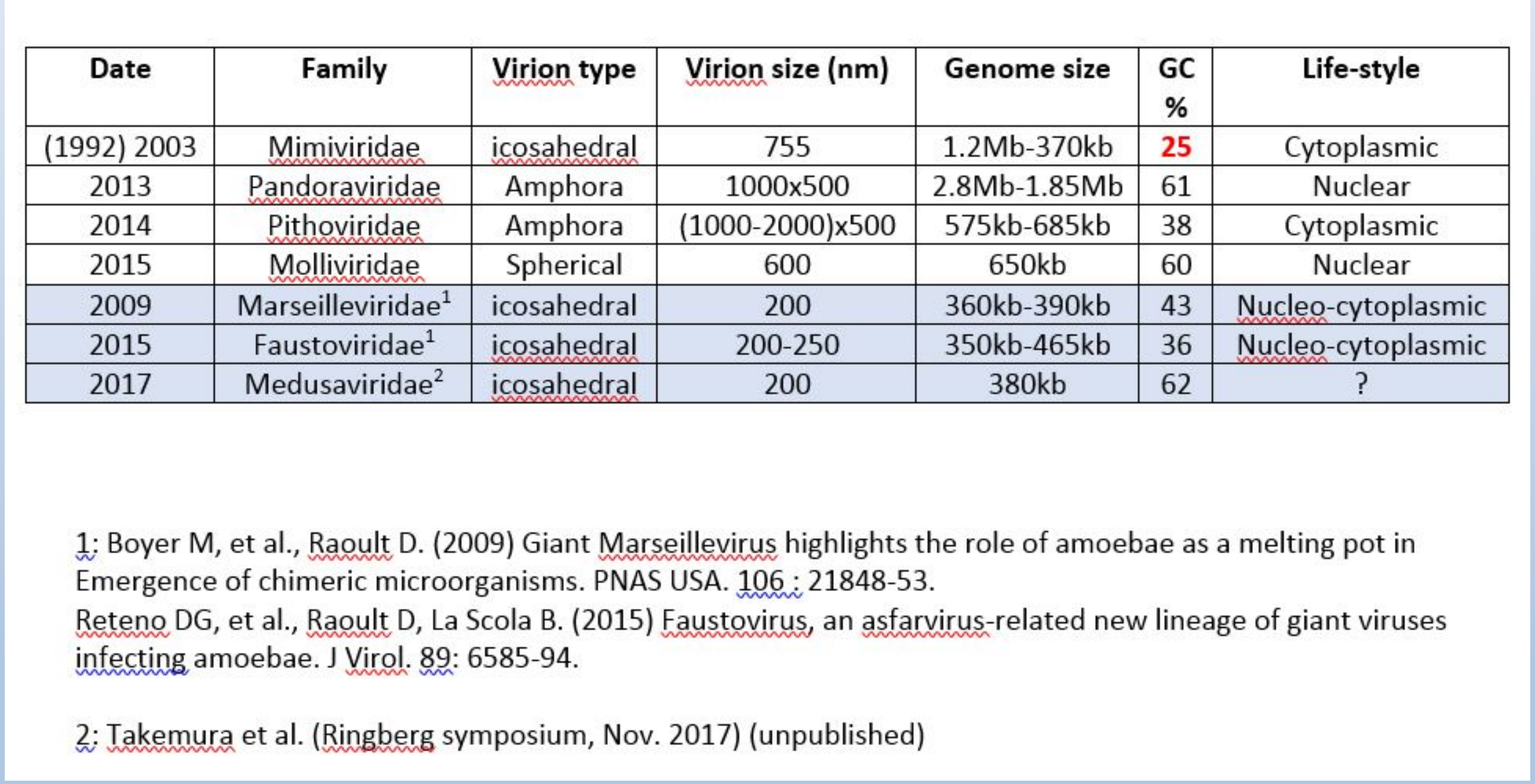
Diversity & Evolution of the emerging *Pandoraviridae* Family

Jean-Michel Claverie

Matthieu Legendre, Chantal Abergel, *et al.*

Giant viruses: a short history

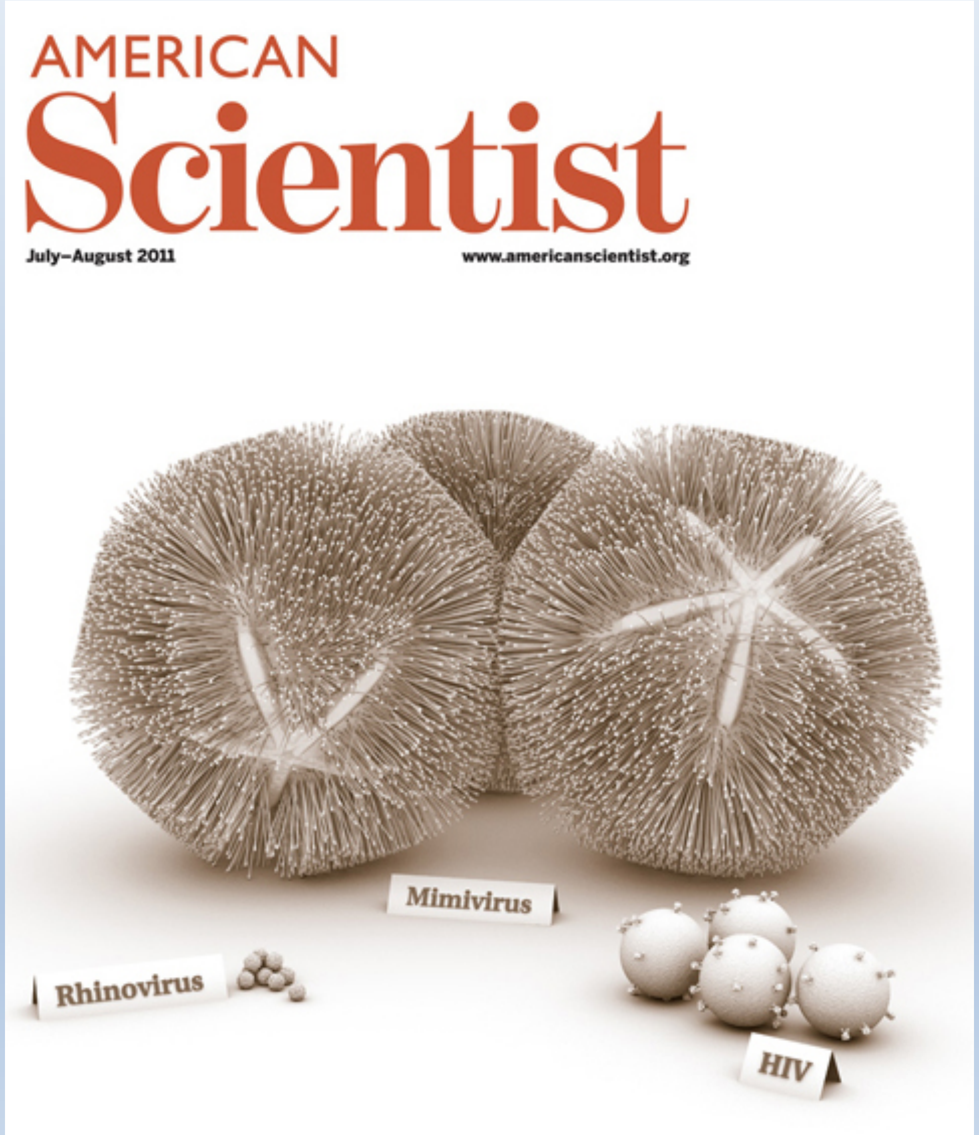


Date	Family	Virion type	Virion size (nm)	Genome size	GC %	Life-style
(1992) 2003	Mimiviridae	icosahedral	755	1.2Mb-370kb	**25**	Cytoplasmic
2013	Pandoraviridae	Amphora	1000x500	2.8Mb-1.85Mb	61	Nuclear
2014	Pithoviridae	Amphora	(1000-2000)x500	575kb-685kb	38	Cytoplasmic
2015	Molliviridae	Spherical	600	650kb	60	Nuclear
2009	Marseilleviridae[1]	icosahedral	200	360kb-390kb	43	Nucleo-cytoplasmic
2015	Faustoviridae[1]	icosahedral	200-250	350kb-465kb	36	Nucleo-cytoplasmic
2017	Medusaviridae[2]	icosahedral	200	380kb	62	?

1: Boyer M, et al., Raoult D. (2009) Giant Marseillevirus highlights the role of amoebae as a melting pot in Emergence of chimeric microorganisms. PNAS USA. 106 : 21848-53.
Reteno DG, et al., Raoult D, La Scola B. (2015) Faustovirus, an asfarvirus-related new lineage of giant viruses infecting amoebae. J Virol. 89: 6585-94.

2: Takemura et al. (Ringberg symposium, Nov. 2017) (unpublished)

Why call them « giant » viruses?

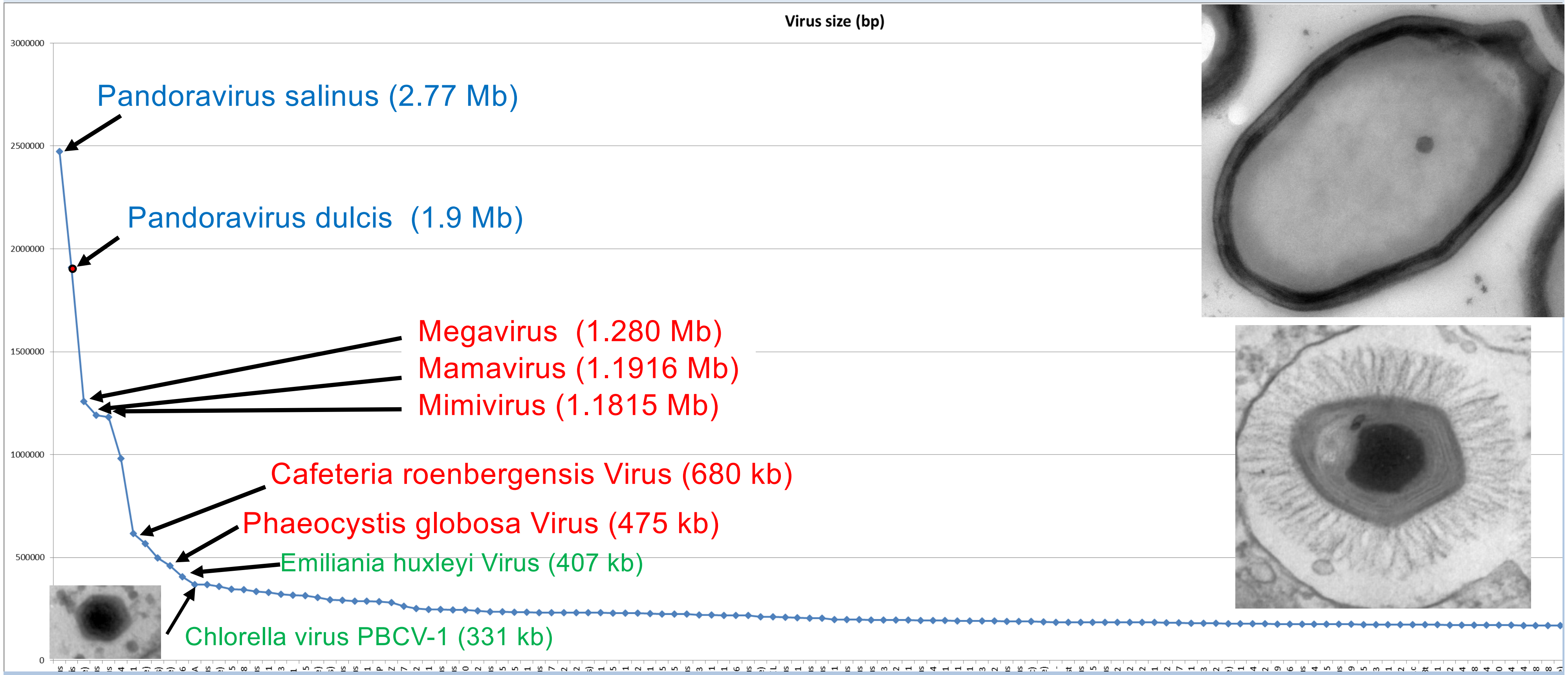


AMERICAN Scientist

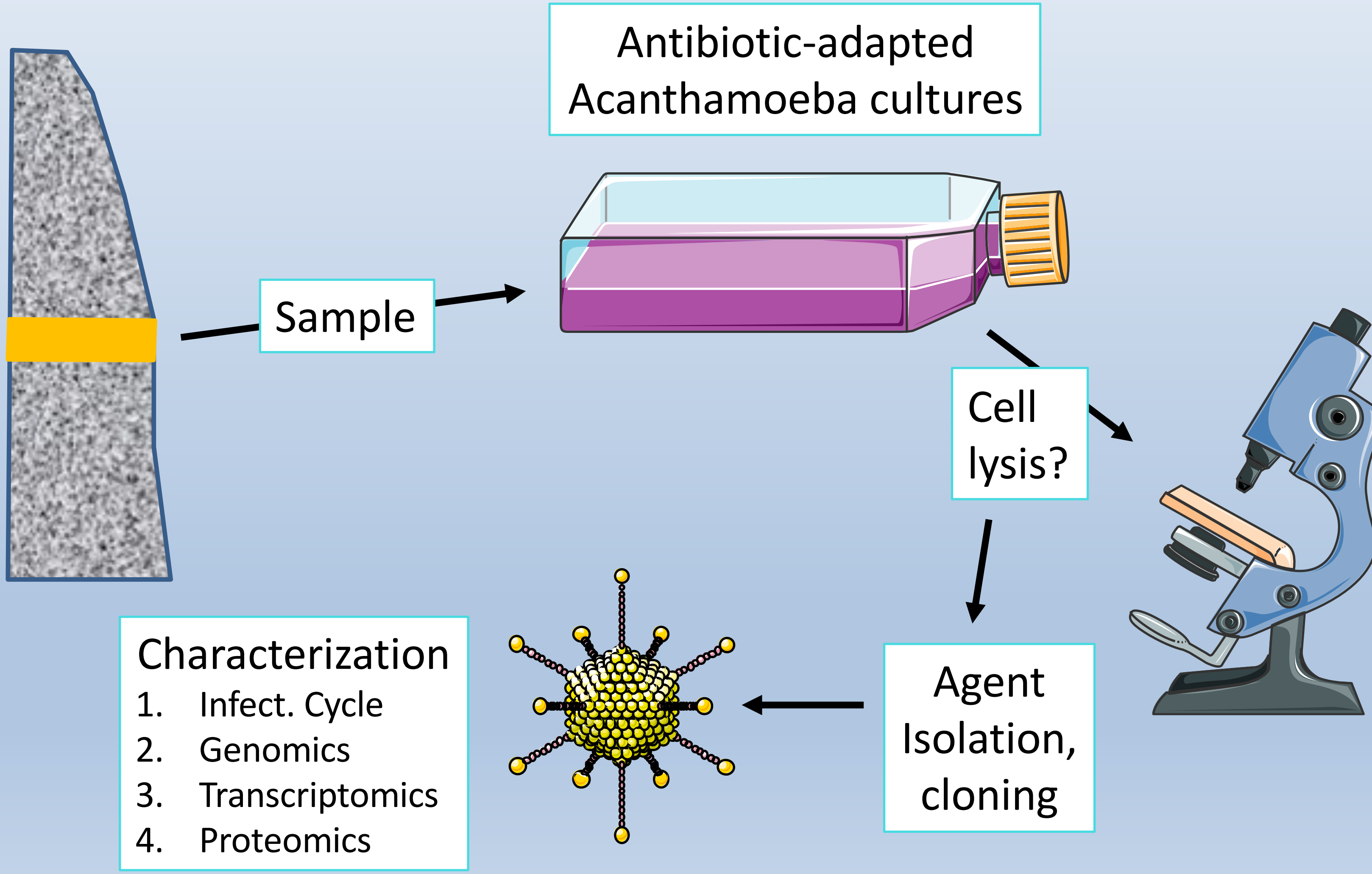
July–August 2011

www.americanscientist.org

Why call them « giant » viruses?



Protocol: looking for Amoeba-killing viruses



Sample

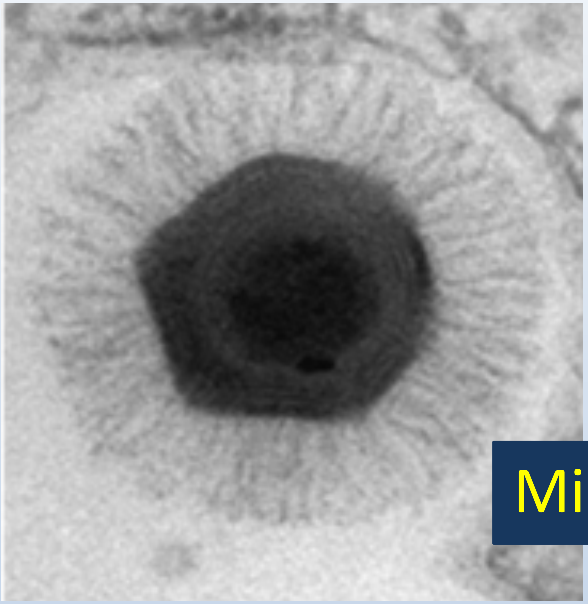
Antibiotic-adapted Acanthamoeba cultures

Cell lysis?

Agent Isolation, cloning

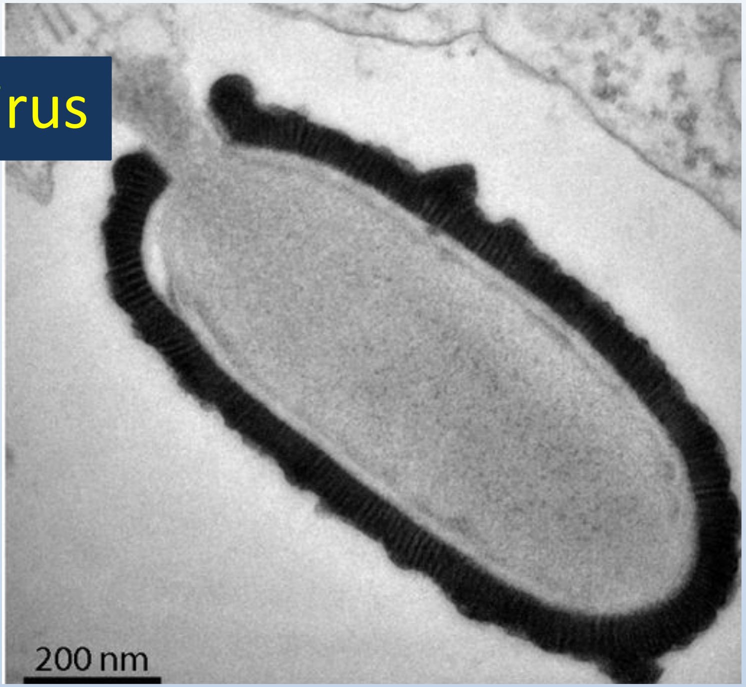
Characterization

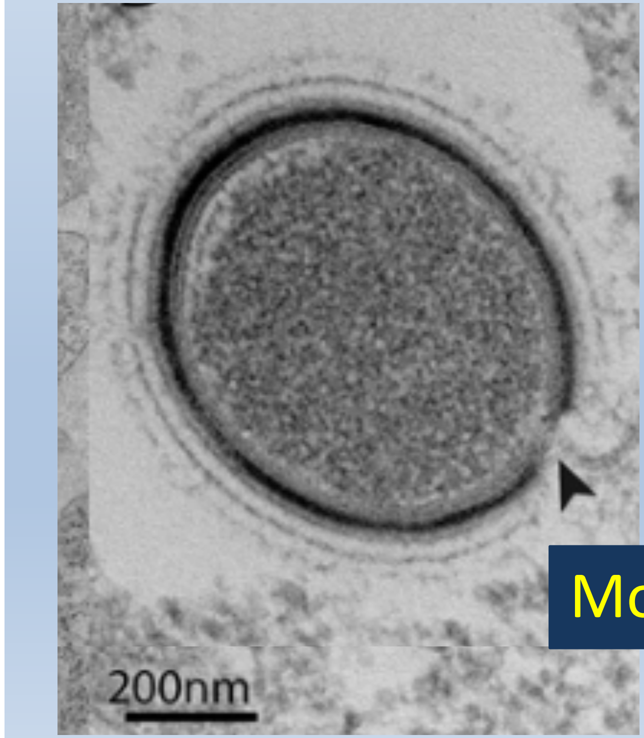
1. Infect. Cycle
2. Genomics
3. Transcriptomics
4. Proteomics



