

Supplemental Material

Title: Widespread Horizontal Gene Transfer from Double-stranded RNA Viruses to Eukaryotic Nuclear Genomes

Huiquan Liu^{1,2}, Yanping Fu², Daohong Jiang^{1,2} * , Guoqing Li^{1,2}, Jiatao Xie², Jiasen Cheng², Youliang Peng³, Said A. Ghabrial⁴, and Xianhong Yi²

1, State Key Laboratory of Agricultural Microbiology, Huazhong Agricultural University, Wuhan 430070, Hubei Province, P R China

2, The Provincial Key Lab of Plant Pathology of Hubei Province, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, 430070, Hubei Province, P R China

3, Department of Plant Pathology, China Agricultural University, Yuanming-Yuan West Road No. 2, Haidian District, 100193 Beijing, People's Republic of China

4, Department of Plant Pathology, University of Kentucky, 201F Plant Science Building, 1405 Veterans Drive, University of Kentucky, Lexington, KY 40546-0312, USA

18 **Table S1. Summary of candidates for non-retroviral integrated RNA viruses (NIRVs)**

19

Database	Organism	Accession no. and position (length, bp)	Best matched virus ^a	Related gene	aa identity	E value ^a	Adjacent TE ^b	ORF disruption ^c
Eukaryota genomic	Gastropods <i>Aplysia californica</i> (California sea hare)	GL006980.1: 51763..52047(285)	Cryptosporidium parvum virus 1	CP	64%	3e-24	DNA/Sola, NonLTR/Nimb	internal stop codons
GSS	Plants <i>Brassica oleracea</i> (Wild mustard)	BH939664: 32..693(662)	Carrot cryptic virus	CP	45%	6e-47		frameshift
	<i>Solanum tuberosum</i> (Potato)	EI814115: 1..138(138)	Raphanus sativus cryptic virus 1	RasR6-3(CP)	52%	5e-05		
GSS	<i>Vigna unguiculata</i> (Cowpea)	EI930635: 3..218(216)	Fragaria chiloensis cryptic virus	RNA3(CP)	41%	1e-05		
HTGS	<i>Lotus japonicus</i> (Lotus)	AP010106.1: 56301..56798(498)	Rose cryptic virus 1	RNA3 (CP)	47%	4e-17	NonLTR/L1	internal stop codon
Eukaryota genomic	Fungi <i>Vanderwaltozyma polyspora</i> DSM 70294(Budding yeast)	AAZN01000409.1: 77491..77733(243) †	Penicillium stoloniferum virus F	CP	35%	2e-05		
Eukaryotic genomic	<i>Penicillium marneffei</i> ATCC 18224 (Penicillium)	ABAR01000142.1: 119..2156(2038)*	Saccharomyces cerevisiae virus L-A	CP	39%	1e-109		
Eukaryotic genomic		ABAR01000272.1:1..1099(1099)	Saccharomyces cerevisiae virus L-A	RdRp	52%	1e-71		
Eukaryotic genomic	<i>Candida parapsilosis</i> CDC317 (Budding yeast)	CABE01000013.1: 736306..738291(1986)*	Saccharomyces cerevisiae virus L-A	CP	35%	1e-99		
Refseq_ genomic	<i>Schizosaccharomyces pombe</i> (Fission yeast)	NC_003421.2 37110..38374(1265)	Saccharomyces cerevisiae virus L-BC	CP	23%	3e-18	DNA/Polinto	internal stop codons, frameshift
GSS	Nematode <i>Strongyloides ratti</i>	CZ541007	Magnaporthe oryzae virus 2	RdRp	35%	7e-05		internal stop codons
		CZ541027	Leishmania RNA virus 1 - 1	RdRp	32%	5e-16		internal stop codons

GSS	Protozoa	<i>Entamoeba terrapinae</i> M	AM686088; AM680125	Saccharomyces cerevisiae virus L-A	RdRp	61%	3e-61
GSS			AM668999	Saccharomyces cerevisiae virus L-A	RdRp	68%	4e-27
GSS			AM686087; AM667464	Saccharomyces cerevisiae virus L-A	CP	42%	4e-47
GSS			AM677306; AM678833	Saccharomyces cerevisiae virus L-A	CP	50%	2e-53
GSS			AM670785	Saccharomyces cerevisiae virus L-A	CP	47%	2e-32
GSS			AM675473	Saccharomyces cerevisiae virus L-A	RdRp	38%	2e-36
GSS			AM677307; AM676129	Saccharomyces cerevisiae virus L-A	CP	33%	4e-25

20 ^a Best matched virus and E value were generated by using totivirus-related sequences as BLASTx queries against the non-redundant (NR) protein
21 database.

22 ^b The class/subclass of transposable element (TE) or repeat is indicated: DNA, DNA transposon; LTR, LTR retrotransposon; Non-LTR, non-LTR
23 retrotransposon. See genome map of TEs in Figure S5 in the supplemental material.

24 ^c The types of ORF disruption (frameshift, stop codon, etc) are listed.

25 † This totivirus-related sequence has been reported by Frank and Wolfe (17).

26 * The indicated two totivirus-related sequences have been reported by Taylor and Brueen (60).

