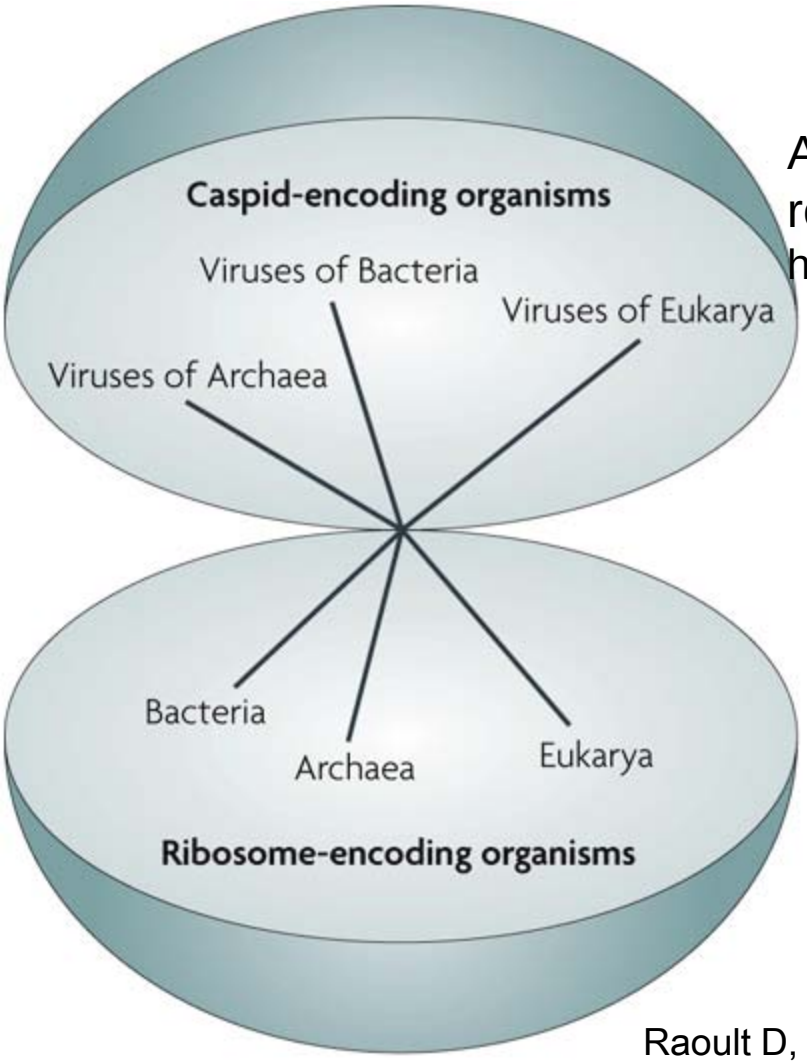
The Virus World, its evolution, evolution of antiviral defense, and the role of viruses in the evolution of cells

Eugene V. Koonin
National Center for Biotechnology Information, NIH, Bethesda

KITP, Santa Barbara , February 17, 2011

What is a virus?



A **virus** is a small infectious agent that can replicate only inside the living cells of organisms.
http://en.wikipedia.org/wiki/Virus

Viruses and virus-like agents *possess*:

- genomes
- very often –though not always – capsids that encase the genome

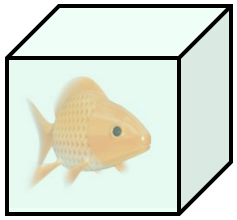
but lack:

- functional translation machinery
- membranes with transport/secretion systems

Raoult D, Forterre P. Redefining viruses: lessons from Mimivirus. Nat Rev Microbiol. 2008; 6(4):315-9

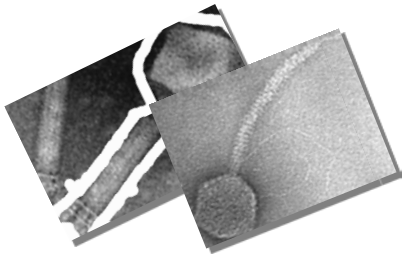
Nature Reviews | Microbiology

Viruses are the dominant entities in the biosphere – physically and genetically – as shown by viral metagenomics – virome studies



1 cm^3 of seawater contains 10^6-10^9 virus particles

Suttle, C.A. (2005) *Nature* **437**:356



There are millions of diverse bacteriophage species in the water, soil, and gut

Edwards and Rohwer (2005) *Nat. Rev. Microbiol.* **3**:504

- *Viruses are the most abundant biological entities in the biosphere: there are 10-100 virus particles per cell*
- *The pangenomes of viruses and cellular organisms have [at least] comparable complexities*

High Frequency of Horizontal Gene Transfer in the Oceans

Lauren D. McDaniel,[1]* Elizabeth Young,[1] Jennifer Delaney,[1] Fabian Ruhnau,[2] Kim B. Ritchie,[3] John H. Paul[1]

Microbes rely on mutation and the processes of horizontal gene transfer (HGT; conjugation, transformation, and transduction) to acquire new traits. Gene transfer agents (GTAs) discovered in the purple nonsulfur bacterium *Rhodobacter capsulatus* (formerly *Rhodopseudomonas capsulata*) are host-encoded viruslike elements that package random fragments of the host chromosome and are found in the genome of almost every sequenced member of the α-Proteobacteria order Rhodobacterales (*1*). To test whether GTAs are natural vectors of gene transfer, we grew nine strains of marine α-proteobacteria containing putative GTA cassettes (table S1) and screened them for the production of GTA-like particles.

Both *Roseovarius nubinhibens* ISM and the isolate *Reugeria mobilis* 45A6 reproducibly produced putative GTA particles during stationary phase growth. We then generated genetically marked donor strains of *R. nubinhibens* and *R. mobilis* containing the transposon Tn5. GTA production in these marked donor strains was equivalent to that of the wild-type strains. To document gene transfer frequencies, we subjected wild-type strains or natural communities from a range of environments to treatment with donor strain GTAs and documented the rates of GTA-mediated gene transfer of kanamycin resistance (fig. S1). In the coral reef environment, spontaneous kanamycin resistance was 4.6×10^{-4}, whereas the GTA-mediated frequency was significantly higher at 2.5×10^{-2} ($P = 0.028$, Student's t test).

For this experiment, both spontaneous mutants and GTA treatments were examined for the presence of the Tn5 streptomycin kinase gene. A total of 47% of the GTA-treated viable colonies but none of the spontaneous revertants contained the gene. That 53% of the putative transductants did not contain the gene is not suprising because these may have contained only the kanamycin resistance gene (*npt*II) and not the flanking streptomycin kinase gene.

The recovery of the streptomycin kinase sequence, which is ~1000 base pairs (1 kbp) from the active site of the kanamycin resistance gene, suggested that up to 1 kbp of the central region of Tn5 was transferred. This is consistent with extracted DNA from the GTAs, which ranged from about 500 to 1000 bp in length (fig. S3). No spontaneous double antibiotic (kanamycin and streptomycin) resistance was detected, and the GTA-mediated frequency of 1.06×10^{-4} was significantly higher ($P = 0.023$). The Tn5 streptomycin kinase sequence was recovered in 1 in 10 viable double antibiotic–resistant strains, suggesting that modifications, truncations, or rearrangements may have occurred, as in natural transformation (*2*).

Similar frequencies of transfer were observed among differing environments (Table 1), demonstrating that cultivated GTAs transduce natural communities of marine bacteria. The 16*S* ribosomal RNA sequences examined showed that the majority of natural GTA recipients were most similar to marine *Flavobacterium* or *Flexibacter* strains (table S2), consistent with the prior reports of abundant *Flavobacterium* in marine systems (*3*).