Mimivirus

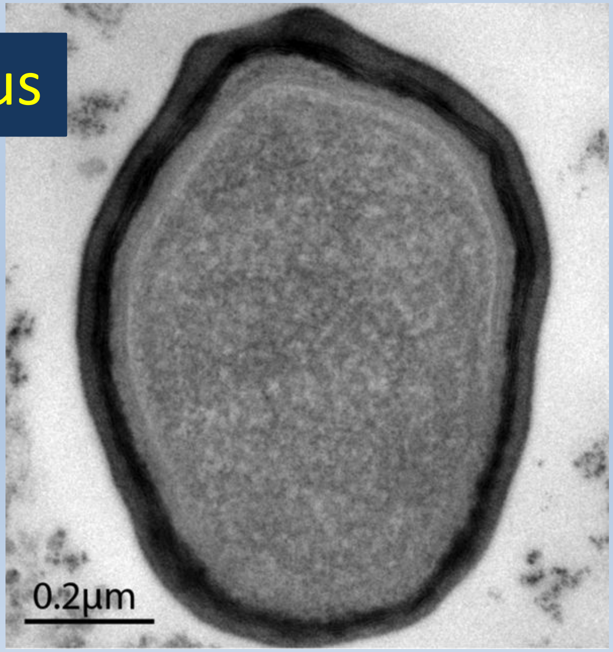
Pithovirus



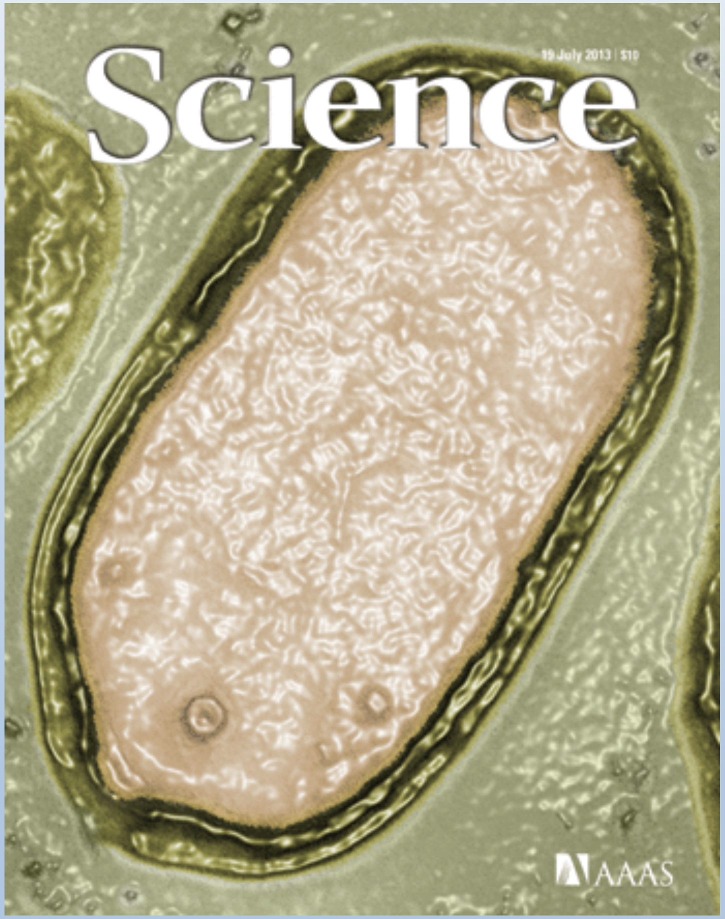


Mollivirus

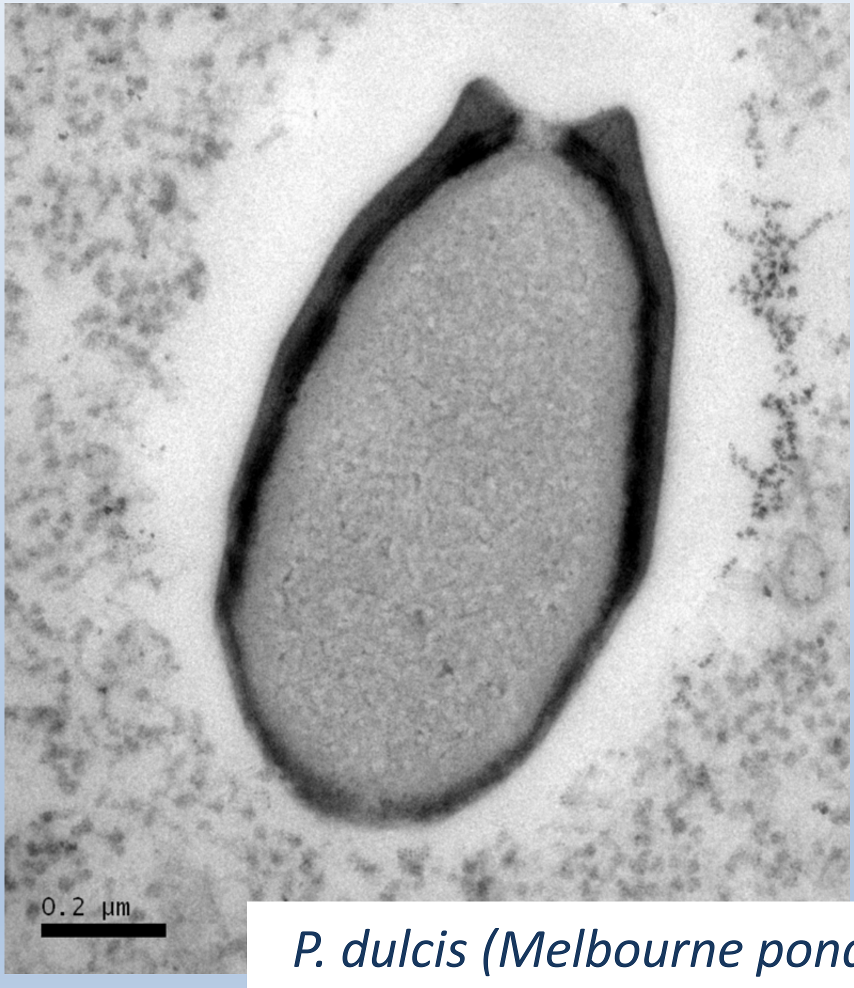
Pandoravirus



2013: Pandoravirus salinus & P. dulcis



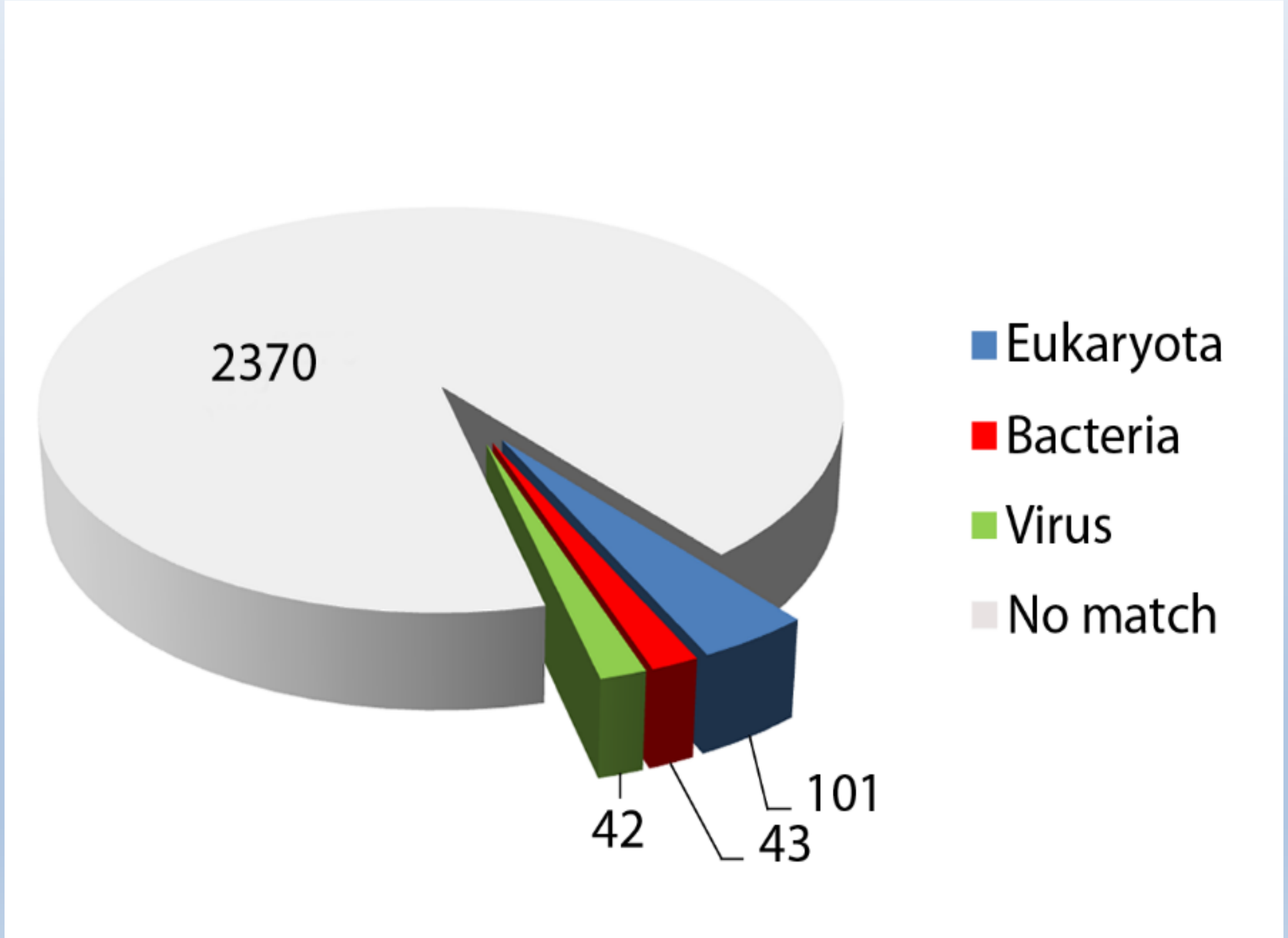
P. salinus (Chilean coast)



P. dulcis (Melbourne pond)

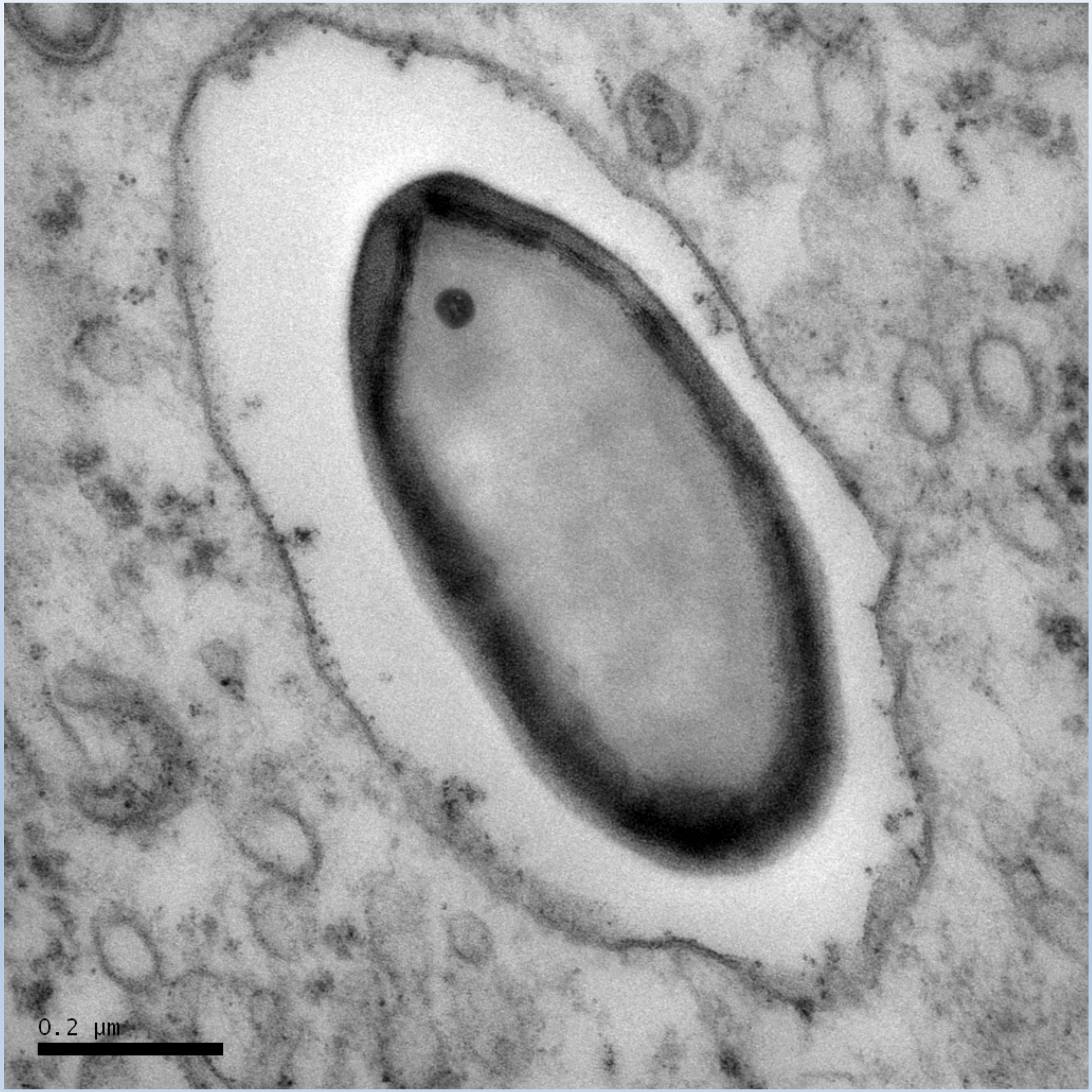
Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. Philippe, et al., Claverie, Abergel (2013). *Science* 341: 281-6

94% of the genes encode ORFans !

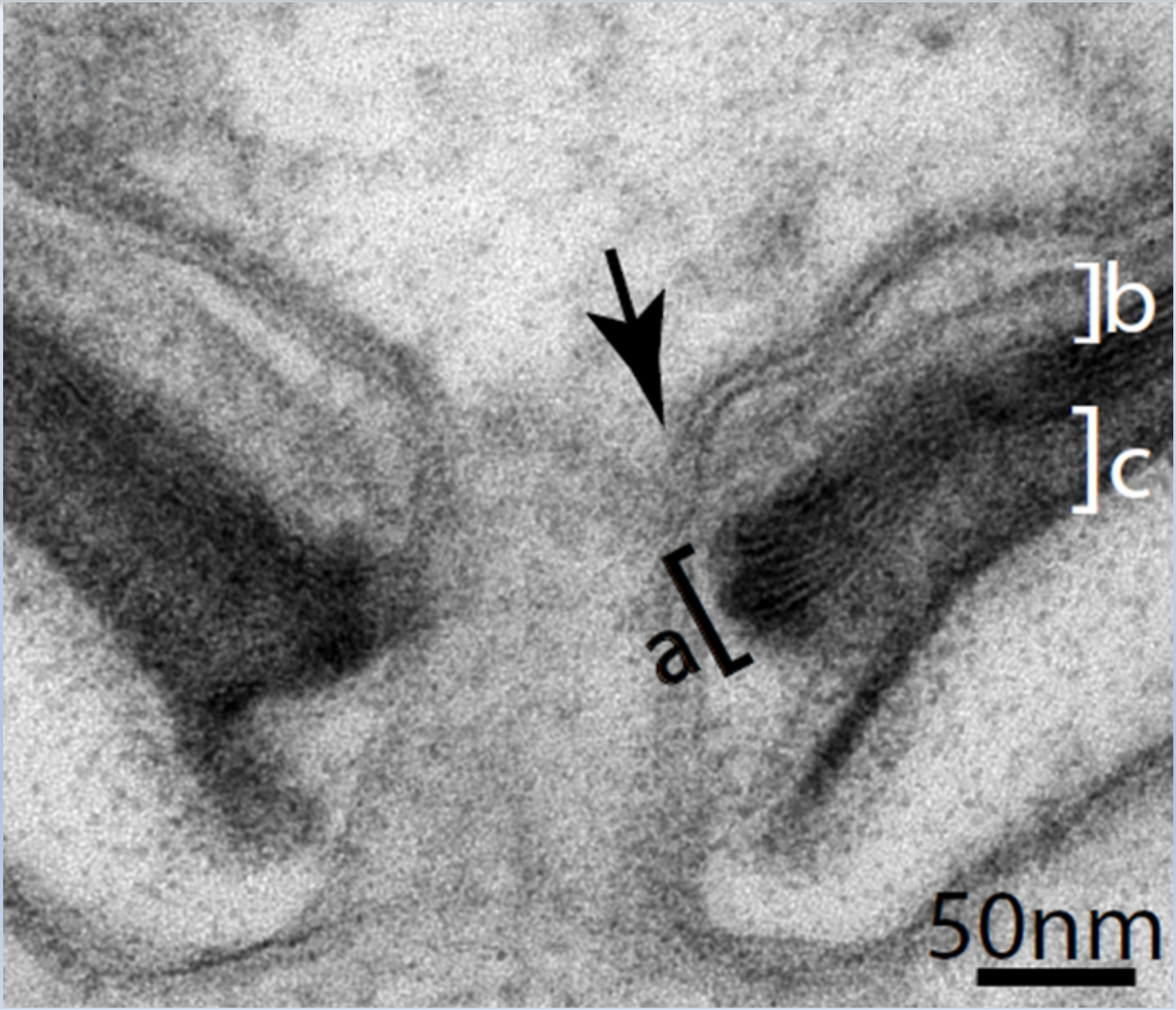


Pandoravirus:
Infectious cycle

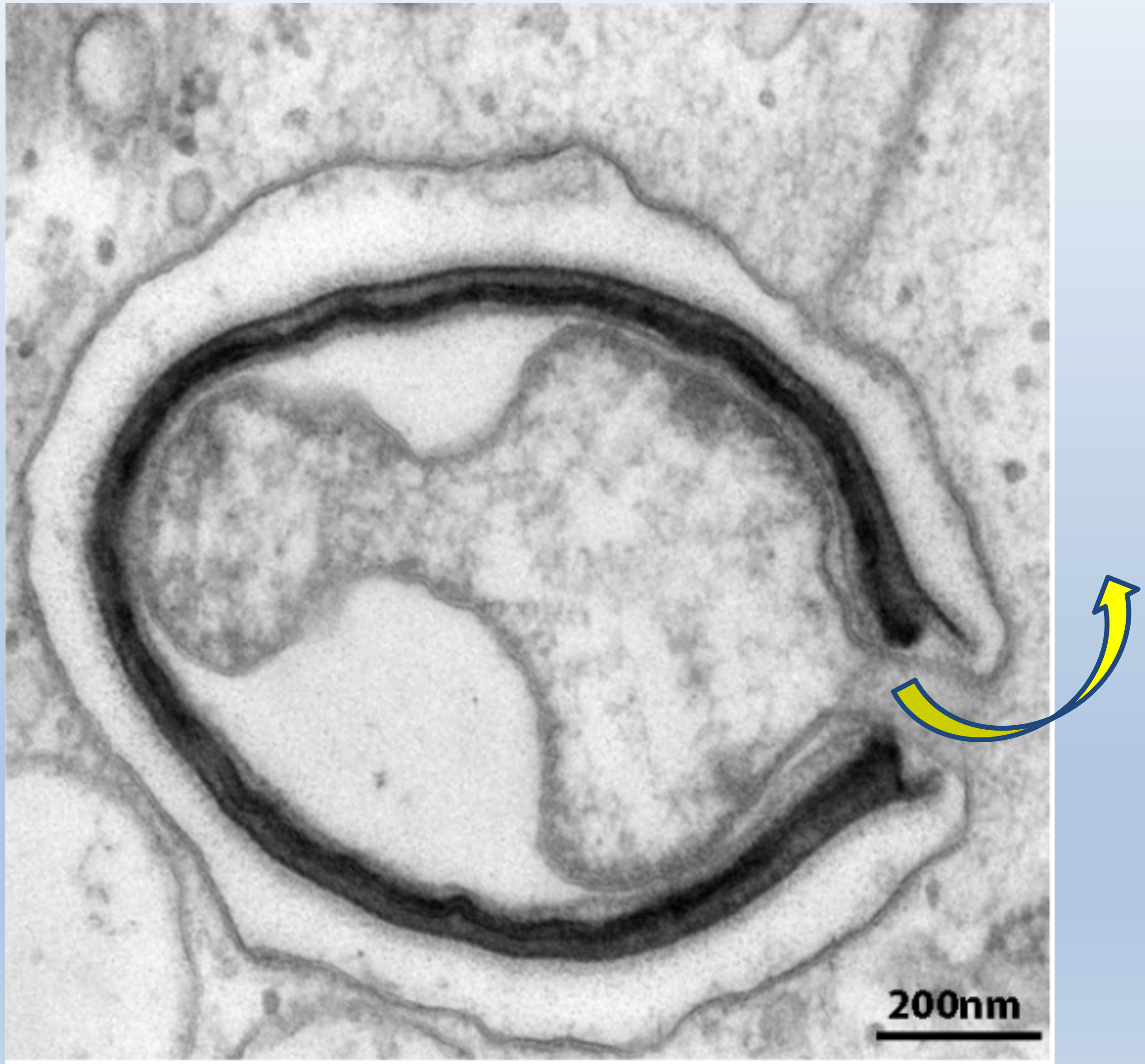
Step 1: phagocytosis



Step 2 : membrane fusion

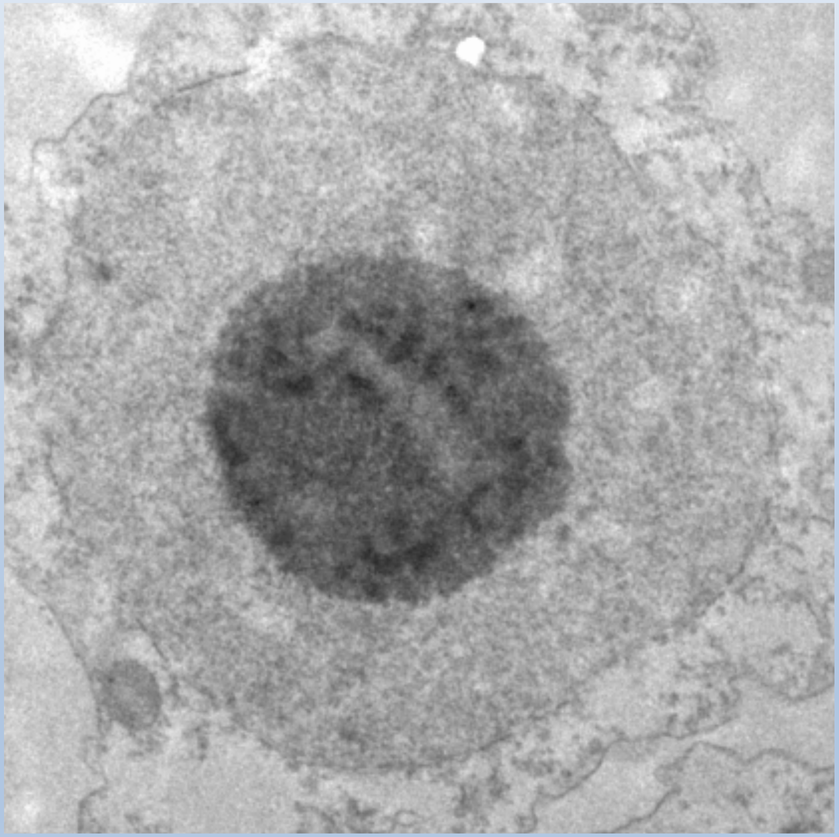


Step 3 : « downloading »