27 CP, capsid protein; RdRp, RNA-dependent RNA polymerase; aa, amino acid.

28 **Table S2. List of genomic survey sequences which were assembled to generate the consensus**
 29 **sequences (contigs).**

Organism	Consensus sequence	No	Accession no. of genomic survey sequences (GSSs)				
<i>Nicotiana tabacum</i>	Contig-1	17	ET873515	ET934781	ET934863	ET967037	
			ET967098	FH297731	FH297797	FH339569	
			FH339650	FH390292	FH390372	FH504911	
			FH504985	FH635597	FH639246	FH642791	
			FH642879				
	Contig-2	7	ET940016	FH002080	FH157022	FH894314	FH001811
			FH157103				
	Contig-3	12	FH208174	FI048635	FI002313	FH232372	FH208234
			FH687998				
			FI009877	ET709315	FH396656	FH339993	FH543953
	Contig-4	9	FH687920				
			FH984307	FH939624	ET766028	FH124202	FH610390
FH215682							
<i>Zea mays</i>	Contig-1	5	FH939683	FH463206	FH610469		
			CL993907	CC669533	CG156006	CG437968	CG080070
<i>Medicago truncatula</i>	Contig-1	4	CR296971	CR499030	CG966070	CG969600	
<i>Strongyloides ratti</i>	Contig-1	4	CZ547029	CZ547019	CZ540847	CZ539206	

30

31 Genomic survey sequences (GSSs) sharing an alignment of at least 97% identity and 50 bp overlap
 32 are assembled to generate the consensus sequences (contigs) by running the CAP3 Sequence
 33 Assembly Program (Huang and Madan. Genome Res.1999, 9:868-877).

34

35

36

37

38

39

40

41

42 **Figure S1. Multiple alignment of *ILR2* of *Arabidopsis thaliana*, *gem* of *Festuca pratensis* and**
43 **capsid protein gene of SsPV-S.** The DNA alignment was derived from aligned amino acid
44 sequences. The region of predicted *ILR2* intron is marked with a red rectangular box. *F.p_gem*, the
45 *gem* of *Festuca pratensis*; *SsPV-S_CP*, capsid protein gene of *Sclerotinia sclerotiorum* partitivirus
46 S.

47 **Figure S2. Schematic representation of regions with similar sequences for the junctions**
48 **between some non-retroviral integrated RNA viruses (NIRVs) and host sequences.** Colored
49 rectangular boxes with arrowheads indicate open reading frames (ORFs). Green rectangular boxes
50 indicate repeated sequences in eukaryotic genomes detected by BLASTn searches. Blue arrows
51 indicate primers which were used to amplify the junctions between NIRVs and host sequences. Red
52 arrows indicate primers which were used to detect the transcripts of viral related homologs by
53 RT-PCR. Gray sectors connect corresponding homologous regions and the % nt identity is indicated.
54 Red bars represent the matched regions of sequences that are identified by BLASTn; the sequence
55 types and % nt identity are indicated: GSS, genomic survey sequences; EST, expressed sequence
56 tags, WGSs, whole-genome shotgun sequences.

57 **Figure S3. Multiple alignments of the amino acid sequences of some non-retroviral integrated**
58 **RNA viruses and their related viruses.** The regions of RdRp conserved motifs of partitiviruses (A)
59 or totiviruses (B) are marked with red solid-line boxes. FsV-1, *Fusarium solani virus 1*
60 (NP_624350); RNPV1-W8, *Rosellinia necatrix partitivirus 1-W8* (YP_392480); VfpV-1, *Vicia*
61 *faba partitivirus 1* (ABJ99996); Dh, *Debaryomyces hansenii* CBS767; Ps, *Pichia stipitis* CBS 6054;
62 Et, *Entamoeba terrapinae* M; Cp, *Candida parapsilosis* CDC317; Pm, *Penicillium marneffei* ATCC
63 18224; ScV-L-A, *Saccharomyces cerevisiae virus L-A* (NP_620495).

64 **Figure S4. Binary rooted trees of some non-retroviral integrated RNA viruses (NIRVs) (A, C)**
65 **and their related viruses (B, D) with Ka/Ks ratios plotted on branches.** These trees were
66 obtained by a neighbor-joining analysis of a codon sequence alignment under the Maximum
67 Composite Likelihood substitution model. Branch lengths are not to scale. (A) and (B), rooted tree

68 of capsid protein-like sequences of NIRVs and partitiviruses respectively. (C) and (D), rooted tree
69 of RNA-dependent RNA polymerase-like sequences NIRVs and partitiviruses respectively. A ratio
70 of less than 1 indicates purifying selection to conserve protein sequence. Note that the Ka/Ks ratios
71 are averaged over sites.

72 **Figure S5. Schematic representation of endogenous non-retroviral dsRNA-like elements and**
73 **flanking transposable elements in eukaryotic genomes.** Yellow rectangular boxes indicate
74 non-retroviral integrated RNA viruses: arrowheads within boxes indicate sequences retaining the
75 reading frames of viral genes; ψ within boxes indicate sequences containing frameshifts and stop
76 codons compared with viral genes. Green rectangular boxes indicate repetitive or transposable
77 elements annotated by BLAST. Blue rectangular boxes indicate repeats annotated and classified by
78 Censor. The Censor server automatically classifies all known repeats and adds the classification to
79 the report. The class/subclass of repeat is indicated: DNA, DNA transposon; LTR, LTR
80 retrotransposon; Non-LTR, non-LTR retrotransposon. CP, capsid protein; RdRp, RNA-dependent
81 RNA polymerase. * Two genomic contigs of *Ixodes scapularis*, which are homologous to each
82 other, one contig (ABJB010791923) contains the viral homologous sequence whereas the other
83 (ABJB010911717) does not. Gray sectors connect corresponding homologous regions.

