Supplemental Material

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

Huiquan Liu[1,2], Yanping Fu[2], Daohong Jiang[1,2] * , Guoqing Li[1,2], Jiatao Xie[2], Jiasen Cheng[2], Youliang Peng[3] , Said A. Ghabrial[4], and Xianhong Yi[2]

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

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

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	*Aplysia californica*(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	*Brassica oleracea* (Wild mustard)	BH939664: 32..693(662)	Carrot cryptic virus	CP	45%	6e-47		frameshift
		Solanum tuberosum (Potato)	EI814115: 1..138(138)	Raphanus sativus cryptic virus 1	RasR6-3(CP)	52%	5e-05		
GSS		*Vigna unguiculata* (Cowpea)	EI930635: 3..218(216)	Fragaria chiloensis cryptic virus	RNA3(CP)	41%	1e-05		
HTGS		*Lotus japonicus*(Lotus)	AP010106.1: 56301..56798(498)	Rose cryptic virus 1	RNA3 (CP)	47%	4e-17	NonLTR/L1	internal stop codon
Eukaryota genomic	Fungi	*Vanderwaltozyma polyspora* DSM 70294(Budding yeast)	AAZN01000409.1: 77491..77733(243) †	Penicillium stoloniferum virus F	CP	35%	2e-05		
Eukaryotic genomic		*Penicillium marneffei* 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		*Candida parapsilosis* CDC317 (Budding yeast)	CABE01000013.1: 736306..738291(1986)*	Saccharomyces cerevisiae virus L-A	CP	35%	1e-99		
Refseq_ genomic		*Schizosaccharomyces pombe* (Fission yeast)	NC_003421.2 chrIII: 37110..38374(1265)	Saccharomyces cerevisiae virus L-BC	CP	23%	3e-18	DNA/Polinto	internal stop codons, frameshift
GSS	Nematode	*Strongyloides ratti*	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	*Entamoeba terrapinae* 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	AM677306;	Saccharomyces cerevisiae virus L-A	CP	42%	4e-47
GSS			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

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

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

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

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

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

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

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

Organism	Consensus sequence	No.	Accession no. of genomic survey sequences (GSSs)
Nicotiana tabacum	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 FH001736
	Contig-3	12	FH208174 FI048635 FI002313 FH232372 FH208234 FH687998 FI009877 ET709315 FH396656 FH339993 FH543953 FH687920
	Contig-4	9	FH984307 FH939624 ET766028 FH124202 FH610390 FH215682 FH939683 FH463206 FH610469
Zea mays	Contig-1	5	CL993907 CC669533 CG156006 CG437968 CG080070
Medicago truncatula	Contig-1	4	CR296971 CR499030 CG966070 CG969600
Strongyloides ratti	Contig-1	4	CZ547029 CZ547019 CZ540847 CZ539206

