Ancient single-stranded DNA viruses in eukaryote genomes

Journal Club in bioinformatics
Aleksander Sudakov
05.03.12

References

- Widespread Horizontal Gene Transfer from Circular Single-stranded DNA Viruses to Eukaryotic Genomes
  Huiquan Liu, Yanping Fu, Bo Li, Xiao Yu, Jiatao Xie, Jiasen Cheng, Said A Ghabrial, Guoqing Li, Xianhong Yi, and Daohong Jiang
  BMC Evol Biol. 2011; 11: 276
- Sequences from Ancestral Single-Stranded DNA Viruses in Vertebrate Genomes: the Parvoviridae and Circoviridae Are More than 40 to 50 Million Years Old
  Vladimir A. Belyi, Arnold J. Levine, and Anna Marie Skalka
  J Virol. 2010 December; 84(23): 12458–12462.

Introduction

- Integrated transcribing RNA viruses (retroviruses) comprise 8-10% of human and mouse genomes
- Integrated retroviral genes play important roles in host
- Until recently integration of non-retroviral viruses has not been found in vertebrate genomes
- ssDNA viruses are smallest viruses known to infect eukaryotes
- Anelloviridae, Parvoviridae, Circoviridae infect vertebrates
- Geminiviridae and Nanoviridae infect plants

ssDNA viruse proteins

- Dependovirus and Parvovirus: 4-6kb genome

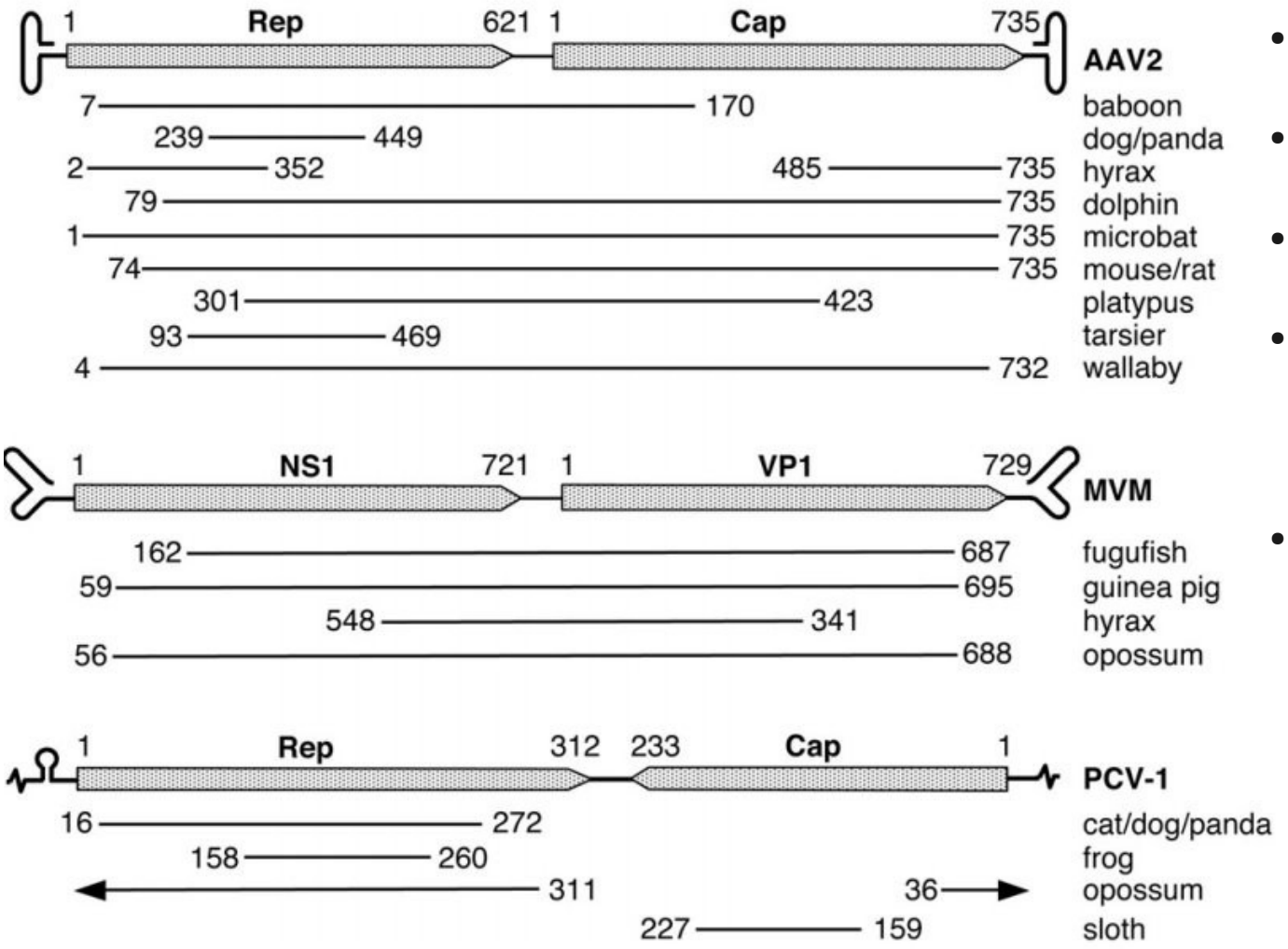
  Rep (NS1) - virus encoded replication initiation protein

  Cap (VP1) – capsid protein
- Circovirus: 2kb genome

  Rep – replicase homologous to Parvoviridae

  Usually contains helicase domain and replication protein domain
- Rep is also found in bacterial plasmids and phages