84
85

Figure S1

ILR2 1 ATCGCCTCTGAATCTTCAACTCACAAGCGAGCTT-----GAGAA 42
F.p_gem 1 ATCTCGTGAAGA CAATGCCTCCATTGAGAGCCGCCTCGCTGCAGCCAAAGCTGCCGTCCGCGAA--CTTGGCTTGAAGACCAATTC-----CTGAG 93
SsPV-S_CP 1 ATGTCTTCCGCTCCGCTCGCAAGGTCTCCA CT CAGGAGAAAACCTCCTCT---CAGTCTGGAAAGAACCTGCCTCCAAGAA GTCTGGAAAGAAATCTGCCCCCCCGAGTATTCTGT CGATACTCCATCCGAA 132

ILR2 43 AAGACTTGTCTTAGTITC-----CTT----- 63
F.p_gem 94 CAAACAGGACTGACCTCG-----CTCGACGAGCTTGACCCCAAGCAGTT CGACGATTGGCC---CCCTATGCTCCAAAGAGCTGAAGCCGTCCACCTCGGAGCCCAACCCCTCGAAA GAGCTCACACGC 219
SsPV-S_CP 133 CAATCGGATGTTGAATCCGACATCCTCTCCGAAGCTCGAAGTCTCTGACGCCGATGACGCCTCCCGTCCCCTGCCCGCTTCAAAGTGAAGAAAGGTACAACACGCAACAAGAA GAAGGAAAAATCTGTGCCTCC 267

ILR2 64 -----TCTAGTTT CACCCGAAAGCAAAATCGTGTACCCAACGCAAGGTATCTCACACTACTACCTTCTCTGCCAC 132
F.p_gem 220 TCTGACGTGACACCTCCGATGACTACACCCCTGAGTCCGCCTCTGACATG---CTCATGCCTTATTTAGGCCTGAACCTTAGGTATGCTCC-----CGCCAGCCCAATCGCGCTACGCCCTCTAGCCAT 345
SsPV-S_CP 268 TCCGCTGGATCATCA-----ATCCACCCTACATGATGATGCTCACCGCCCTCAACTTTTCTCTCAGCAAAATCGAA-----CATGTACCATCTCAACTATACTCCAAAGCTGTTGG 378

ILR2 133 ATGATG-----GTCCATGCTATGACATTACACTTTGCGATAATTTT GATTTCAAACGGGC AATCCCAACTATCACCCCTACATCCTCCGCCTCTACTGTGCTGTTCTCTTTTGGATCCAAGTGTCTTAGA 258
F.p_gem 346 ATGATGGACTACATCGTCCACCTCATCAACGACAATTTATCGGACAATTTCTATTTCAAGA GAAGCTGTCGCGACTACACCCCTACATCCTCCGCCTCTACTATGCGTATTTATTTGGATCCAAGTGCCTCCG 480
SsPV-S_CP 379 TC GATGTACGCCGCTCTCGACGCTATGCATGATCTCGTTGGTGATAACGCTTCTCTTCGTGCTTTCTGTCCCTACTATCACGTGCCCCTGTCCAATAATTTACTATGATATTACTTCATCATCAAGTCTCCG 513

ILR2 259 GCTGAAACGATGTGAATGATCTTACTGATGTTCAACACCGGTTCTTAAACCGGTTCTTGGACAACCATCCTTGGAAACTCTCGTGTTC CCGGCCCTCTCCTCGGACTCTTCAAGACTCTTTGCTCTTCGCAG 393
F.p_gem 481 GCTGCCCGT CACGTCGGTGTCTT GAGGACCAAGAATATCAGTTT T GATGCGTTTCTAGATGCTTACCCTCTAGAGTGGTTACCATGTCCA GCCCTTGTGTTTATATTCAAA ACTTTGTGCTCTCCAG 615
SsPV-S_CP 514 GCCAATCAAGTCCCAACAATCTCTCTCAACCTGATTTT CAGTTTCTGCTTTT T T GATGCAATTTTGTCTCGAAGAACTCCTGTGCGCGTCCCCTCGTTCCTTTTCCAGACCTCGCCGCTTCAA 648

ILR2 394 CCGGAGTTTCTCATAATCATAATGGGAAA GTGATCCTCGCATCCCTGCCAACGAGCCCGGGCTCGAGCTCACGCGTTCATGGACATCTTCTT GAAAGCCACTTCTTGGCTAACGTCCTGGCATCTTCGCC 528
F.p_gem 616 CCAAGATTTCCACC-----TACGGTAAAGTTTACCCGCCCTCCCTTCACTCCCCGCCCAACCGCCCGCAGAGTTCATGCATGAACACGTGCAAAATTTCTTTGTTCCCAATCCC CGGGATCATCGCC 744
SsPV-S_CP 649 CCTGATGGAAATCGT-----TTTAACTGGTTGTTCTCTCACTAACAAC---TACGGTCCCGCAACGCTCTAATTCACCCCGTGTGAACGCACAACTACCGCACTTCCCAACTACTCAGATGATCTCA 774