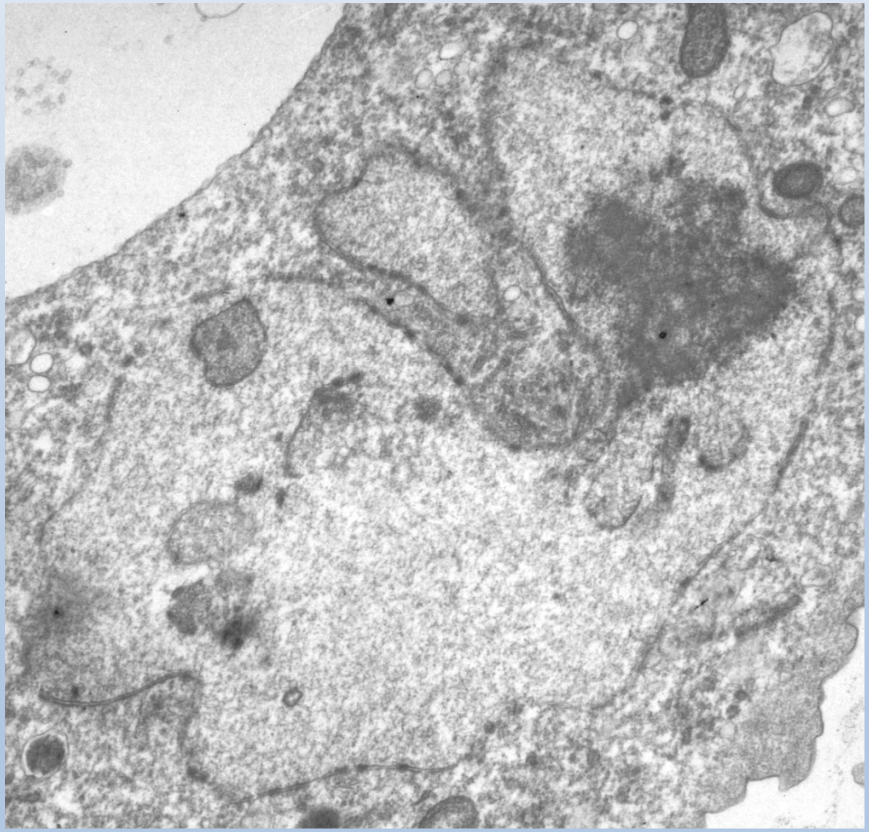


Step 4: Early nuclear phase?

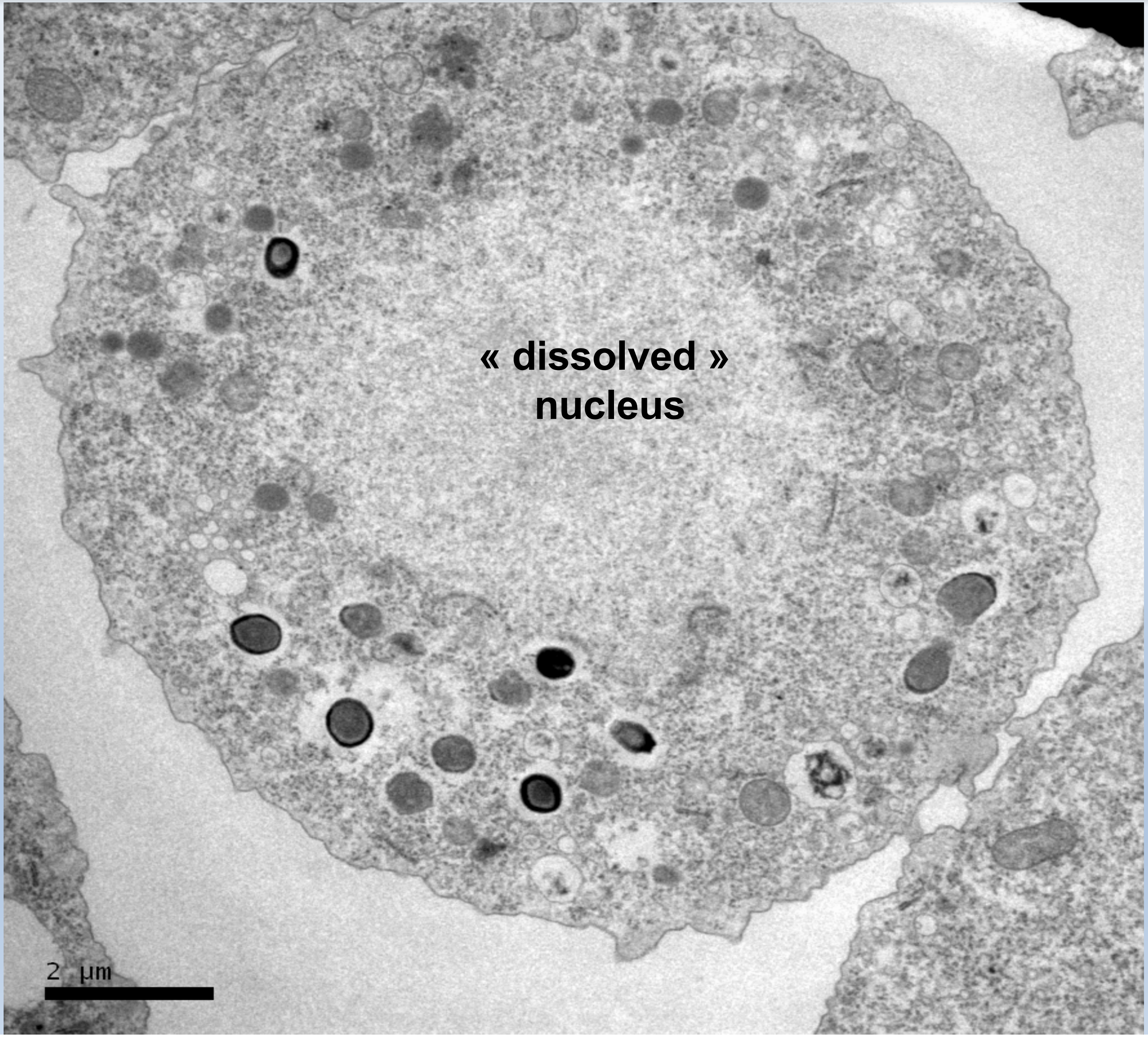
Healthy Acanthamoeba cell



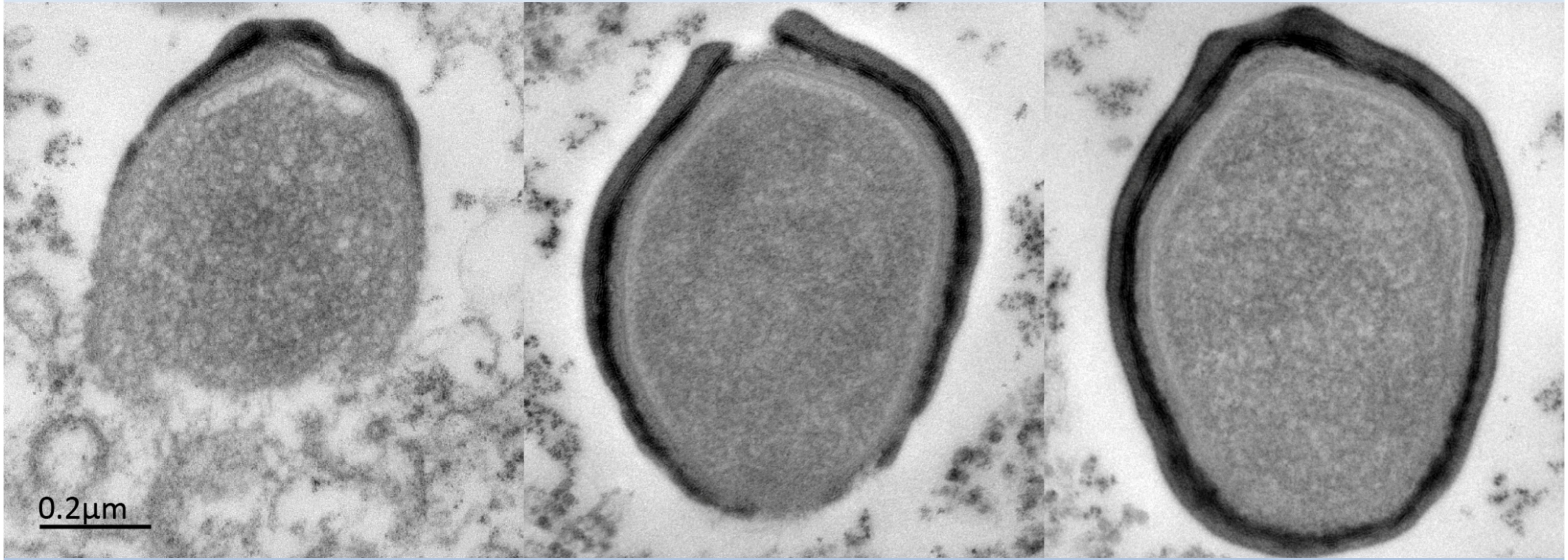
Infected cell (3h p.i.)



Step 5: Particle formation

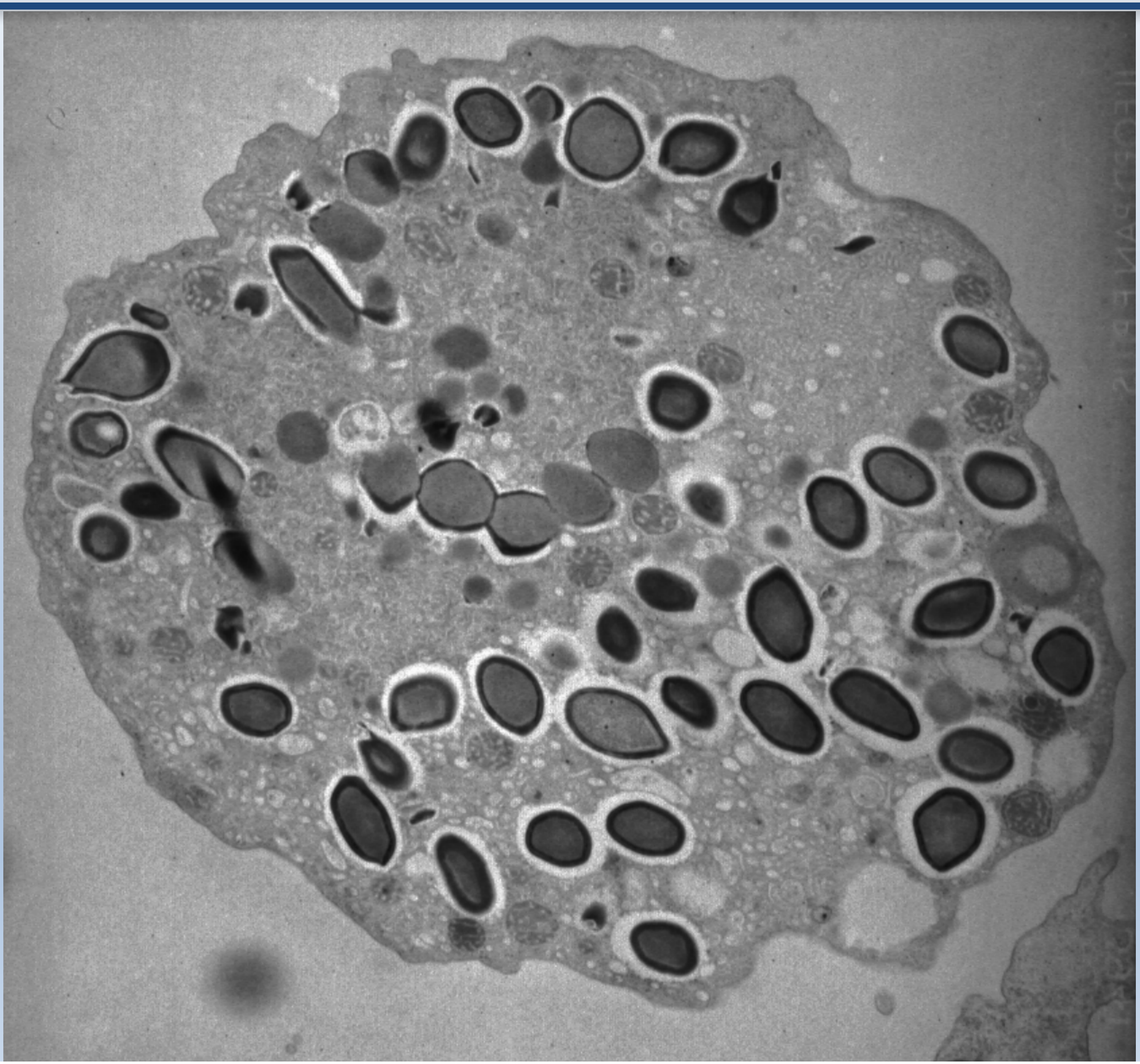


Particle formation: “knitting”



No division

End of cycle



Despite their huge genome Pandoraviruses are nucleus-dependent

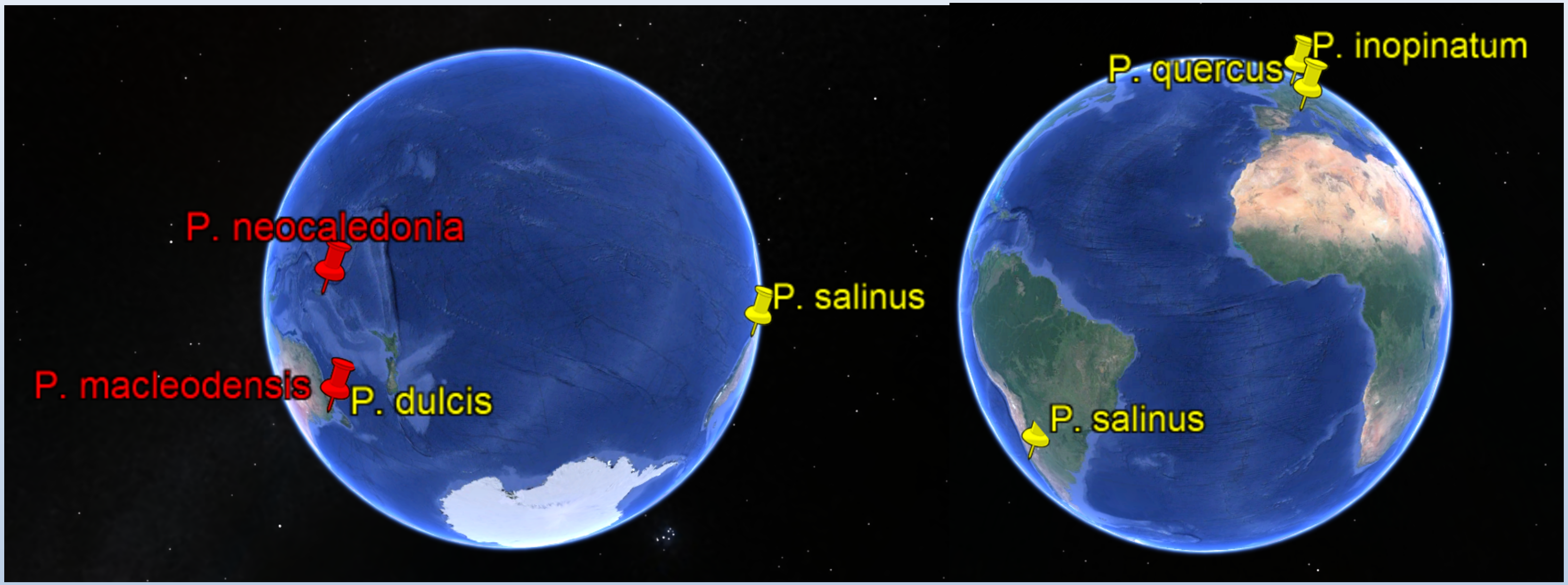
EM: Cell nucleus is quickly modified after the infection

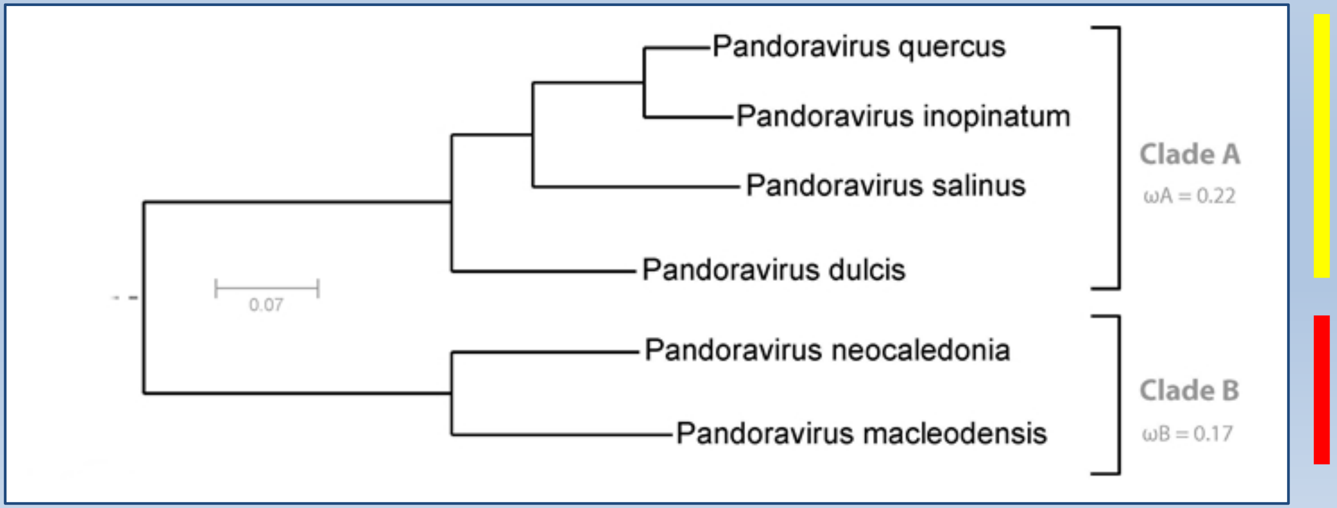
Transcriptome:
At 10% (7.5%-13%) of the genes exhibit spliceosomal introns (U2-dependent, GT-AG)
(*These introns are short (<200 nt), more than one third remain in phase with the flanking exons).*

Proteome:
The particles do not incorporate any transcription machinery
102 “core proteins” common to all isolates.

