
----------------------------------------ENV GENE----------------------------------------
CTACCTGTGACATTAGGTGTGCCCCCTTTTTGGTGTCTTGTGCCAAAGGGGGGAGAGAGTCTAGGGATTTGATTTCATTCGAACTTTGCA
TTGCTTTGCTGATTTGCTTTGAACTTGGTTTGAGGGGGAGCTTTGTTTGCTATCGTGTGTGAGACATGATCTATCCTATGTGAGATTAG
ATTTATTTCCATCATATGCTGTGAGTCATATGTCATATATCTCTATCCTTTTTATTCTCATATTATTATCCATGCAAATCCGGTATTGT
CATCAATCCACCAAAAAGGGGGGAGATTGTTGAGGCATAATTTATGCCTAAGTAATTTTGGTGATTGATGACAGTACCGTACAGGACTAA
TCGTGTGTGTCAAGGTTTCAGGCAACATTCGTCAACGGCACAAGACGACACGACTCCTCTCTTCGGGAACGGAAACACGGCGCTATCCT

AGATTCTCTTCATTTGAGTCATAGGAAAGCCGTACTATTAAGAGGGGATCCGTAGTGGAAAGGTTTGGGTGGAATCTATCTTTGCACAC
GCACACCTCTATTTTCTCCCTTTCCTTTATCCTTGGAGCGGCCCTCGCCGGTTTGTCCTTTGCGTCTCGGCAAAATGGTCCCCAGCGGT
AGTACCGCTGGTTCCAGCGGTAGTACCGCTGGACTGCCAGCGGTAGTACCACTGCAGCCAGCGGTAGTACCGCTGCAGCCAGCGGTATT
ACCGCTGGCTGGCGGTAGTACCGCCCCTGGTCAGCGGTAGTACCGCTCCAGGTGCCCTGTTCCACCTACCTAGCGGTAGTTCGTGGACG
GACCCTTTTGCGAAGACTTTCTCGGCGGTAGTAGTGTTTATGCACTACCAAGGGCCAGCGGTAGTACCGCTGGTGCCAGCGGTAGTACC
GCTGGAAGCCCCAGCGGTAGTACCGGTGCATGCAGCGGTAGTACCGCCAGGCAGCGGTAGTACCGCTCCTTCCCAGCGGTAGTACCGCT

GGATCTCGGGCAGAGAGTGGGAAAACGGTTTGAATTTCCCCCCACTATATAAAGGGTTCTTCTAGCTGAGGAACCCTATCTCTTACCT
CTCTAAGCTCCATTGTTGCTCCACAAGCTTAAAAGTGCCCGATCTCTCTCCCTAGCCAATCAAACTTGTTGATTCTTTCGGGATTGGTT