ILR2 529 CTCTTGAAGACTTAAACCGCTTTTACC CAATA CCTCCGCTATCCTCAAAGAAGGCAG---GC CATTCCGTTACCCAAATACCTGCT-----TCTAACTTCGGTCAAAAGACCTTTGGTCT 645
F.p_gem 745 CTCTCGCCACC TAAACAGTATCATCAACA CCGCCGACCTCGCAACGCTGTGTTTCCCGCAAAGGAAACACATCCTGTGACCGCAACGCCGCCAGGCCACTGTTTTTGGCCACTTCTTTCCCAT 878
SsPV-S_CP 775 CTCTTAATCTCTTTGGTGCT-----TCCAACGCTGCCCTCTTACAGCTATGAT---ACCGCAGGACAAATGGAAACA-TTCACTTTC---GCCGCTGGTGAACCATCGCTGGTTTGCCTTATGGTGCTGG 896

ILR2 646 CCATGCAAGCAGGACAGAGGCTGAAAATGGTCACTCGTTTTACCTGGCTTACAGTACCCTTGTGAAGCCGATCAAAGCCTGATGAGGCGTTCCTGATTGCTACAGCAATTTAACTTTCAGGTCACTTCTGC 780
F.p_gem 879 CGCTGCAGAACGCAACAATTTTAAAAGTGGAAGCCTGTCTCCTCTGGCTCCA GTACCCTGCGAA GCCGATCAACGGCTTAAAGAACTTTTCTGAGGATATGAATCGTTCAACTTCCCTGCCAATTGCC 1013
SsPV-S_CP 897 ATTTGCTACTGATGCTGCTGCATCGA CCAAACCTTCGCGCCCAAGCAATTACGTTCTTGGCAACATGACACCGCCATCCATAGGAAACTCTTCTATCAGTGAAGAAATGAGCGTCCAACAATGAACCG 1031

ILR2 781 CGCTGATAA CCTCGAAAAGATCTCTTCTTTCTCCACATGAACACAGCATGGCTTGGTTCAATCAAGTAAAAGGAGTTGCGGATGATGTTGCTGCATCCTTTGAAAGATCCGGCACCTT GCTGACTGCTCTCC 915
F.p_gem 1014 AAAGGATGATTTACGATTTCTTTCAAGCTTTT GAGTATGAACGGTGACCTTTCTGGTTCGCCAGTCCGCGCTGCCAAGCCGACCGTTTAAACGCTCTGGCACATTAGCTGACTGTCCCC 1148
SsPV-S_CP 1032 TGCTAAGATTCC-----CCACGCACTTACTGTGGCTAGACGGTCAATTTCACTGGTTTACAAAGCCATTGGAGCTGTCAAGGCTAAGTATTTCTGGTGGATCTACTACACTCGCTAATATCAATCC 1160

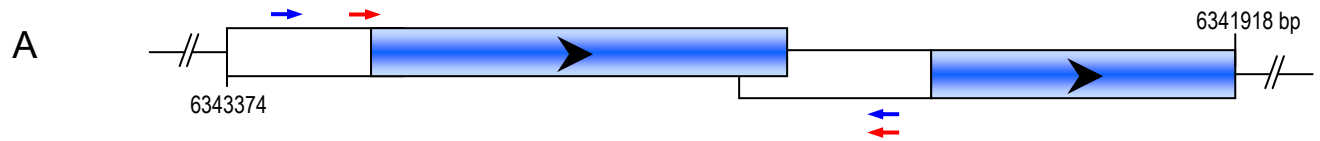
ILR2 916 ACATGGGTTGGTGTAAACCAAGTATGGTCTGTTCTCTACC CGGAACATCTTCCAGAATCACCA AATTGATTTGCTGATAAGCGTGCAACCTATGAGTTCGGTTACCAGCTCAAA-----AGCACAGTTCC 1044
F.p_gem 1149 AACCGGGATAGTCTCAAACAGATTTGCGTTGAGTACATCGCCCCCGTACCGACGTAACCGCCCACACACAACGCCGACCCCTCTCGCTGTCCCATTCACCATCAAGCTCCAT-----TCCACAGCCA 1277
SsPV-S_CP 1161 TTC AACCCGCTCTTCCGCGTCTGTTGAAACTACTGTGCTGTGCCGCGCTGCTATTCTCAAGCAACCGCC-----TGGTATGAAGGTATCCCTCAACTCTTAACTAACTGTTAACACTCGGCTCCG 1286

ILR2 1045 CAACCTCCCTCCCCTTGGCAAGCATTGCCGCTTCTCTCAGACATATACAGGATGTTTCTTAAACATCCCTTCTTCCGGACGTTTGGCTGAAAACTCTTGATCATGGCCGTTCTGG-----AAATAAG 1173
F.p_gem 1278 GAACCTTCTCGCTATCACAAGTATGACCTGCACCGCCAGACCCACATCCGATGTTTCCGACCCACCGTATTTCCGGCAGTTTGAAGATCATCGTGGCTGC---GGCCATTCTGG-----GAAATCCG 1403
SsPV-S_CP 1287 TTCATCTCTGAAACAGATCTAAAGATTTGGTACCTTCCCTGCTACTTGT---AATATTATTACAGGTGAAACTTCAACCGCTGCAACGCTTCGTCCGTTTCTCGT---GGCGATTCTTCTGTGCTGCACCGAA 1415

