

Challenges in classifying newly discovered viruses (cf. giant viruses)

Jean-Michel.Claverie@univ-amu.fr

Structural & Genomic Information Laboratory (IGS)
Mediterranean Institute of Microbiology (IMM)

CNRS - Aix-Marseille University

3 short stories

- 1- Handling the unknown (dark matter)
- 2- Issues with attempted deep taxonomy
- 3- Dubious classifications in Phycodnaviridae & Mimiviridae

Different sizes, morphologies, genomes

Viruses have nothing in common, except the way they propagate their genomes

Megavirus chilensis

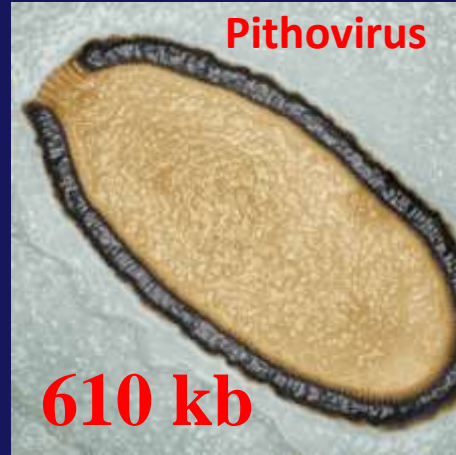


1,2 Mb



Pandoravirus

2,8 Mb



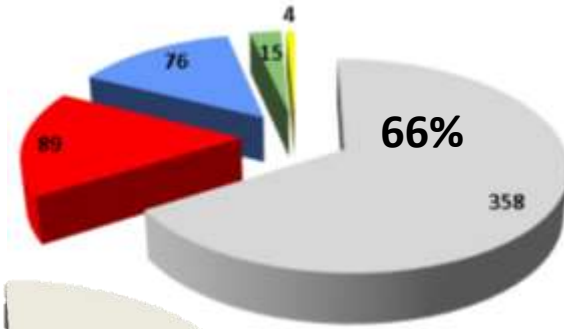
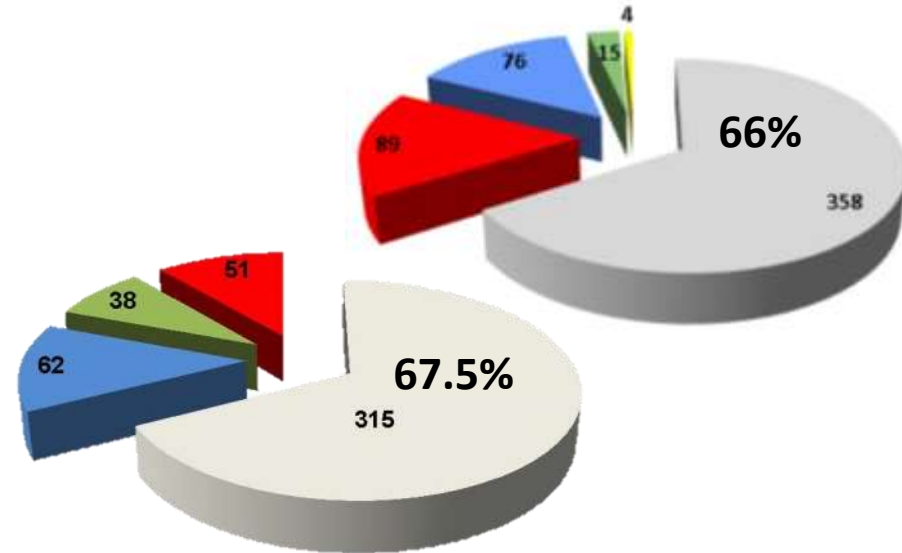
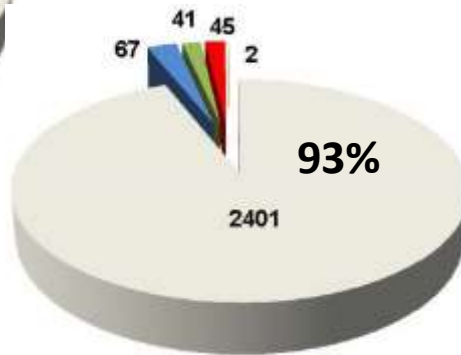
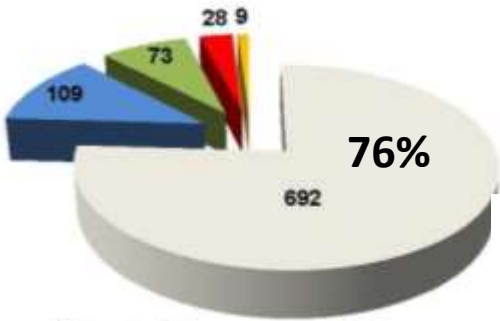
Pithovirus

610 kb



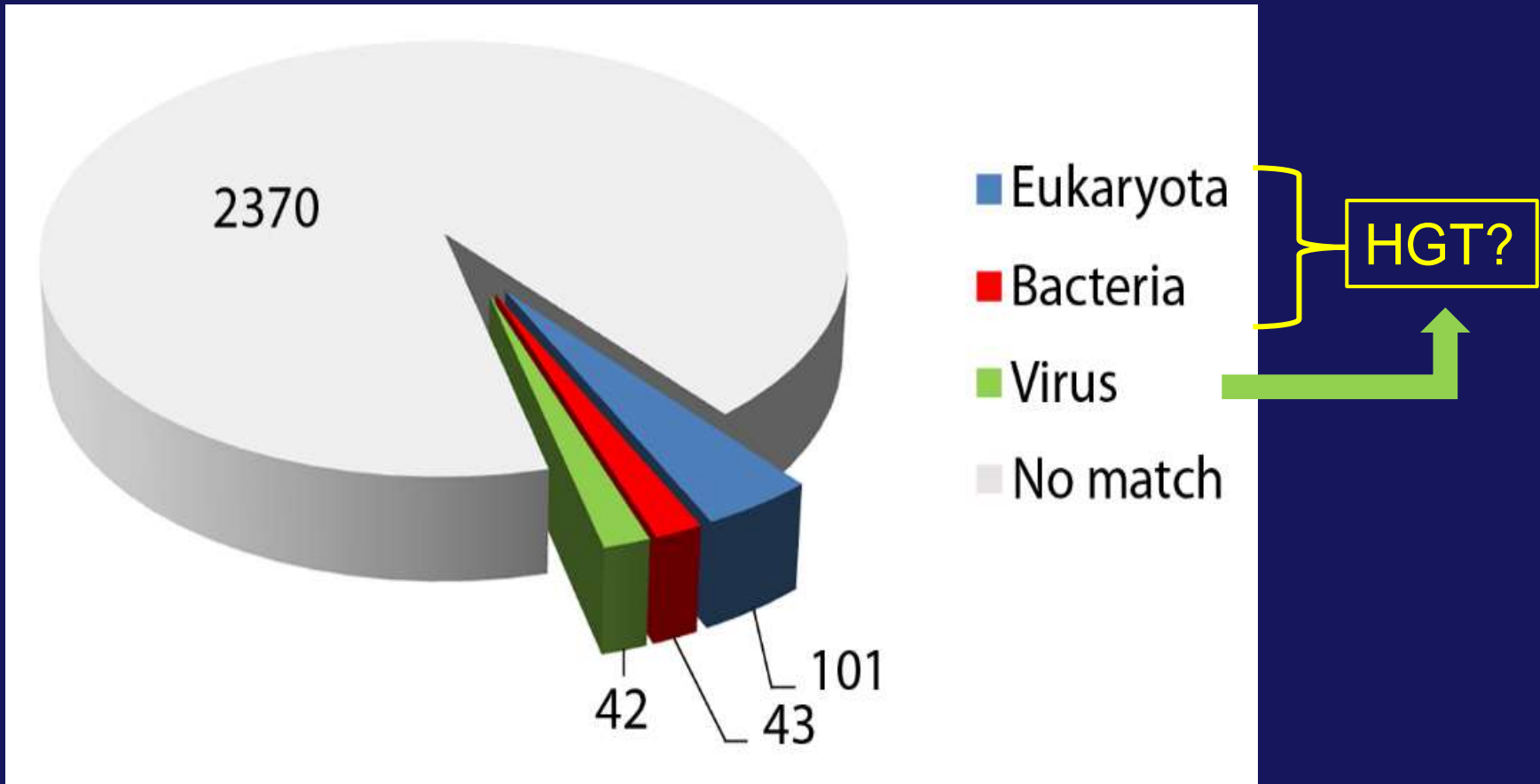
Mollivirus

650 kb



For each new member >2/3 ORFans

93% ORFans: guess what that is !