GAGAAGGCCTAGATCCACACTTTCACCAAGAGAAAAGTTGATTCCCCCACCTATCCCTTGCGGATCTTGTTACTCTTGGGTGTTTGAGC
ATCCTAGACGGTTGAGGTCACCTCGAAGCCATATTCCATTGTGGTGAAGCTTCGTGGTCTTGTTGGGAGCCTCCAAGCTTTGTGTGGAG
TTGCCCCAACCTTGTTTGTAAAGGTTCGGTCGCCGCCTTCAAGGGCACCTATAGTGGAATCACGGTACCTTGCATCGTGCGAGGGCGTG
AGGAGAATACGGTGGCCTTAGTGGCTTTTTGGGGAGCATTGTGCCTCCACACCGCTCCAACGGAGACGTACTTCCTATCAAAGGGAAGG
AACTTCGGTAACACATCCTCGTCTCCATCGGTTCCACTTGTGGTTATCTCTAACCTTTACTTTGTATTTGCTTTTCCTTGTGACAAAAT
CTTACTTGTCTTAATAGGTCTTGTTGATAGCTTCTATAGGGGTCACCTCTTCGTCATATTATTTGTGAAATCGTATAGTGTTTACCTTA
ACTTGTTAAGATTAATTAAAAAGTGGTCGTTCTCTATTCACCCCCCCCCCTCTAGCCAACCATATCGATCCTTTCA