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

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

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

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

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

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

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

Figure S1

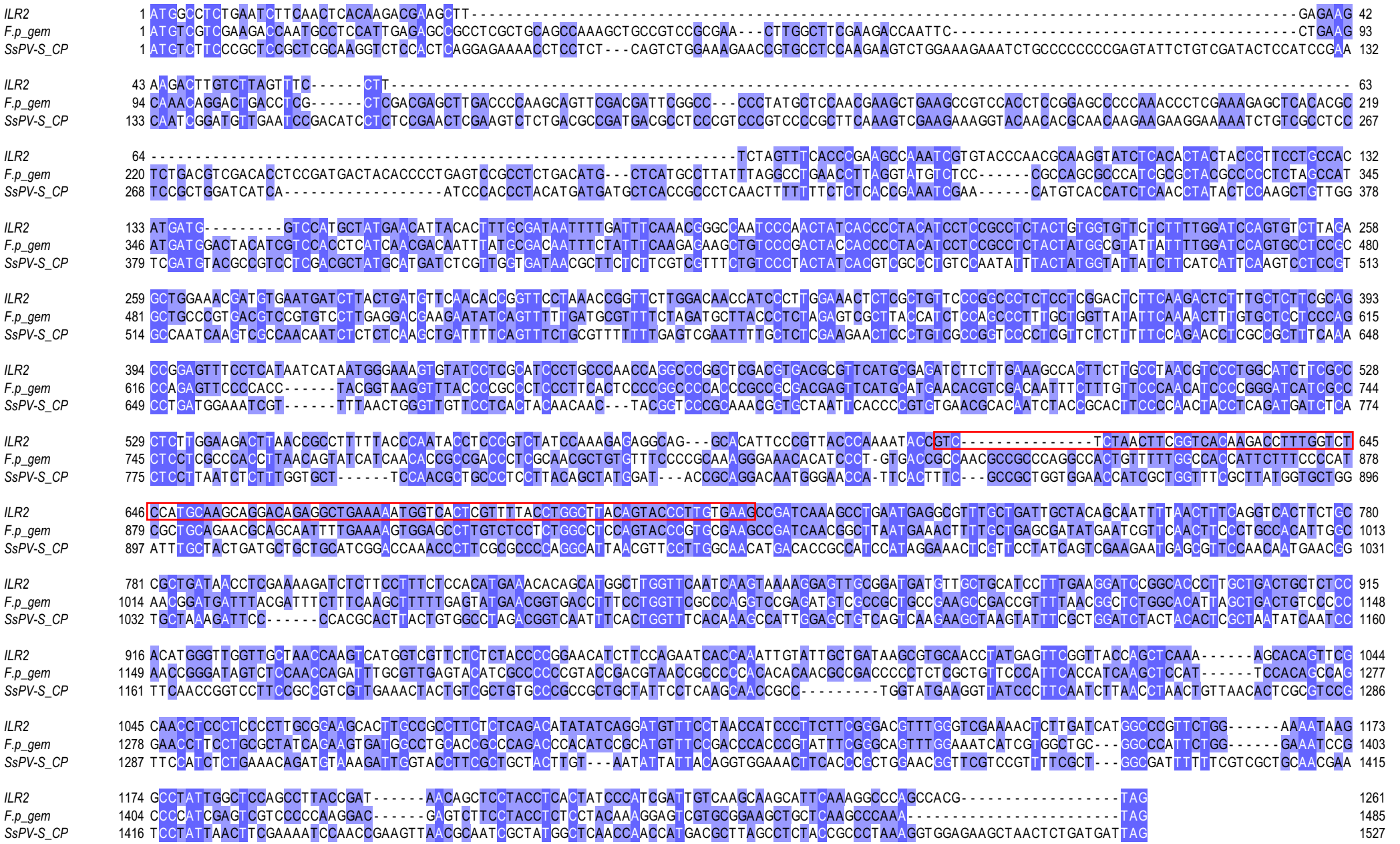
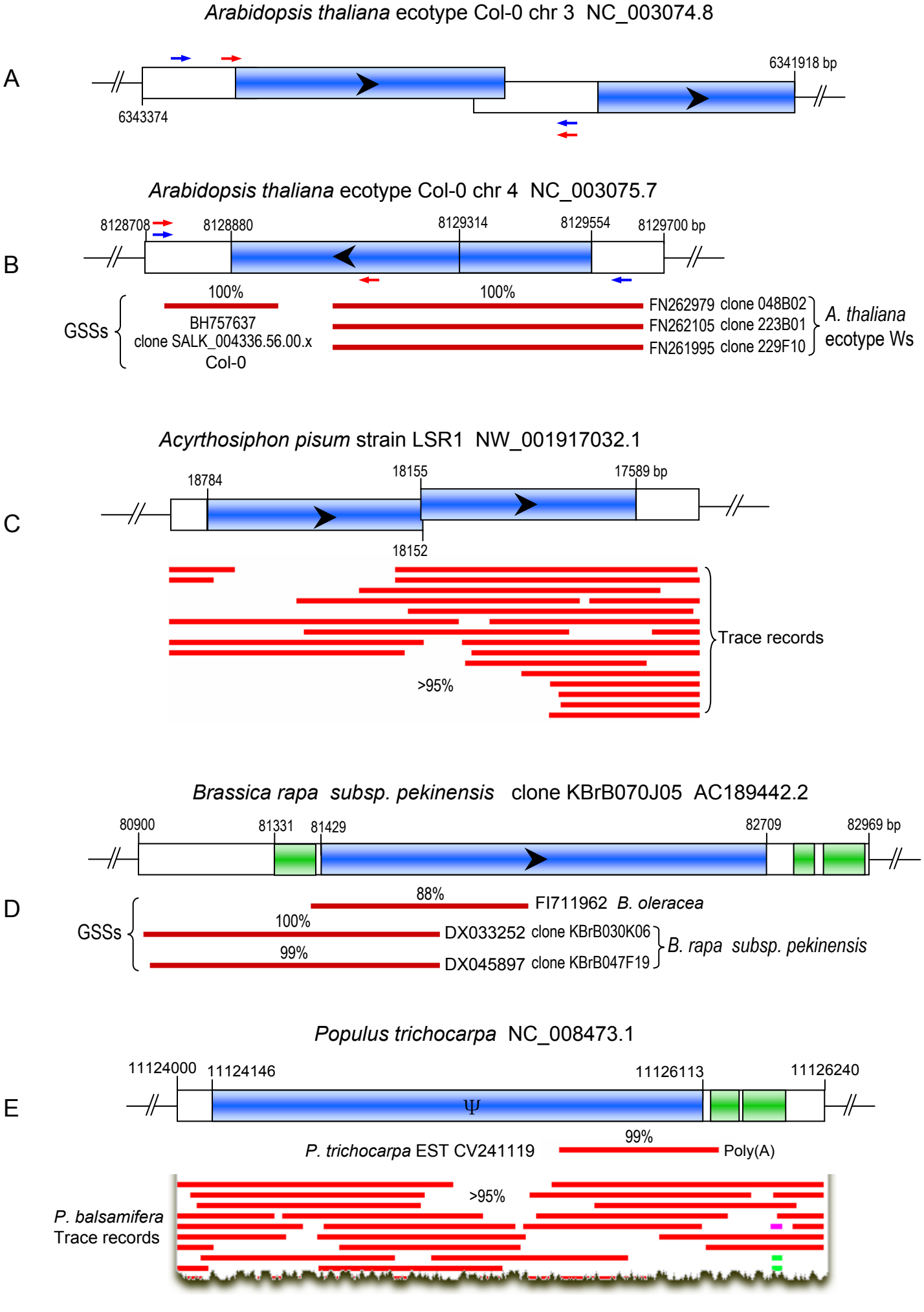