Going back to the basics: A. Lwoff (how to discriminate cellular organisms from viruses)

Viruses are defined by negative properties:

- 1) Not visible by light microscopy
- 2) Not retained by the Chamberland filter
- 3) Not cultivable
- 4) No energy production
- 5) No translation (no ribosome)
- 6) No division

Lwoff A. (1957). The concept of virus. *Journal General Microbiology* 17, 239–253.

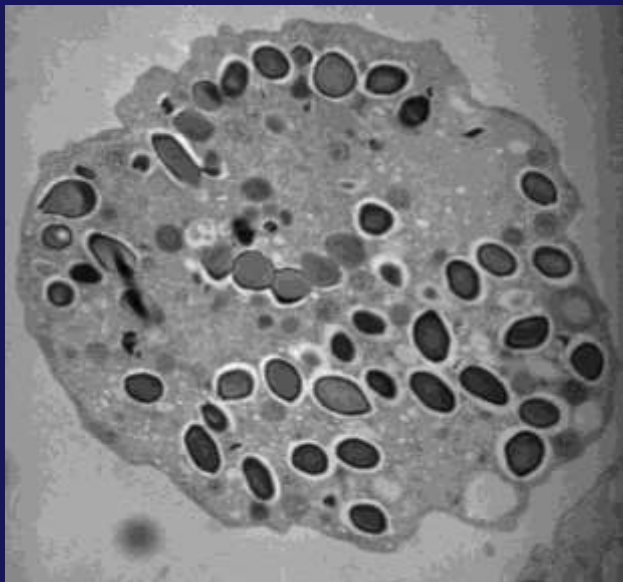
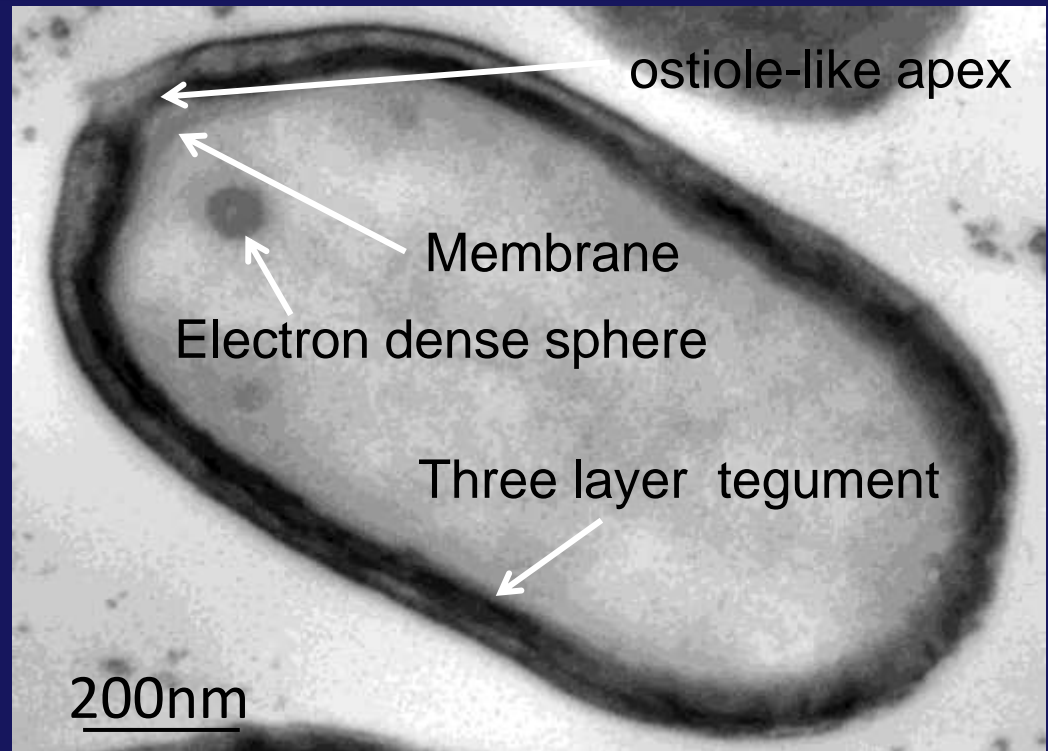
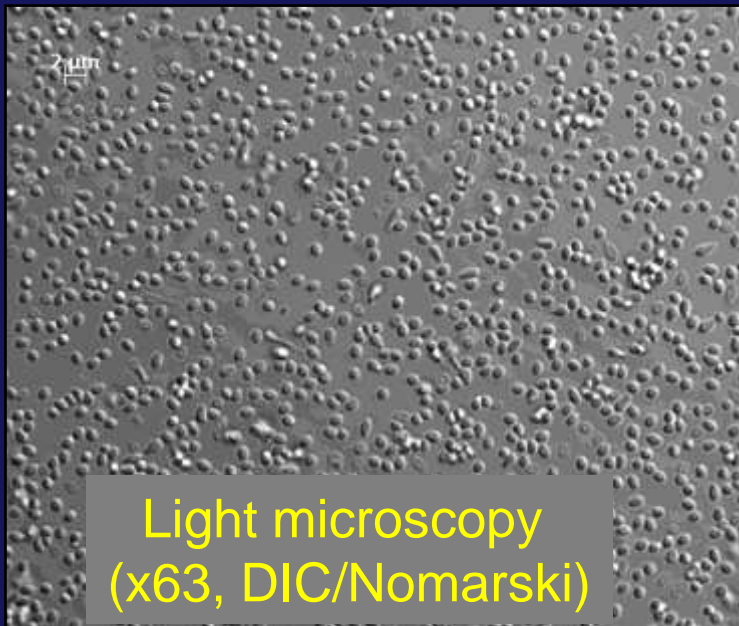
Lwoff A, Tournier P (1966). The classification of viruses. *Annual Reviews Microbiology* 20, 45–74.