```

Sireviruses – What we know now



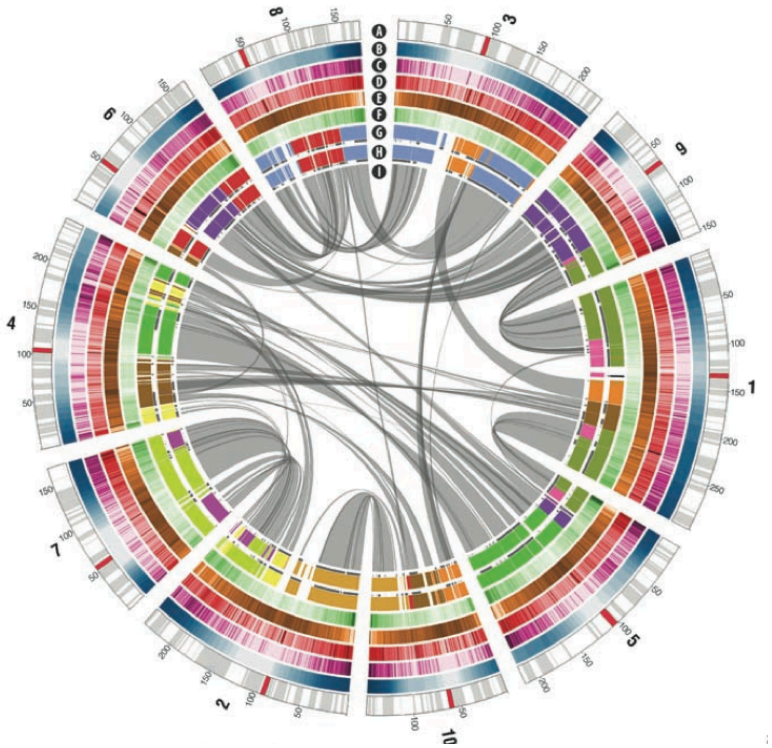
- Ancient genus with vastly divergent members
- Short highly conserved motifs in key non-coding domains that are critical for the life cycle of LTR-RTNs
- Astonishing conservation – hosts diverged 140-150 mya
- Only LTR-RTN genus with such an intriguing genome structure

The MASiVE algorithm

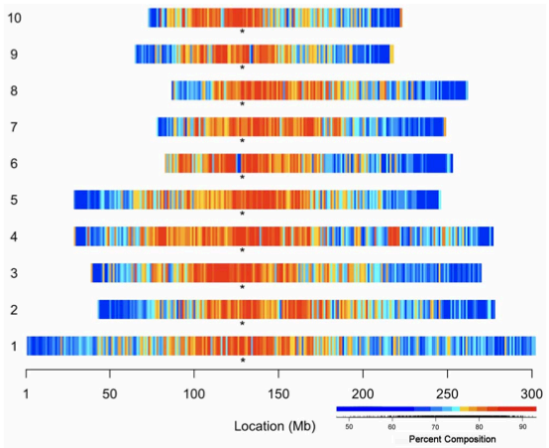


Darzentas et al. 2010 - Bioinformatics

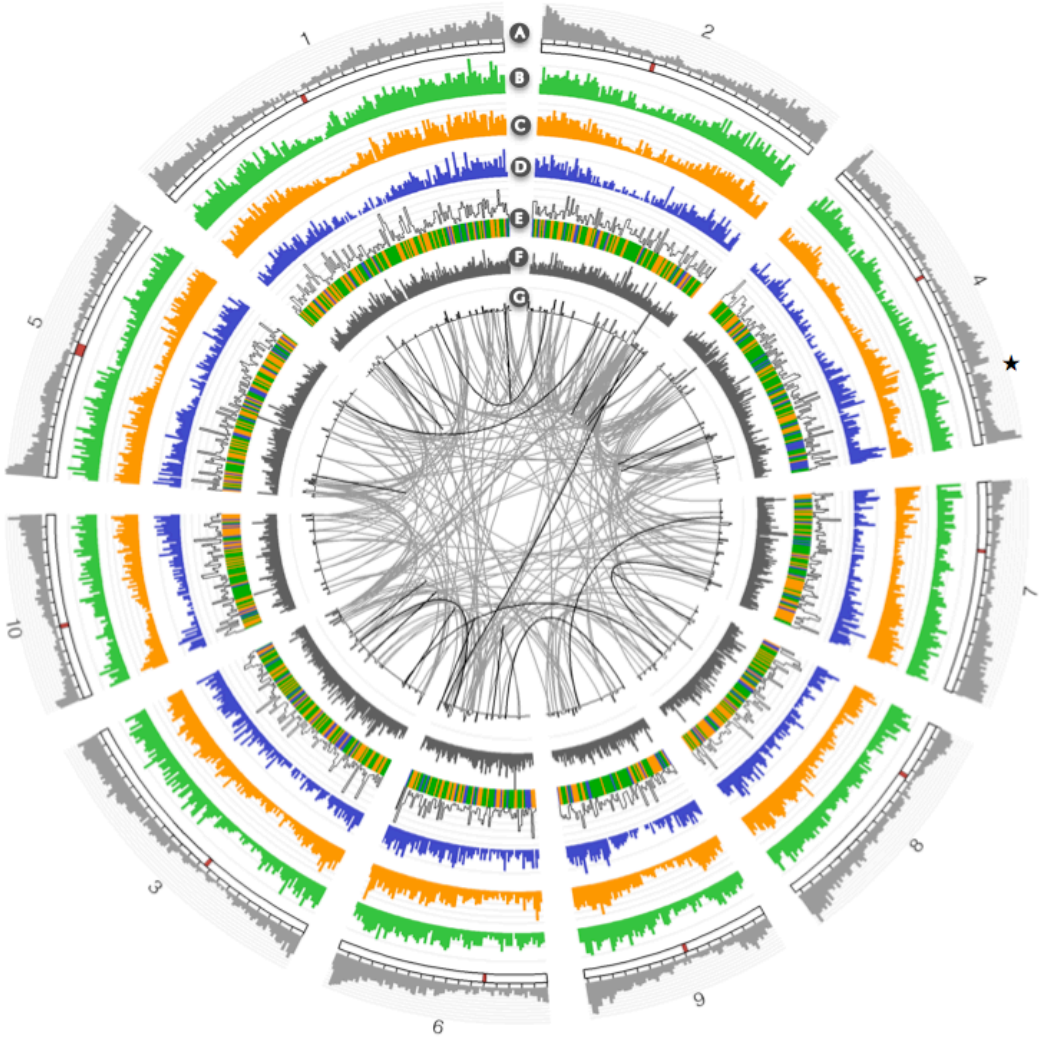
Sirevirus infiltration patterns in maize



Schnable et al. 2009 – Science



Baucom et al. 2009 – Plos Genetics