Figure S2


A
Arabidopsis thaliana ecotype Col-0 chr 3 NC_003074.8
6341918 bp
6343374
B
Arabidopsis thaliana ecotype Col-0 chr 4 NC_003075.7
8128708
8128880
8129314
8129554
8129700 bp
100%
100%
GSSs
BH757637
clone SALK_004336.56.00.x
Col-0
FN262979 clone 048B02
FN262105 clone 223B01
FN261995 clone 229F10
A. thaliana
ecotype Ws
C
Acyrthosiphon pisum strain LSR1 NW_001917032.1
18784
18155
17589 bp
18152
Trace records
>95%
D
Brassica rapa subsp. pekinensis clone KBrB070J05 AC189442.2
80900
81331
81429
82709
82969 bp
88%
FI711962 B. oleracea
GSSs
100%
DX033252 clone KBrB030K06
99%
DX045897 clone KBrB047F19
B. rapa subsp. pekinensis
E
Populus trichocarpa NC_008473.1
11124000
11124146
11126113
11126240
Ψ
99%
P. trichocarpa EST CV241119
Poly(A)
P. balsamifera
Trace records
>95%

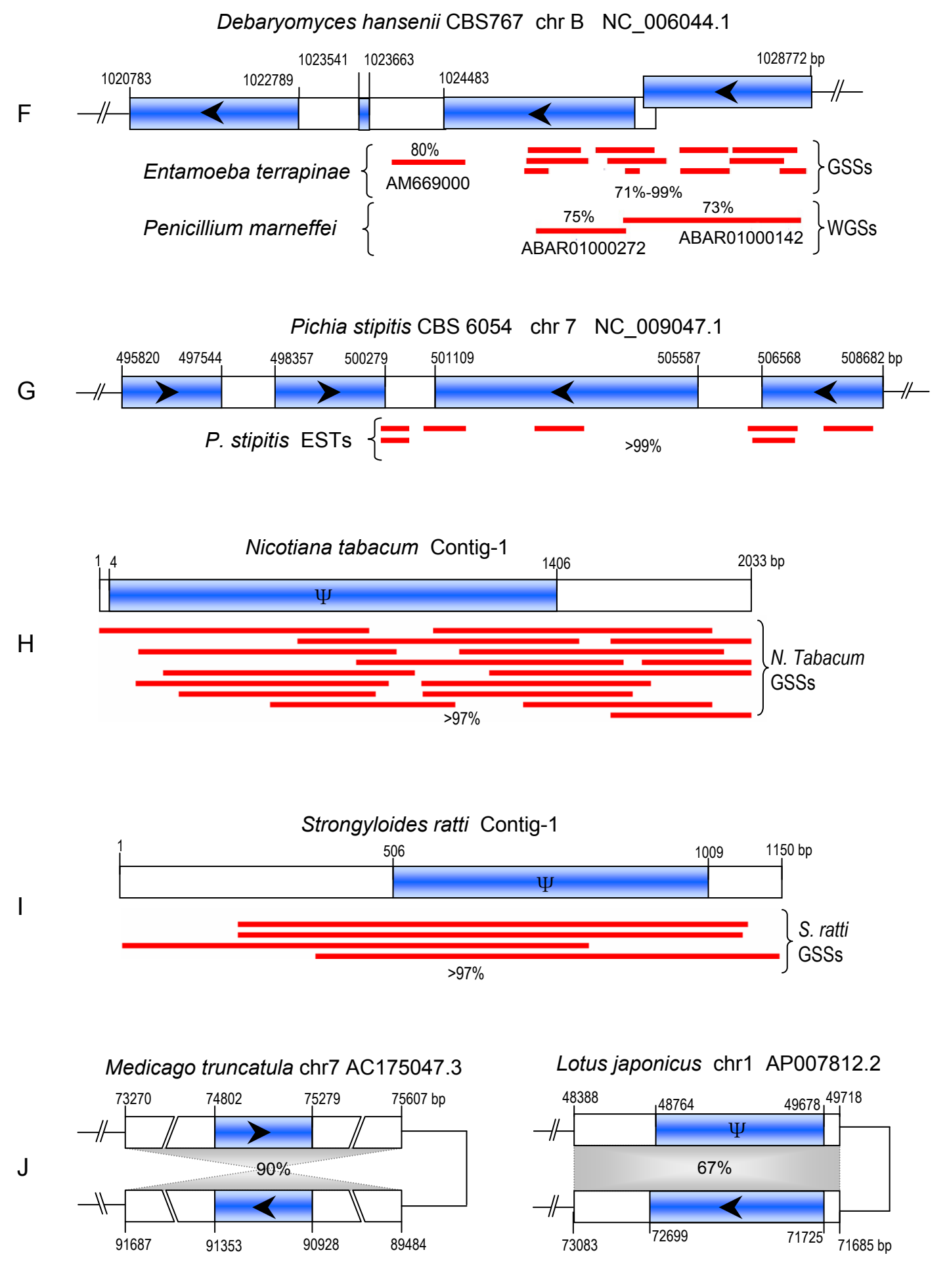

F
Debaryomyces hansenii CBS767 chr B NC_006044.1
1020783
1022789
1023541
1023663
1024483
1028772 bp
Entamoeba terrapinae
80%
AM669000
71%-99%
GSSs
Penicillium marneffei
75%
ABAR01000272
73%
ABAR01000142
WGSs
G
Pichia stipitis CBS 6054 chr 7 NC_009047.1
495820
497544
498357
500279
501109
505587
506568
508682 bp
P. stipitis ESTs
>99%
H
Nicotiana tabacum Contig-1
1
4
1406
2033 bp
Ψ
N. Tabacum
GSSs
>97%
I
Strongyloides ratti Contig-1
1
506
1009
1150 bp
Ψ
S. ratti
GSSs
>97%
J
Medicago truncatula chr7 AC175047.3
73270
74802
75279
75607 bp
90%
91687
91353
90928
89484
Lotus japonicus chr1 AP007812.2
48388
48764
49678
49718
Ψ
67%
73083
72699
71725
71685 bp

Figure S3


FsV-1_RdRp
D.grimshawi_AAPT01020539
R.prolixus_ACPB01059984
A.pisum_NW_001917032
RNPV1-W8_RdRp
VfPV-1_RdRp
A.pisum_NW_001916518
3
4
5
6
7
8