The (formally) required experimental evidence (how to discriminate cellular organisms from viruses)

Viruses are defined by negative properties:

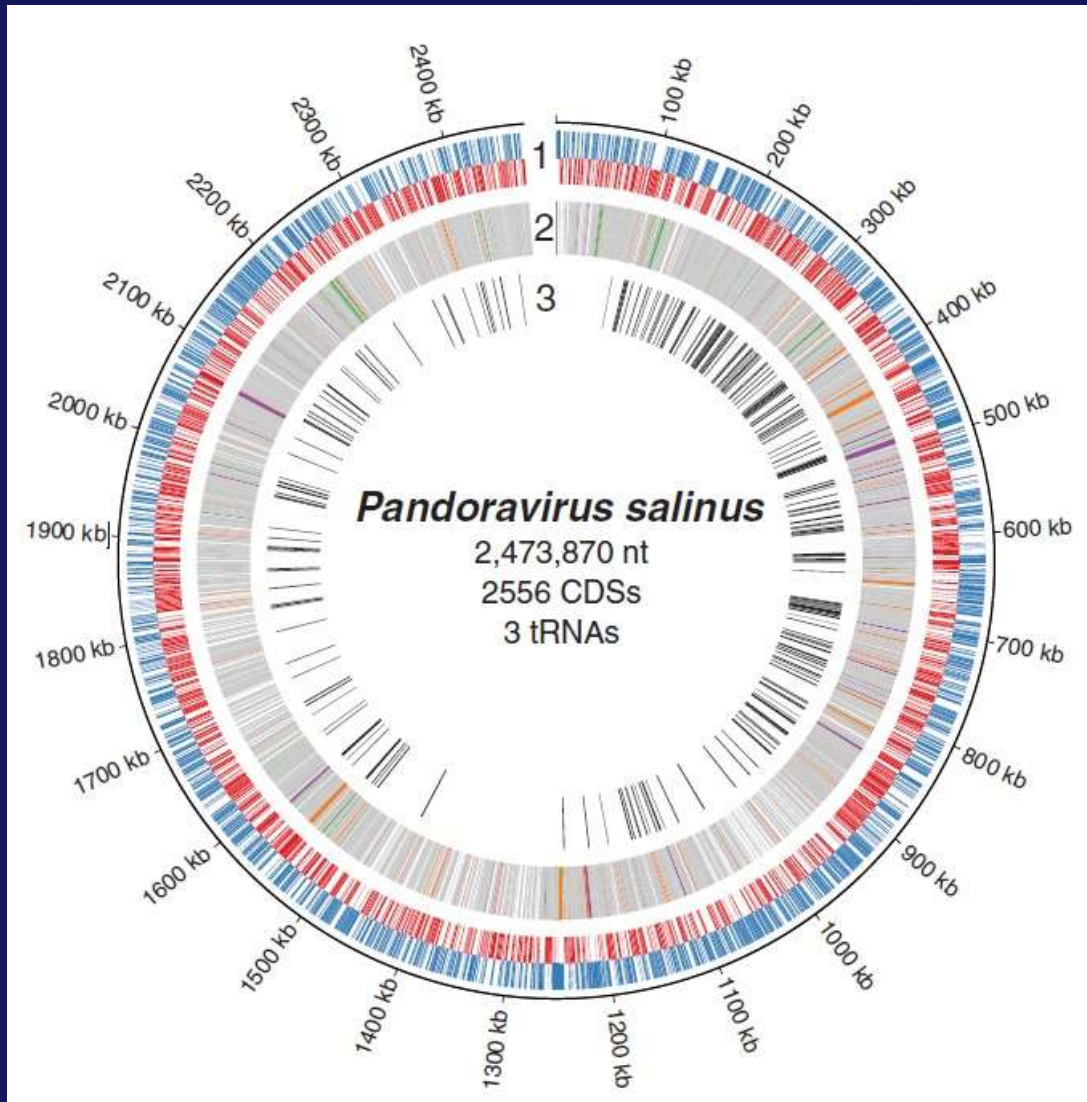
- ~~1) Not visible by light microscopy~~
- ~~2) Not retained by the Chamberland filter~~
- 3) Not cultivable (cell dependent) → observation
- 4) No energy production → whole genome
- 5) No translation (no ribosome) → whole genome
- 6) No division → observation

Documenting a new « life form »



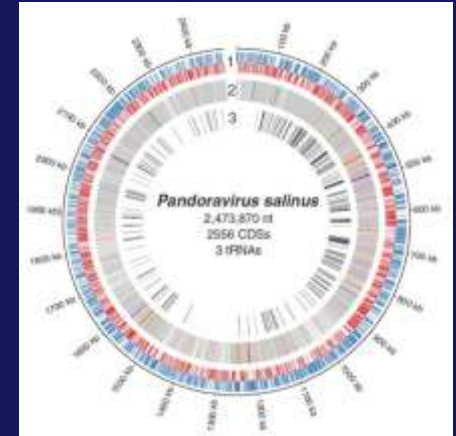
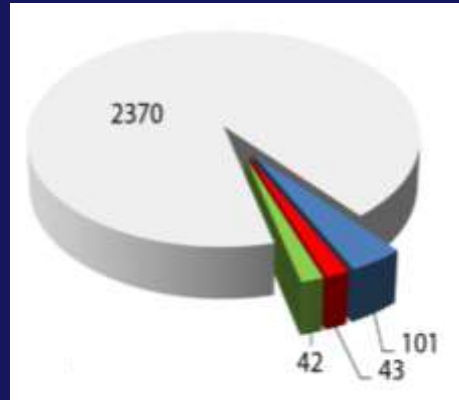
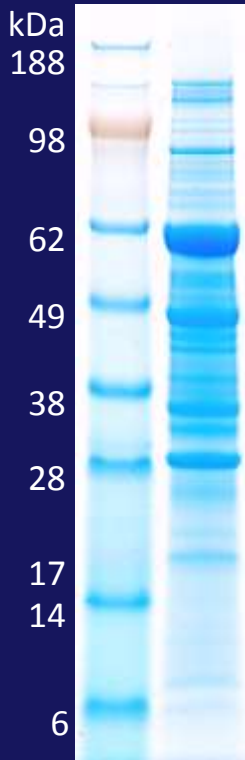
← Intracellular (*Acanthamoeba*)
replication

Pandoravirus salinus: 2.8 Mb, 62%GC, 2556 CDS, 3 tRNA



- No ribosomal protein
- No division apparatus (FtsZ)
- No ATP production pathways
- This must be a **virus**
- **But:** No trace of Major Capsid Protein

Genetic code? -> Proteomic validation



~ 200 predicted proteins found in the particle
83% of them have no database homolog !

-> Pandoraviruses use the standard genetic code

Particle proteomics

MS-MS spectrometry (Y. Couté, C. Bruley, J. Garin, Grenoble)

1- Conclusion

1- « If the material is not available, this is not Science »
(*dixit* George Garrity)

2- Nomenclature rules should hold to the challenge of future totally unexpected discoveries (rigorous AND flexible)

3- Criteria, methods, and level of required evidence probably cannot be the same for all virus families

2- Issues with deep taxonomy attempts

Yutin and Koonin *Biology Direct* 2013, **8**:25
<http://www.biologydirect.com/content/8/1/25>