Structure and organization of Parvoviridae and Circoviridae genomes



- Dependovirus: adenoassociated virus 2 (AAV2)
- Parvovirus: minute virus of mice (MVM)
- Circovirus: porcine circovirus 1 (PCV-1)
- Horizontal lines beneath the maps indicate the lengths of similar sequences that could be identified by BLAST
- The numbers indicate the locations of amino acids in the viral proteins where the sequence similarities in the endogenous insertions start and end

Belyi et al., 2010

Methods: genome screening

- Query: peptide sequences from ssDNA viruses
- Target databases:
  nr
  refseq_genomic
  NCBI genomes
  wgs (whole genome shotgun)
  gss (genome survey sequence)
  htgs (high throughput genomic sequence)
- TBLASTN search
  e-value < 1e-5
  matches extracted along with 1kb flanking sequence and screened against nr database
- Genomic sequences from host genomes that unambigously matched viral proteins were considered as candidate enogenous viral sequences

Methods: avoid contamination

- Rule out chimeric clones or misassembly from contaminated sequences
- Search candidates and flanking sequences against NCBI Trace archive and WGS database (cutoff 95% identity)
- Examine junctions between endogenous viral sequences and cellular sequences
- Validate by PCR

Endogenous viral inserts

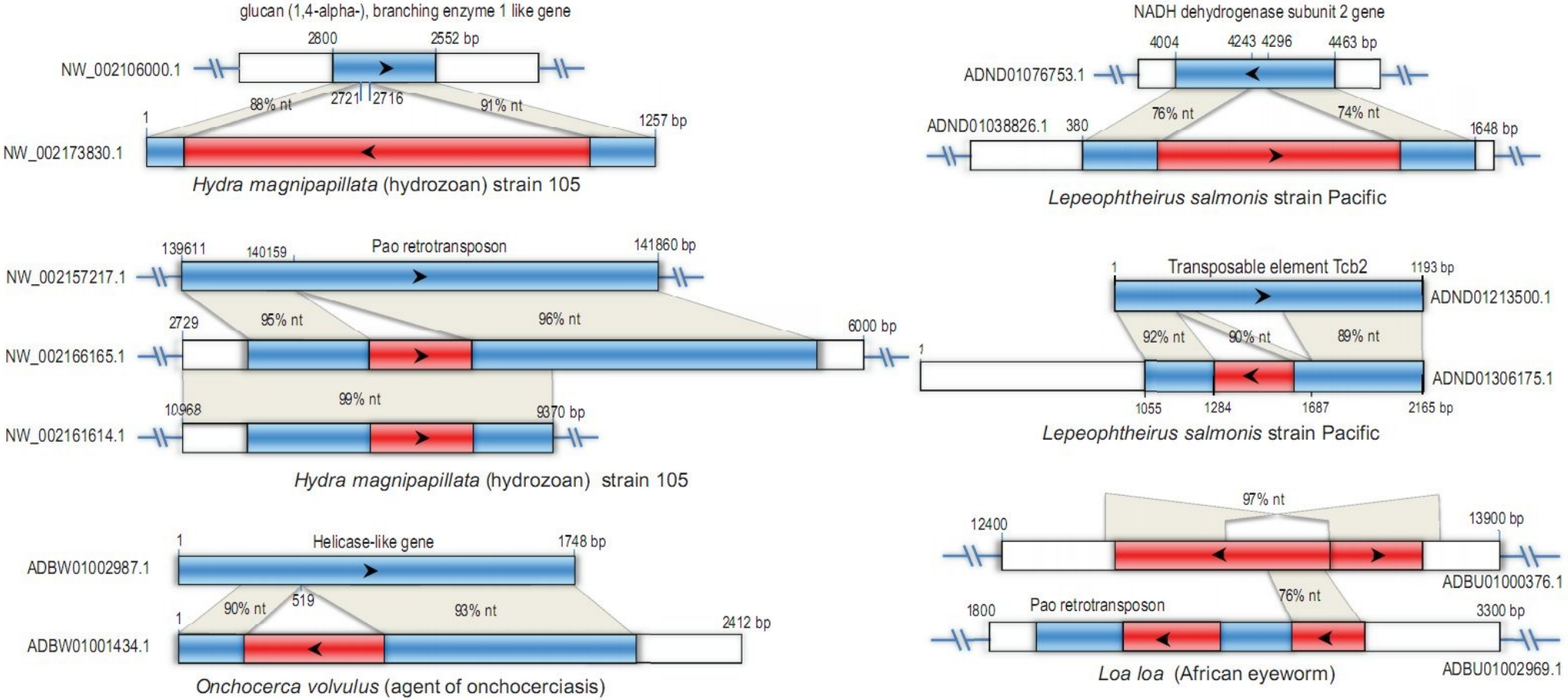


Figure 3 Genomic comparisons showing the endogenous viral sequences inserted into coding regions of host genes. Rectangular boxes with arrowheads indicate genes (Red, viral Rep-like genes; blue, host genes). Gray sectors connect corresponding homologous regions and the % nucleotide (nt) identity scores are indicated.