- No standard Major Capsid Protein
- No DNA packaging ATPase
- No DNA repair enzyme

6 isolates from 6 distant locations

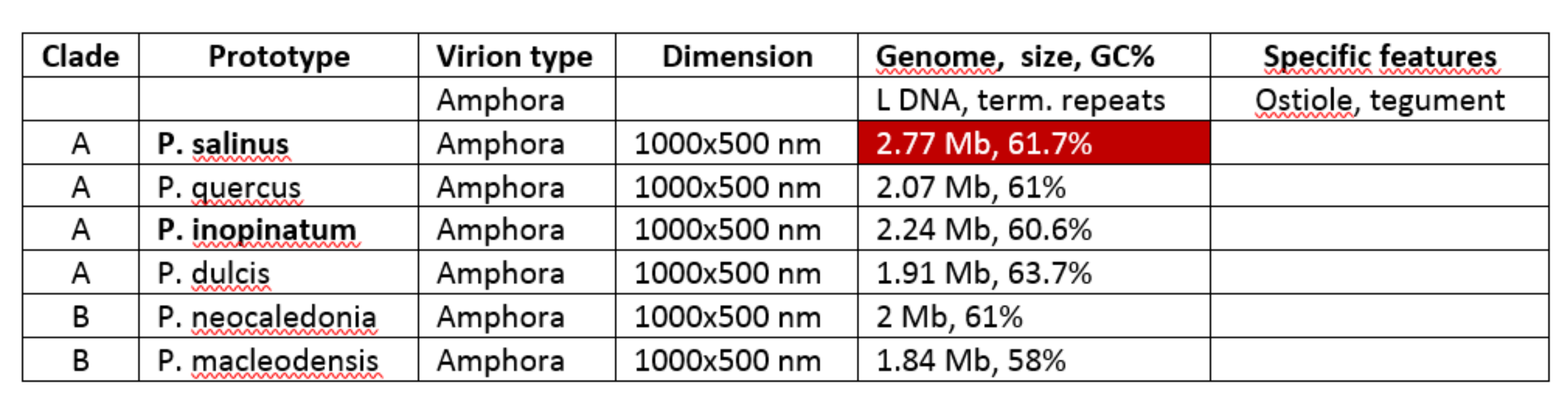




From *Pandoravirus dulcis* to *P. macleodensis*

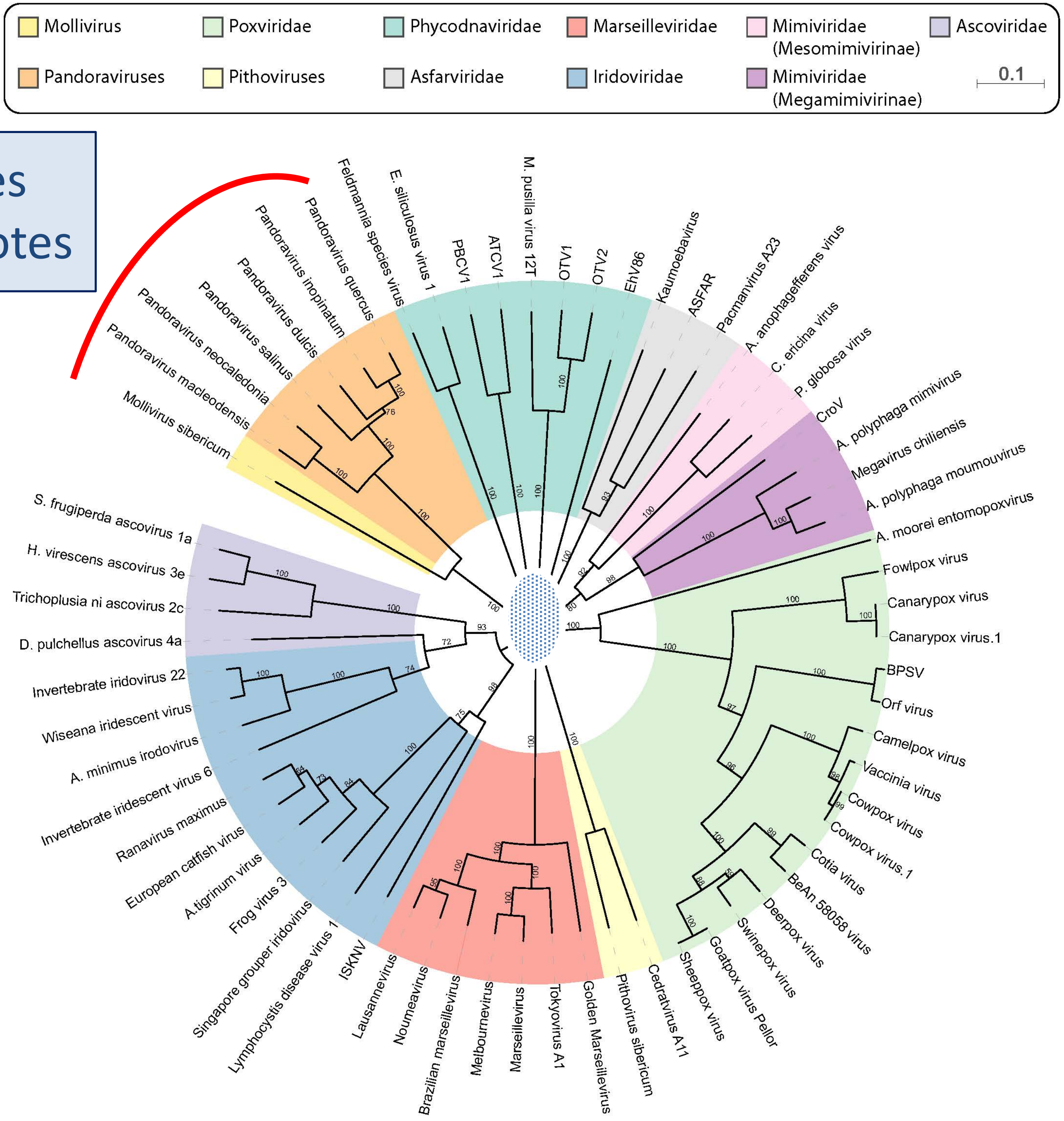
	P. salinus	*P. inopinatum*	*P. quercus*	*P. dulcis*	*P. macleodensis*
P. inopinatum	73%				
P. quercus	74%	88%			
P. dulcis	70%	71%	72%		
P. macleodensis	54%	54%	55%	55%	
P. neocaledonia	54%	54%	54%	55%	76%

The Pandoraviridae today



Clade	**Prototype**	**Virion type**	**Dimension**	**Genome, size, GC%**	**Specific features**
		Amphora		L DNA, term. repeats	Ostiole, tegument
A	**P. salinus**	Amphora	1000x500 nm	2.77 Mb, 61.7%	
A	P. quercus	Amphora	1000x500 nm	2.07 Mb, 61%	
A	**P. inopinatum**	Amphora	1000x500 nm	2.24 Mb, 60.6%	
A	P. dulcis	Amphora	1000x500 nm	1.91 Mb, 63.7%	
B	P. neocaledonia	Amphora	1000x500 nm	2 Mb, 61%	
B	P. macleodensis	Amphora	1000x500 nm	1.84 Mb, 58%	

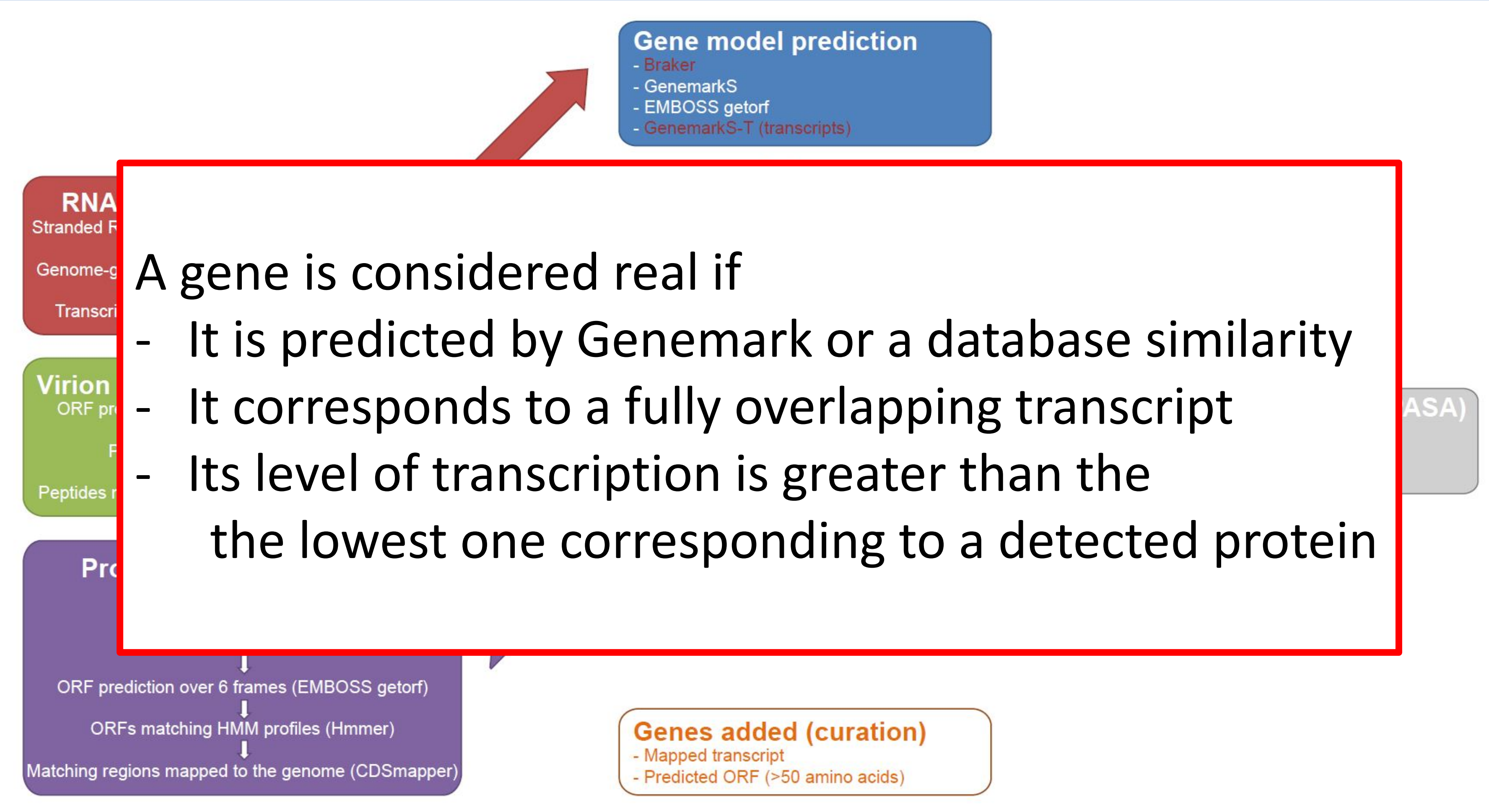
Large DNA viruses infecting eukaryotes



Gene content-based cladistic tree of large DNA viruses

A stringent reannotation: are ORFans real?

Compensate high GC% - induced artefacts with additional information

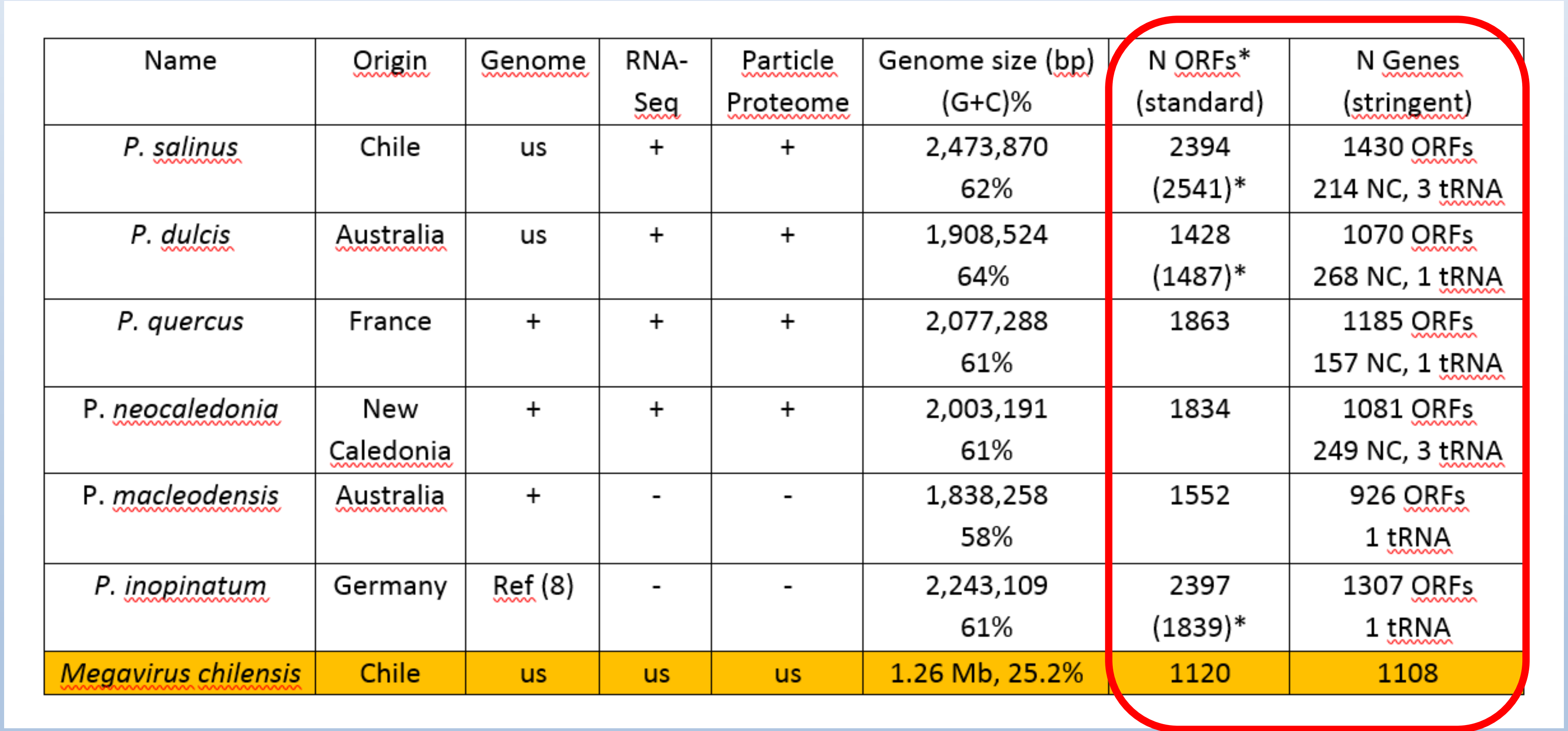

Gene model prediction
- Braker
- GenemarkS
- EMBOSS getorf
- GenemarkS-T (transcripts)
RNA
Stranded R
Genome-g
Transcri
Virion
ORF pr
P
Peptides r
Pro
ORF prediction over 6 frames (EMBOSS getorf)
ORFs matching HMM profiles (Hmmer)
Matching regions mapped to the genome (CDSmapper)
ASA)
Genes added (curation)
- Mapped transcript
- Predicted ORF (>50 amino acids)

A gene is considered real if

- It is predicted by Genemark or a database similarity
- It corresponds to a fully overlapping transcript
- Its level of transcription is greater than the the lowest one corresponding to a detected protein

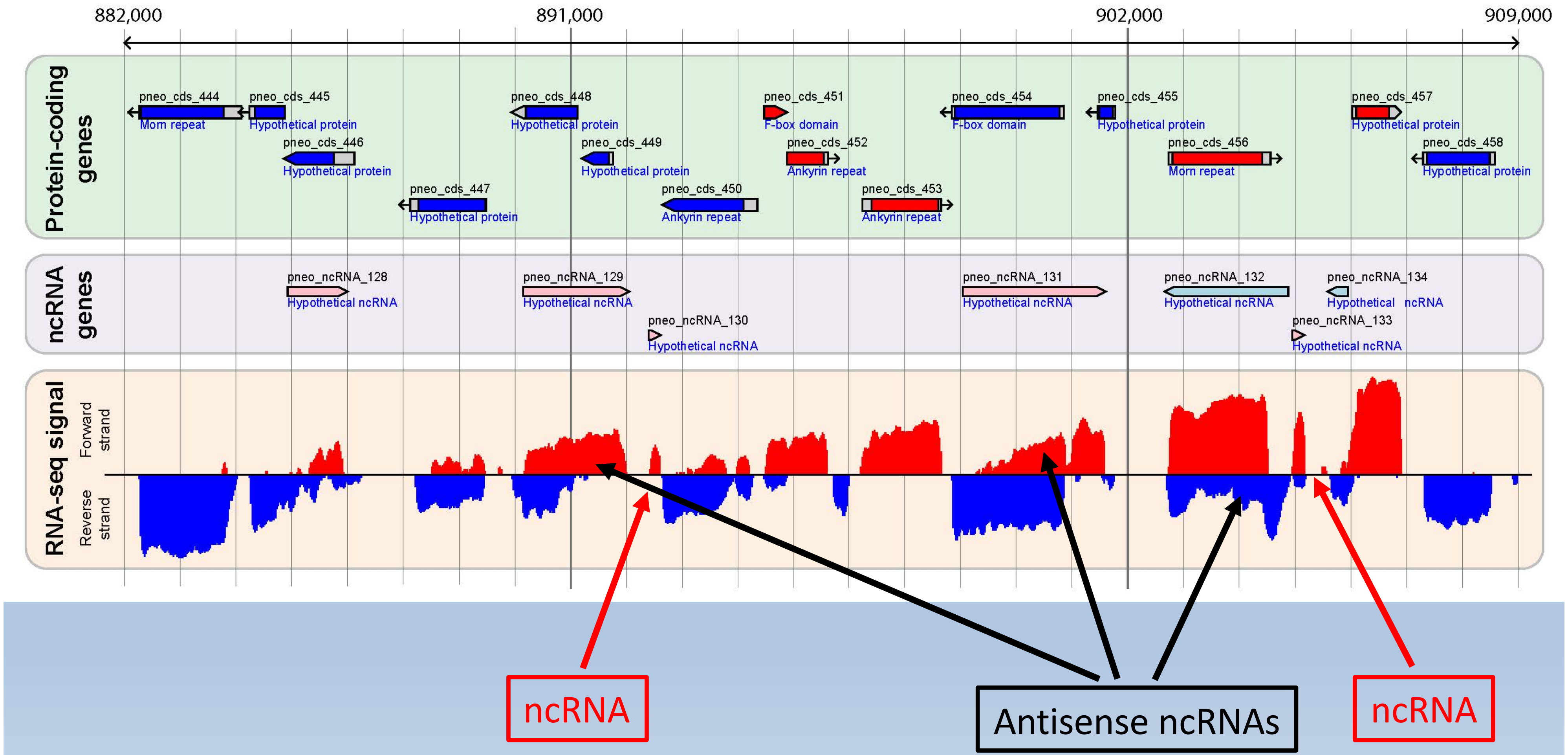
A stringent reannotation:

up to 44% less protein-coding genes



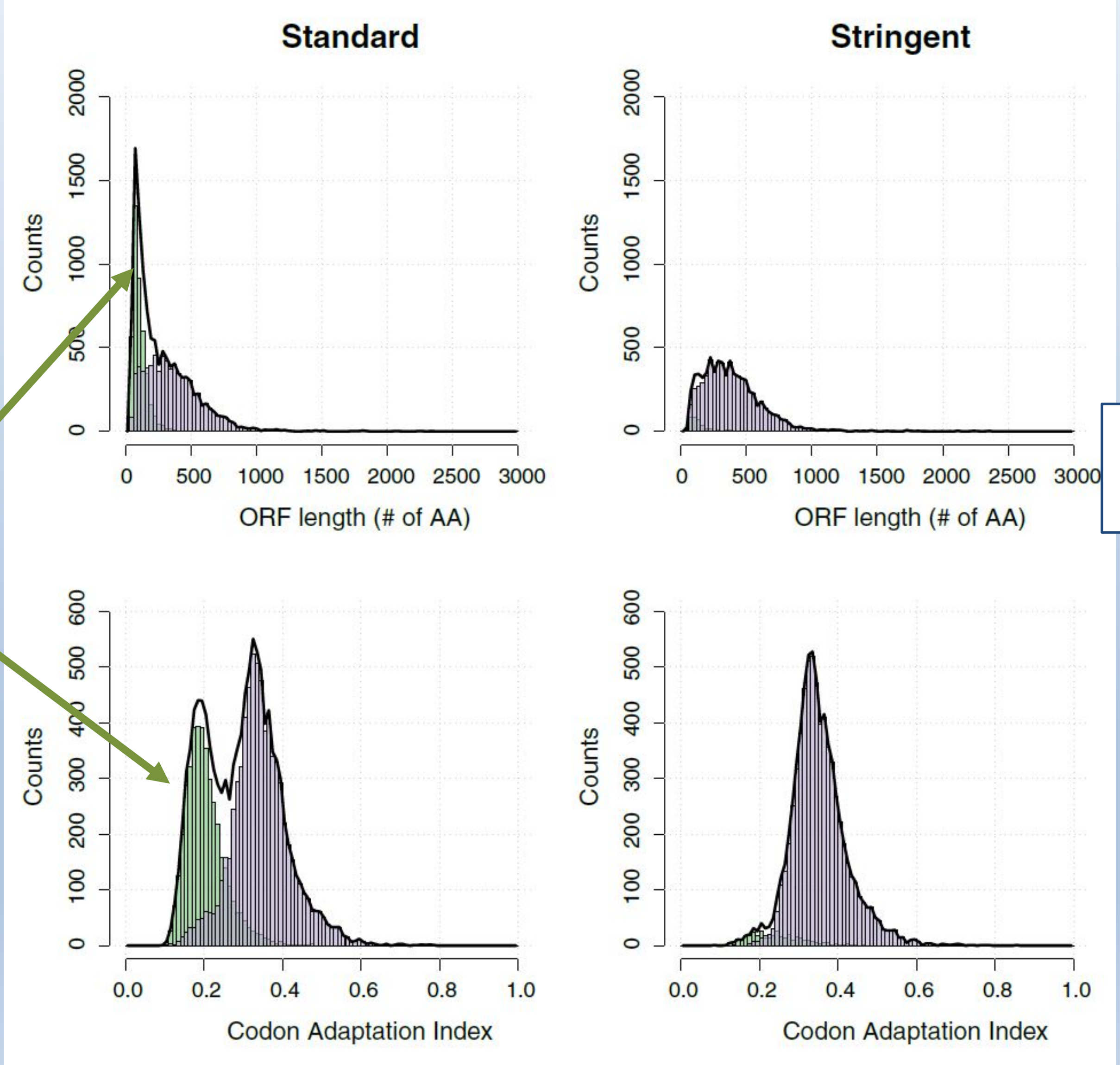
Name	Origin	Genome	RNA-Seq	Particle Proteome	Genome size (bp) (G+C)%	N ORFs* (standard)	N Genes (stringent)
P. salinus	Chile	us	+	+	2,473,870 62%	2394 (2541)*	1430 ORFs 214 NC, 3 tRNA
P. dulcis	Australia	us	+	+	1,908,524 64%	1428 (1487)*	1070 ORFs 268 NC, 1 tRNA
P. quercus	France	+	+	+	2,077,288 61%	1863	1185 ORFs 157 NC, 1 tRNA
P. *neocaledonia*	New Caledonia	+	+	+	2,003,191 61%	1834	1081 ORFs 249 NC, 3 tRNA
P. *macleodensis*	Australia	+	-	-	1,838,258 58%	1552	926 ORFs 1 tRNA
P. inopinatum	Germany	Ref (8)	-	-	2,243,109 61%	2397 (1839)*	1307 ORFs 1 tRNA
Megavirus chilensis	Chile	us	us	us	1.26 Mb, 25.2%	1120	1108