BIOLOGY DIRECT

DISCOVERY NOTES

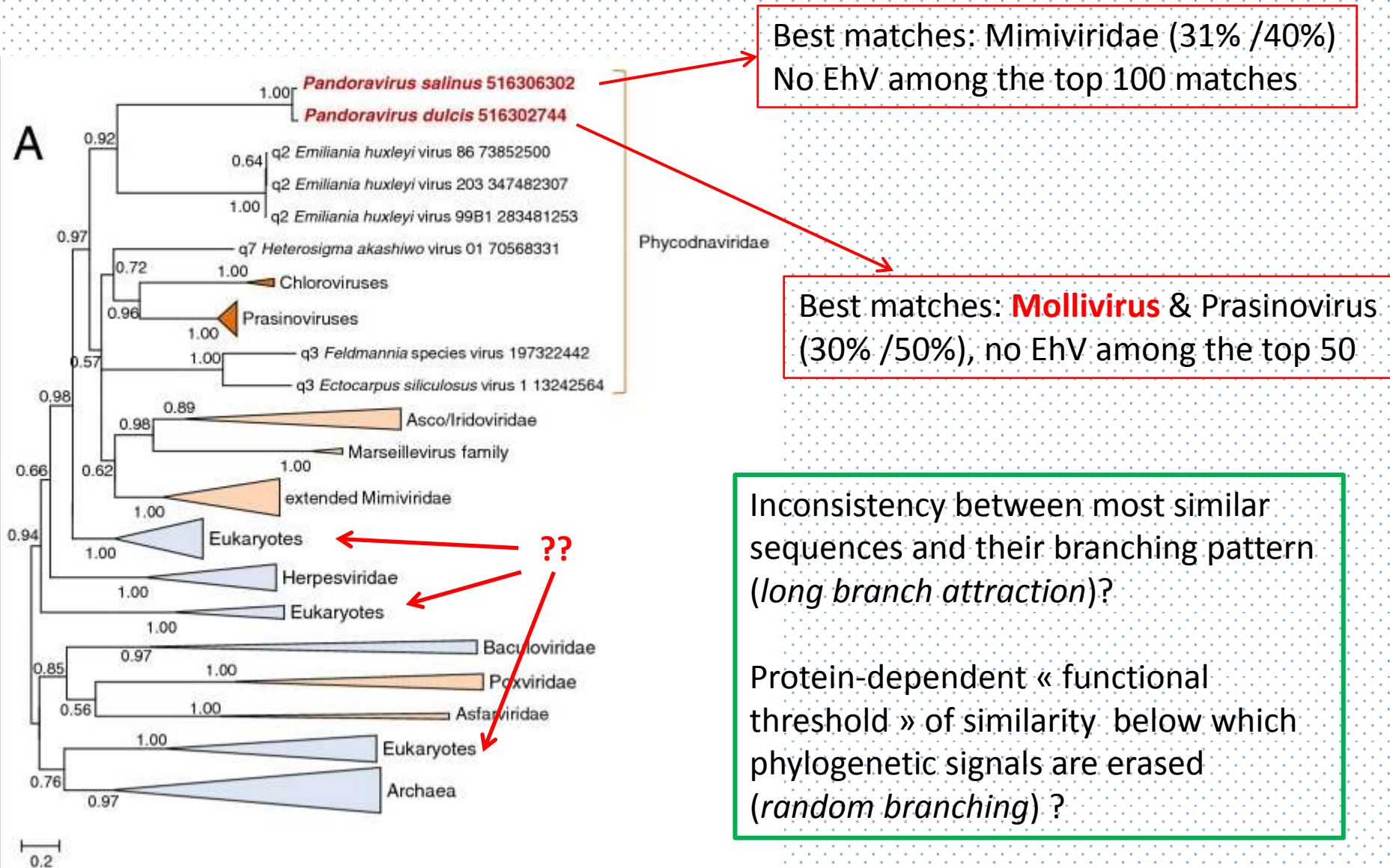
Open Access

Pandoraviruses are highly derived phycodnaviruses

Natalya Yutin and Eugene V Koonin*

Are trees reliable? (below 30% ID)

DNA polymerase



Initial results (2013)

- 16 other « NCLDV core genes » homologs
 - 11 best matches are in cellular organisms ?
(strange for « viral » core genes)
 - 5 viral matches: in Prasinoviruses (2), Phaeovirus (1), Coccolithovirus (1) , and Marseillevirus (1)
- Classifying Pandoravirus on the basis of 5 remote homologs of « viral core genes » (over >2500 ORFs) ?

Before / After

Table 1 The ancestral NCLDV genes represented in Pandoraviruses

Gene/NCVOG	<i>P. dulcis</i> genes	<i>P. salinus</i> genes	Presence in the 7 NCLDV families ^a	Best hit for Pandoraviruses ^b / % identity/alignment length
D5-like helicase-primase/NCVOG0023	516302795	516304338	7	<i>Bathycoccus</i> sp. RCC1105 virus BpV2 (Phycodnaviridae) /33/579

516304338: 37% ID over 60%, score [314-306]: **Yellow Stone LV**, Bathicoccus BpV2V

DNA or RNA helicases of superfamily II (COG1061) (A18hel)/NCVOG0076	516302732	516304266	5	<i>Ectocarpus siliculosus</i> virus 1 (Phycodnaviridae) /35/238
---	-----------	------------------	---	---

516304266: 35% ID over 28%, score [164-136]: **Mollivirus**, Esv-1

A32-like packaging ATPase/NCVOG0249	516302762, 516303626	516306303 , 516303793, 516305958, 516303807, 516305953	7	<i>Ostreococcus tauri</i> virus 2 (Phycodnaviridae) /45/247
-------------------------------------	----------------------	---	---	---

516306303: 44% ID over 55% , score [224-211]: **Yellow Stone LV**, all prasinoviruses
 516303793: 31% ID over 70%, score [117-106]: **Yellow Stone LV** , small Megaviridae
 516305958: 38% ID over 59%, score [185-171]: **Yellow Stone LV** , all prasinoviruses
 516303807: 29% ID over 67%, score [117-102]: **Yellow Stone LV** , all prasinoviruses
 516305953: no match, no A32-like domain ?

Before / After

Table 1 The ancestral NCLDV genes represented in Pandoraviruses

Gene/NCVOG	<i>P. dulcis</i> genes	<i>P. salinus</i> genes	Presence in the 7 NCLDV families ^a	Best hit for Pandoraviruses ^b / % identity/alignment length
pfam04947, Poxvirus Late Transcription Factor VLTF3 like (A2L)/NCVOG0262	516302769, 516303263	516304304 , 516305311	7	<i>Emiliana huxleyi</i> virus 202 (Phycodnaviridae) /34/264

516304304: 35% ID over 33% , score [134-125]: EhV202, **Mollivirus**, EsV-1, Chlorovirus, no other EhV
 516305311: 35% ID over 53%, score [102-99]: ACTV_Br0604L, **Guillardia Theta**, Chloroviruses

cd00127, DSPc, Dual specificity phosphatases (DSP); Ser/Thr and Tyr protein phosphatases/NCVOG0040	516303124, 516303141	516304931 , 516304951	3	<i>Lausannevirus</i> (Marseillevirus family) /41/149
--	----------------------	------------------------------	---	--

516304931: 43% ID over 80% , score [120-111]: **Mollivirus**, Marseilleviridae
 516304951: 51% ID over 84%, score [127-108: **Mollivirus**, Marseilleviridae

5 viral matches: **Yellow Stone Virus (2)**, ~~Prasinoviruses (2)~~, **Mollivirus (2)**, ~~Phaeovirus (1)~~, ~~Marseillevirus (1)~~, ~~Coccolithovirus (1)~~

All these « best-matching » cases are fluctuating, borderline, and compatible with

- old HGTs from ancestors of known viruses
- recent HGTs from unknown viruses from known families
- all scenarios in between

RNA polymerases ?