Bousios et al. 2011 – The Plant Journal

Sirevirus phylogenetic diversity in maize

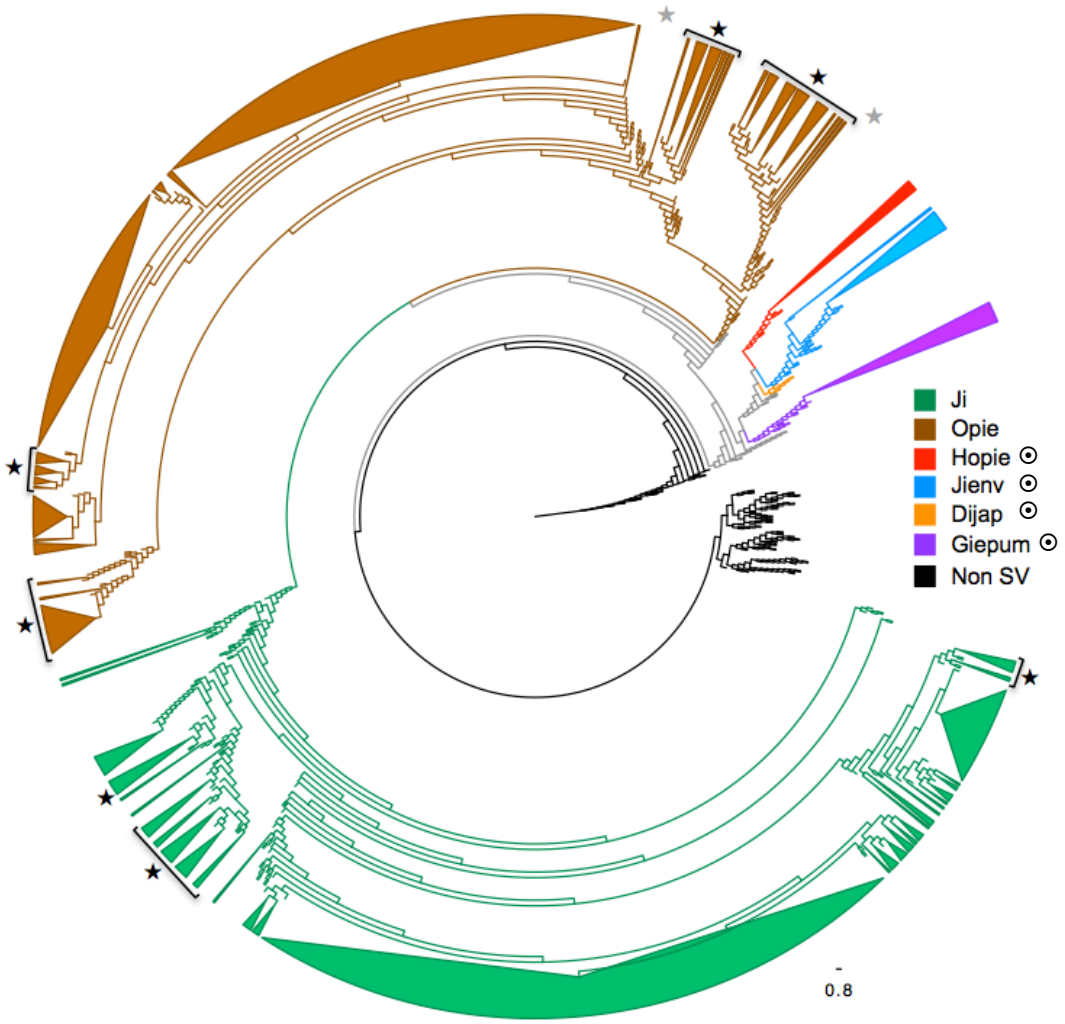


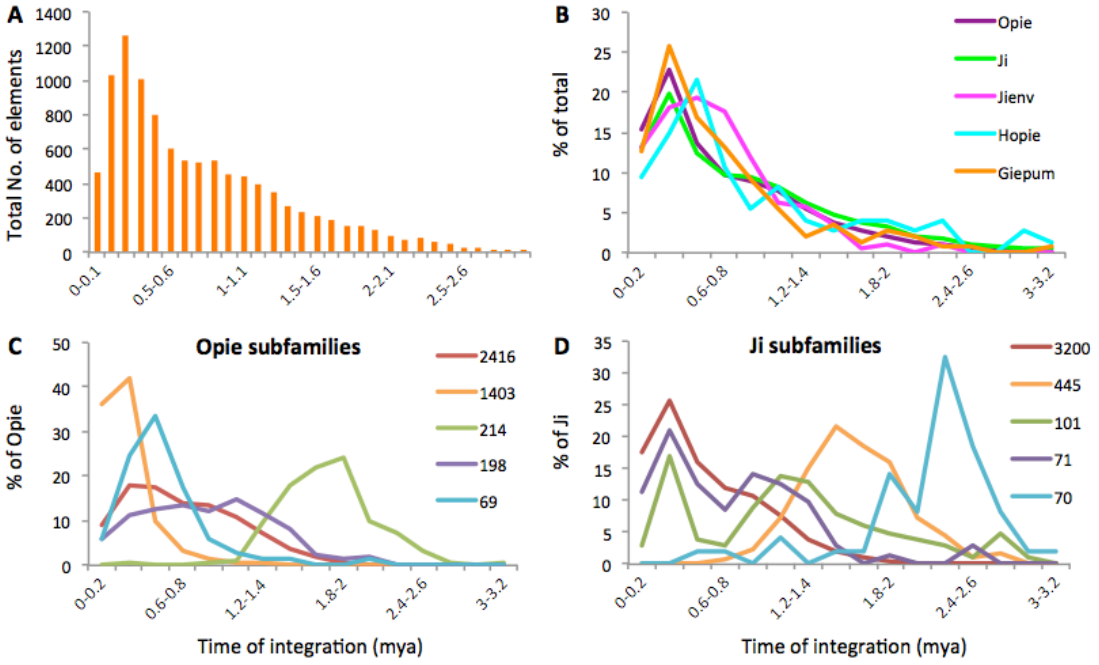
Table 1. Properties of the Sirevirus families identified in the maize genome

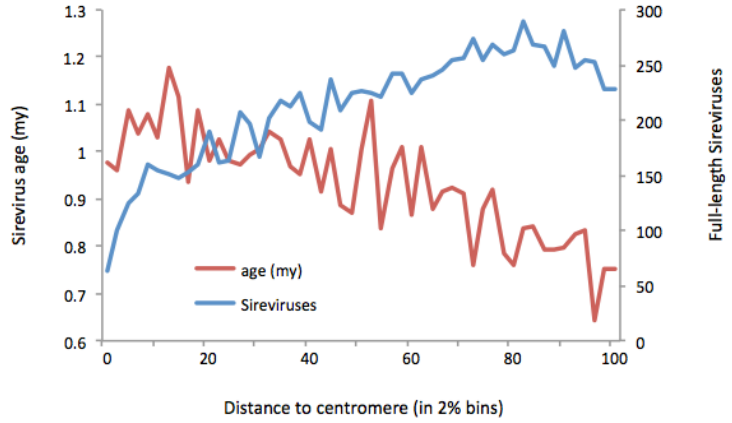
Family[a]	FL	solo	frag	FL:solo	FL:frag	solo:frag	avg age (my)	avg length FL	avg length LTR
Opie	5310 +1780[b]	2028	9826	2.6	0.5	0.2	0.90	9117	1254
Ji	4865 +772[b]	2421	11377	2.0	0.4	0.2	0.94	9519	1271
Jienv[c]	175 +175[b]	103	469	1.7	0.4	0.2	0.71	12123	1534
Giepum[c]	143	180	698	0.8	0.2	0.3	0.76	12666	1469