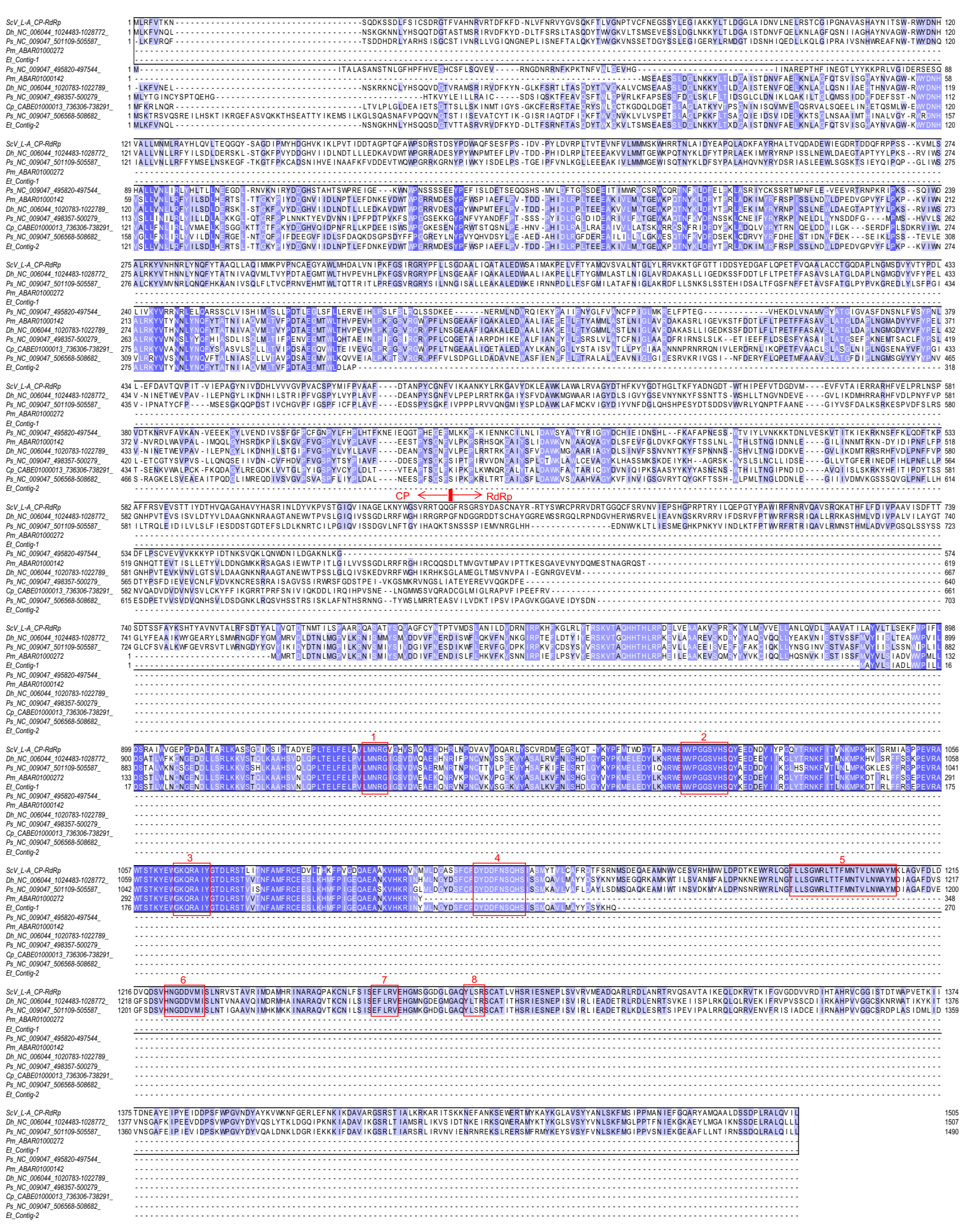

ScV_L-A_CP-RdRp
Dh_NC_006044_1024483-1028772_
Ps_NC_009047_501109-505587_
Pm_ABAR01000272
Et_Contig-1
Ps_NC_009047_495820-497544_
Pm_ABAR01000142
Dh_NC_006044_1020783-1022789_