R. nubinhibens contains both a GTA and an inducible prophage (*4*). Transmission electron microscopy (TEM) demonstrated that *R. nubinhibens*–induced prophage preparations contained tailed phage (*4*), whereas GTA particles were nontailed (fig. S2A), resembling the GTA of *Silicibacter pomeroyi* (*5*). In contrast to the GTA particles, the purified prophages of *R. nubinhibens* had no gene transfer activity. Additionally, maximal expression of the *R. nubinhibens* GTA terminase gene cooccurred with maximal GTA production (fig. S4). TEM of GTAs of *R. mobilis* revealed tailed viral particles (fig. S2B).

GTA dose, or multiplicity of infection (MOI), was linearly correlated with increased resistance to antibiotics (MOI range from 0.01 to 10, $R^2 = 0.9593$), which enabled extrapolations of gene transfer frequencies to natural systems (*6*).

GTAs from *R. nubinhibens* ISM show a wide host range and interspecific gene transfer under ecologically relevant conditions. Environmental gene transfer frequencies ranging from 6.7×10^{-3} to 4.7×10^{-1} (Table 1) are 1900 to 459 million times the frequency for transformation (*2*) and 650,000 to 31 million times the frequency of transduction previously measured in the marine environment (*7*). These results suggest a genomic flexibility in marine microbial populations that facilitates their adaptation to changing environmental conditions.

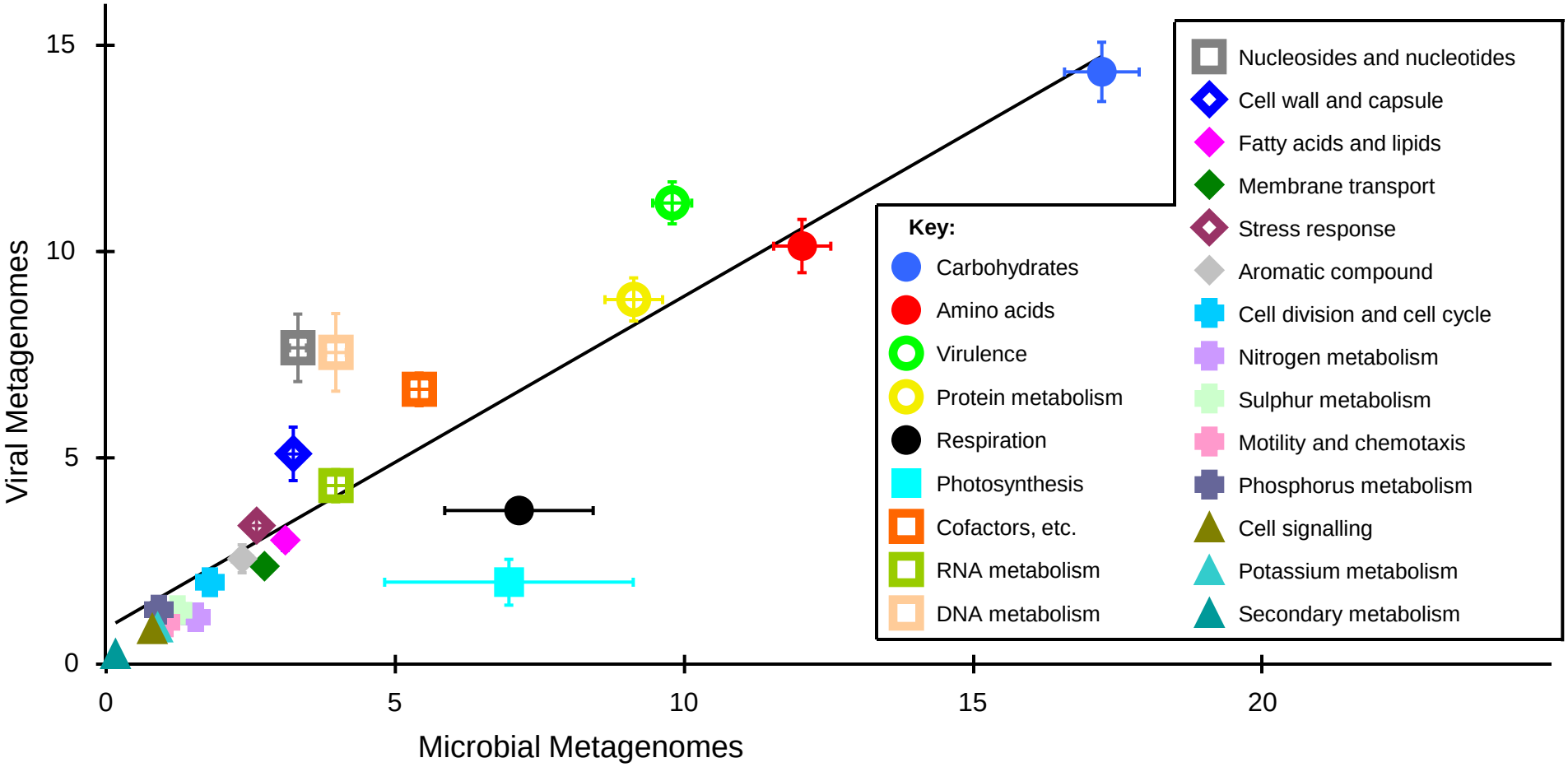
Table 1. Frequencies of transfer of marker genes to both cultured and natural communities. N/A indicates not applicable; BDL, below detection limit.

Environment	Avg. spontaneous	Range	Avg. GTA-	Range	Number

References and Notes

1. A. S. Lang, J. T. Beatty, *Trends Microbiol.* **15**, 54 (2007).
2. H. G. Williams, J. Benstead, M. E. Frischer, J. H. Paul, *Mol. Mar. Biol. Biotechnol.* **6**, 238 (1997).
3. T. Woyke *et al.*, *PLoS ONE* **4**, e5299 (2009).
4. Y. L. Zhao *et al.*, *Appl. Environ. Microbiol.* **76**, 589 (2010).
5. E. J. Biers *et al.*, *Appl. Environ. Microbiol.* **74**, 2933 (2008).

Mean % of sequences with matches to major functional categories



Most of the viromes might not even consists of typical viruses but rather of pseudovirus particles that carry microbial genes (GTAs)

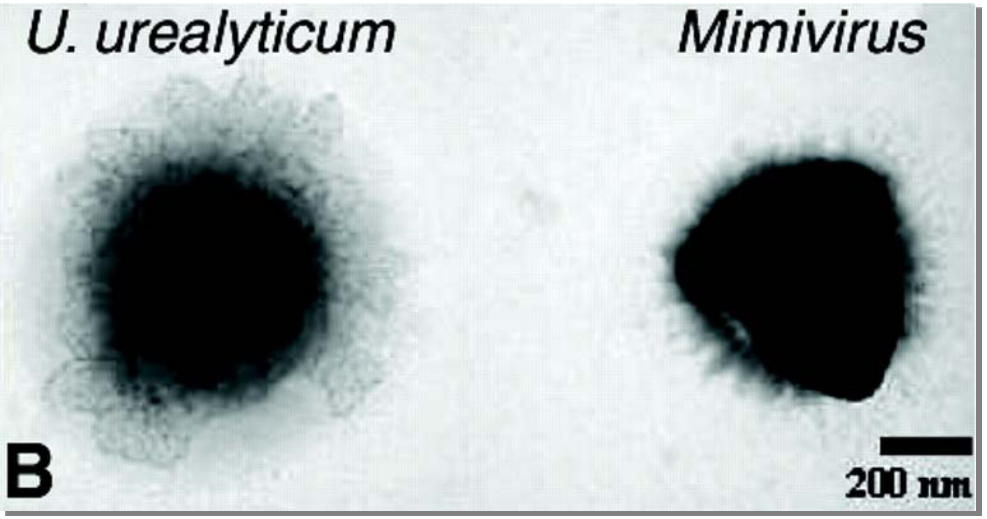
Kristensen, Mushegian, Dolja, Koonin Trends Microbiol 2010