Hopie[c]	74 +31[b]	149	478	0.5	0.2	0.3	1.03	11696	1675
Dijap[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

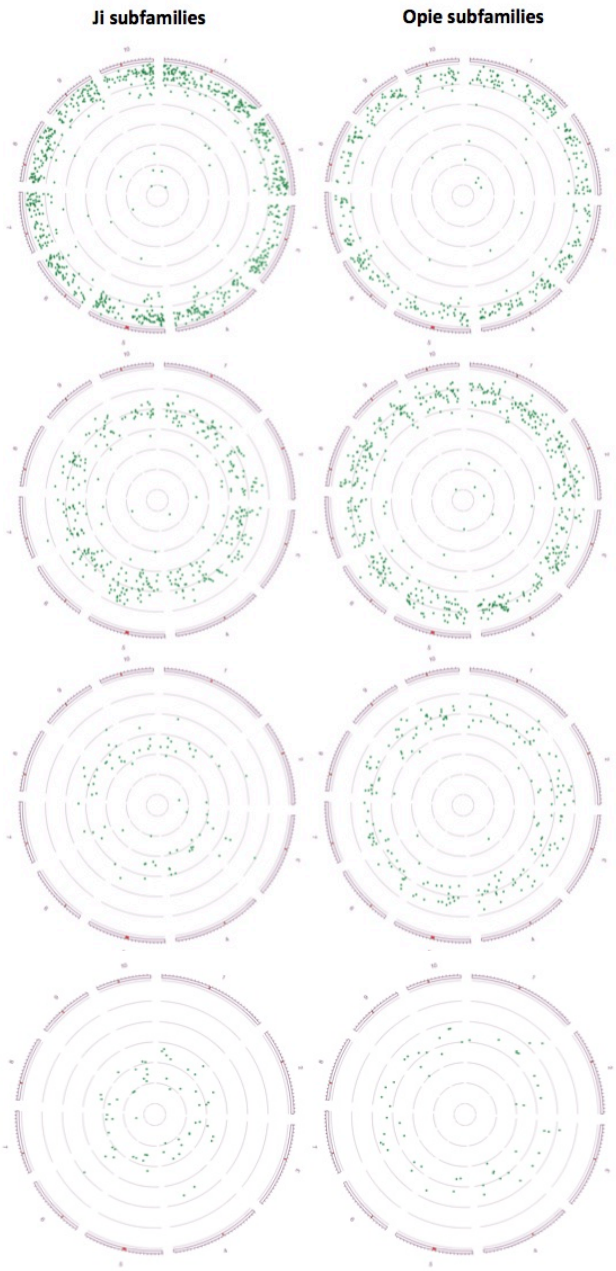
- Phylogenetic tree of maize Sireviruses based on the RT gene

Bousios et al. 2011 – The Plant Journal

Temporal activity of Sireviruses in the maize genome

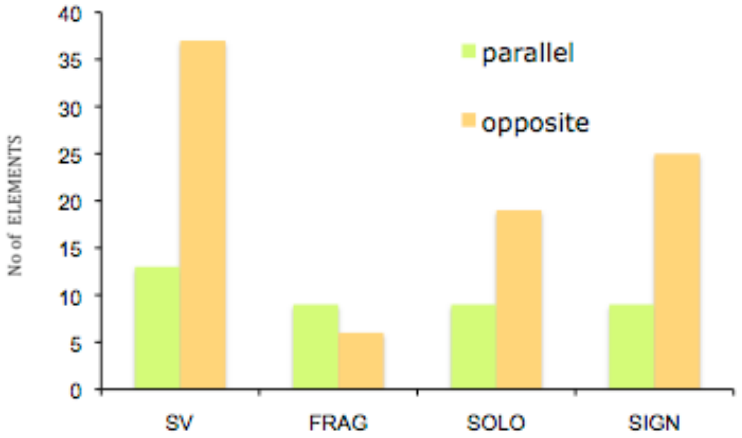


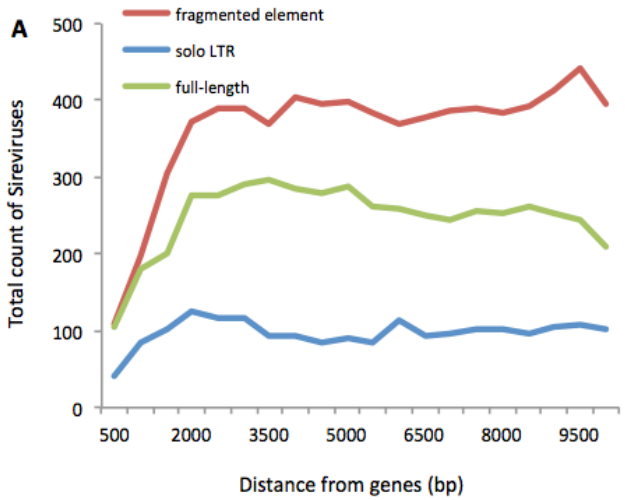




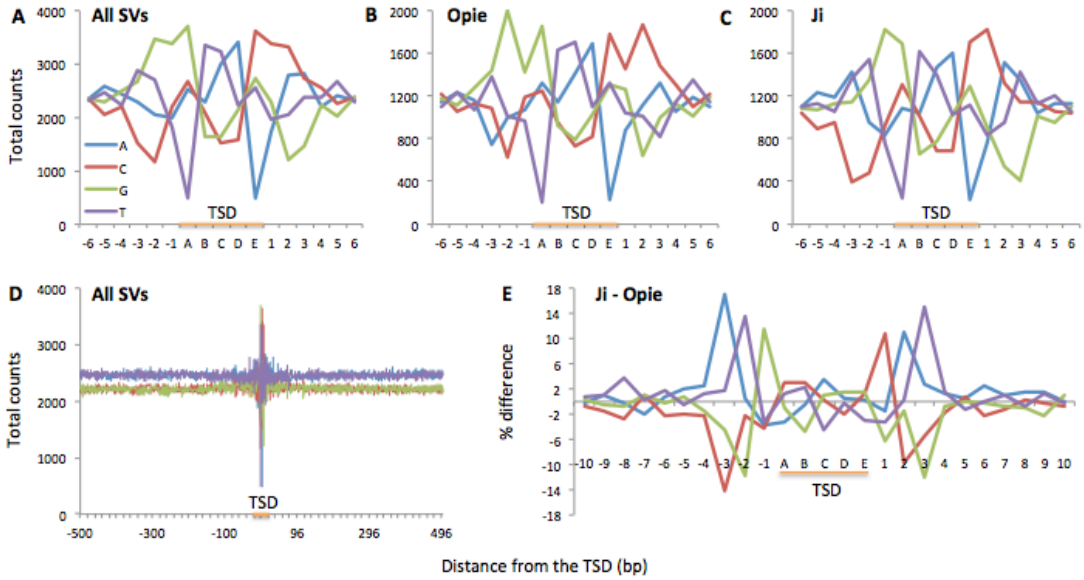
Bousios et al. 2011 – The Plant Journal

Sireviruses integration biases in the maize genome





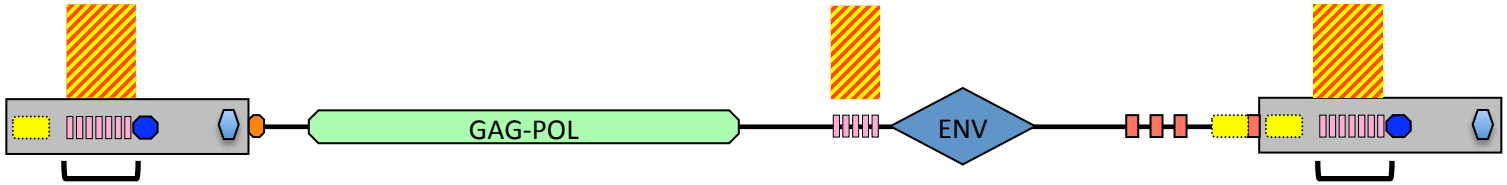
within and near genes…

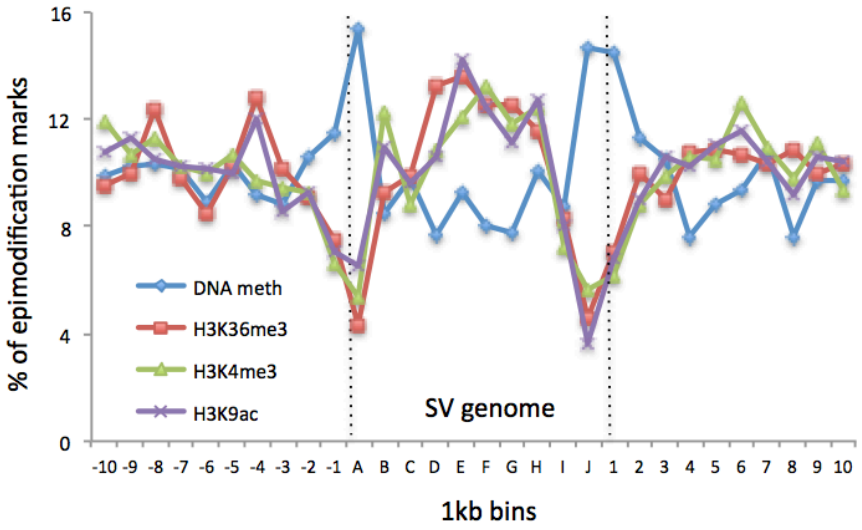