Ps_NC_009047_498357-500279_
Cp_CABE01000013_736306-738291_
Ps_NC_009047_506568-508682_
Et_Contig-2
CP
RdRp
1
2
3
4
5
6
7
8

Figure S4

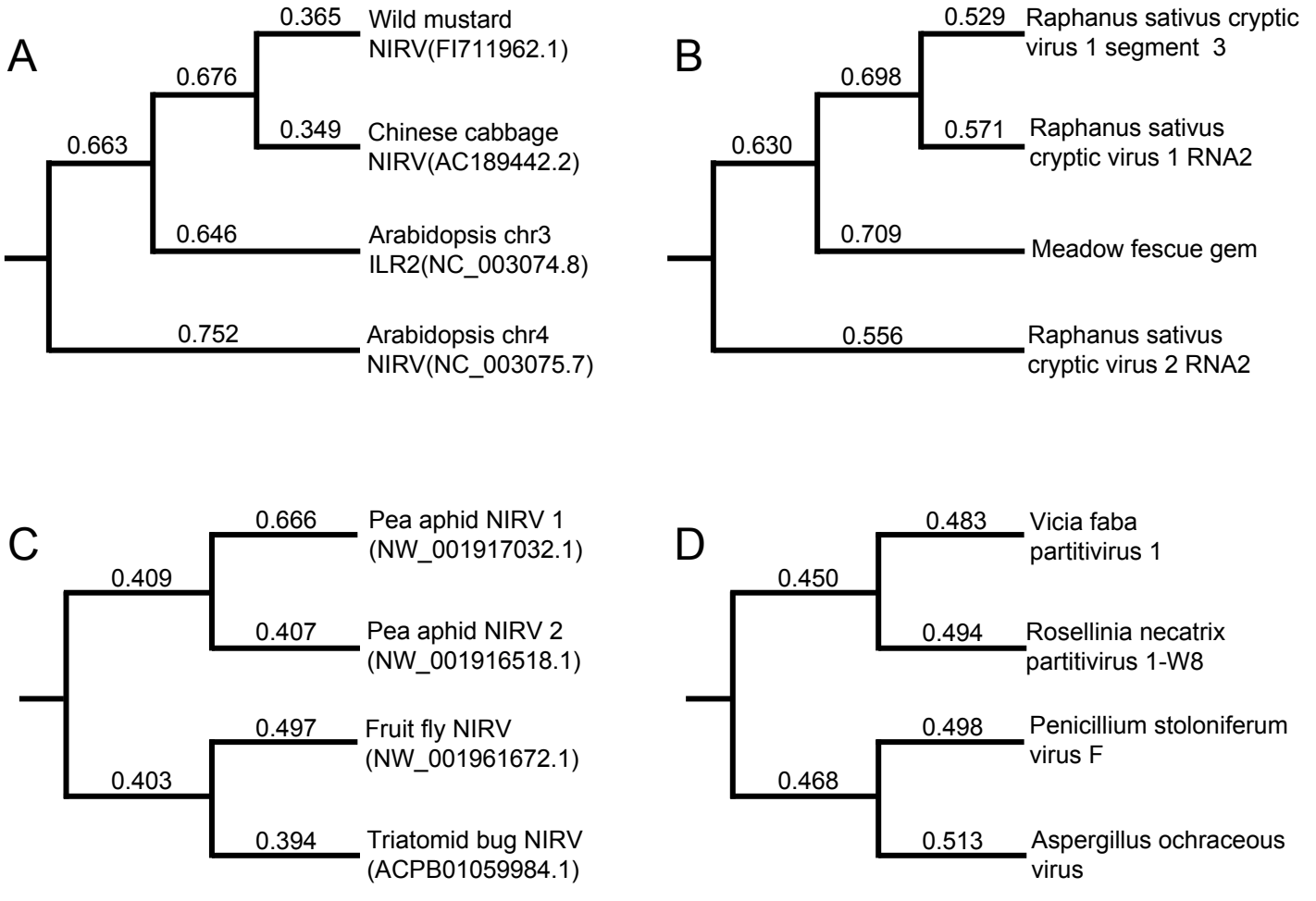
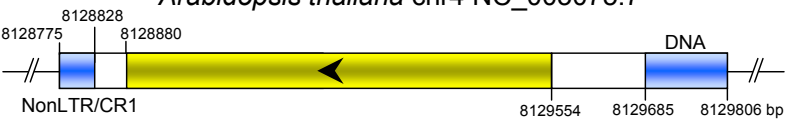