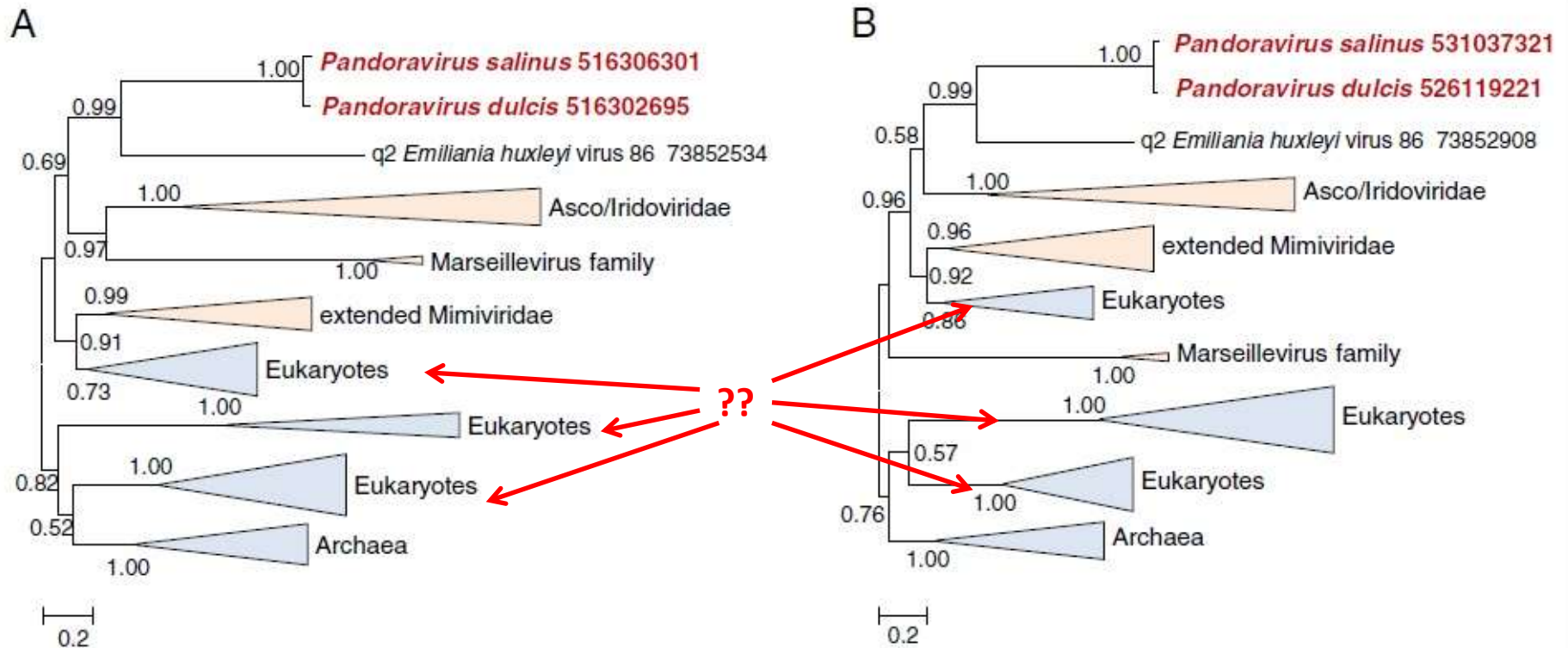


Figure 3 Maximum-Likelihood trees of DNA-directed RNA polymerase. **A**, alpha subunit. **B**, beta subunit. The designations are as in Figure 2.

516306301 (RPB1) : 37% ID over 26%, score [323-224]: various Fungi, Eukaryota, No virus

531037321 (RPB2): 37% ID over 85%, score [718-467] *Encephalitozoon cuniculi*, Eukaryota, No virus

Pandoraviruses are ~~most likely~~ unrelated to « phycodnaviruses »

This paper has the merit of raising 4 essential questions:

- 1) The danger of classifying new viruses on the basis of a predetermined reference gene set
 - « *ad hoc* » selection of genes ($C_{(5, 40)} = 6.58 \cdot 10^5$)
- 2) Status of homologous/orthologous genes unclear (domain sharing, random matching, HGT)
 - *HGT and non-orthologous replacements are impossible to dismiss*
- 3) First members of new families might be at risk of being classified on the basis of the minority of genes acquired by HGT (cladistics!)
- 4) Lack of clear rules by which to classify viruses in existing families (groups) : total % shared gene, similarity threshold among a family-based reference gene set, virion morphology, replication scenario, host, disease type, etc.

Additional remarkss

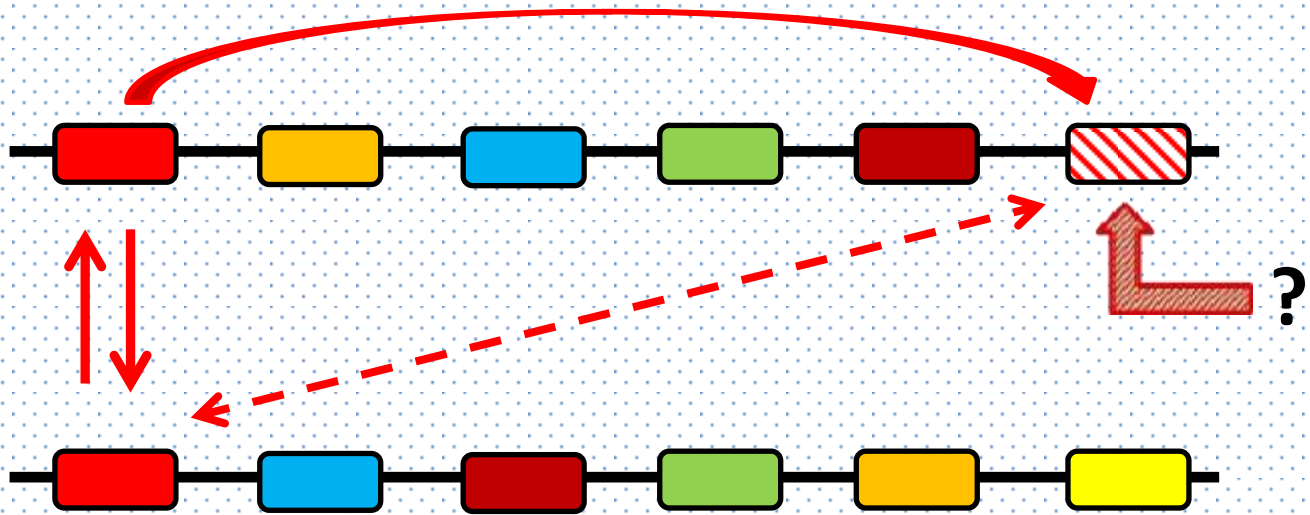
- «Core» genes are not «sacred» one-copy genes:
duplicated RPB, Packaging ATPases, etc ...
facilitating «core» gene exchanges
- Homology to « ancestral » core genes does not imply that they are « ancestral » in a given virus genome
- Significant similarity can be reached by chance (+ Bonferroni correction)

Known duplications of NCLDV «core genes»

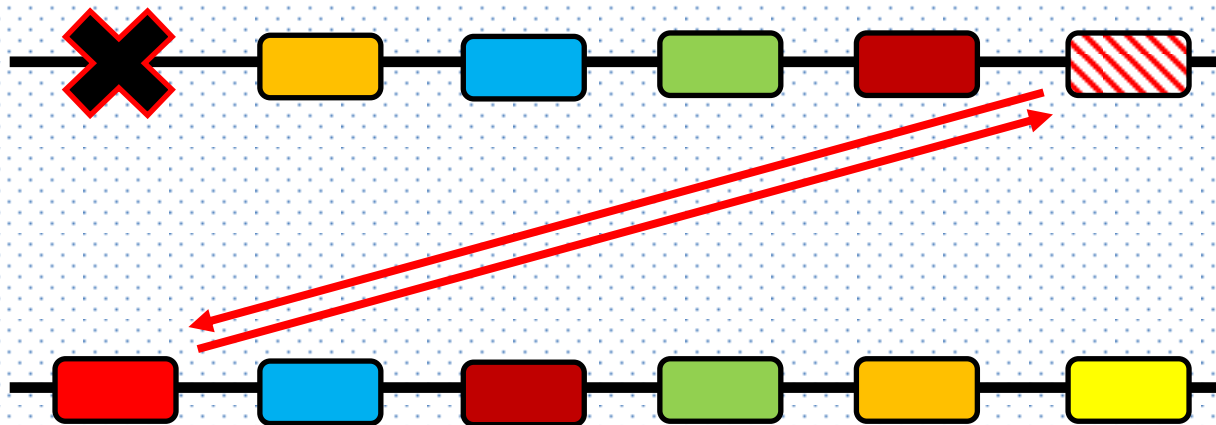
- DNA-dependent RNA Pol second largest subunit (Rpb2):
 - PgV, CeV, OLPV, AaV
- Packaging ATPase (VV32-like) :
 - PgV, OLPV
- Ribonucleotide Reductase (small sub.):
 - CeV
- DNA-dependent RNA Pol largest subunit (Rpb1):
 - AaV (AaV_242, AaV_320) (Aureococcus anophagefferens virus)