LncRNA: mostly antisense, a few others



157 to 268 LncRNAs

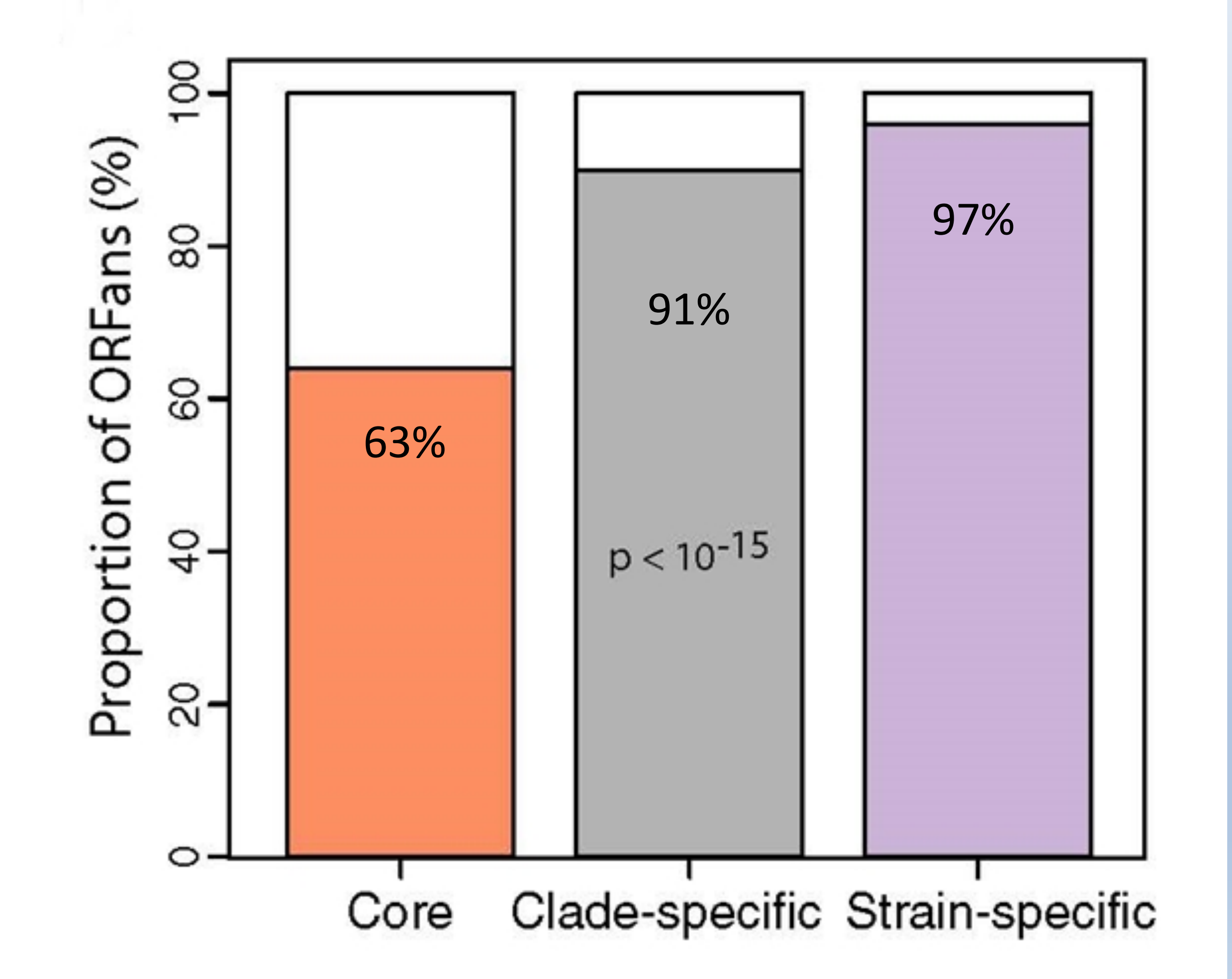
Stringent annotation: a healthier starting point



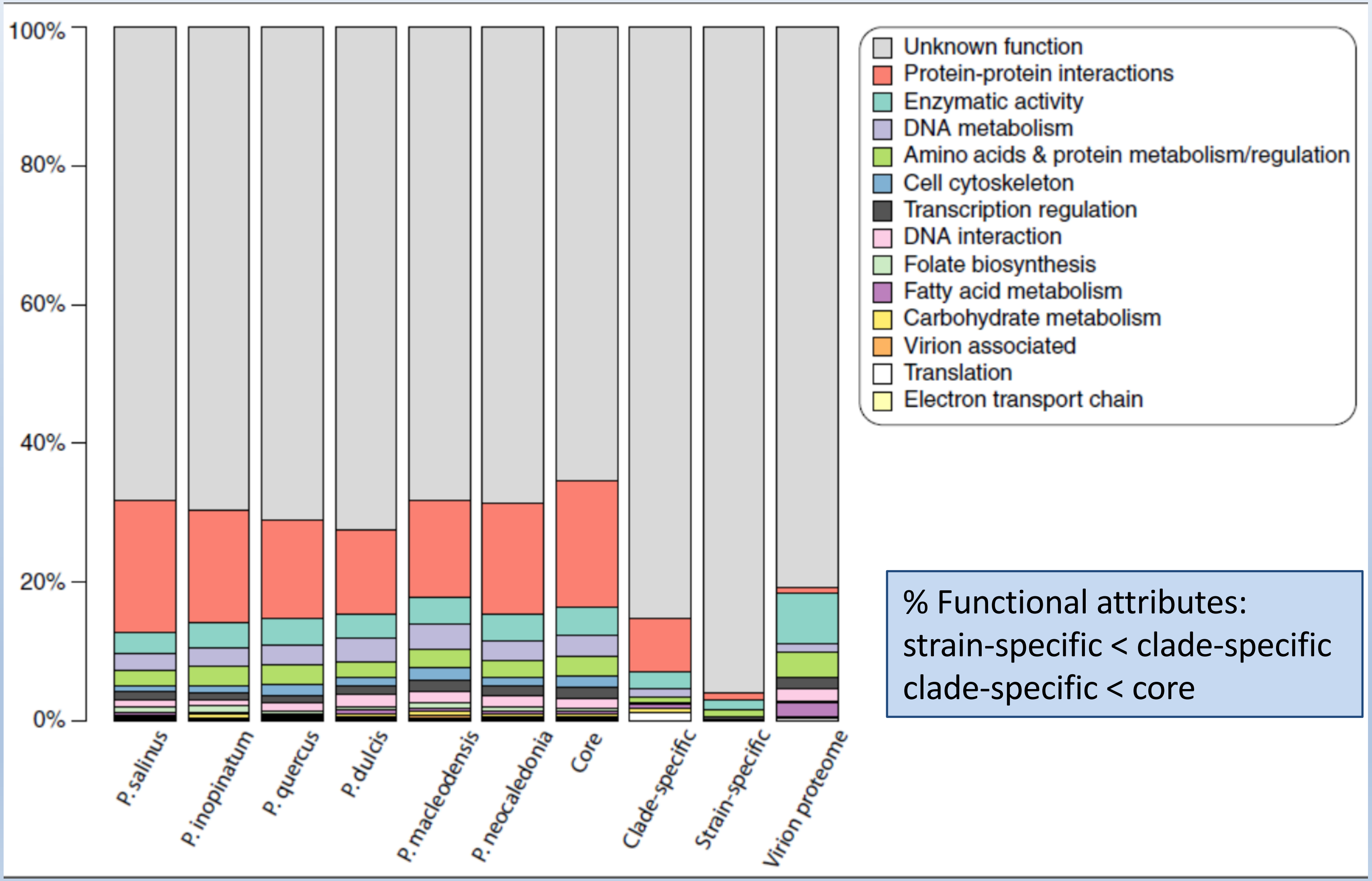
Strictly ORFans

Family ORFans: 67% -73%

Stringent annotation: proportion of ORFans

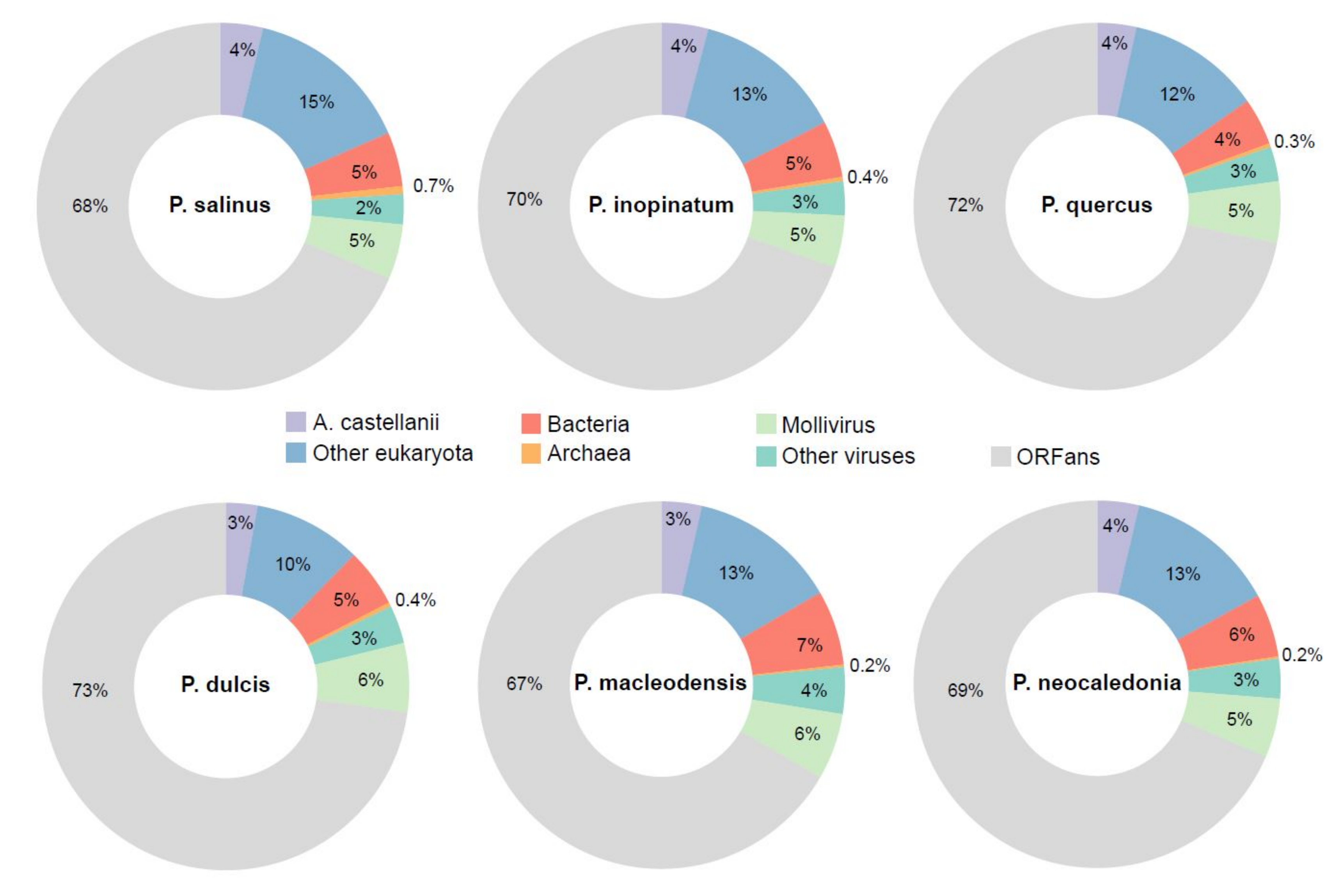


Stringent annotation: functional analysis



% Functional attributes:
strain-specific < clade-specific
clade-specific < core

Stringent annotation: still 70% of family ORFans

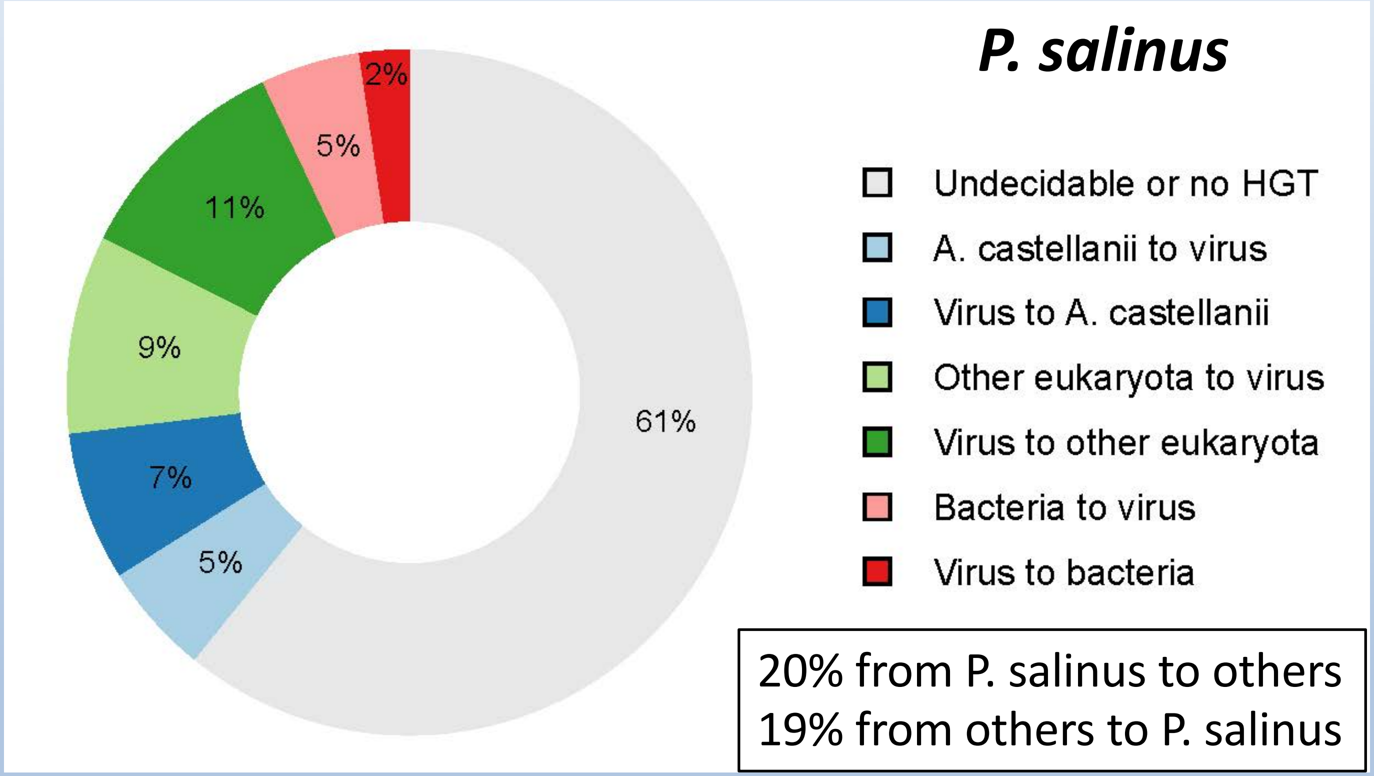


What could explain

- the uniquely large genome of Pandoraviruses ?
- the large proportion of anonymous proteins
- the large proportion of ORFans ?

- a huge frequency of gene gain through HGT ?
- a huge frequency of gene duplication ?
- a hugely complex ancestor ?
- anything else ?

HGTs: contributed at most 15% of the gene content (at least) 6%

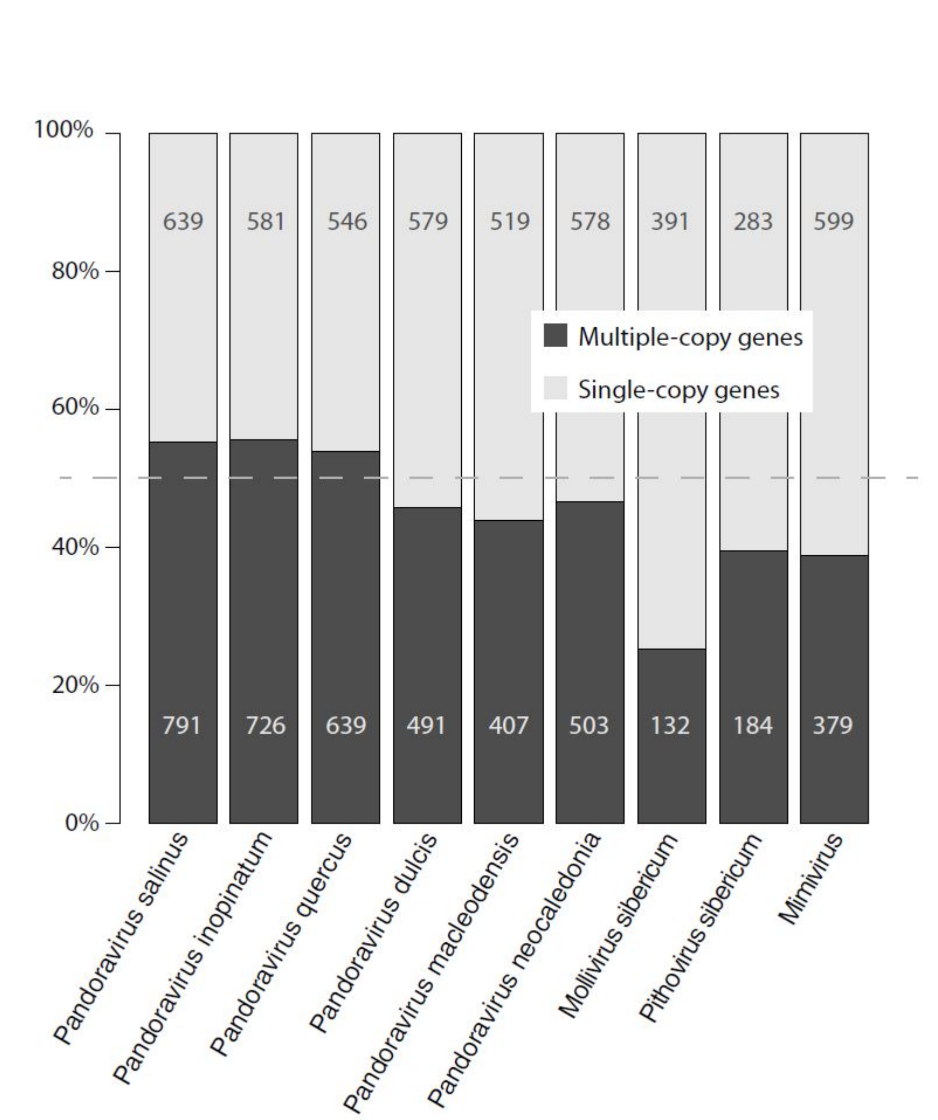


20% from P. salinus to others
19% from others to P. salinus

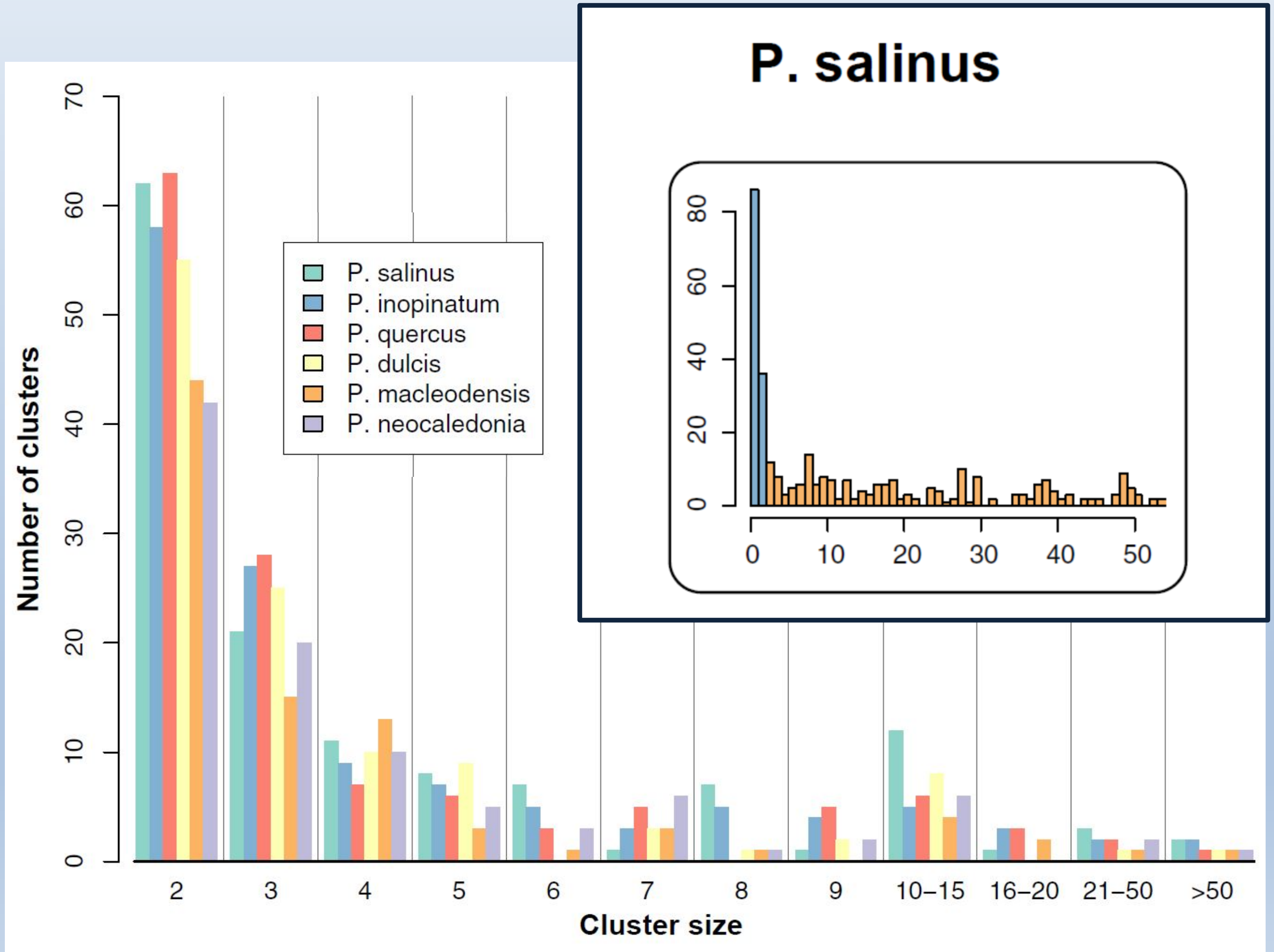
Nothing special compared to other large dsDNA viruses