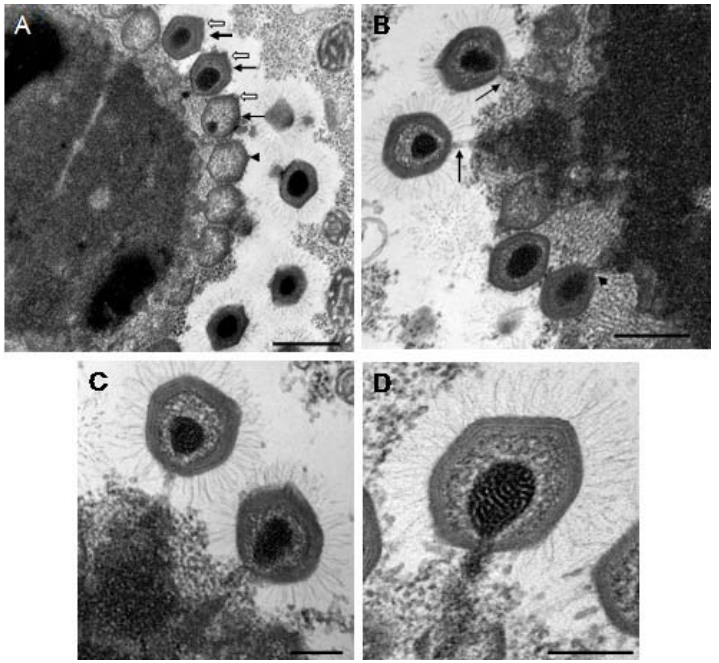
Comparative genomics shows that viruses that cause human diseases belong to families that evolved hundreds of millions or even billions years ago:

Viruses accompany the evolving cellular life throughout its history and might even predate it

The largest, most complex viruses: NCLDV (Nucleo-Cytoplasmic Large DNA viruses of eukaryotes) – this is where the smallpox virus belongs

Some viruses are comparable to cellular life forms in size and genetic complexity!





Mimivirus genome (~1.2 Mbp, ~1,000 genes) is twice as large as that of *Mycoplasma genitalium* (580 kbp; ~500 genes)

Raoult et al. Science. 2004 Nov 19;306(5700):1344; Suzan-Monti et al. PLoS ONE. 2007 Mar 28;2(3)

The largest, most complex viruses: the Nucleocytoplasmic Large DNA Viruses (NCLDV) (this is where the smallpox virus AND the mimivirus belong)

6 families of NCLDV…and counting

	#	*size, kb*	*hosts*
-poxviridae	**26**	**[134-360]**	**vertebrates, insects**
-asfarviridae	1	[170]	vertebrates, protists(?)
-iridoviridae	8	[103-212]	vertebrates, insects, protists(?)
-ascoviridae	4	[119-174]	insects
-phycodnaviridae	9	[155-407]	algae, haptophytes, stramenopiles
-mimiviridae	2	[1181-1200]	amoebozoa, algae(?)
-[Marseille virus]	1	[368]	amoebozoa – new family?

Iyer, Aravind, Koonin,
Common origin of four diverse families of large eukaryotic DNA viruses. J. Virol. 2001, 75: 11720
Iyer et al. **Evolutionary genomics of nucleo-cytoplasmic large DNA viruses.** Virus Res. 2006 Apr;117(1):156

The case for the monophyly (common origin) of NCLDV

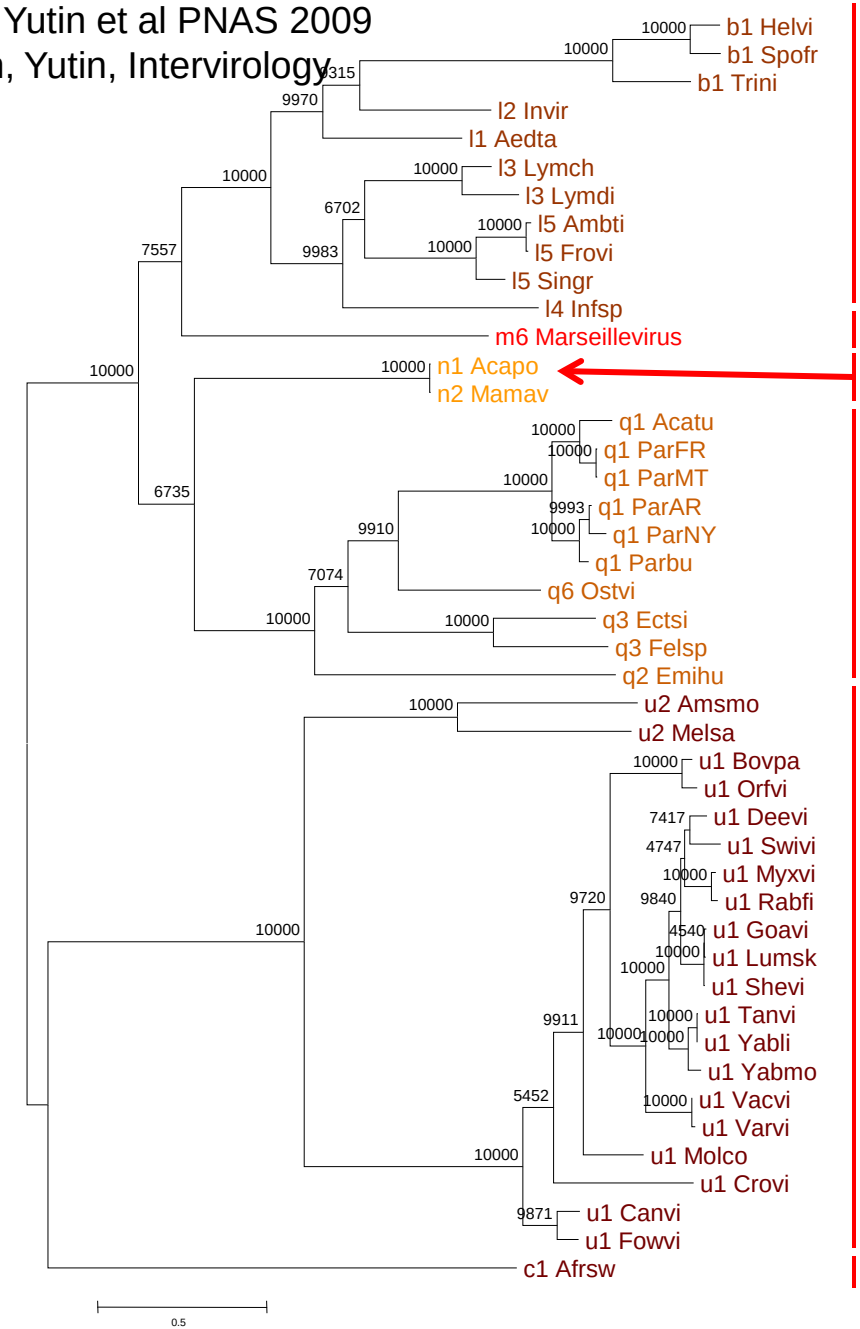
9 universally conserved hallmark genes (vaccinia gene names):
-primase (D5-N)
-helicase (D5-C)
-DNA polymerase (E9)
-packaging ATPase (A32)
-Major capsid protein (D13, non-capsid in poxviruses)
-Thiol-oxidoreductase (E10)
-Helicase (D6, D11)
-S/T protein kinase (F10)
-Transcription factor VLTF2 (A1)

47 genes mapped to the last common ancestor of NCLDV
by maximum likelihood – all main functional systems represented

Iyer et al. J. Virol. 2001; Virus Res. 2006; Yutin et al. Virol J. 2009

Phylogeny of NCLDV based on concatenation of 4 universal genes: primase-helicase, DNA polymerase, packaging ATPase, VLTF2 transcription factor

Boyer, Yutin et al PNAS 2009
Koonin, Yutin, Intervirology 2010



Ascoviridae
&
Iridoviridae

Marseillevirus
Mimiviridae

Phycodnaviridae

Poxviridae

Asfarviridae

HOST
Animals
+ diverse protists

Amoebozoa, Algae, animals(?)