Proper and misleading use of the reciprocal best match rule (RBM)

A
Ortholog
+ Paralog



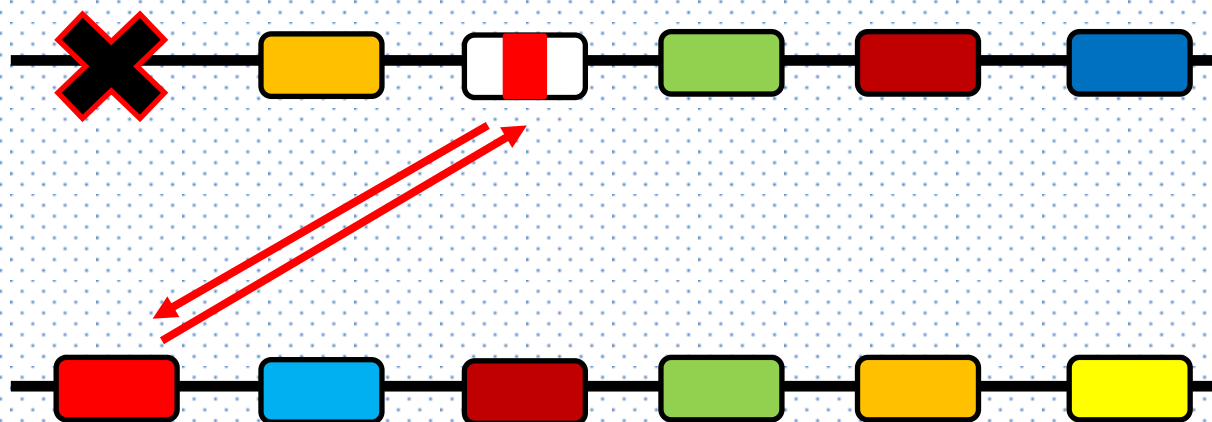
B
False Ortholog



Misleading use of the reciprocal best match rule (RBM)

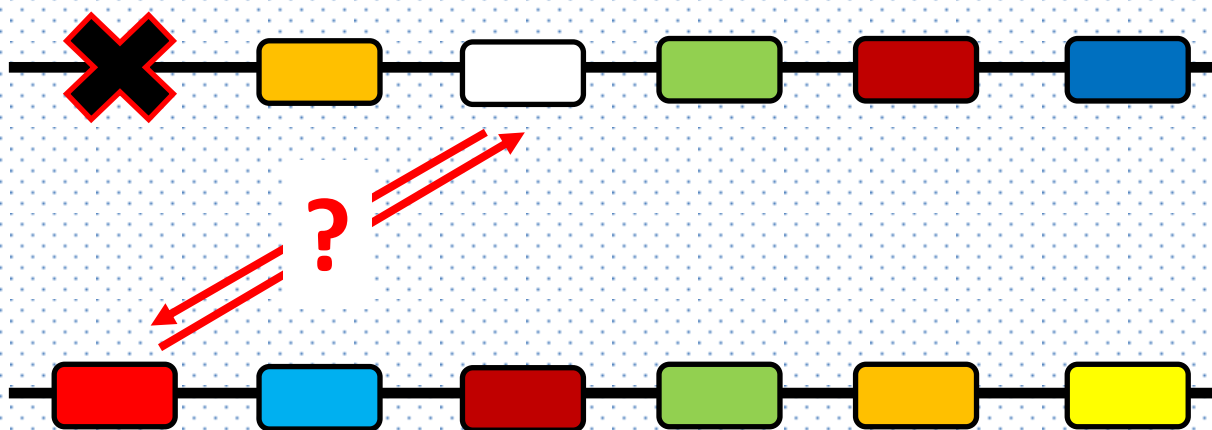
C

Domain-induced
« orthology »

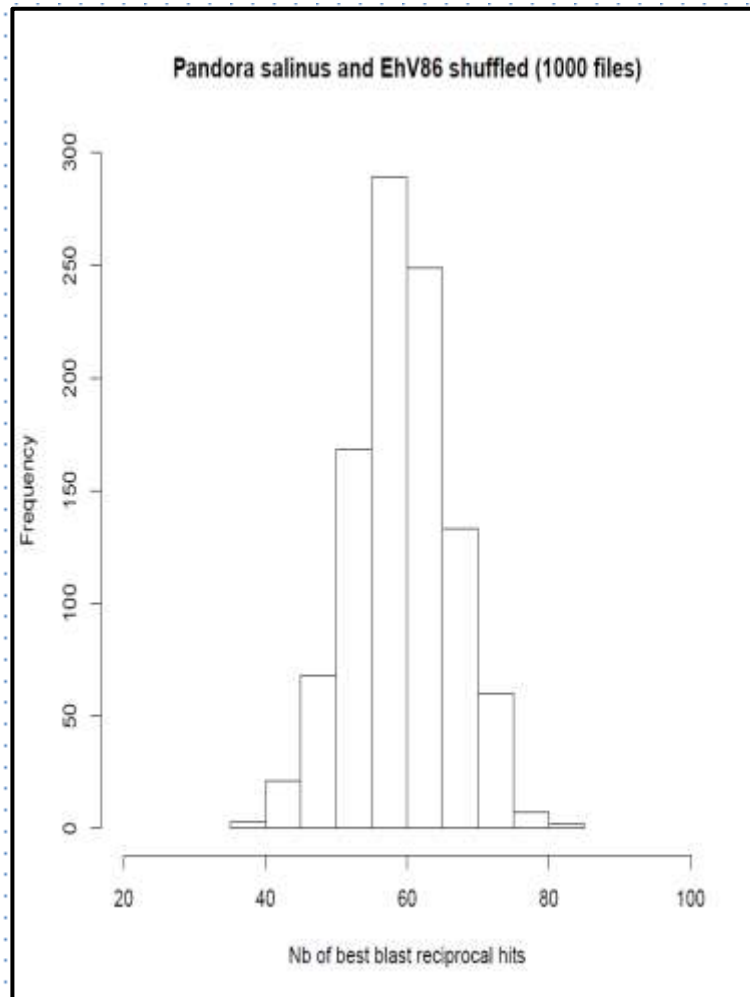


D

Random match



Statistics of best BLAST hits between *P. salinus* and shuffled EhV86



$36 < \text{hits (RBM)} < 84$
For each individual run !