Liu et al., 2011

Organization and transcription maps of endogenous viral Rep-like proteins

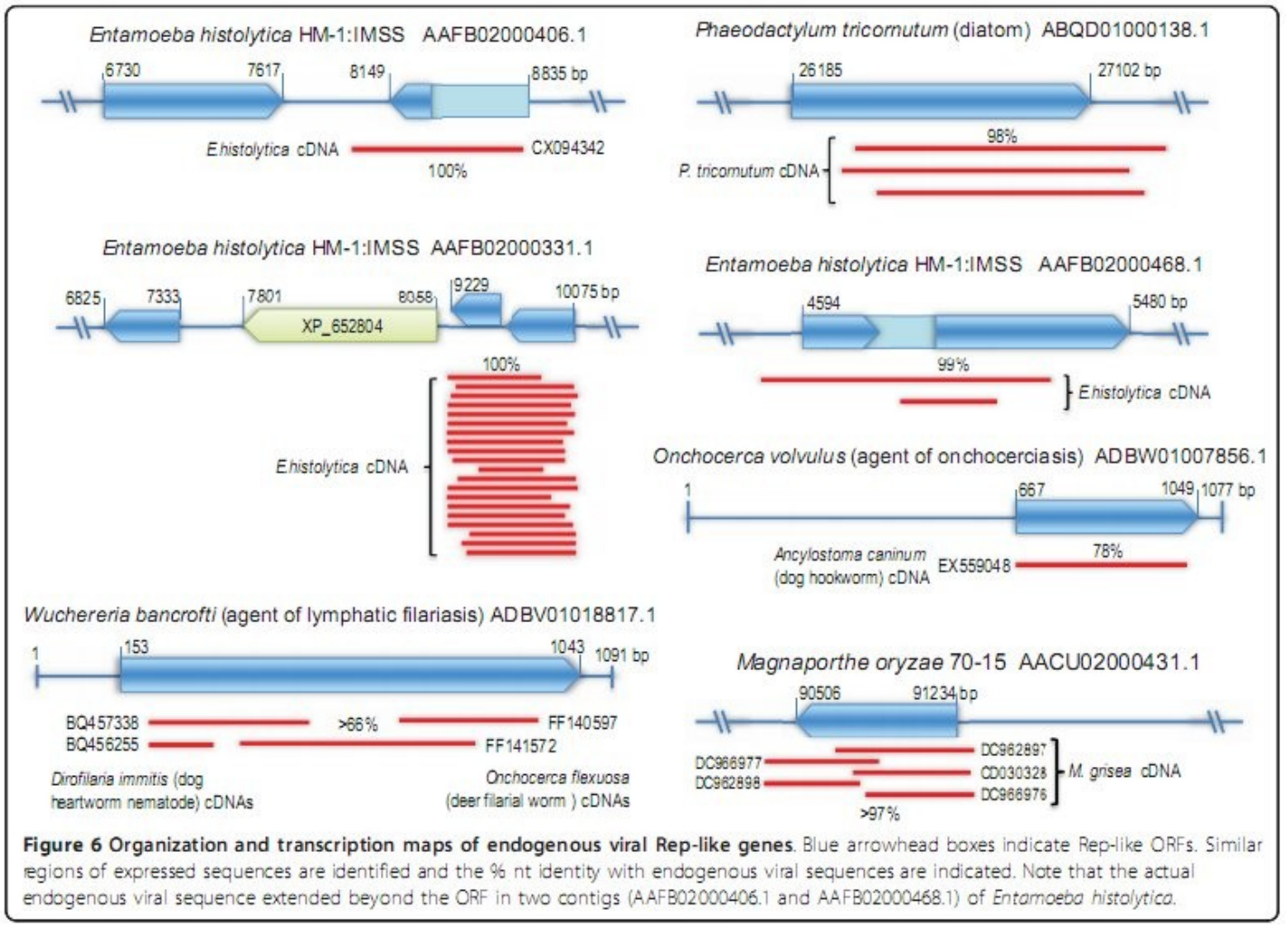


Figure 6 Organization and transcription maps of endogenous viral Rep-like genes. Blue arrowhead boxes indicate Rep-like ORFs. Similar regions of expressed sequences are identified and the % nt identity with endogenous viral sequences are indicated. Note that the actual endogenous viral sequence extended beyond the ORF in two contigs (AAFB02000406.1 and AAFB02000468.1) of *Entamoeba histolytica*.

Liu et al., 2011

Phylogenetic analysis

- Synthenic locations in host genomes
- Indirect age estimation:
- Putative peptides of endogenous viral sequences were obtained according to BLASTX hits and manual editing
- Alignments were compared with a neutral model of genome evolution, numbers of stop codons and frameshifts converted into expected genomic drift