Chlorophytes, Haptophytes, stramenopiles

Animals

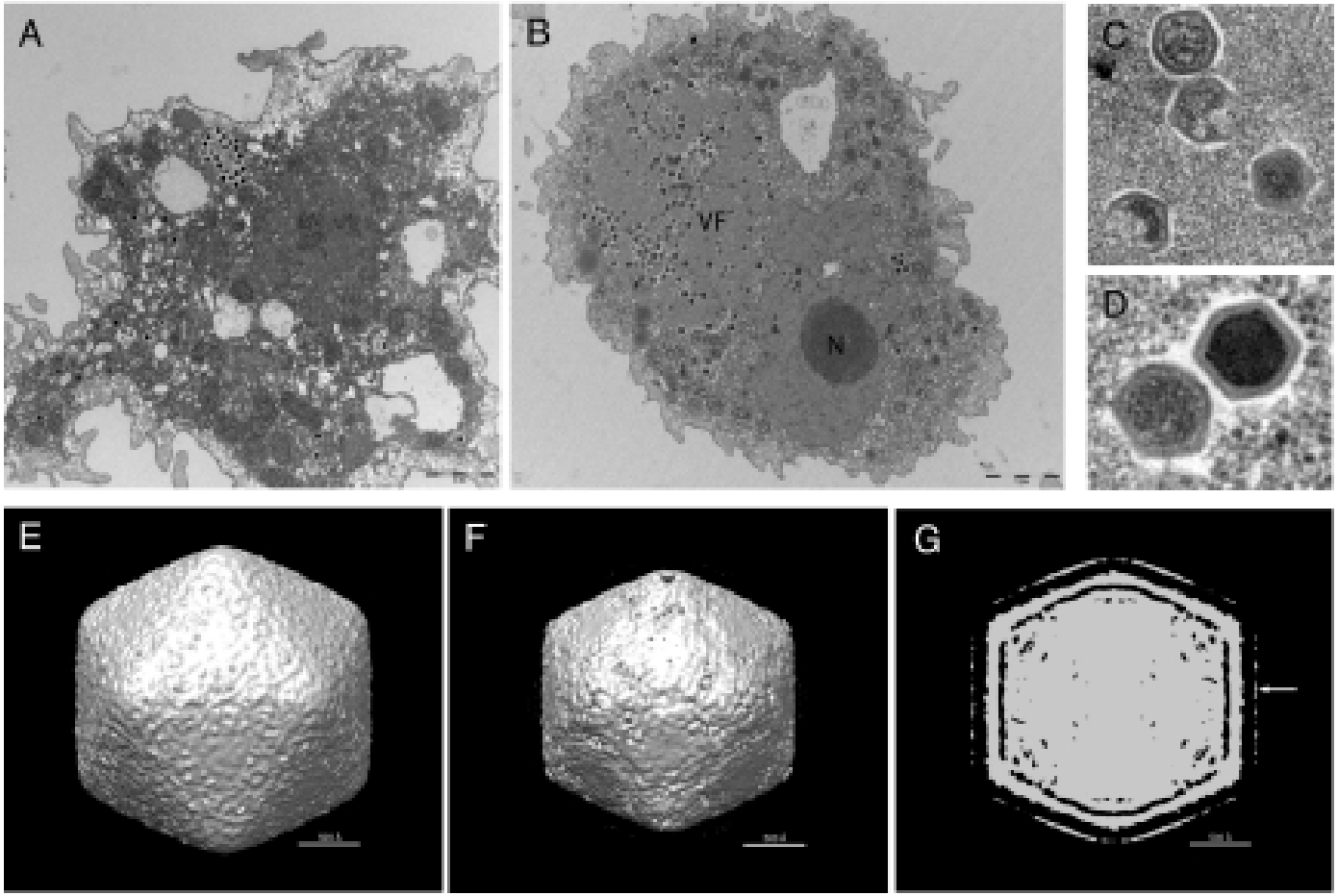
Animals, haptophytes

- **Divergence of NCLDV most likely antedates divergence of eukaryotic supergroups**
- **Alternative/addition: extensive horizontal transfer of viruses**

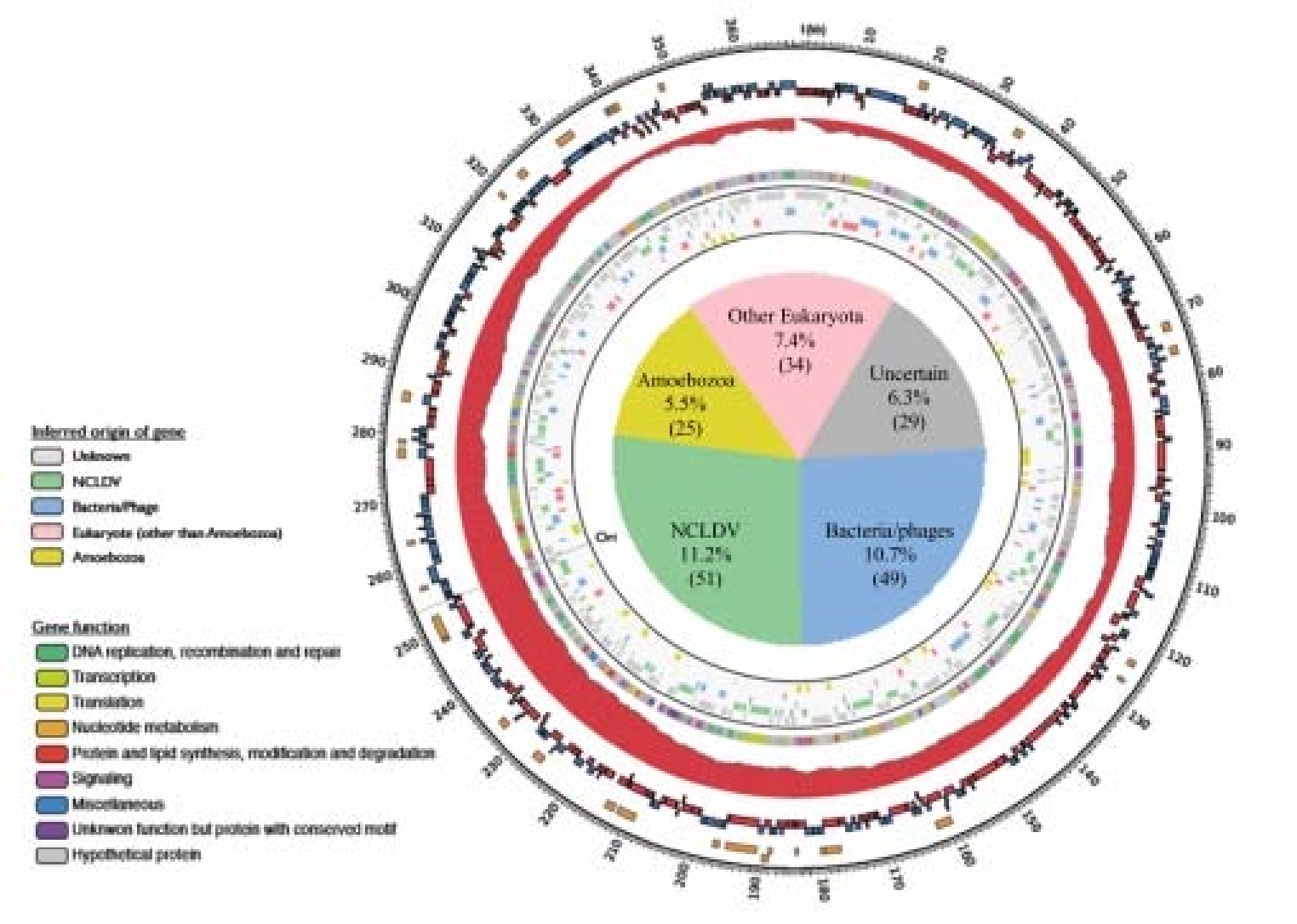
Boyer M, Yutin N, Pagnier I, Barrassi L, Fournous G, Espinosa L, Robert C, Azza S, Sun S, Rossmann MG, Suzan-Monti M, La Scola B, Koonin EV, Raoult D
PNAS 2009;106(51):21848-53.

Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms.

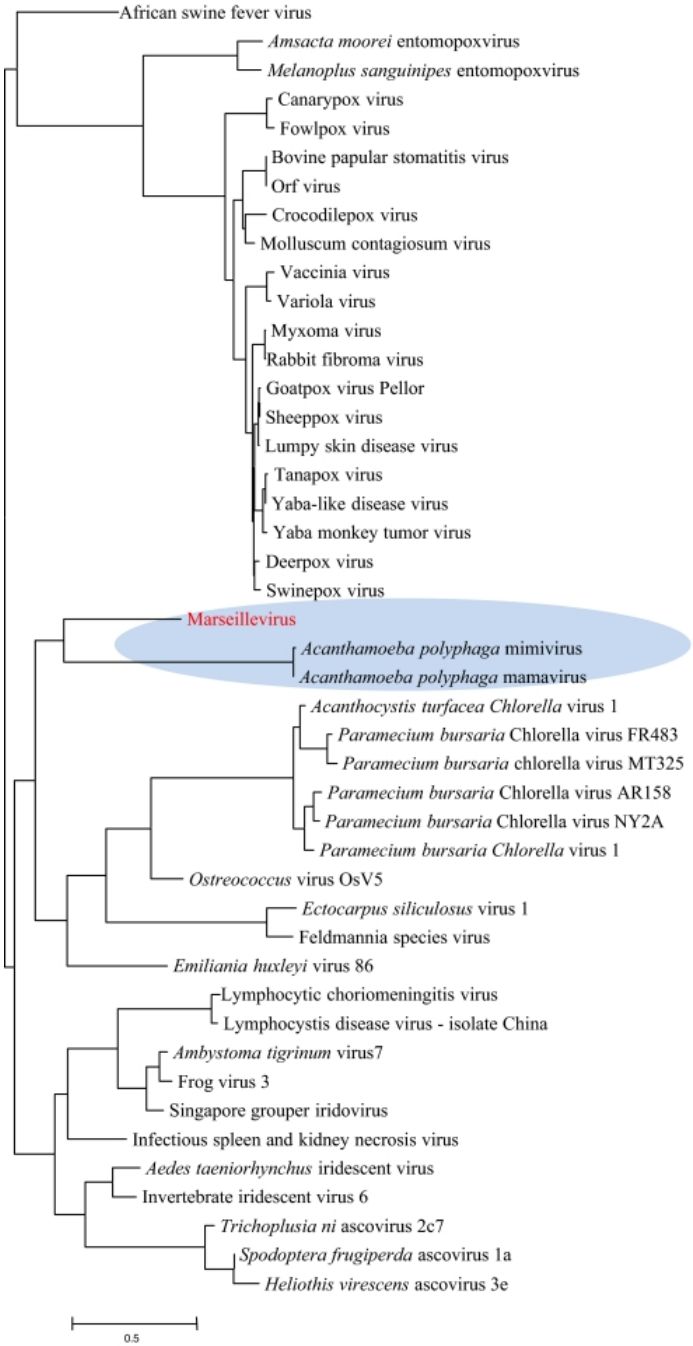
Giant viruses such as Mimivirus isolated from amoeba found in aquatic habitats show biological sophistication comparable to that of simple cellular life forms and seem to evolve by similar mechanisms, including extensive gene duplication and horizontal gene transfer (HGT), possibly in part through a viral parasite, the virophage. We report here the isolation of "Marseille" virus, a previously uncharacterized giant virus of amoeba. The virions of Marseillevirus encompass a 368-kb genome, a minimum of 49 proteins, and some messenger RNAs. Phylogenetic analysis of core genes indicates that Marseillevirus is the prototype of a family of nucleocytoplasmic large DNA viruses (NCLDV) of eukaryotes. The genome repertoire of the virus is composed of typical NCLDV core genes and genes apparently obtained from eukaryotic hosts and their parasites or symbionts, both bacterial and viral. We propose that amoebae are "melting pots" of microbial evolution where diverse forms emerge, including giant viruses with complex gene repertoires of various origins.


A
B
VF
N
C
D
E
F
G

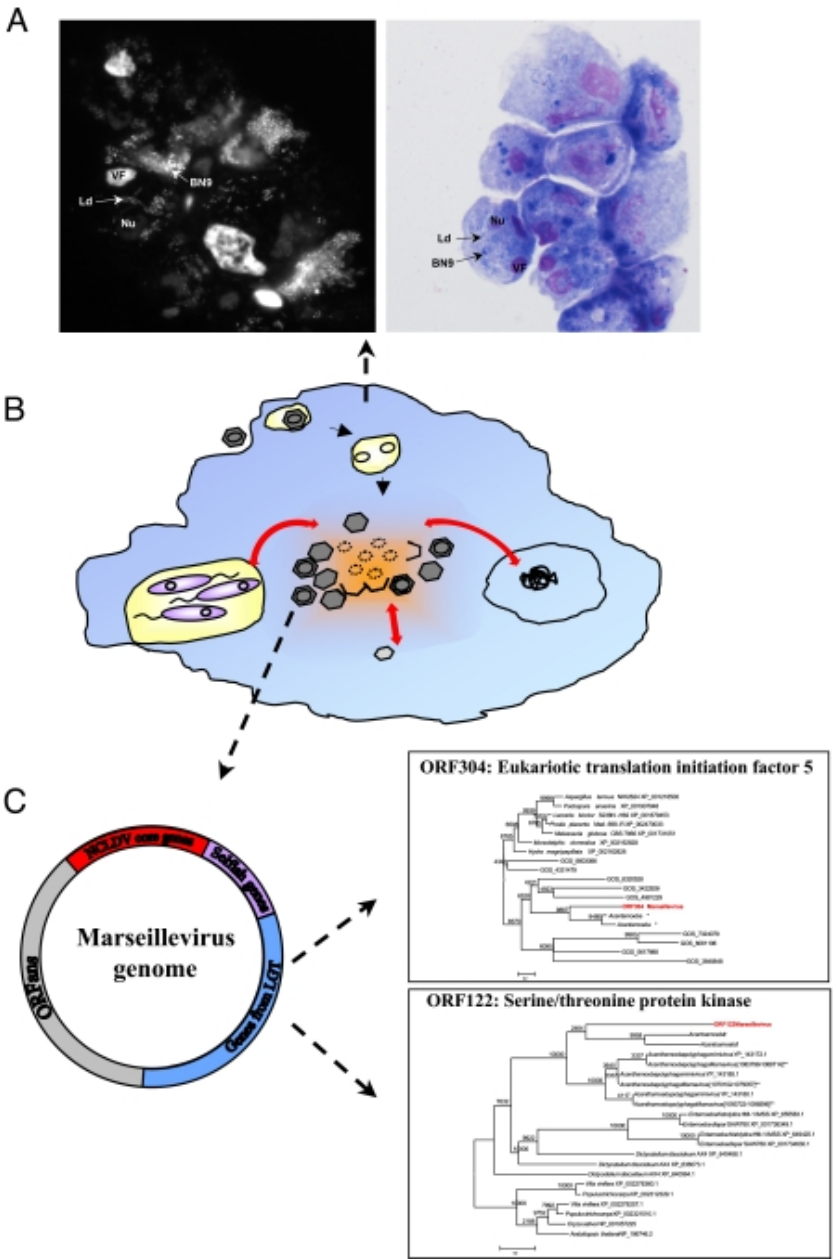
The mosaic composition of the Marseille virus genome