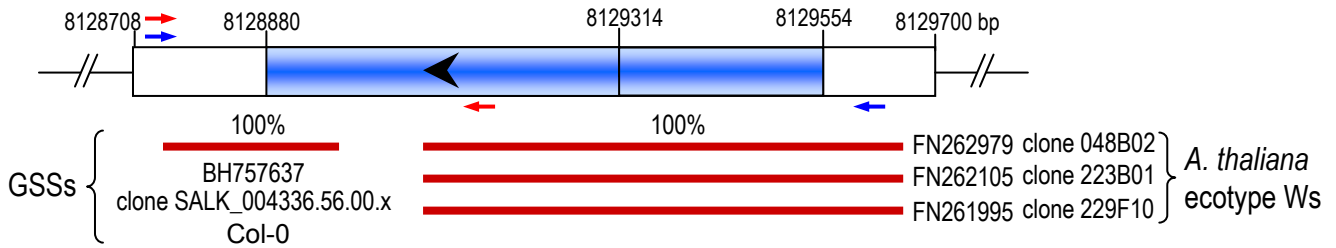
ILR2 1174 GCCTATTGGCTCCAGCCTTACCGAT-----AACAGCTCTACCTCACTATCCATCGATTTGTCAAGCAAGCATTCAAAGGCCCAAGCCAG-----TAG 1261
F.p_gem 1404 CCGCATCGAGTCGTCCCCAAGGAC-----GAGTCTTCTACCTCTCTCTACAAAGGAGTCGTGCGAAGCTGCTCAAGCCCAA-----TAG 1485
SsPV-S_CP 1416 TCTATTAACTTCGAAAATCAAACCGAAGTTAACGCAATCGCTATGGCTCAACCAACCA TGACGCTTAGCTCTACCGCCCTAAAGGTGGAGAAGCTAACTCTGATGATTAG 1527

Figure S2

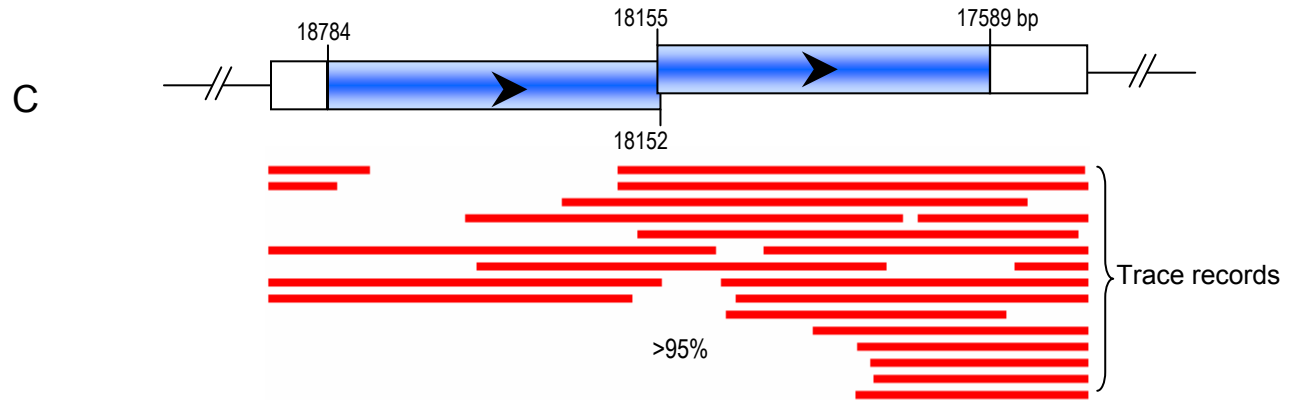
Arabidopsis thaliana ecotype Col-0 chr 3 NC_003074.8



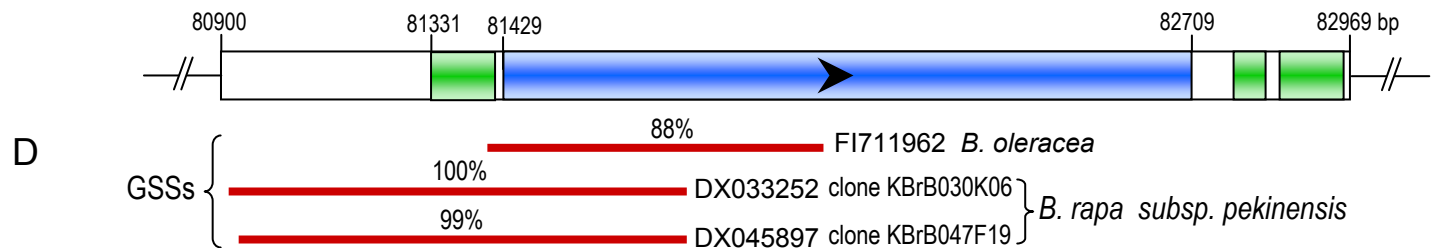
Arabidopsis thaliana ecotype Col-0 chr 4 NC_003075.7