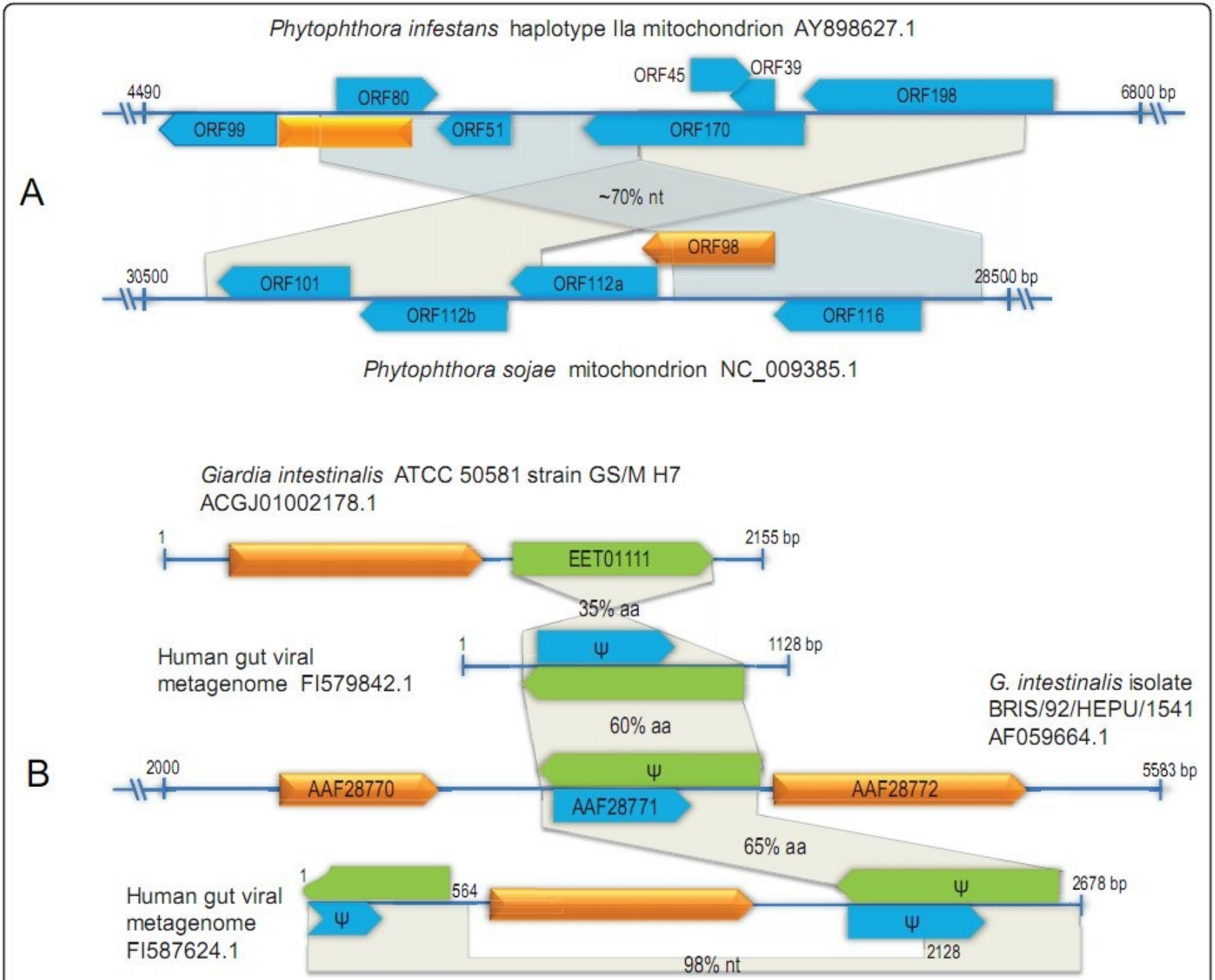
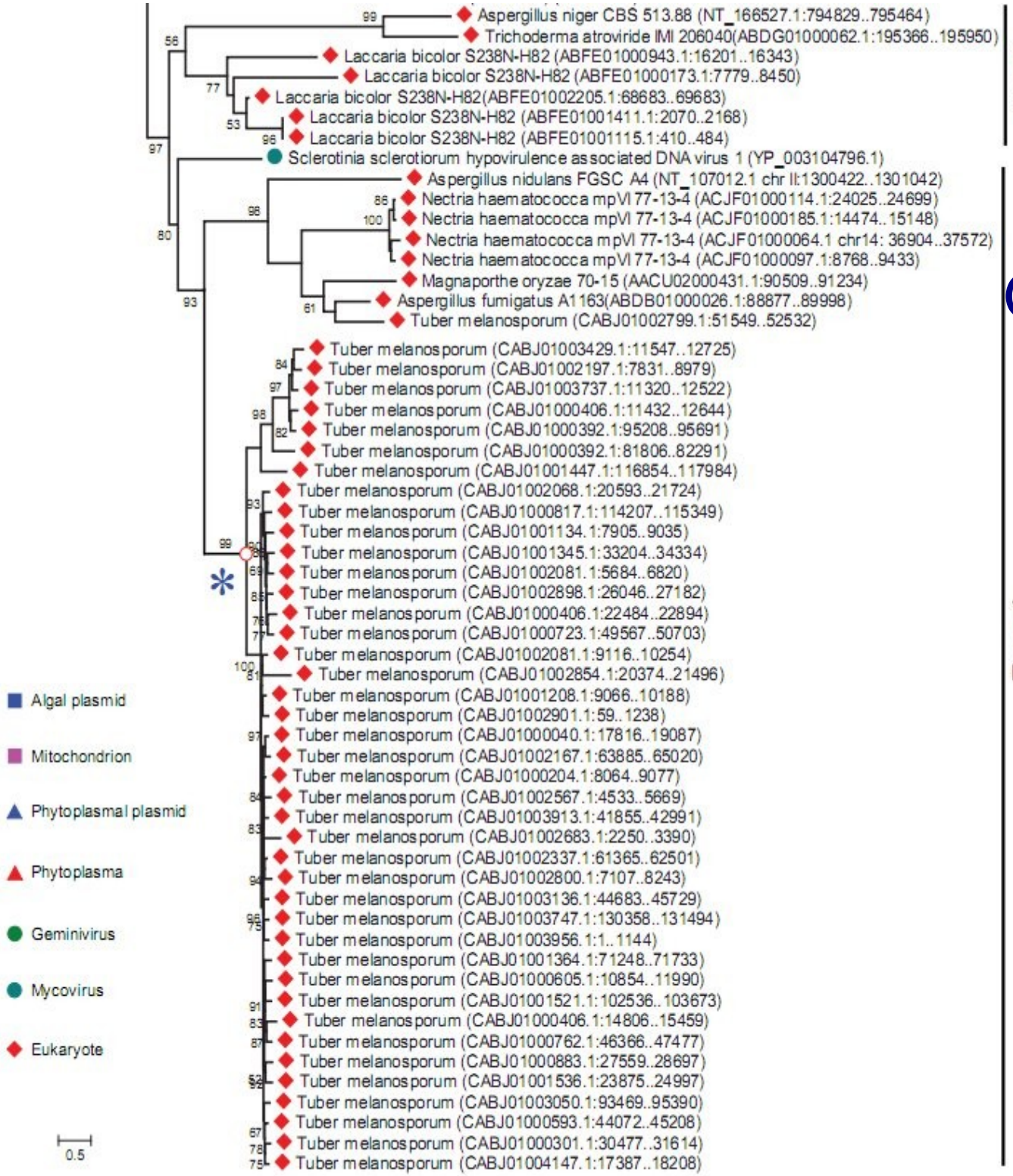