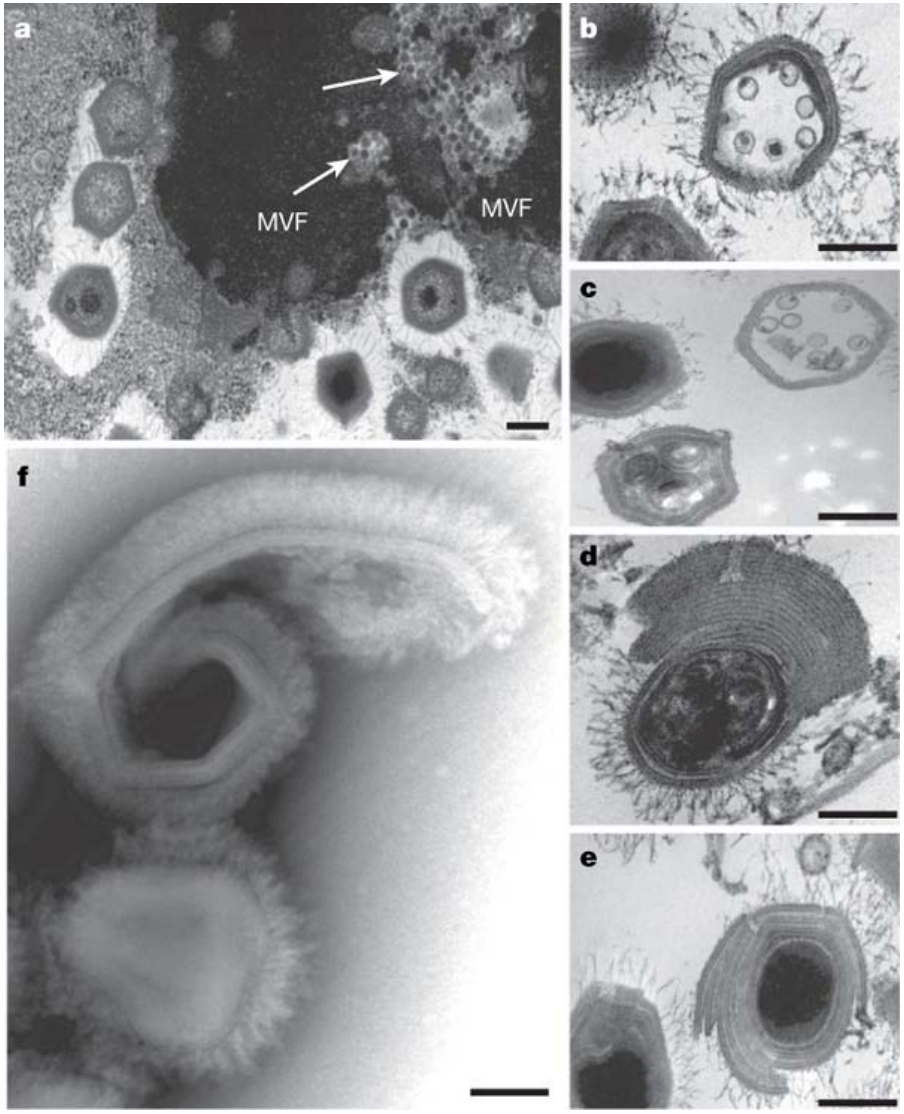
Gene content tree: intervirus gene transfer

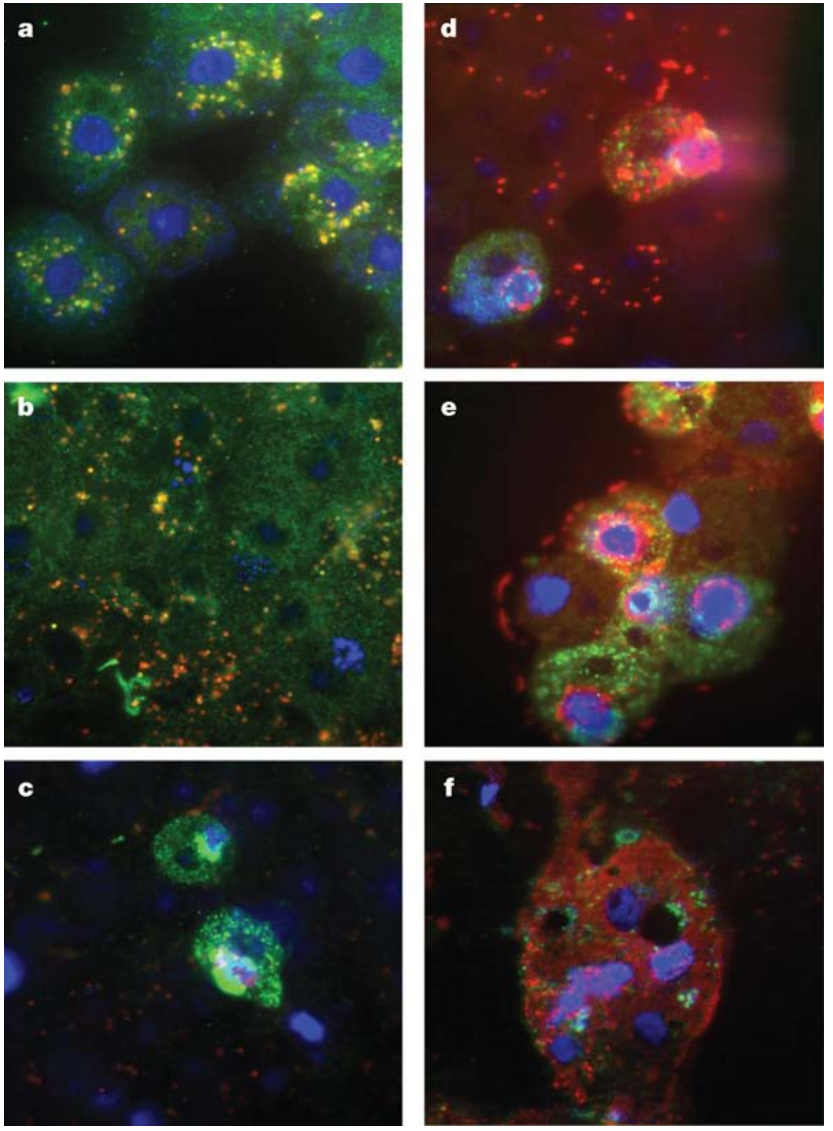


Amoeba as a melting pot for HGT between viruses and bacterial endosymbionts



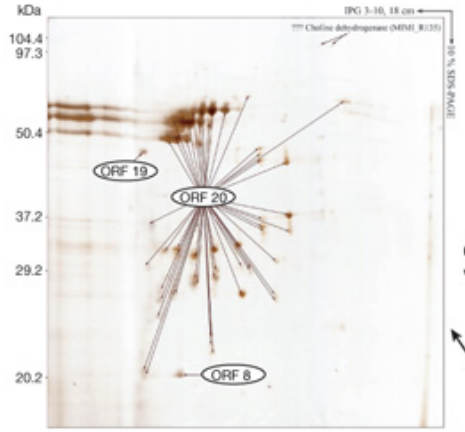
There are really weird creatures out there:Some NCLDV host their own parasites