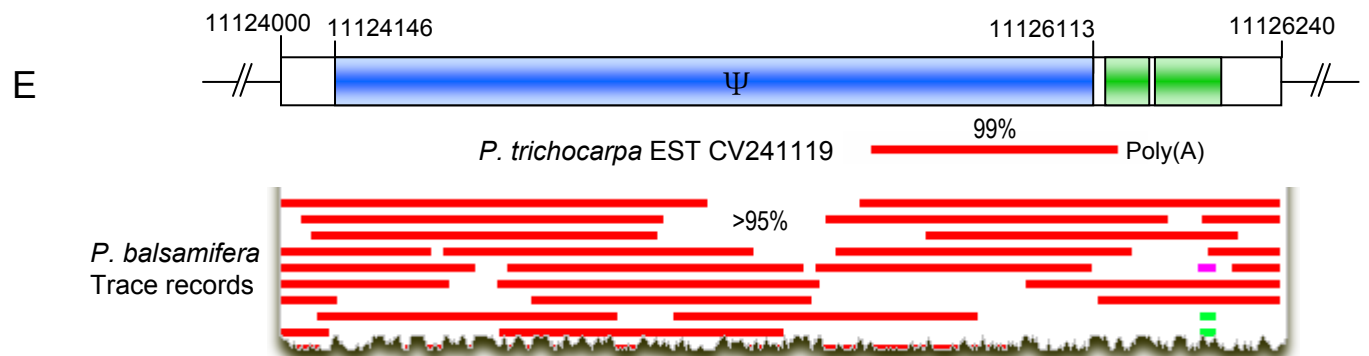
Acyrtosiphon pisum strain LSR1 NW_001917032.1



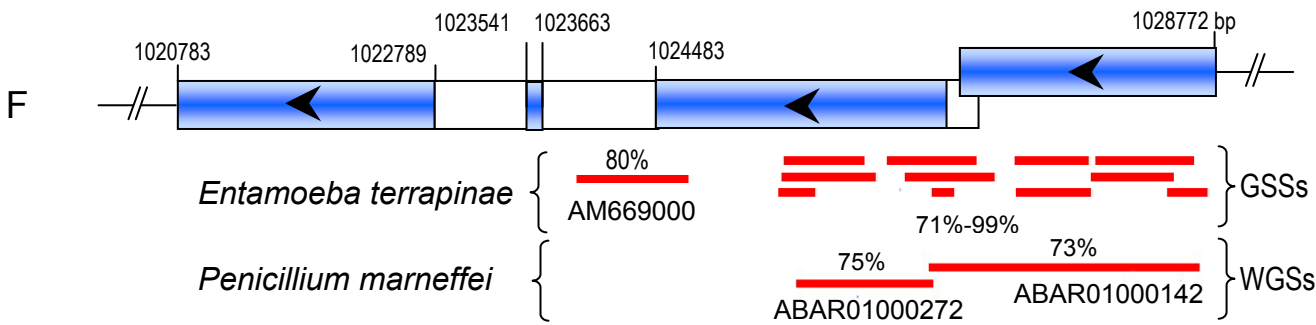
Brassica rapa subsp. *pekinensis* clone KBrB070J05 AC189442.2



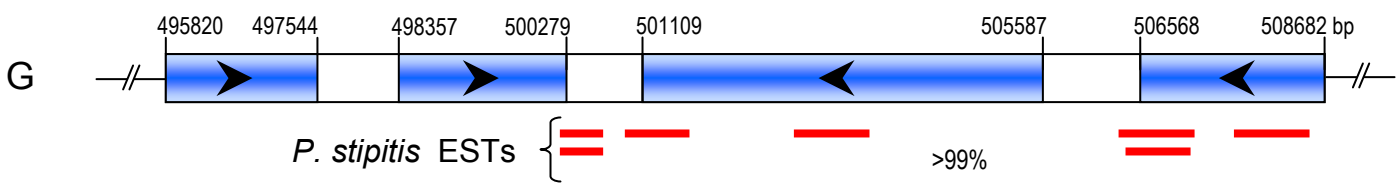
Populus trichocarpa NC_008473.1



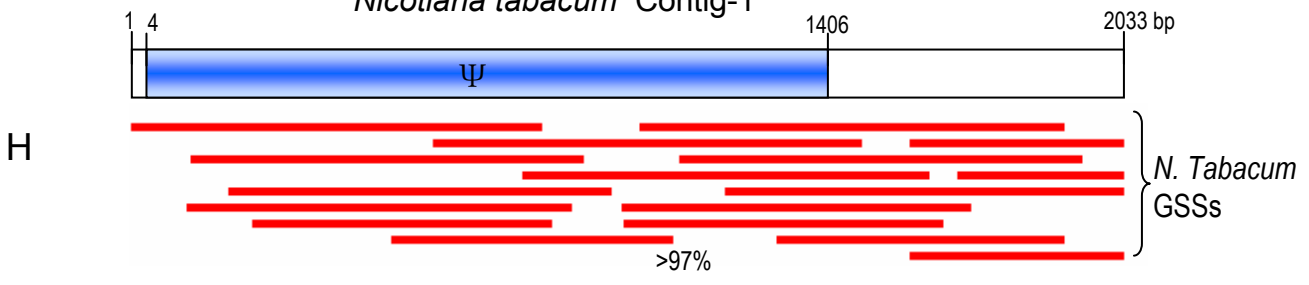
Debaryomyces hansenii CBS767 chr B NC_006044.1