3- Dubious classification in «Phycodnaviridae» & Mega/Mimiviridae

Family level: present status

(recognized/listed by ICTV)

- *Phycodnaviridae*
 - *Chlorovirus* (*PBCV-1*, 1995)
 - *Coccolithovirus* (*EhV86*, 2005)
 - *Phaeovirus* (*EsV-1*, 2001)
 - *Prasinovirus* (*MpV-1*, 2010)
 - *Prymnesiovirus* (*PgV 16T*, 2013)
 - *Raphidovirus* (3 genes, *Heterosigma akashiwo virus 1*)
- *Mimiviridae*
 - *Cafeteriavirus* (*CroV*, 2010)
 - *Mimivirus* (*APMV*, 2003)

Phycodnaviridae: present status at NCBI

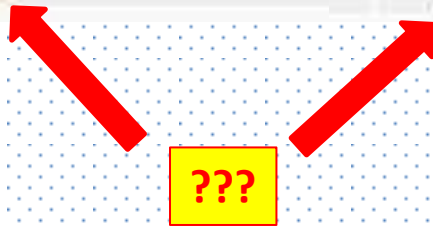
Viral complete genome browser

Phycodnaviridae			
Acanthocystis turfacea Chlorella virus 1	288047 nt	NC_008724	
Aureococcus anophagefferens virus	370920 nt	NC_024697	← Mega/Mimiviridae
Bathycoccus sp. RCC1105 virus BpV1	198519 nt	NC_014765	
Chrysochromulina ericina virus	473558 nt	NC_028094	← Mega/Mimiviridae
Ectocarpus siliculosus virus 1	335593 nt	NC_002687	
Emiliana huxleyi virus 86	407339 nt	NC_007346	
Feldmannia species virus	154641 nt	NC_011183	
Micromonas sp. RCC1109 virus MpV1	184095 nt	NC_014767	
Ostreococcus lucimarinus virus 1	194022 nt	NC_014766	← Now Haptolina!
Ostreococcus lucimarinus virus 2	196300 nt	NC_028091	
Ostreococcus lucimarinus virus 7	182309 nt	NC_028093	
Ostreococcus mediterraneus virus 1	193301 nt	NC_028092	
Ostreococcus tauri virus 1	191761 nt	NC_013288	
Ostreococcus tauri virus 2	184409 nt	NC_014789	
Ostreococcus virus OsV5	185373 nt	NC_010191	
Paramecium bursaria Chlorella virus 1	330611 nt	NC_000852	
Paramecium bursaria Chlorella virus AR158	344691 nt	NC_009899	
Paramecium bursaria Chlorella virus FR483	321240 nt	NC_008603	
Paramecium bursaria Chlorella virus NY2A	368683 nt	NC_009898	
Phaeocystis globosa virus	459984 nt	NC_021312	← Mega/Mimiviridae
Yellowstone lake phycodnavirus 1	178262 nt	NC_028112	} Metagenomics
Yellowstone lake phycodnavirus 2	171045 nt	NC_028110	
Yellowstone lake phycodnavirus 3	171454 nt	NC_028108	

Mimiviridae: present status at NCBI *Viral complete genome browser*

Mimiviridae

Acanthamoeba polyphaga mimivirus	1181549 nt	NC_014649
Acanthamoeba polyphaga moumouvirus	1021348 nt	NC_020104
Cafeteria roenbergensis virus BV-PW1	617453 nt	NC_014637
Megavirus chiliensis	1259197 nt	NC_016072
Megavirus Iba	1230522 nt	NC_020232
Megavirus terra1	1244621 nt	NC_023640
Mimivirus terra2	1168989 nt	NC_023639
Yellowstone lake mimivirus	73689 nt	NC_028104



Partial "complete" metagenomes

LOCUS NC_028108 171454 bp DNA linear VRL 30-OCT-2015

DEFINITION **Yellowstone lake phycodnavirus 3**, complete genome, isolate: 3

-> Not a single polymerase (RNA or DNA!)

LOCUS NC_028104 **73689 bp** DNA linear VRL 30-OCT-2015

DEFINITION **Yellowstone lake mimivirus**, complete genome, isolate: 1 ???

-> No DNA polymerase nor RNA polymerase ?

Zhang,W., Zhou,J., Liu,T., Yu,Y., Pan,Y., Yan,S. and Wang,Y. (2015)

Four novel algal virus genomes discovered from Yellowstone Lake metagenomes

Sci Rep 5, 15131, PUBMED [26459929](https://pubmed.ncbi.nlm.nih.gov/26459929/)

Unannotated "complete" genome

LOCUS NC_023640 1244621 bp DNA linear VRL 06-MAR-2014
DEFINITION Megavirus terra1 genome.

LOCUS NC_023639 1168989 bp DNA linear VRL 06-MAR-2014
DEFINITION Mimivirus terra2 genome.

Not a single annotation?

Fully annotated Mimiviridae
complete genomes in **Genbank**
but not listed in « viral genomes » ?

LOCUS JX975216 1246126 bp DNA linear VRL 16-APR-2014

DEFINITION Megavirus courdo11, complete genome.

LOCUS KF493731 1181042 bp DNA linear VRL 20-NOV-2013

DEFINITION Hirudovirus strain Sangsue, complete genome

LOCUS NC_020104 1021348 bp DNA linear VRL 11-JAN-2013

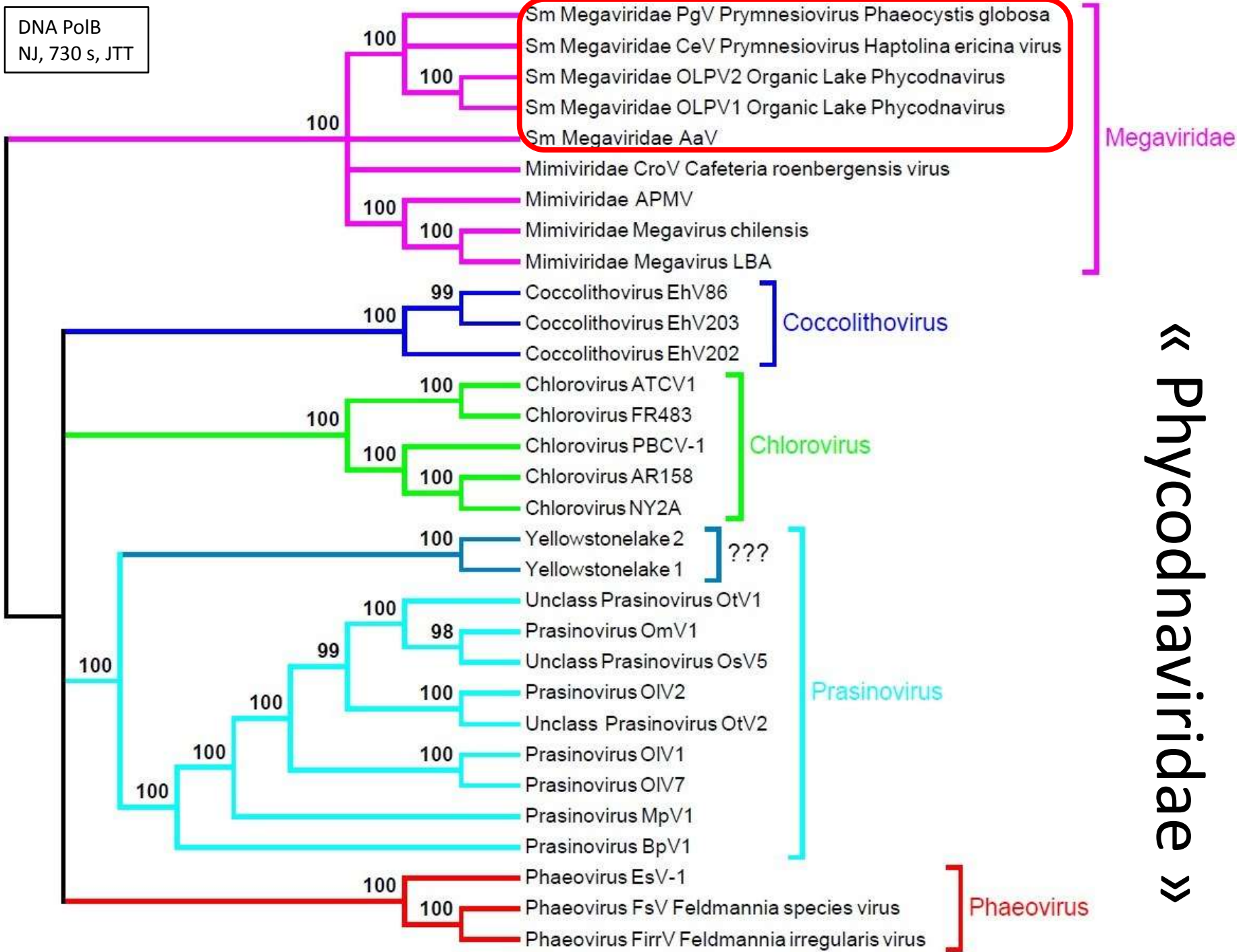
DEFINITION *Ac. polyphaga* moumouvirus, complete genome.