A
0.365
Wild mustard
NIRV(FI711962.1)
0.676
0.349
Chinese cabbage
NIRV(AC189442.2)
0.663
0.646
Arabidopsis chr3
ILR2(NC_003074.8)
0.752
Arabidopsis chr4
NIRV(NC_003075.7)
B
0.529
Raphanus sativus cryptic
virus 1 segment 3
0.698
0.571
Raphanus sativus
cryptic virus 1 RNA2
0.630
0.709
Meadow fescue gem
0.556
Raphanus sativus
cryptic virus 2 RNA2
C
0.666
Pea aphid NIRV 1
(NW_001917032.1)
0.409
0.407
Pea aphid NIRV 2
(NW_001916518.1)
0.497
Fruit fly NIRV
(NW_001961672.1)
0.403
0.394
Triatomid bug NIRV
(ACPB01059984.1)
D
0.483
Vicia faba
partitivirus 1
0.450
0.494
Rosellinia necatrix
partitivirus 1-W8
0.498
Penicillium stoloniferum
virus F
0.468
0.513
Aspergillus ochraceous
virus

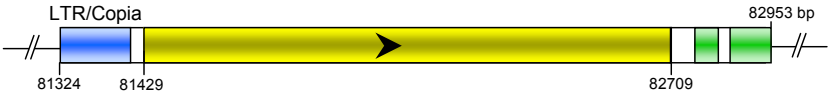
Figure S5

Arabidopsis thaliana chr3 NC_003074.8

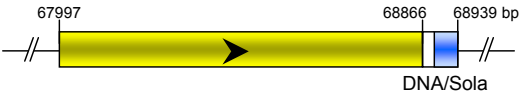
Arabidopsis thaliana chr4 NC_003075.7



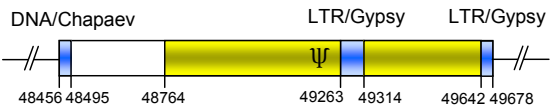
Brassica rapa subsp. pekinensis AC189442.2