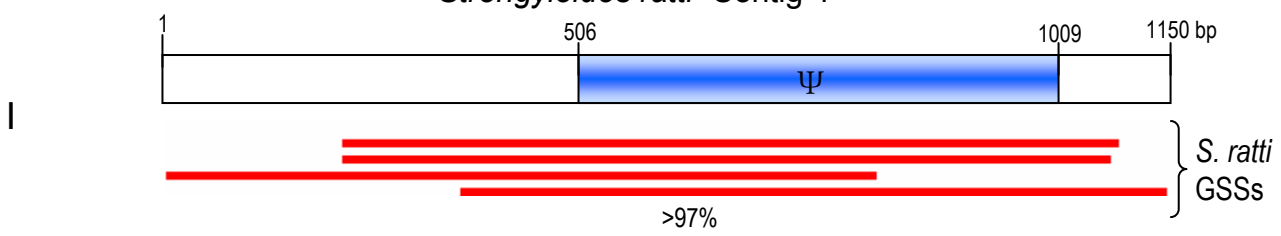
Pichia stipitis CBS 6054 chr 7 NC_009047.1



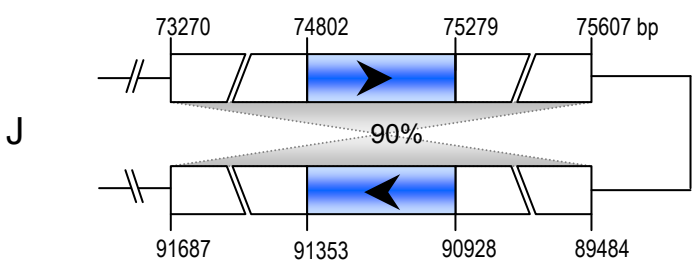
Nicotiana tabacum Contig-1



Strongyloides ratti Contig-1



Medicago truncatula chr7 AC175047.3



Lotus japonicus chr1 AP007812.2

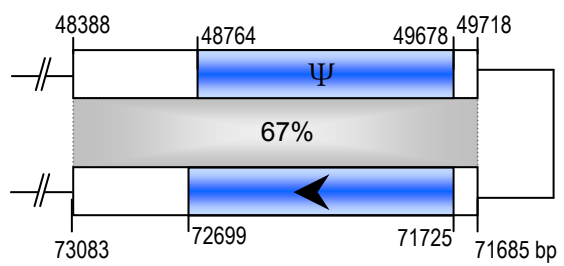


Figure S3



Figure S4

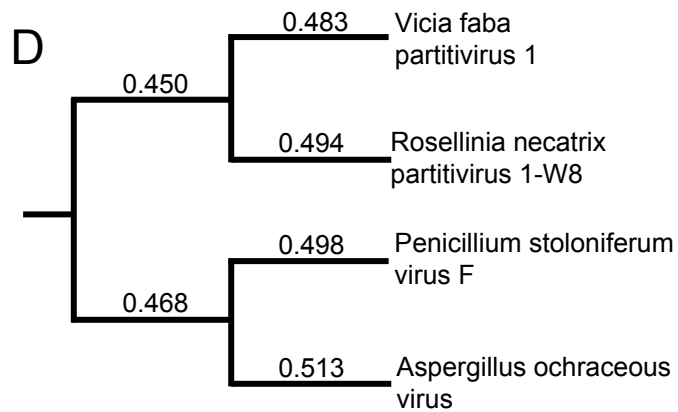
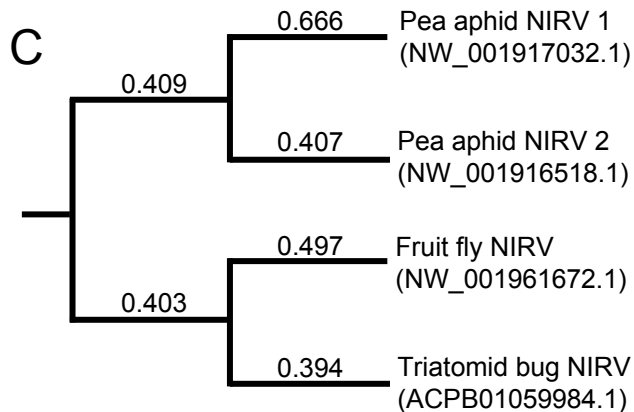
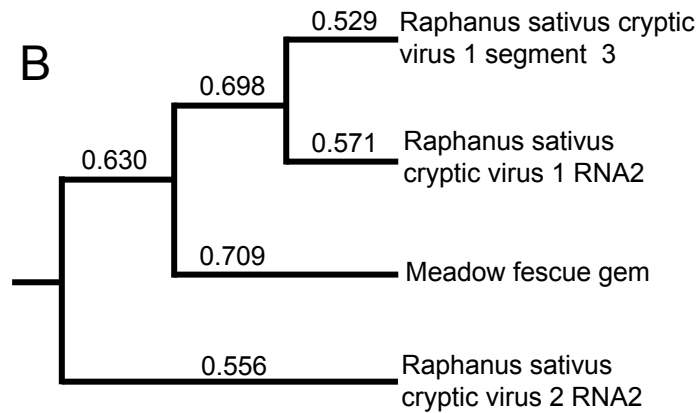
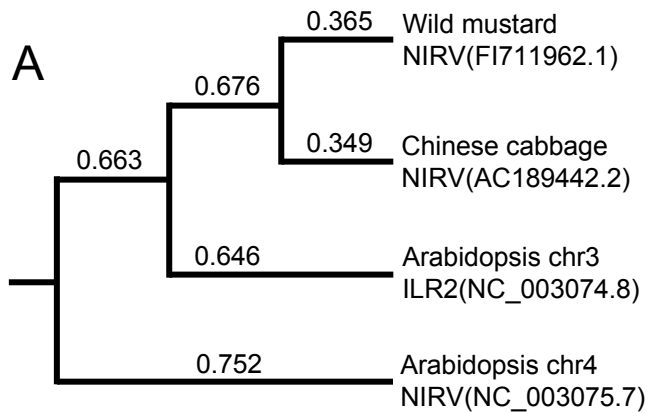
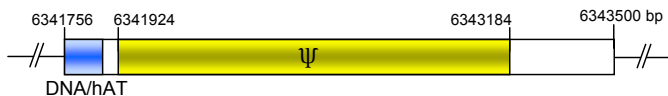
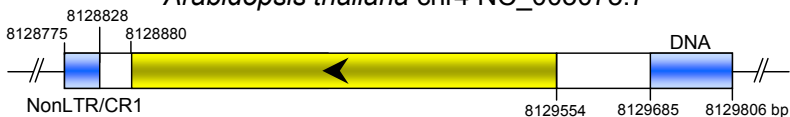