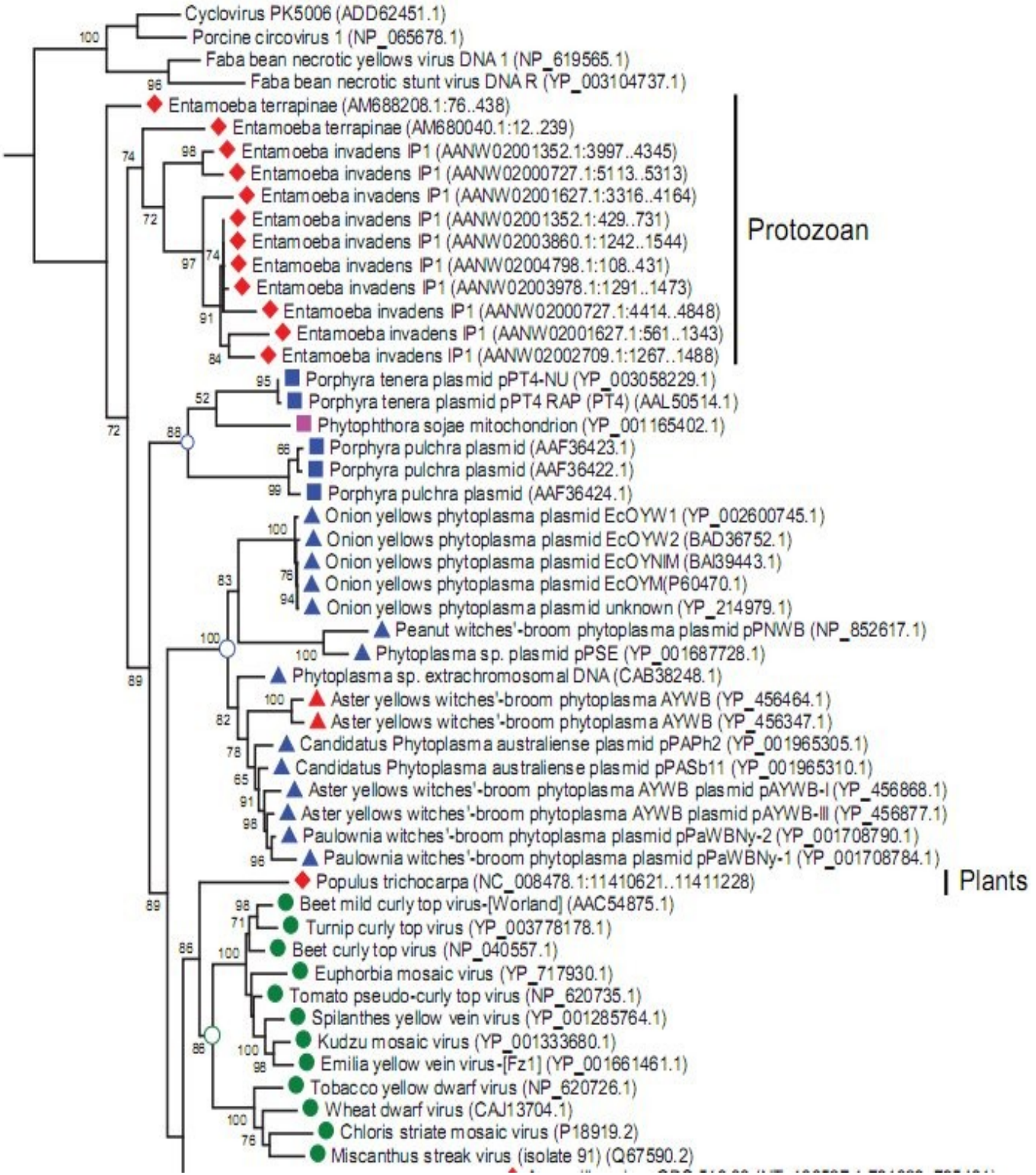
Figure 1 Integrated plasmid or virus-like genes in *Phytophthora* sp. (A) and *Giardia intestinalis* (B). Arrowhead boxes indicate ORFs (orange, Rep-like genes; other colors, unknown genes). Gray sectors connect corresponding homologous regions and the % nucleotide (nt) or amino acid (aa) identity are indicated. The annotated ORF names or accession numbers are indicated. ψ, interrupted ORF.

Liu et al., 2011



Geminiviral rep-like sequences

Liu et al., 2011



Liu et al., 2011

Phylogeny

- Sequences formed three large clades: geminivirus-like, nanovirus-like and circovirus-like
- In each clade endogenous sequences did not fall into established viral families, suggesting that these virus-like sequences may have originated from previously undescribed viral lineages
- There are 42 geminivirus-like Rep genes or remnants interspersed in the genome of Perigord black truffle (Tuber melanosporum), an ectomycorrhizal fungus. All but one are most closely related to each other and formed a distinct clade. They share high (>95%) nucleotide sequence identities with each other and thus allow us to reconstruct a consensus sequence.
- The reconstructed copy contains one interrupted Rep-like open reading frame (ORF), two transposase ORFs (one is interrupted and the other is truncated), and one microsatellite sequence