Duplication analysis

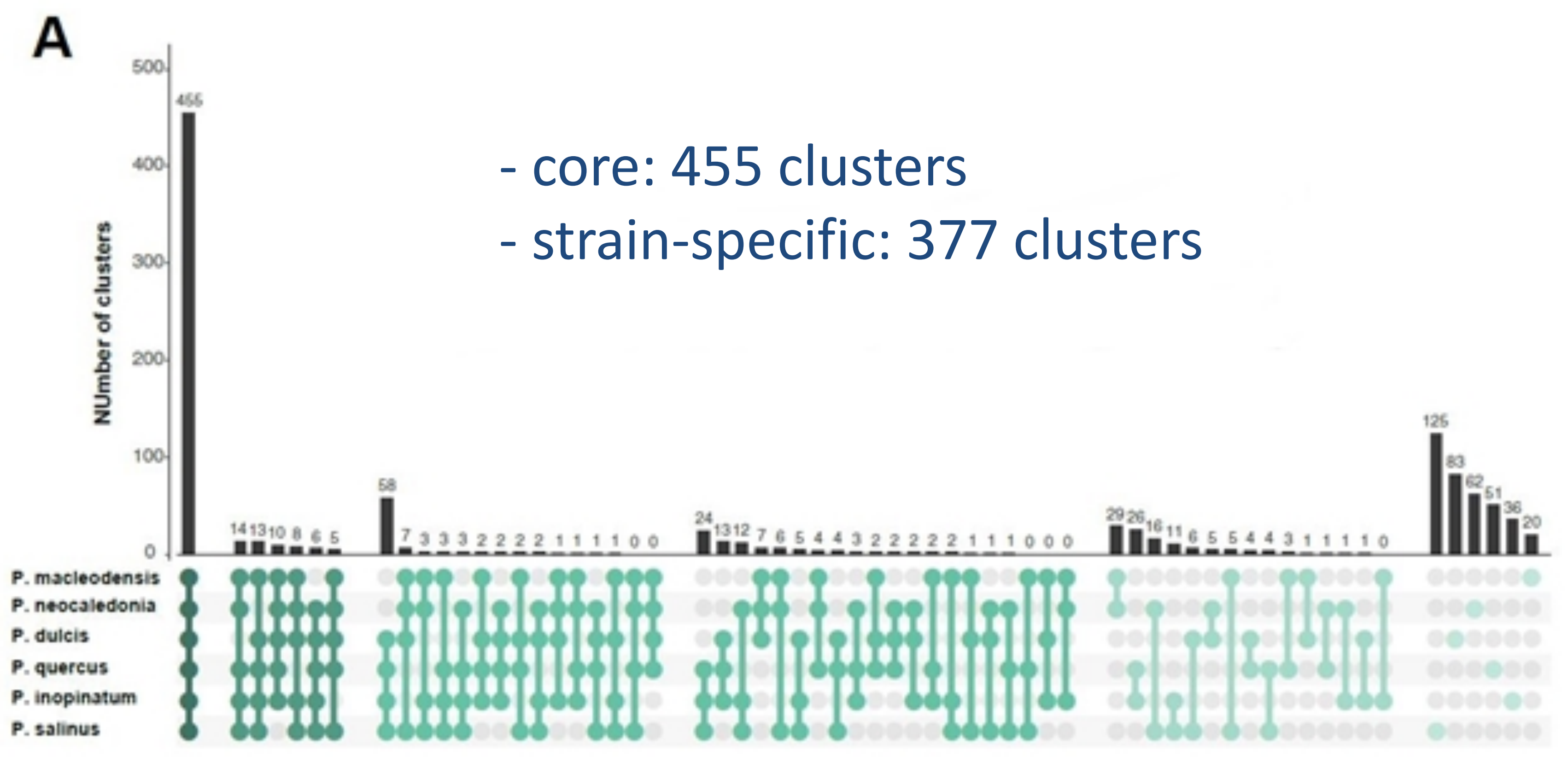
Not so different from Mimivirus (half the size)



Duplications are mostly tandem repeats

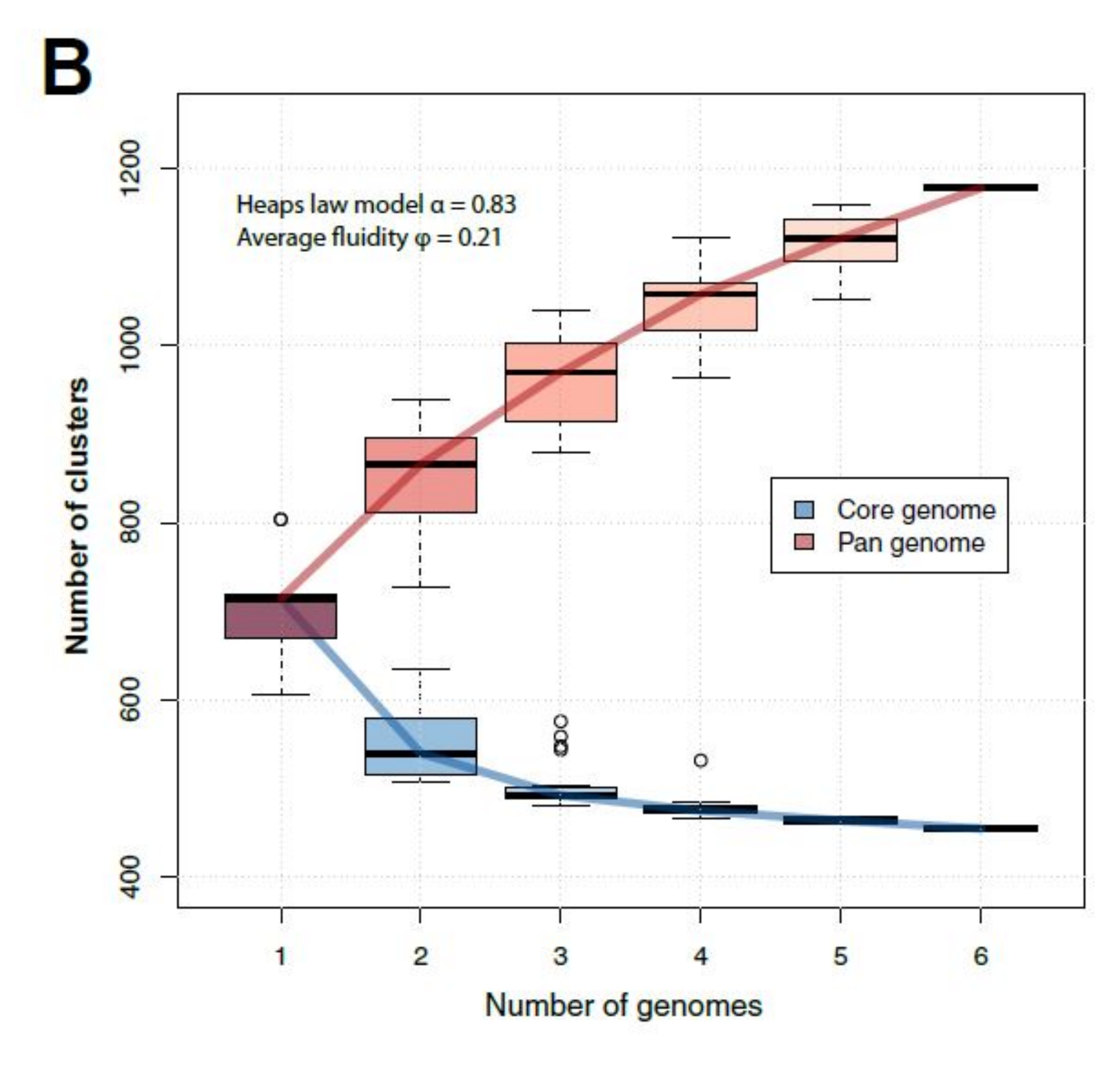


The Pandoravirus genomes are diverse

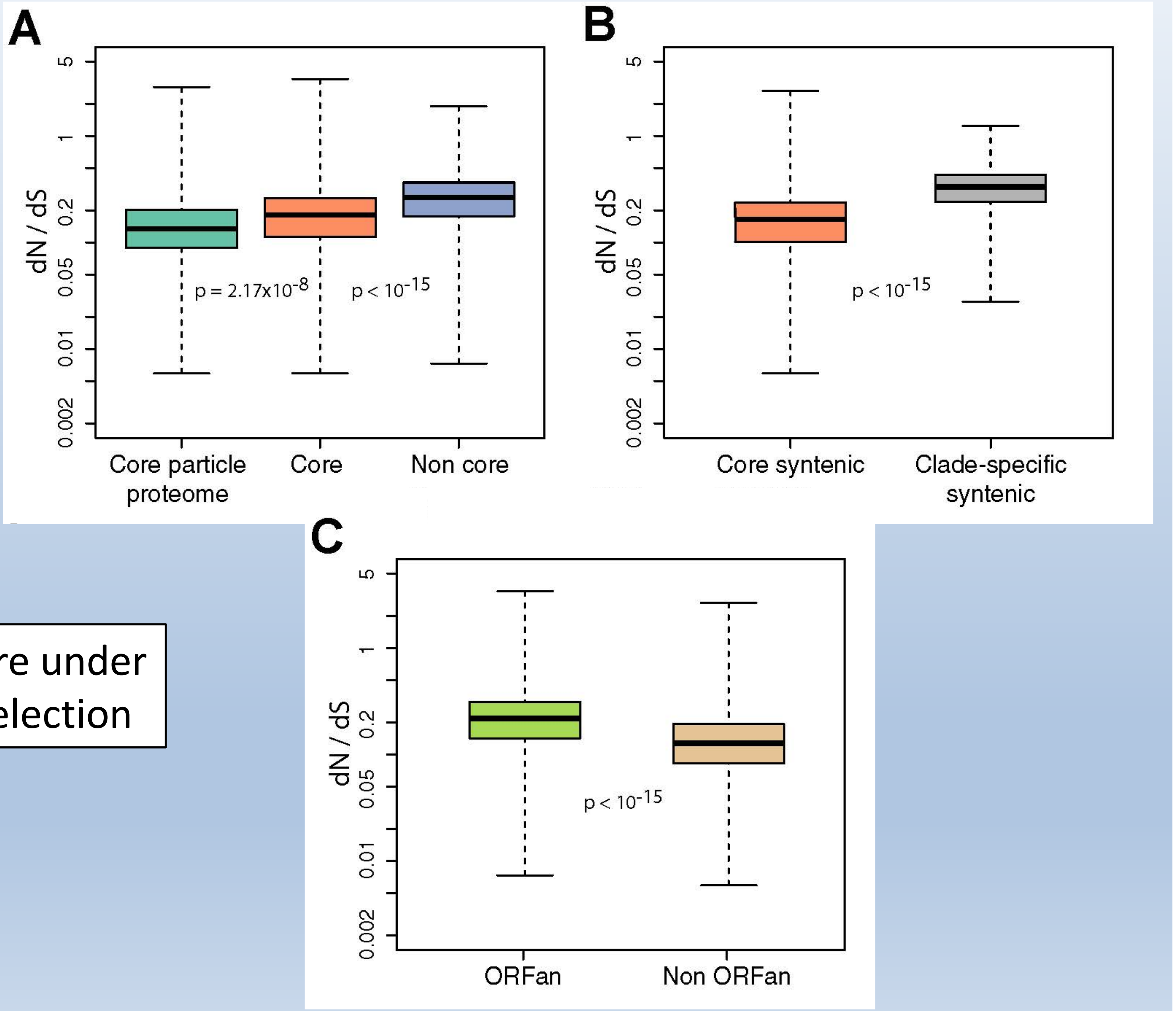


- core: 455 clusters
- strain-specific: 377 clusters

The Pandoraviridae pan genome is … open!

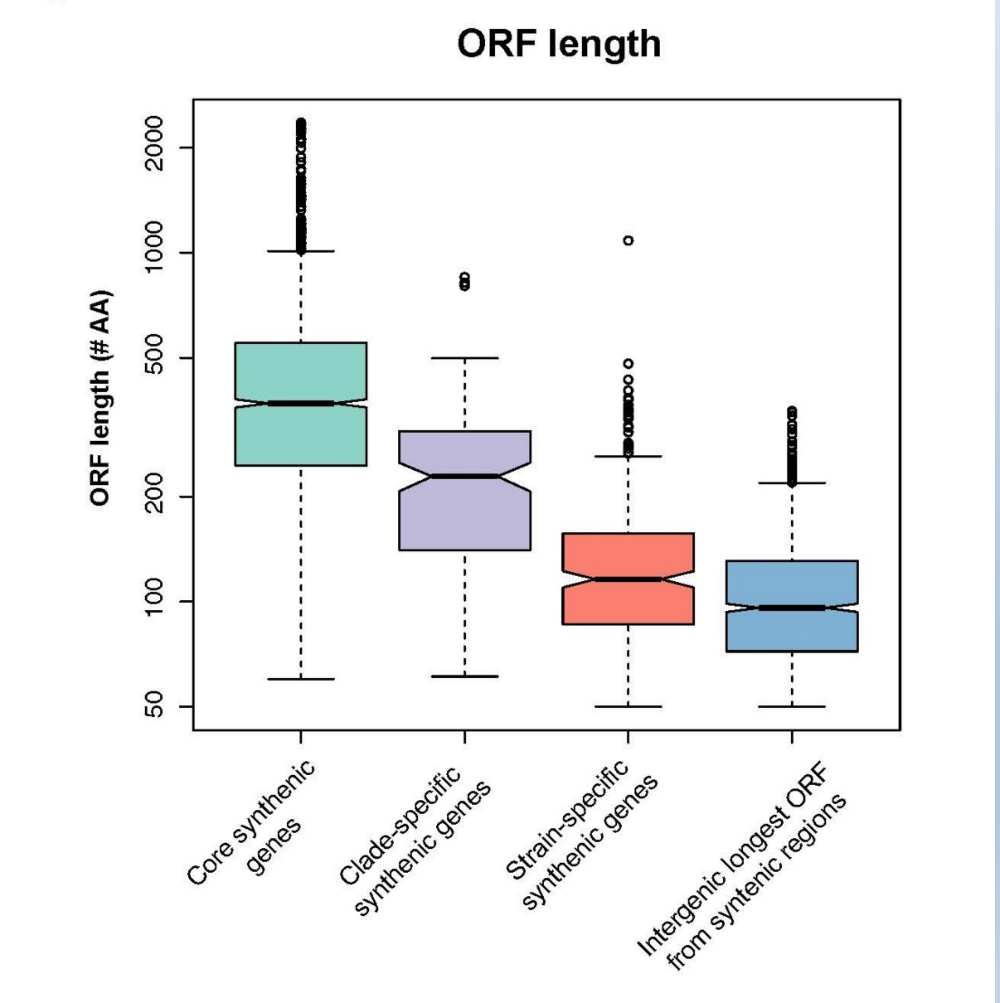


Gene categories: selection pressure

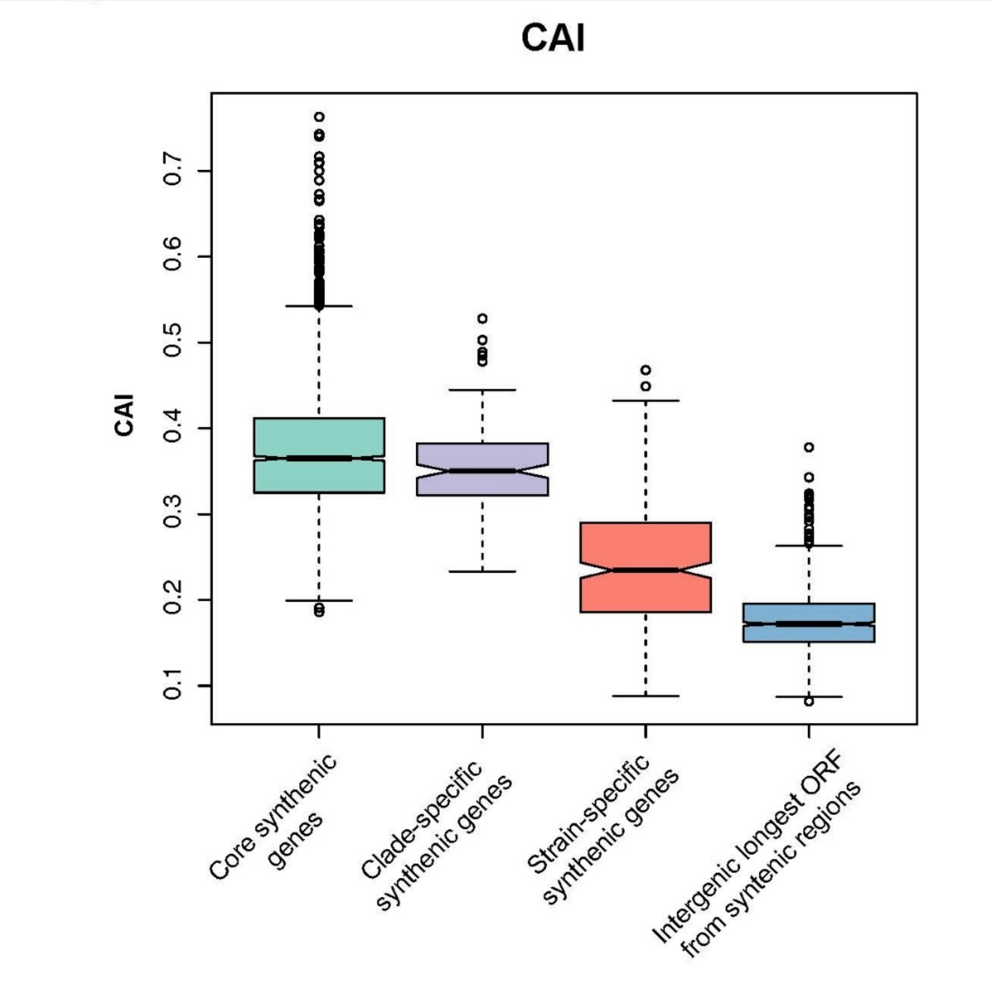


All genes are under purifying selection

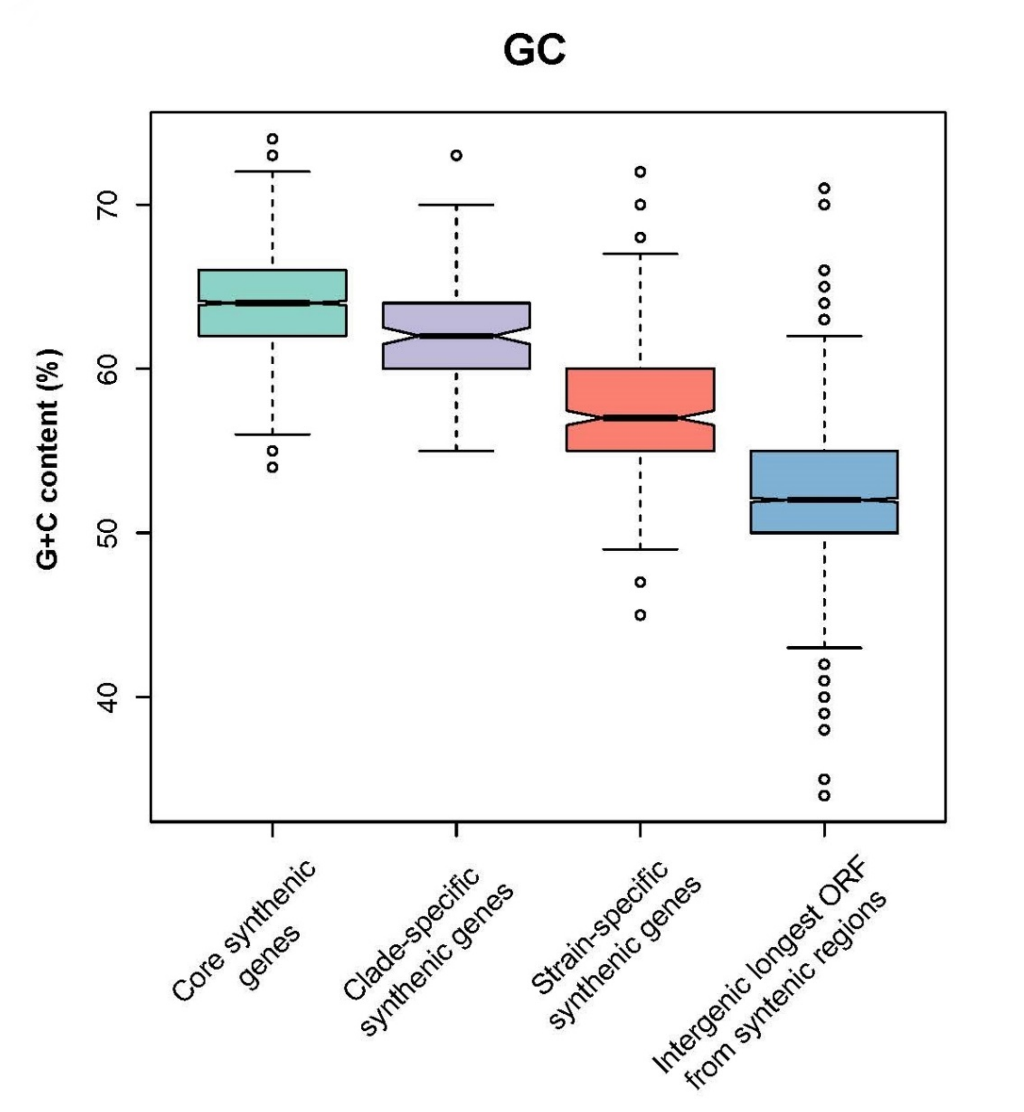
Strain-specific genes: statistical similarity with intergenic regions: 1) ORF length