for a palindromic sequence motif…

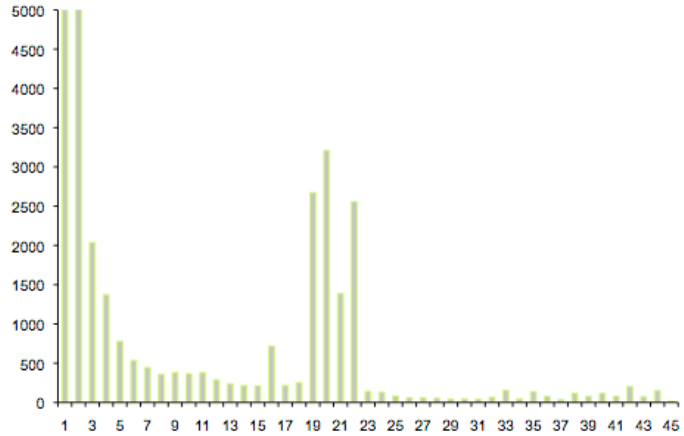
Bousios et al. 2011 – The Plant Journal

Sirevirus LTRs are methylation and recombination hotspots

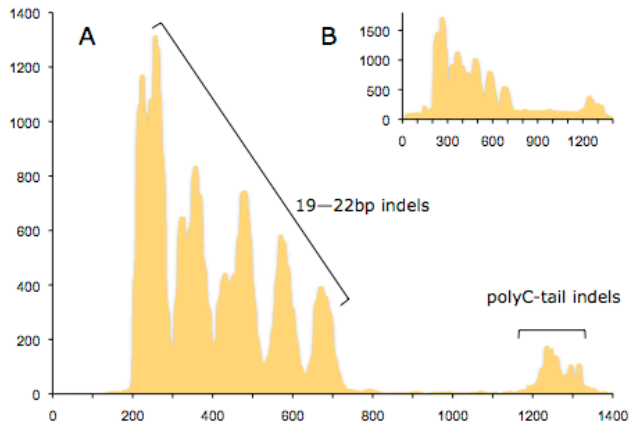




❖ Epigenetic modification profiles of Sirevirus genomes



❖ LTR indels length (bp) distribution



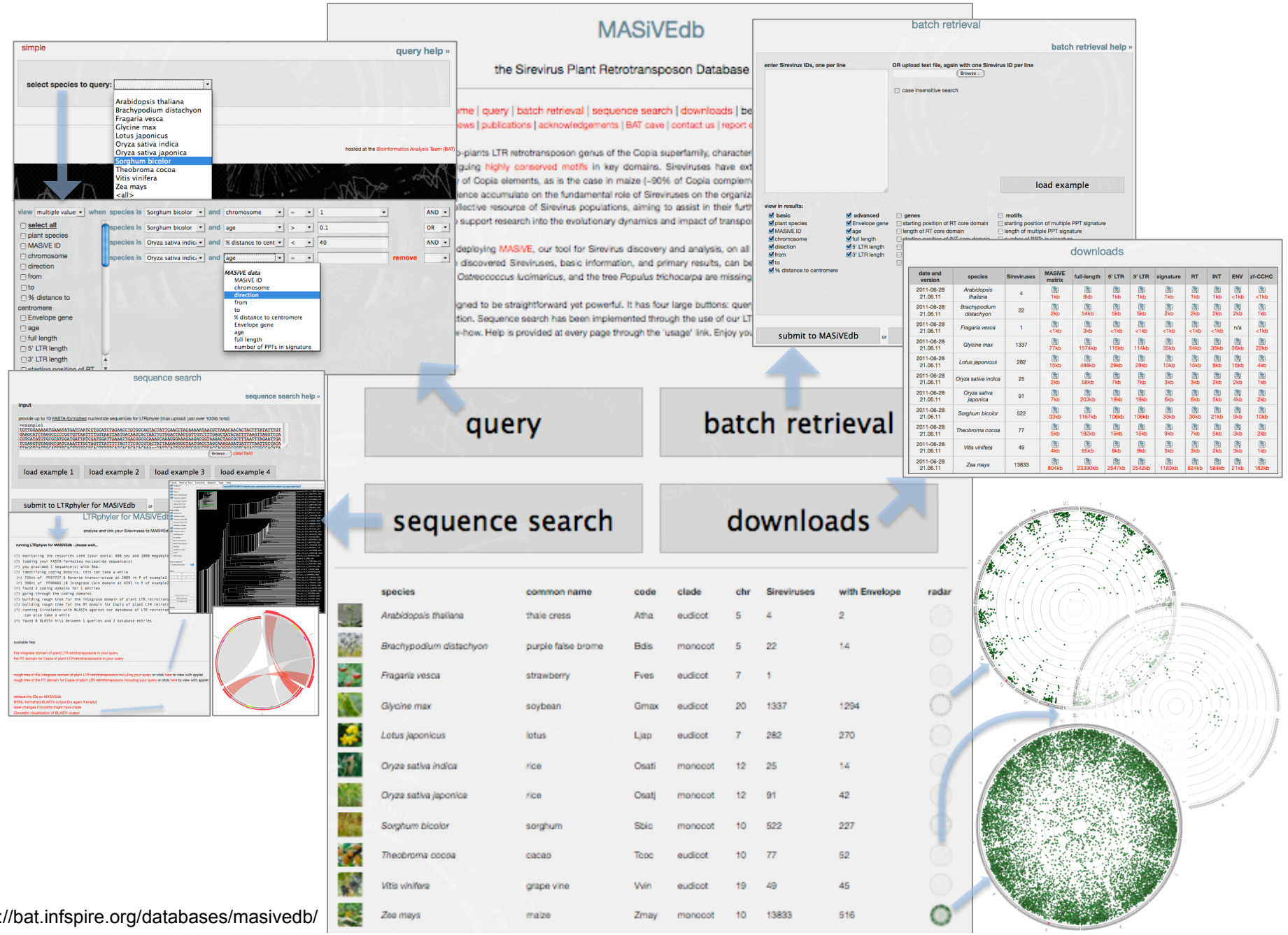
❖ Indels topology on the Sirevirus LTRs

Bousios et al. 2011 – The Plant Journal

The turbulent life of Sireviruses in the maize genome

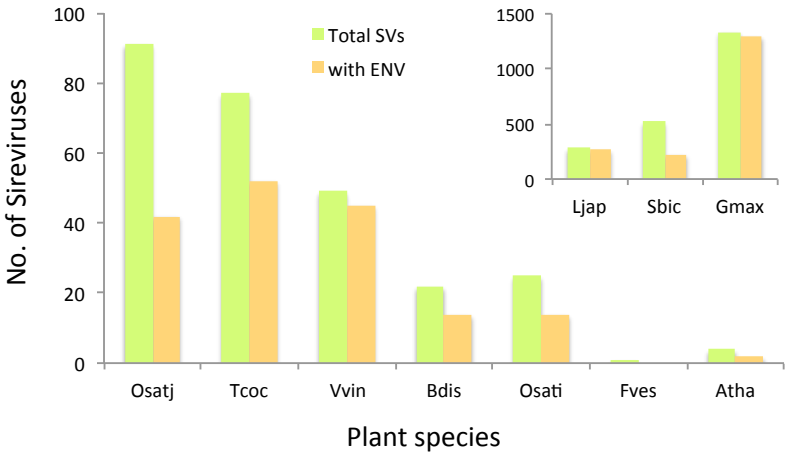
- 21% of the genome – 90% of the Copia complement