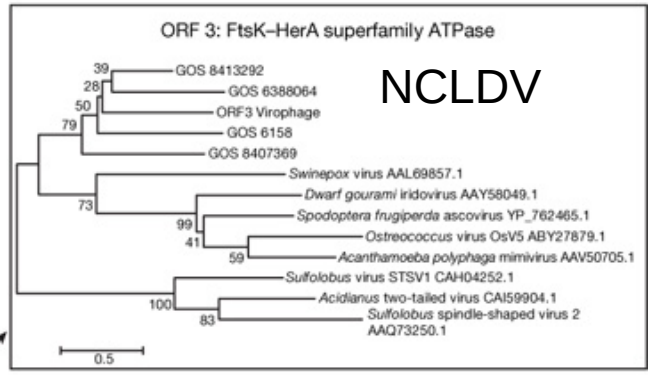


La Scola et al. **The virophage as a unique parasite of the giant mimivirus.** Nature. 2008

Chimeric origin of the virophage genome

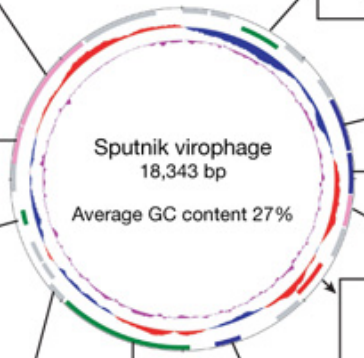


ORF 20: major virion protein



NCLDV

ORF 19: minor virion protein



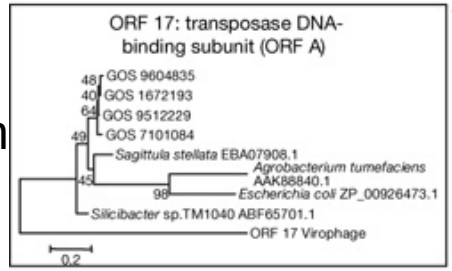
ORF 6: collagen triple-helix-containing protein

ORF 7: collagen triple-helix-containing protein

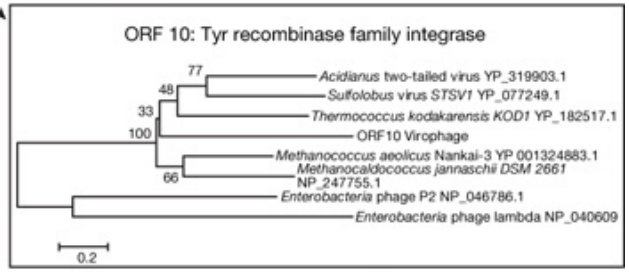
ORF 8: minor virion protein

Mimivirus

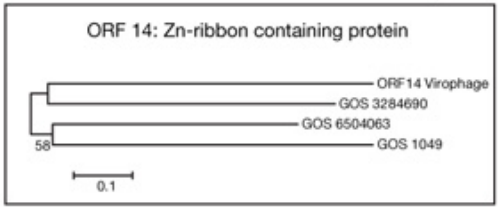
Bacterial transposon



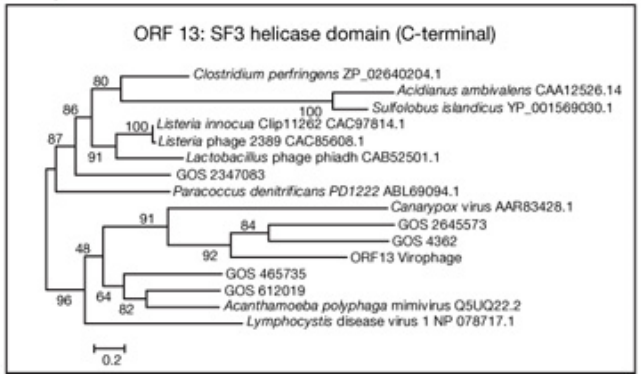
ORF 12: unknown function



Archaeal virus



?environmental only



NCLDV

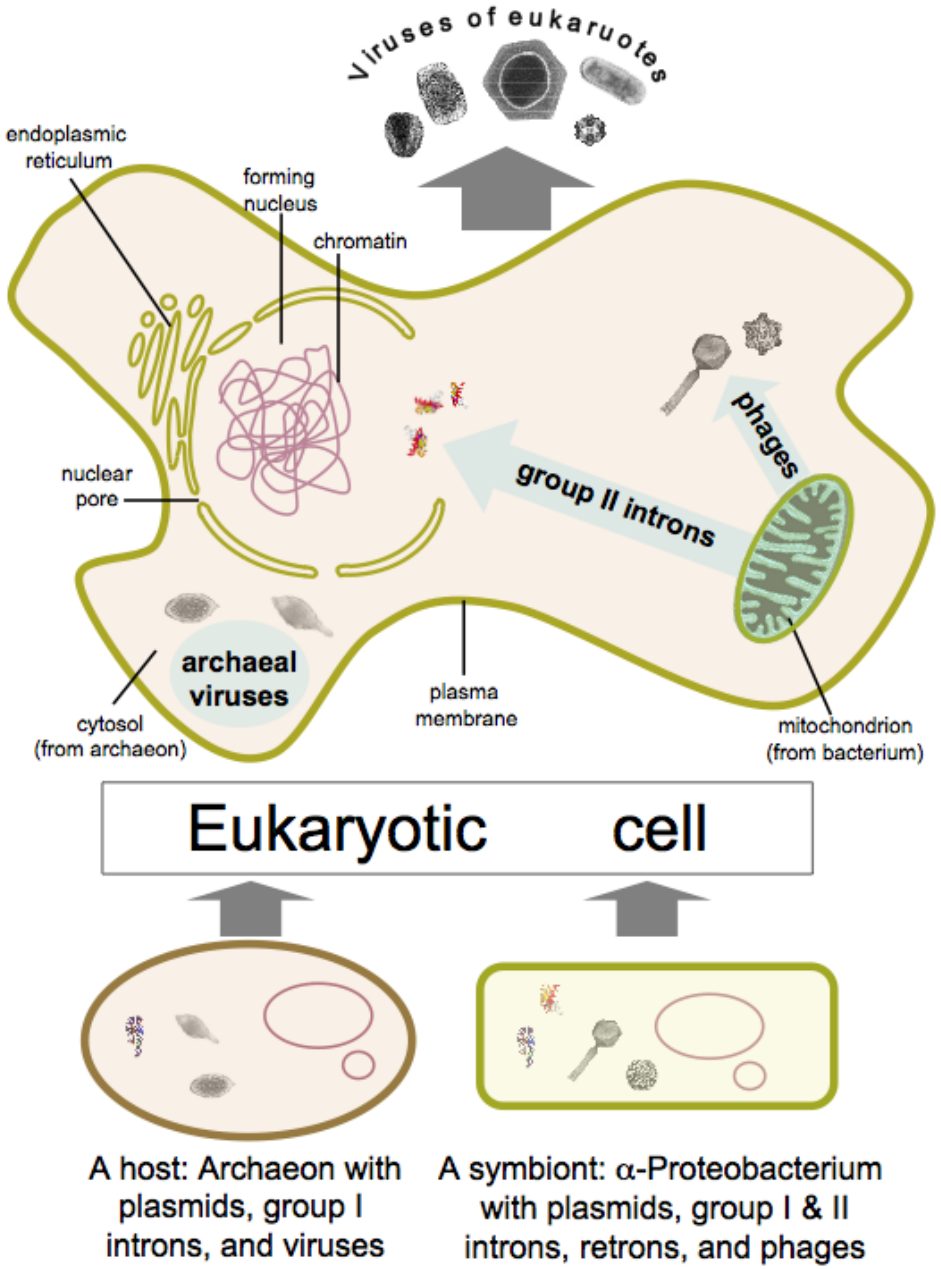
La Scola et al. Nature, 2008; 455:100-4

Hypothesis:
origin of the NCLDV (and other viruses of Eukaryotes) in the the second melting pot of virus evolution – eukaryogenesis