Fungi and lower eukaryotes

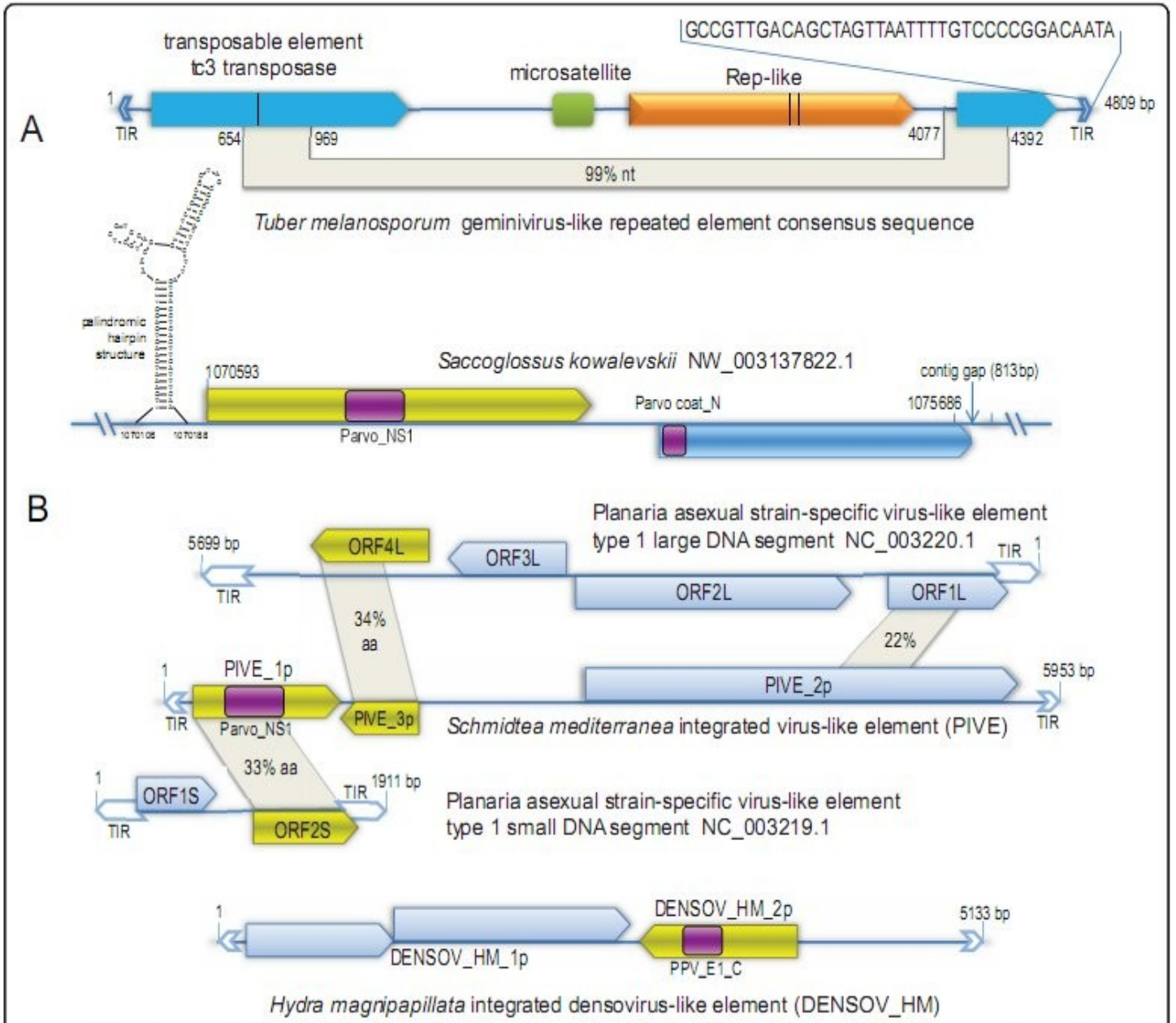
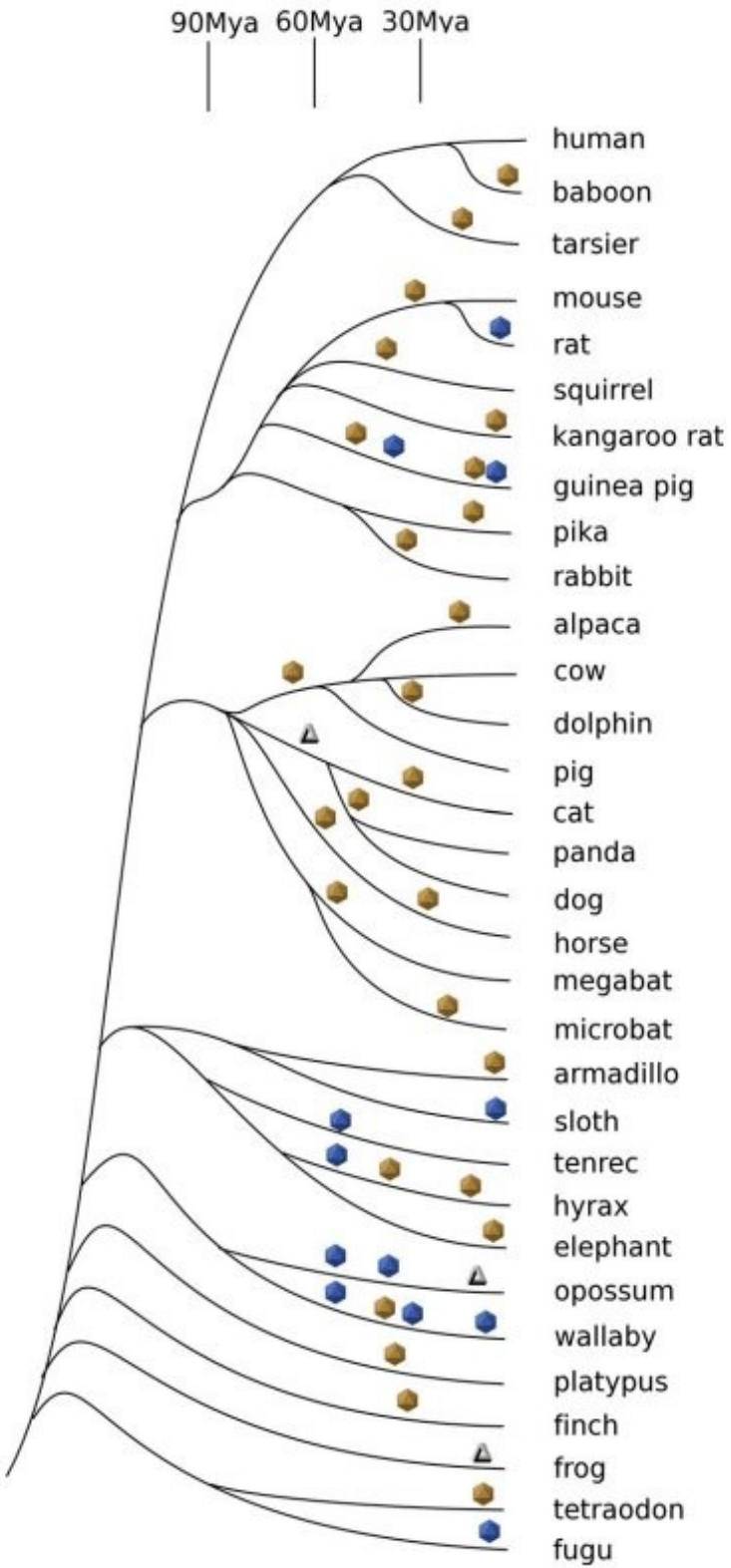