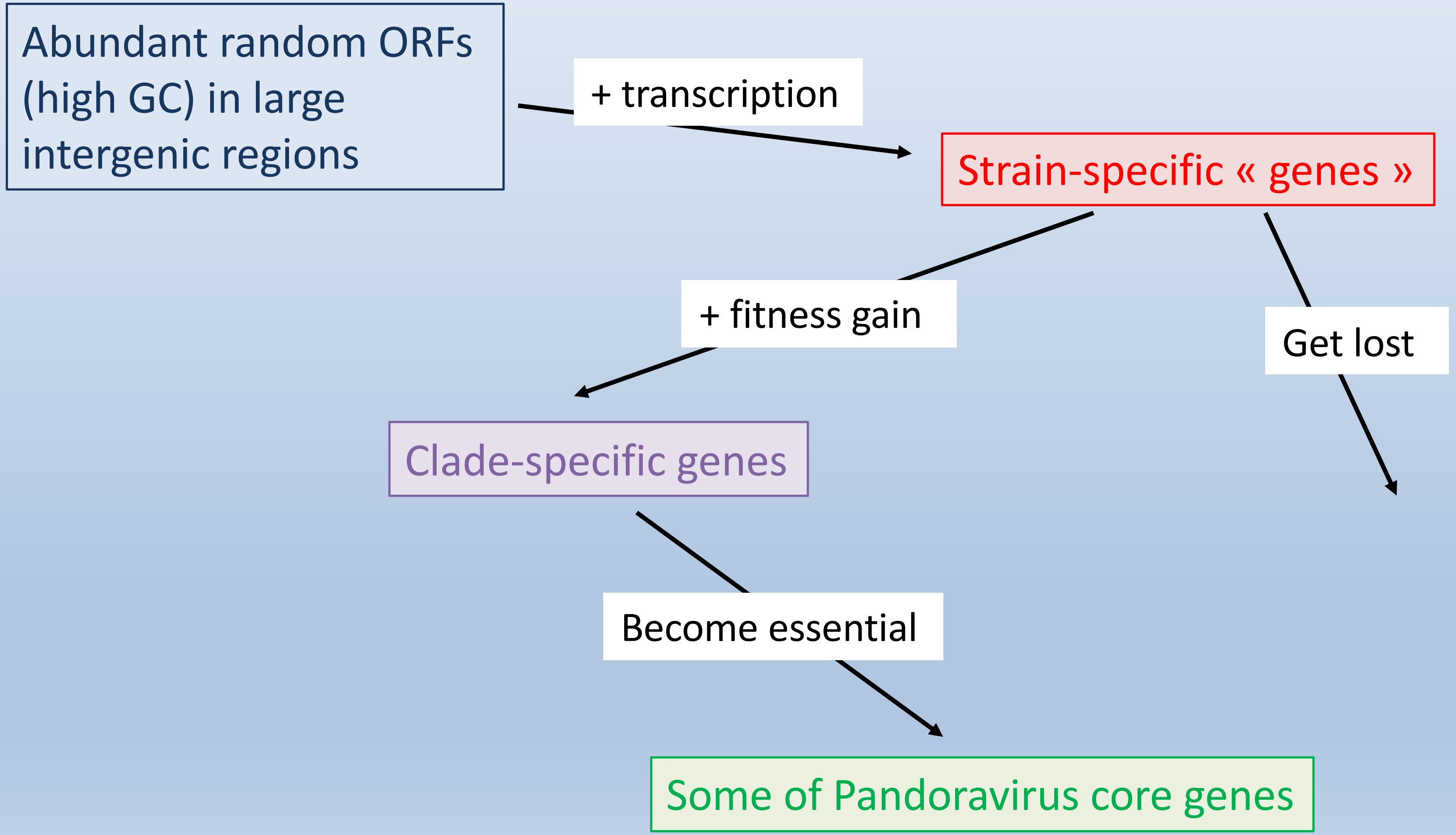
Strain-specific genes: statistical similarity with intergenic regions: 2) Codon adaptation



Strain-specific genes: statistical similarity with intergenic regions: 3) Base composition



The *de novo* gene creation scenario



Abundant random ORFs (high GC) in large intergenic regions

+ transcription

Strain-specific « genes »

+ fitness gain

Get lost

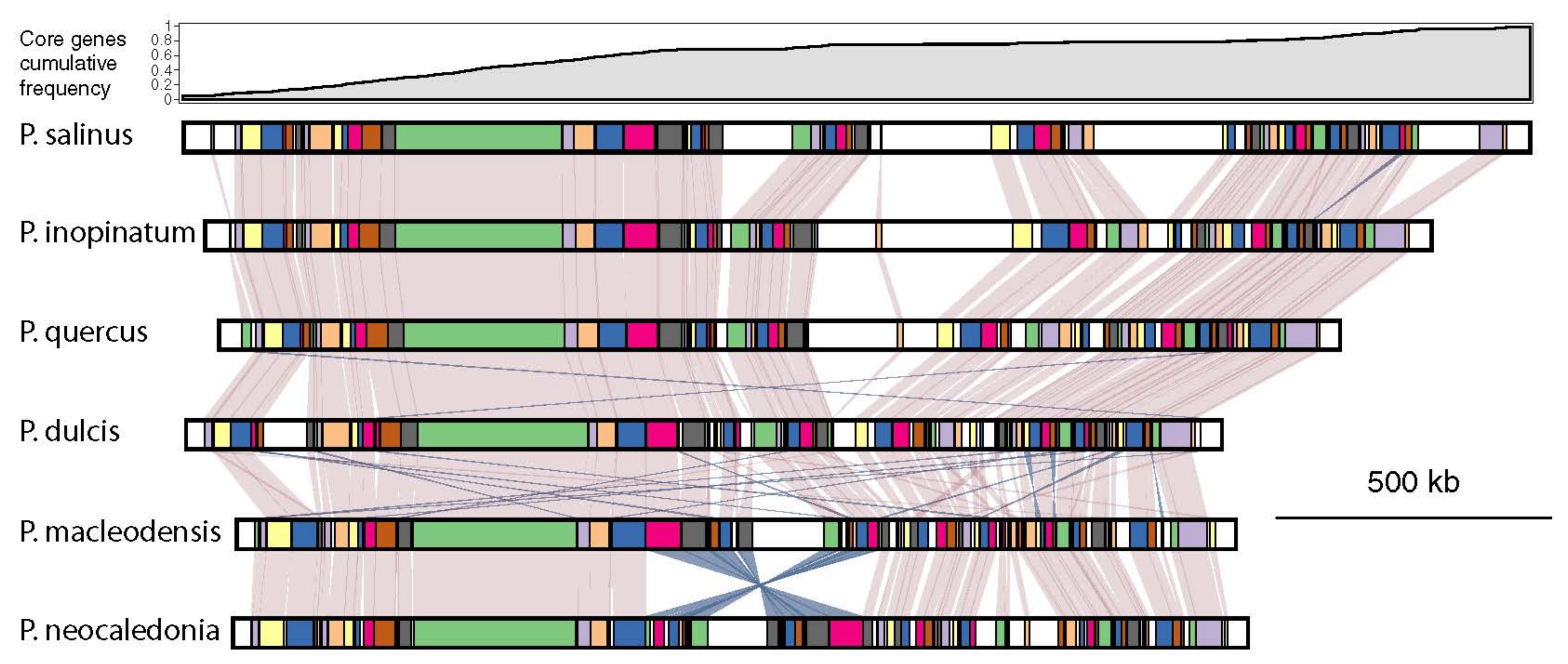
Clade-specific genes

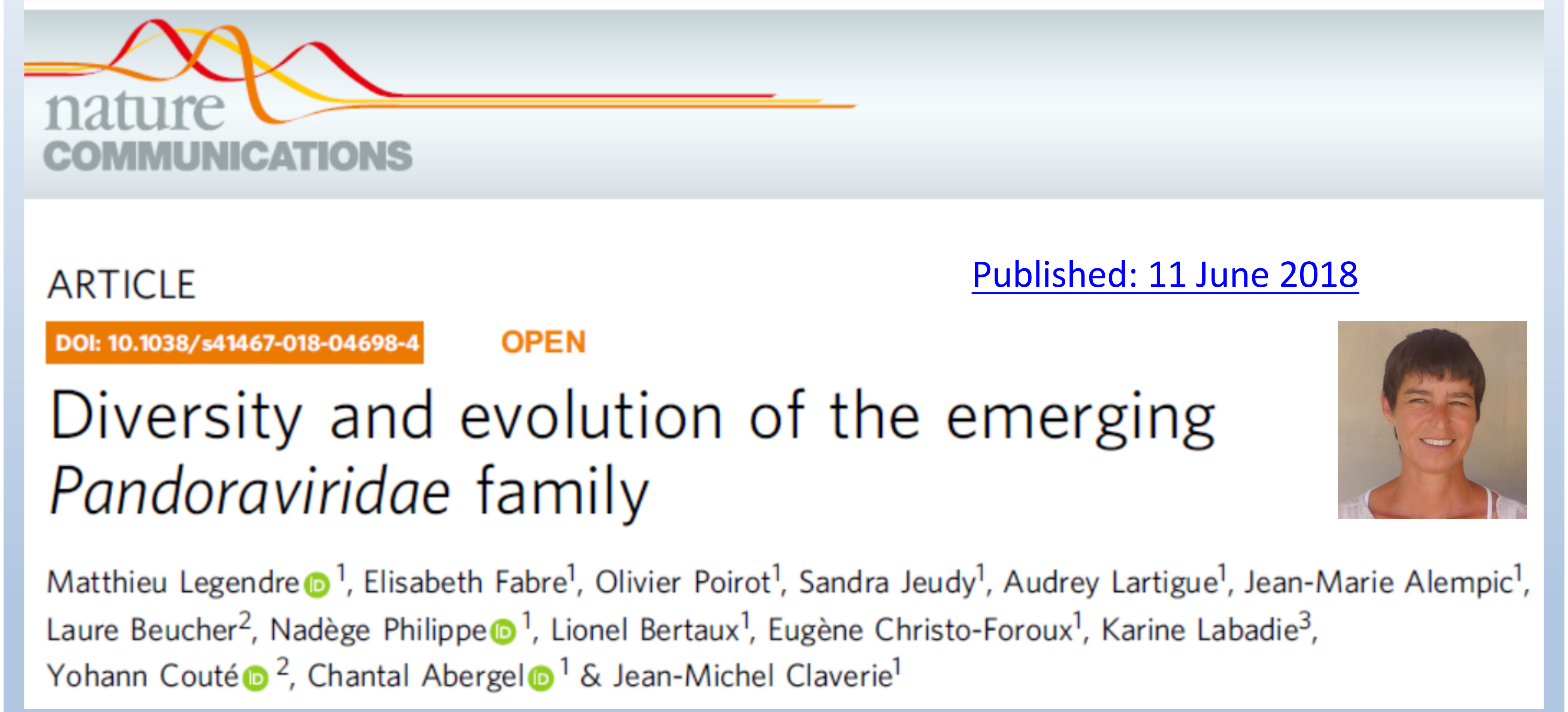
Become essential

Some of Pandoravirus core genes

The *de novo* gene creation scenario

would maintain the overall collinearity





nature COMMUNICATIONS

ARTICLE

Published: 11 June 2018

DOI: 10.1038/s41467-018-04698-4

OPEN

Diversity and evolution of the emerging *Pandoraviridae* family

Matthieu Legendre[1], Elisabeth Fabre[1], Olivier Poirot[1], Sandra Jeudy[1], Audrey Lartigue[1], Jean-Marie Alempic[1], Laure Beucher[2], Nadège Philippe[1], Lionel Bertaux[1], Eugène Christo-Foroux[1], Karine Labadie[3], Yohann Couté[2], Chantal Abergel[1] & Jean-Michel Claverie[1]

Pro/con arguments

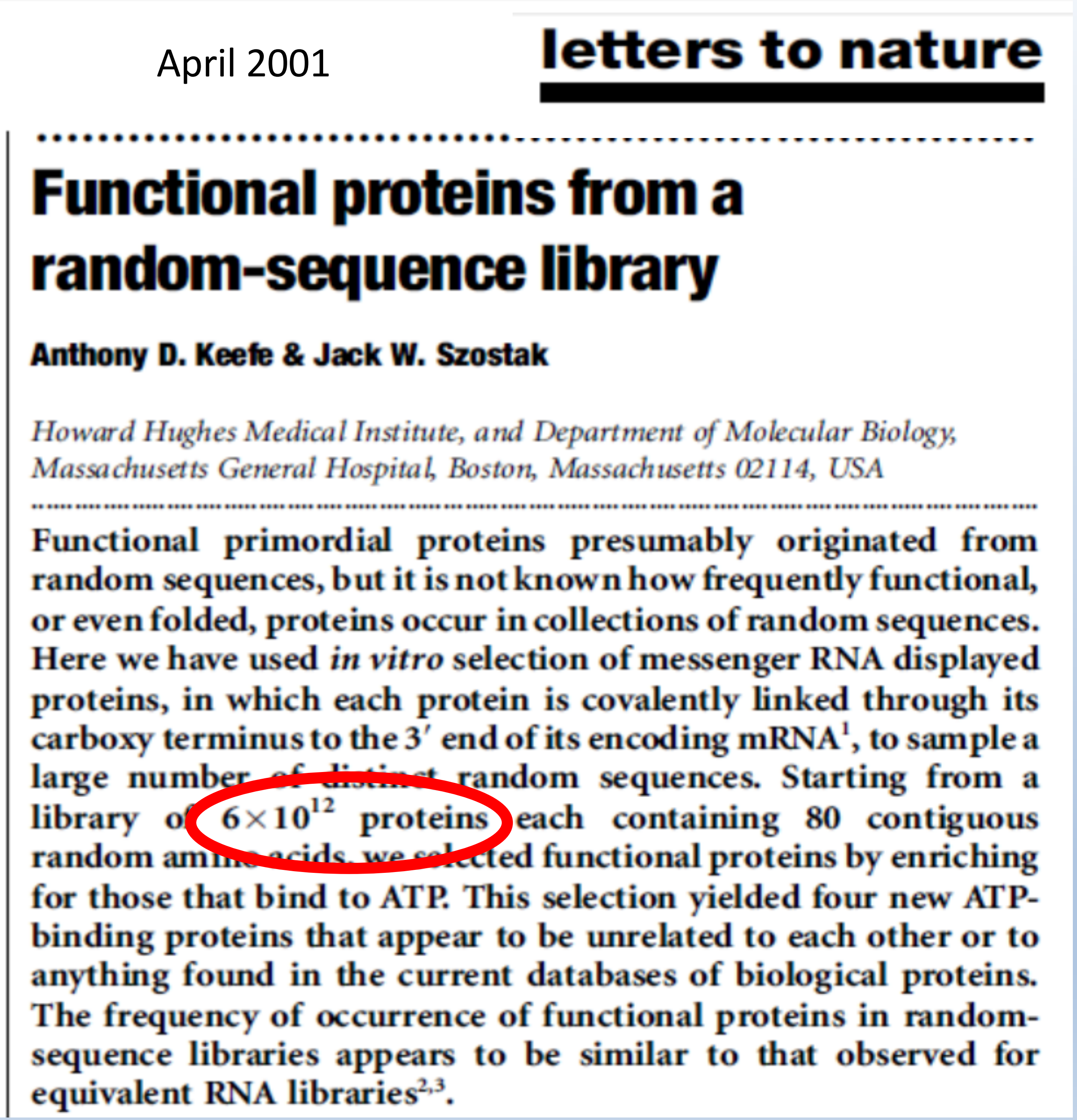
- Random aa sequences have a near zero propensity to fold
- Protein sequences made of a reduced set of aa fold better (high G+C)
- Non-structured proteins are detrimental (aggregates)
- Non-structured proteins make great regulatory components
- Random aa sequences have a 10^{-11} probability to have a function
- Gene without useful functions are quickly eliminated from parasites
- Viruses don’t care about wasting the host’s resources
- No mechanism is known to create « de novo » DNA sequences
- De novo DNA sequences creation had to happen once (!)