A job for ICTV: Phycodnaviridae/Mimiviridae

Three main problems:

- One family embedded in another one
- Genera as distant from each other as different families
- Nomenclature associated to an unwarranted host range

DNA PoIB
NJ, 730 s, JTT



Megaviridae

Cocolithovirus

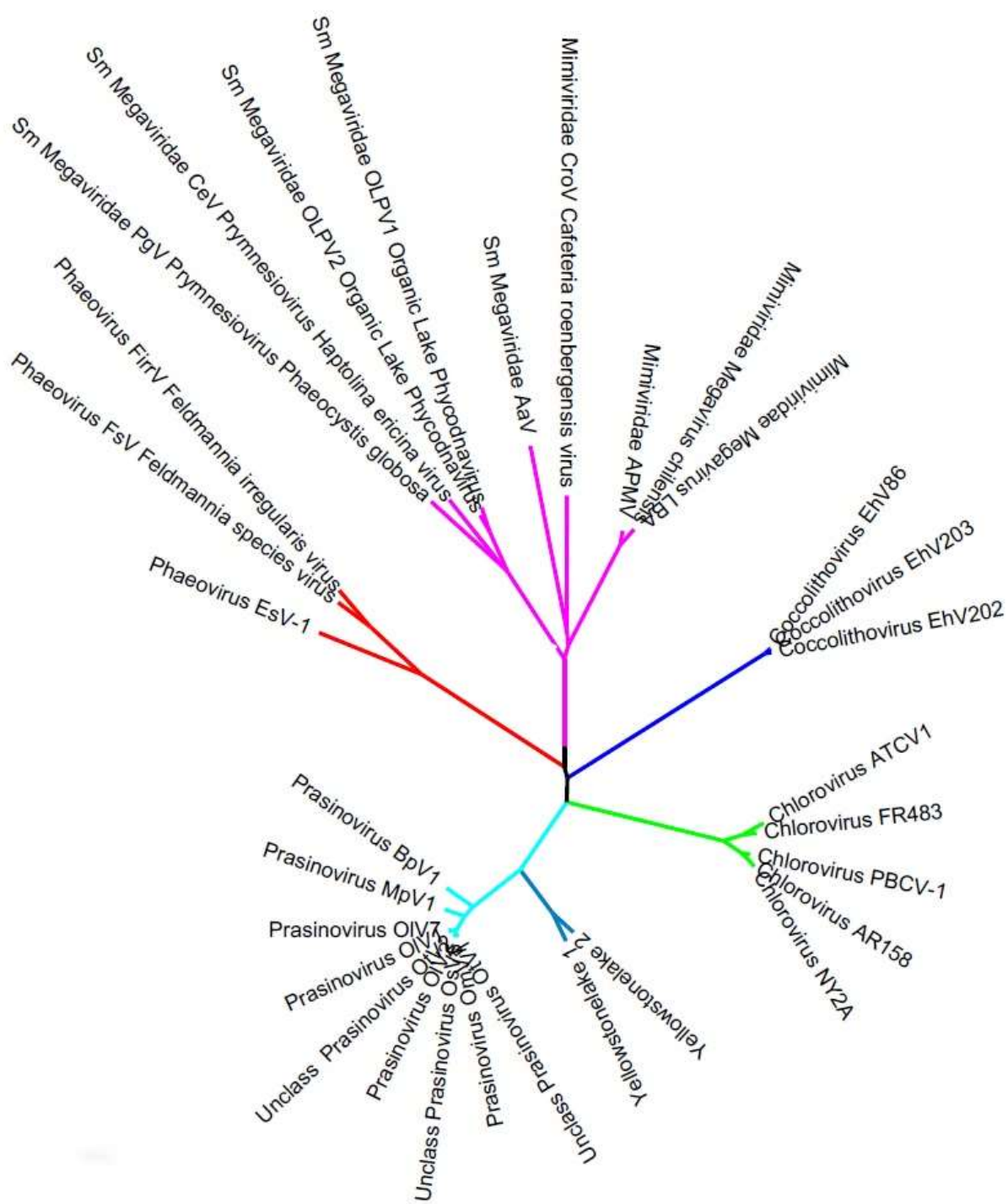
Chlorovirus

???

Prasinovirus

Phaeovirus

« Phycodnaviridae »



Five equidistant clades:

- Mega/mimiviridae**
- Coccolithoviruses**
- Chloroviruses**
- Prasinoviruses**
- Phaeoviruses**

Global features

Virus	Genome Size (kb)	Virion \emptyset (nm)	RNA pol	DNA pol size	DNA pol Intein	MutS7	GC%
Chloroviruses	288-368	190	-	$\cong 900$	-	-	40%
Prasinoviruses	182-196	125	-	$\cong 900$	-	-	45-48%
Phaeoviruses	154-335	120-150	-	$\cong 1,000$	-	-	53%
Coccolithoviruses	405	175	+	$\cong 1,000$	-	-	40%
Sm_Mimiviridae	370-474	150-300	+	$\cong 1,600$	+/-	+	32%
Mimiviridae	730-1,26	300-700	+	$\cong 1,700$	+	+	26-30%

%ID DNA Polymerase B

DNA PolB % ID	Chlorovirus	Prasinovirus	Phaeovirus	Coccolithovirus	SmMimiviridae	Mimiviridae
Chlorovirus	>71	<40	<32	<33	<33	<24
Prasinovirus	<40	>73	<32	<37	<29	<31
Phaeovirus	<32	<32	>44	<34	-	-
Coccolithovirus	<33	<37	<34	>92	<33	-
SmMimiviridae	<33	<29	-	<33	>45	>41
Mimiviridae	<24	<31	-	-	>41	>65

>44% ID is presently the divergence limit within each of these clades

Possible ways out (to discuss among SGs)

- 1) Stop using Phycodnaviridae as a «family» name
- 2) Create 2-3 subfamilies within Megaviridae
 - Mimivirinae -> large ones
 - Unclassified Megaviridae for others (pending more)
- 3) Upgrade
 - Chloro-, Phaeo-, Prasino-, Coccolitho-virus as families ?
 - However using «host names» might become misleading (beware of future host specificities)

Other viral taxonomy problems

- Which objective criteria for families ?
- What minimum knowledge is required?
- Which genes (if any) to use as references ?


Closing remarks: "real" vs. "virtual" viruses

“Artificially created viruses and laboratory hybrid viruses will not be given taxonomic consideration. Their classification will be the responsibility of acknowledged international specialist groups”

- 1) What about metagenomic assembly ?
- 2) What about incomplete genomes
- 3) What about isolated genes?

Should we name and classify viruses that have never been seen and/or isolated ?

The lack of a coherent policy makes « manual »
data mining a nightmare
and automated data mining impossible



**Thank you for
your attention.**

**I am now available
for questions
if any.**



Dr. Chantal Abergel