Figure 5 Genomic organization of ssDNA virus-like transposons in fungi (A) and lower eukaryotes (B). (A) The genomic organization of geminivirus-like transposon in *Tuber melanosporum*. Arrowhead boxes indicate ORFs (orange, Rep-like gene; blue, transposase gene). The black vertical lines in the arrowhead boxes indicate stop codons. Green rectangular box indicates microsatellite sequence. The sequence of terminal inverted repeat (TIR) is shown at the top to the right. (B) The genomic organization and comparison of parvovirus-like transposon with related exogenous planaria virus. Yellow arrowhead boxes indicate Rep-like ORFs. Swallow tails indicate terminal inverted repeats (TIRs). The annotated ORF names are indicated. Purple rectangular boxes indicate protein domains and the domain family names are shown: Parvo_NS1, Parvovirus non-structural protein NS1 (pfam01057); Parvo_coat_N, Parvovirus coat protein VP1 (pfam08398); PPV_E1_C, Papillomavirus helicase (pfam00519). Gray sectors connect corresponding homologous regions and the % nucleotide (nt) or amino acid (aa) identity are indicated. The Planaria asexual strain-specific virus-like element has not been found to integrate in the host genome.

Liu et al., 2011

Phylogenetic tree of vertebrate organisms and history of ssDNA virus integrations



Times of integration of ancestral dependoviruses (yellow icosahedrons), parvoviruses (blue icosahedrons), and circoviruses (triangles) are approximate.