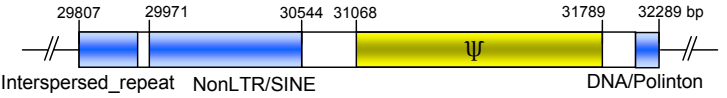
Medicago truncatula AC196856.3



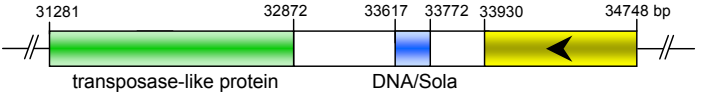
Lotus japonicus chr1 AP007812.2



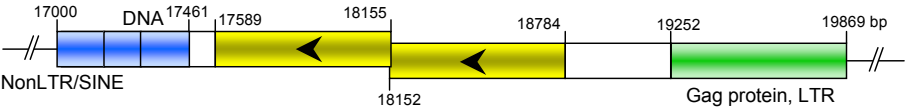
Entamoeba histolytica HM-1:IMSS NW_001915030.1



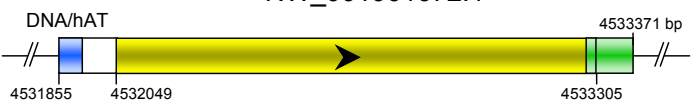
Acyrthosiphon pisum strain LSR1 NW_001916518.1



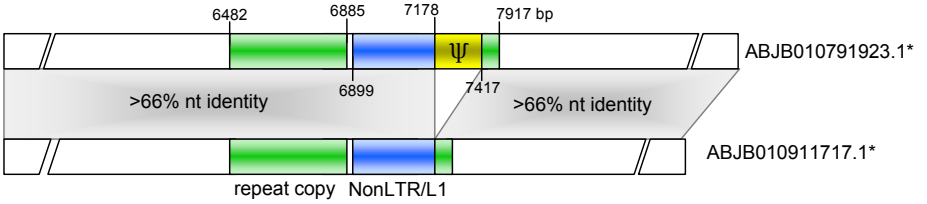
Acyrthosiphon pisum strain LSR1 NW_001917032.1



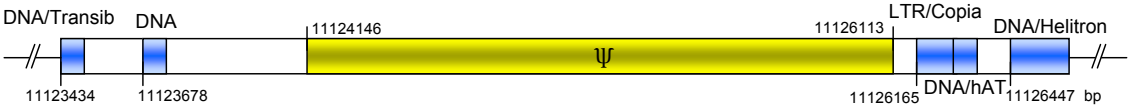
Drosophila grimshawi strain TSC #15287-2541.00 NW_001961672.1



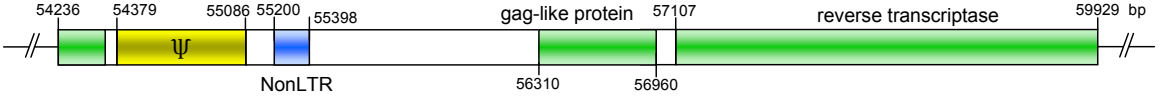
Ixodes scapularis strain Wikel



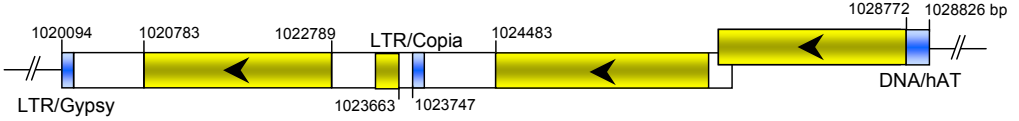
Populus Trichocarpa NC_008473.1



Aedes aegypti strain Liverpool AAGE02000678.1



Debaryomyces hansenii CBS767 chrB NC_006044.1



Pichia stipitis CBS 6054 chr7 NC_009047.1