Figure S5

Arabidopsis thaliana chr3 NC_003074.8



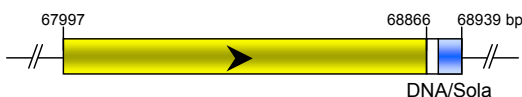
Arabidopsis thaliana chr4 NC_003075.7



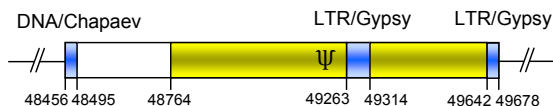
Brassica rapa subsp. pekinensis AC189442.2



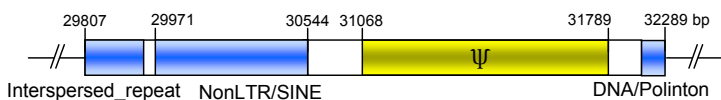
Medicago truncatula AC196856.3



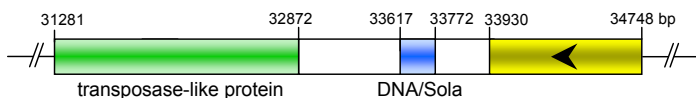
Lotus japonicus chr1 AP007812.2



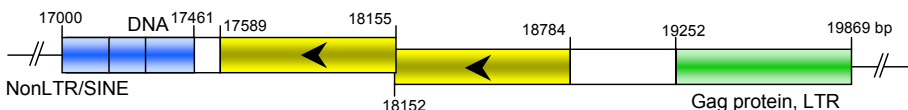
Entamoeba histolytica HM-1:IMSS NW_001915030.1



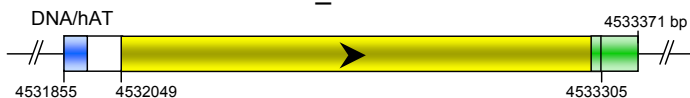
Acyrtosiphon pisum strain LSR1 NW_001916518.1



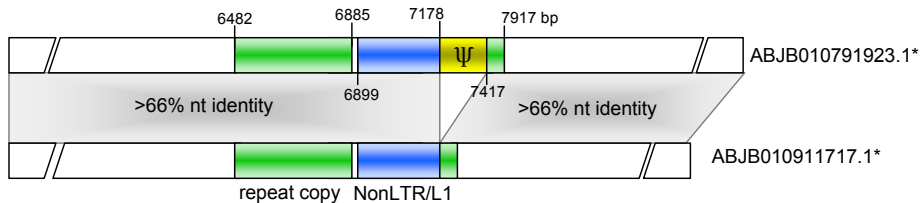
Acyrtosiphon pisum strain LSR1 NW_001917032.1



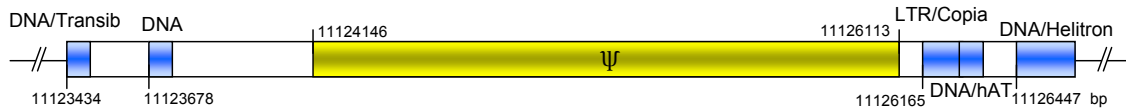
Drosophila grimshawi strain TSC #15287-2541.00
NW_001961672.1



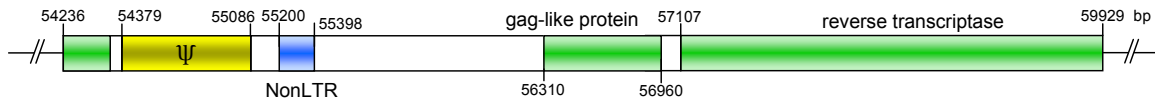
Ixodes scapularis strain Wikel



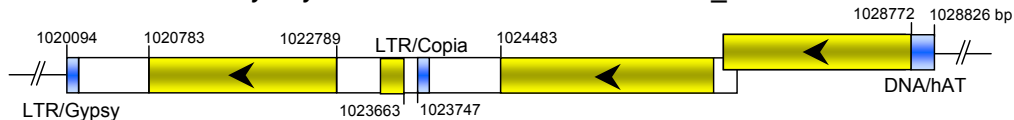
Populus Trichocarpa NC_008473.1



Aedes aegypti strain Liverpool AAGE02000678.1



Debaryomyces hansenii CBS767 chrB NC_006044.1



Pichia stipitis CBS 6054 chr7 NC_009047.1