- Non-translated RNAs are detrimental, for some reasons
- Translation per se is beneficial (even in absence of function)

- Acquisition of function/fitness is much faster than we think it is
- Loss of useless gene is much slower than we think it is

Key statistics

	Mimivirus	Pandoravirus
G+C%	25	61
Bp/gene	1136	1750
Coding %	90	62-68
Max Size Random ORF/kb	90 aa	325 aa



April 2001

letters to nature

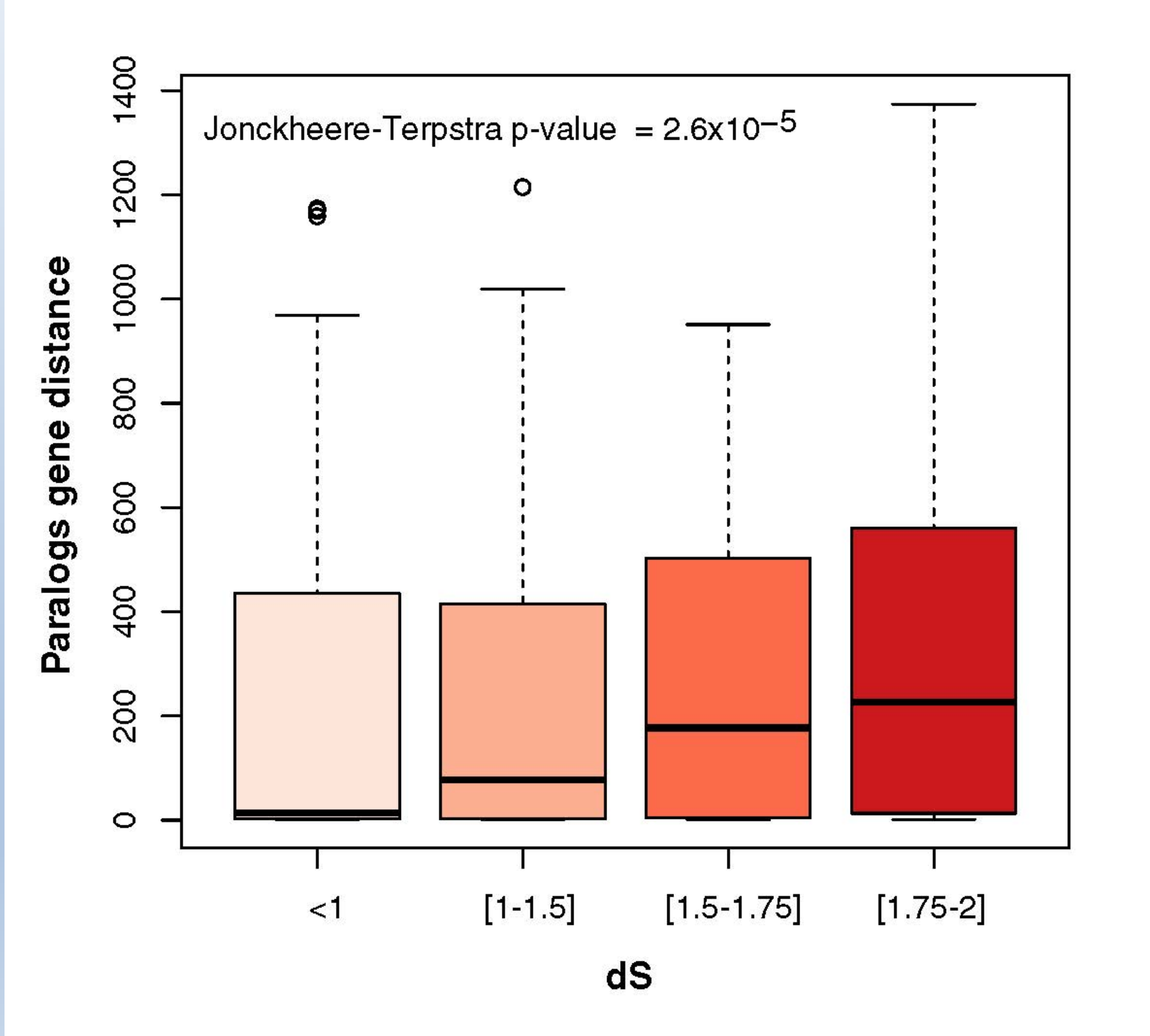
Functional proteins from a random-sequence library

Anthony D. Keefe & Jack W. Szostak

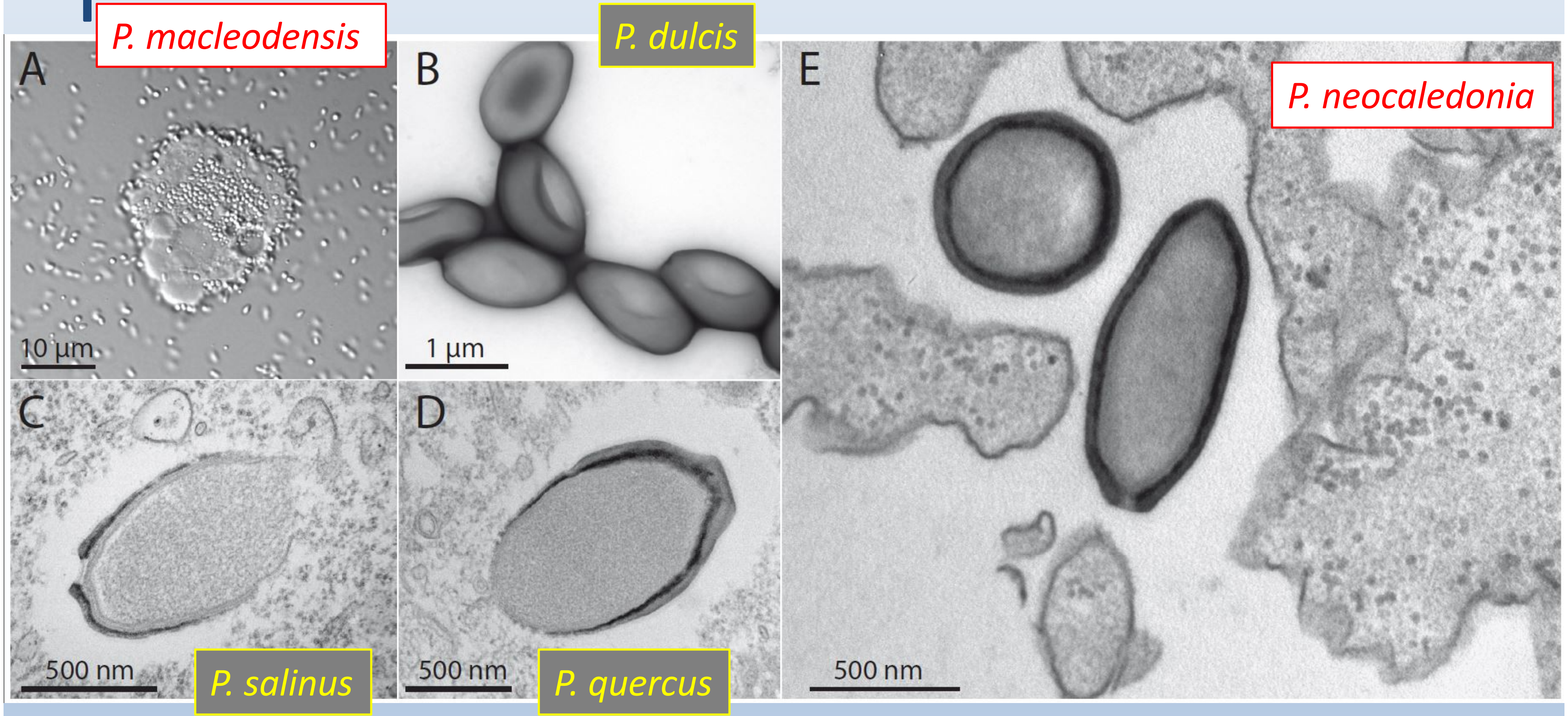
Howard Hughes Medical Institute, and Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA

Functional primordial proteins presumably originated from random sequences, but it is not known how frequently functional, or even folded, proteins occur in collections of random sequences. Here we have used *in vitro* selection of messenger RNA displayed proteins, in which each protein is covalently linked through its carboxy terminus to the 3′ end of its encoding mRNA[1], to sample a large number of distinct random sequences. Starting from a library of 6×10^{12} proteins each containing 80 contiguous random amino acids, we selected functional proteins by enriching for those that bind to ATP. This selection yielded four new ATP-binding proteins that appear to be unrelated to each other or to anything found in the current databases of biological proteins. The frequency of occurrence of functional proteins in random-sequence libraries appears to be similar to that observed for equivalent RNA libraries[2,3].

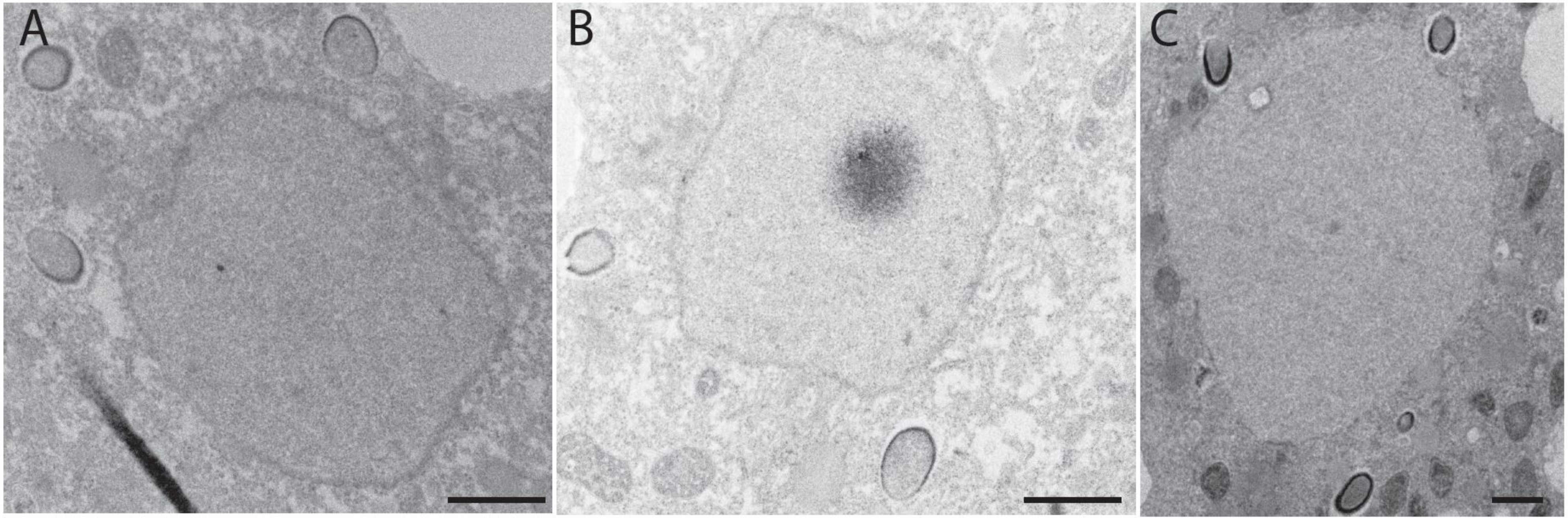
Paralogs divergence and distance correlate



6 isolates looking all the same



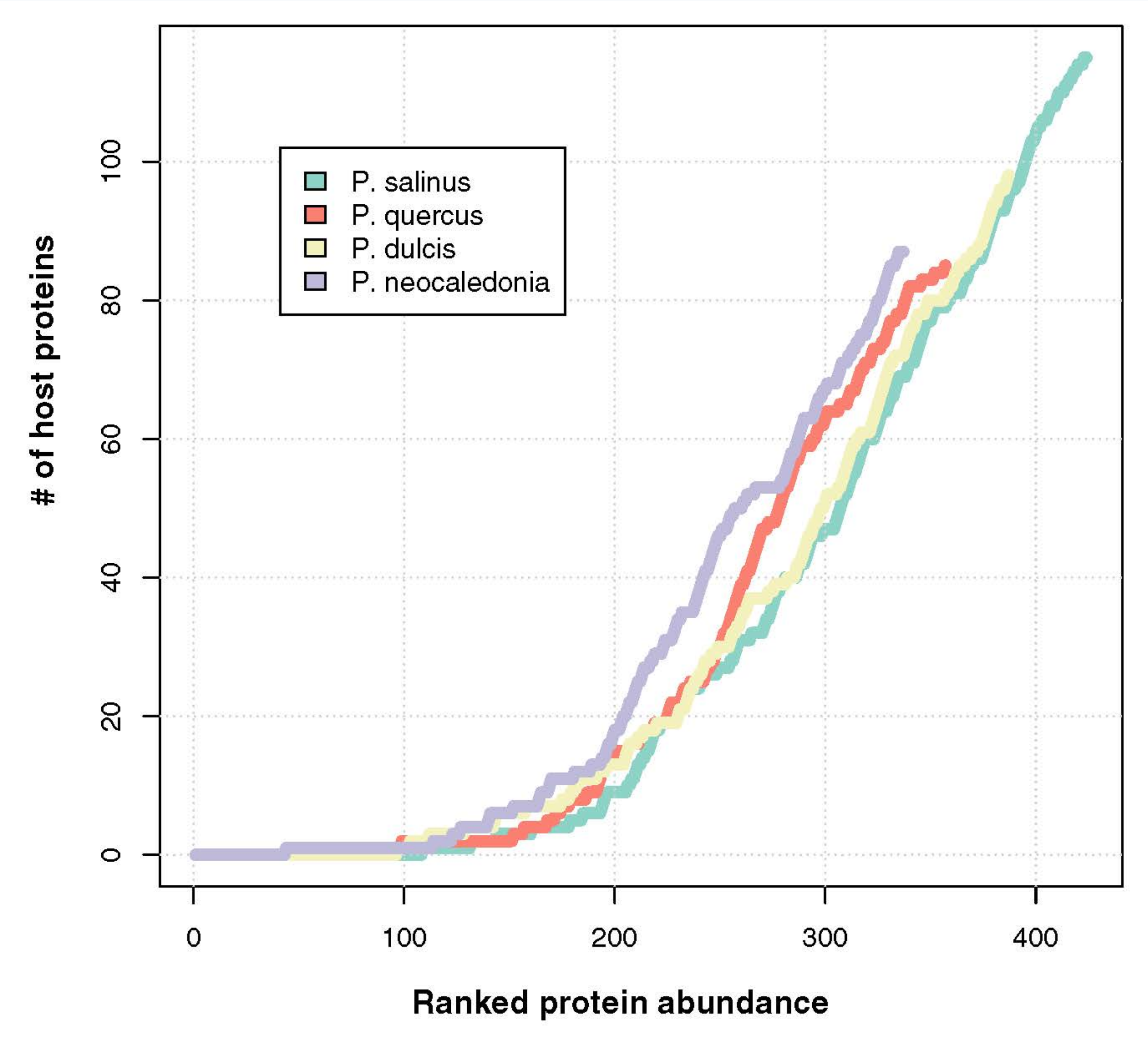
The nucleus is maintained to the end of the Pandoravirus infectious cycle



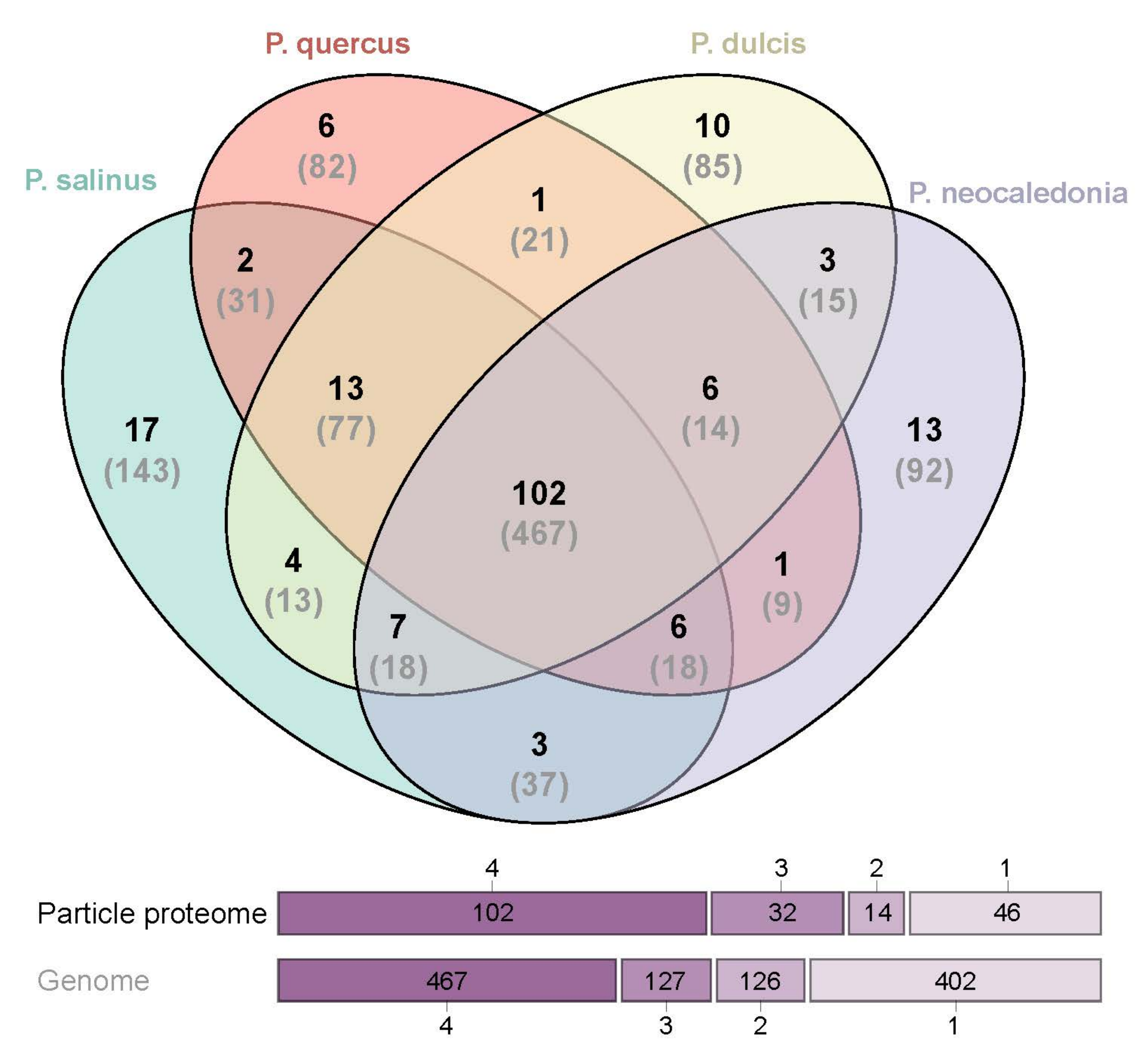
P. neocaledonia

P. salinus

The pandoravirion proteome is fuzzy

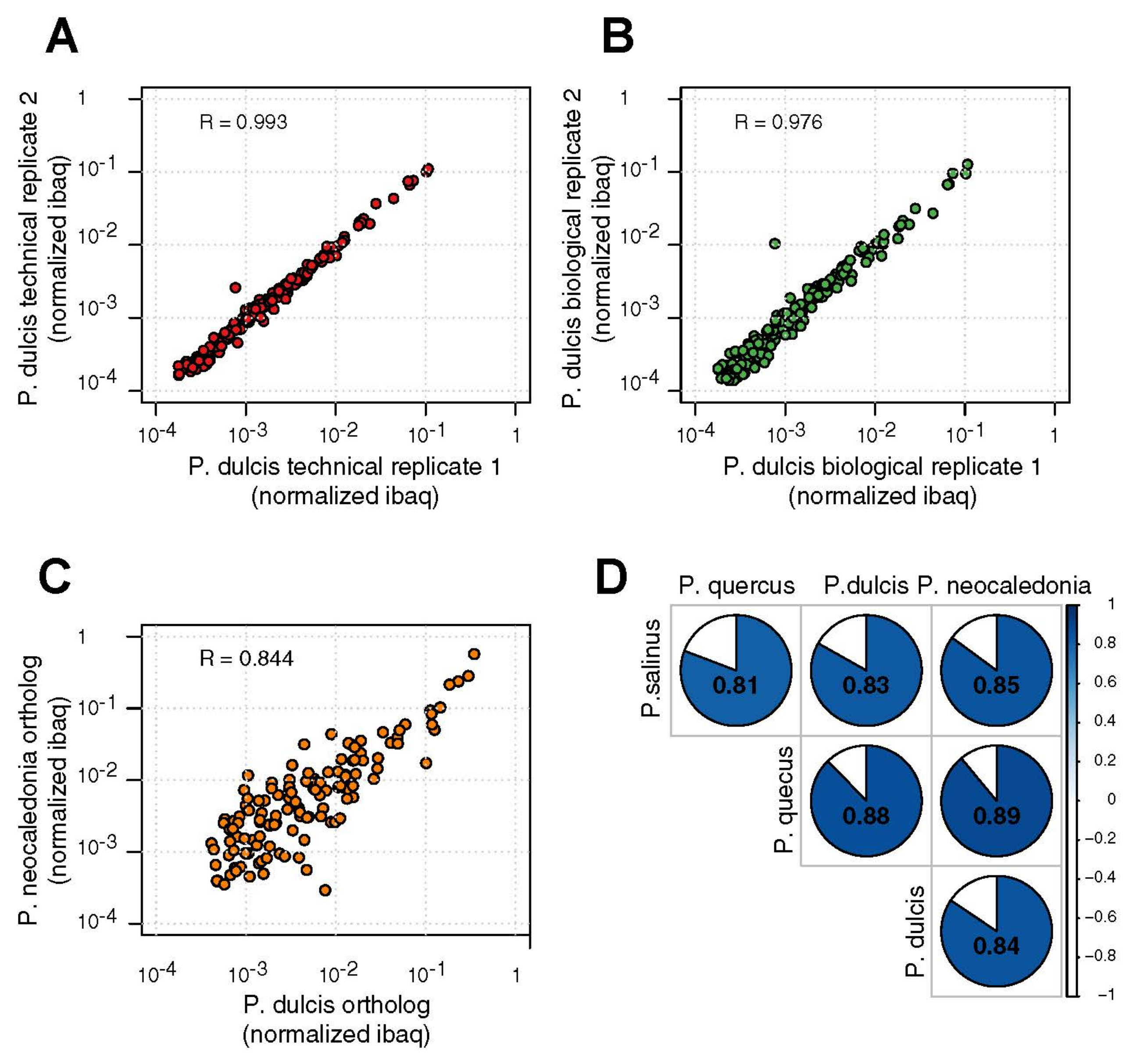


The Pandoravirions are more conserved than the genomes they propagate



52.6 % of core genes
versus
41.6% for the genomes

The Pandoravirus boxes are well conserved


A
R = 0.993
P. dulcis technical replicate 2 (normalized ibaq)
P. dulcis technical replicate 1 (normalized ibaq)
10^{-4}
10^{-3}
10^{-2}
10^{-1}
B
R = 0.976
P. dulcis biological replicate 2 (normalized ibaq)
P. dulcis biological replicate 1 (normalized ibaq)
C
R = 0.844
P. neocaledonia ortholog (normalized ibaq)
P. dulcis ortholog (normalized ibaq)
D
P. quercus
P.dulcis
P. neocaledonia
P.salinus
P. quecus
P. dulcis