- Intense amplification period the past 600,000 years
- Plethora of families with different life and genome characteristics
- Colonize gene-rich areas, mediating gene diversification and the formation of gene islands
- Sirevirus LTRs may be the epicenter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

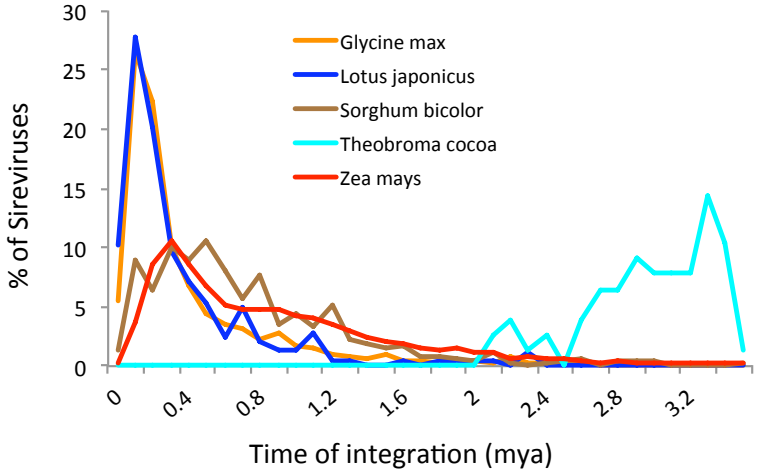
MASiVEdb: the Sirevirus Plant Retrotransposon Database

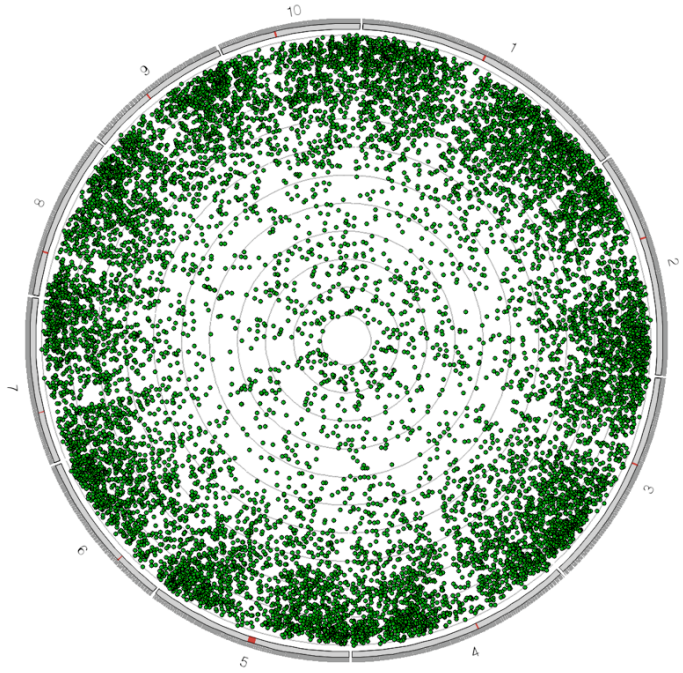


http://bat.infspire.org/databases/masivedb/

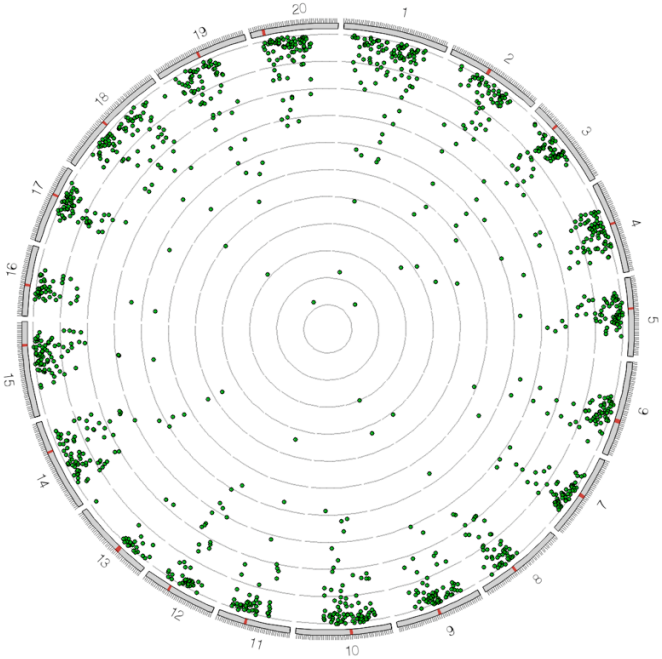
Sireviruses dynamics in other angiosperm genomes







Zea mays



Glycine max

Thank you for your attention…