Belyi et al., 2010

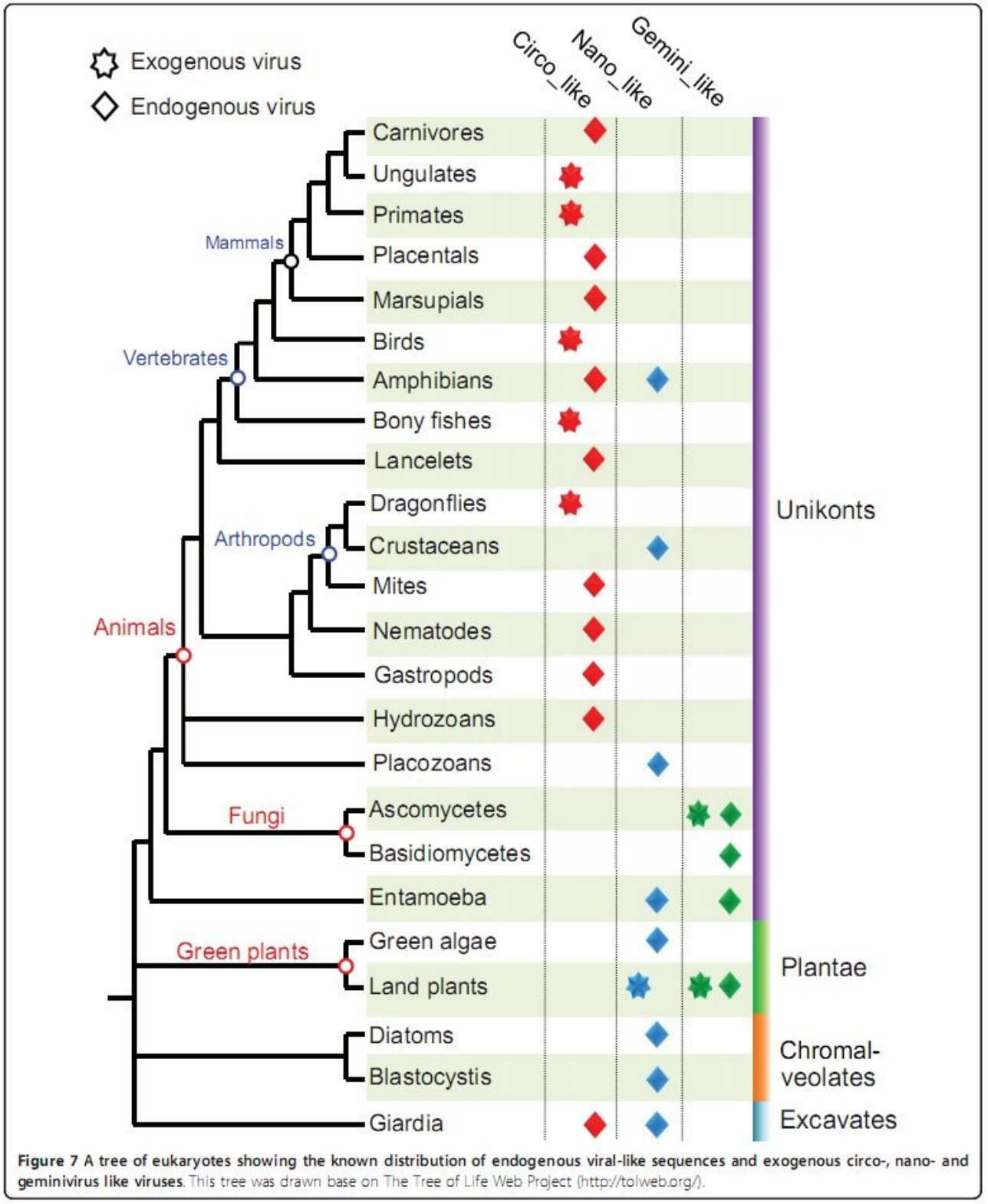


Figure 7 A tree of eukaryotes showing the known distribution of endogenous viral-like sequences and exogenous circo-, nano- and geminivirus like viruses. This tree was drawn base on The Tree of Life Web Project (http://tolweb.org/).

Distribution of endogenous viral-like sequences and exogenous circo-, nano- and geminivirus like viruses

Liu et al., 2011

Table 1 Numbers of endogenous circular ssDNA virus-like sequences in eukaryotic genomes

Organism group	Organism	No. of virus-related genes	
		Rep	Capsid
Plants			
land plants	*Populus trichocarpa* (black cottonwood)	1	
	Nicotiana tabacum (common tobacco)		1
green algae	*Micromonas pusilla* (green algae) CCMP1545	1	
Fungi			
ascomycetes	*Aspergillus nidulans* FGSC A4	1	
	Aspergillus fumigatus A1163	1	
	Aspergillus niger CBS 513.88	1	
	Trichoderma atroviride IMI 206040	1	
	Magnaporthe oryzae 70-15 (rice blast fungus)	1	
	Nectria haematococca mpVI 77-13-4	4	
	Tuber melanosporum Mel28 (Perigord truffle)	42	
basidiomycetes	*Laccaria bicolor* S238N-H82 (Bicoloured deceiver)	5	
Protists			
protozoans	*Entamoeba invadens* IP1	10	
	Entamoeba terrapinae	3	
	Entamoeba histolytica HM-1:IMSS	14	
	Entamoeba dispar SAW760	7	
	Blastocystis hominis Singapore isolate B (sub-type 7)	7	
	Giardia intestinalis ATCC 50581 strain GS/M H7	13	
	Giardia intestinalis isolate BRIS/92/HEPU/1541	2	
diatoms	*Phaeodactylum tricornutum* (diatom)	1	
Animals			
mammals	*Canis lupus familiaris* (dog) *	4	
	Monodelphis domestica (gray short-tailed opossum) *	1	
	Felis catus(domestic cat) *	6	
	Ailuropoda melanoleuca (giant panda) *	12	
	Choloepus hoffmanni (Hoffmann's two-fingered sloth) *		2†
gastropods	*Aplysia californica* (California sea hare)	1	
amphibians	*Xenopus (Silurana) tropicalis*(western clawed frog) *	2	
lancelets	*Branchiostoma floridae* (Florida lancelet) strain S238N-H82	7	
roundworms	*Brugia malayi* (agent of lymphatic filariasis)	1	
	Loa loa (African eyeworm)	10	
	Wuchereria bancrofti (agent of lymphatic filariasis)	3	
	Onchocerca volvulus (agent of onchocerciasis)	5	
crustaceans	*Lepeophtheirus salmonis* (salmon louse) strain Pacific	59	
mites & ticks	*Varroa destructor* (honeybee mite)strain Korean	56	
placozoans	*Trichoplax adhaerens* (placozoan) strain Grell-BS-1999	2	
hydrozoans	*Hydra magnipapillata* (hydrozoan) strain 105	18	
Total	35	302	3

Liu et al., 2011

Results

- Belyi et al, 2010:
  as many as 110 ssDNA virus-related sequences have been integrated into 49 vertebrate genomes during time period ranging from present to over 40 Mya
- Liu et al., 2011:
  discovered enogenous virus-like sequences in at least 35 species among plants, fungi, animals and protists
- Most circovirus-like elements consist of only rep-like sequences
  In the opossum genome cap sequence is adjacent to rep, which suggests that it was derived from viral genomic DNA.

Conclusions

- Genes of small ss DNA viruses have been transferred to broad range of vertebrate genomes
- Some of transferred genes are conserved and functional in host genomes
- Capture and functional assimilation of exogenous viral genes may represent and important force in the evolution of eukaryotes
- Remains unclear why these elements reached fixation (incidental or provided advantages to host?)
- There is an intrinsic limit on how far back in time we can reach to identify ancient endogenous viral sequences