Phage scaffold
(virus hallmark genes)

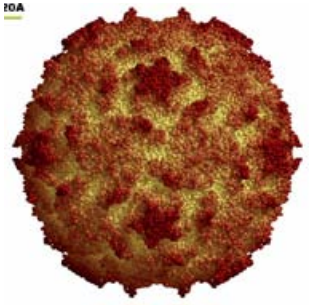
↓

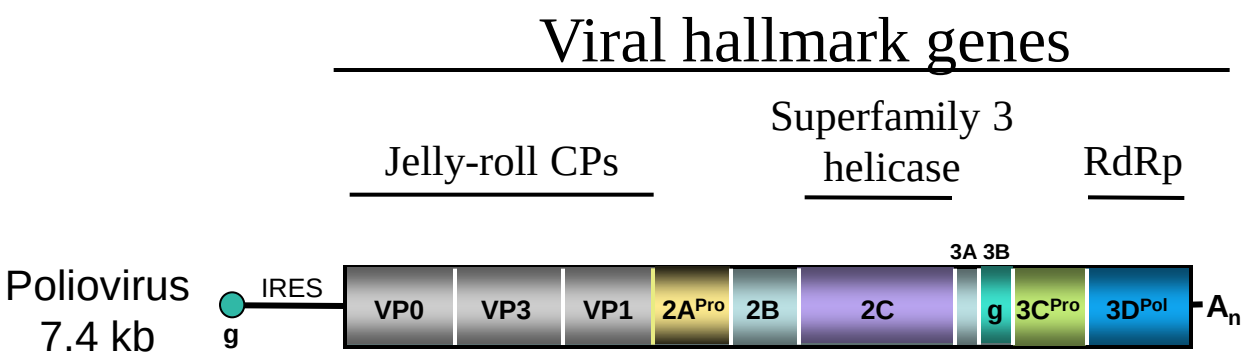
Eukaryotic additions/
displacements



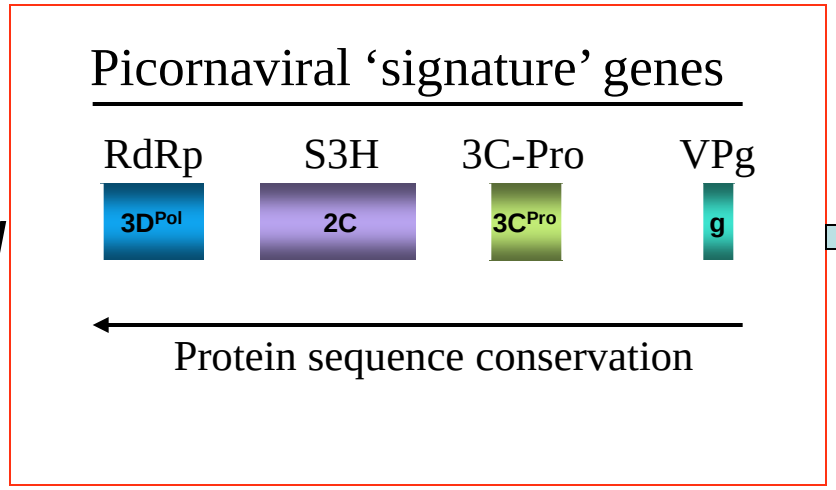
Koonin, Yutin, Intervirology 2010

Some of the smallest viruses (this is where poliovirus belongs): The Big Bang of picorna-like virus evolution





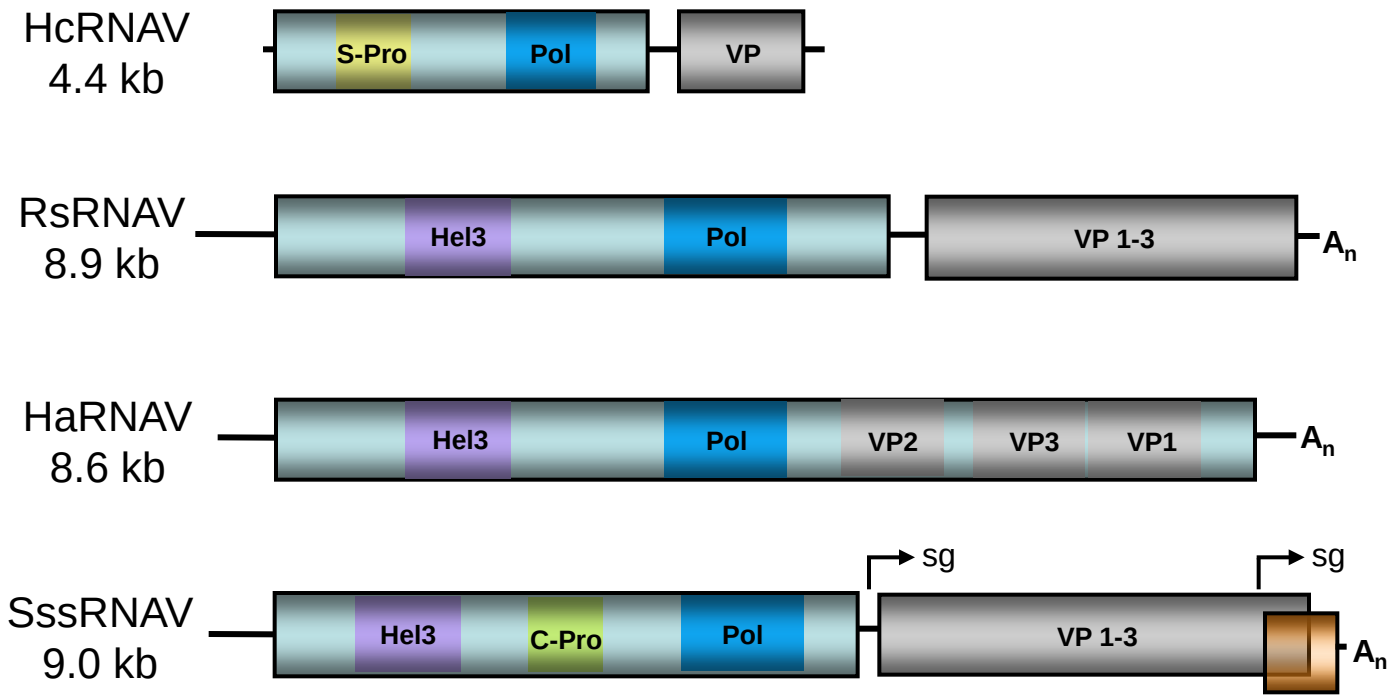
*Picorna-like viruses possess diverse arrays of **hallmark** and unique genes*



Picorna-like viral superfamily

Koonin, Wolf, Nagasaki, Dolja, Nat. Rev. Microbiol. 2008 Dec;6(12):925-39

Marine eukaryotic plankton carries a wealth of positive-strand RNA viruses: **nearly all belong to the picorna-like superfamily**



Lang *et al*. (2004) *Virology* **320**:206

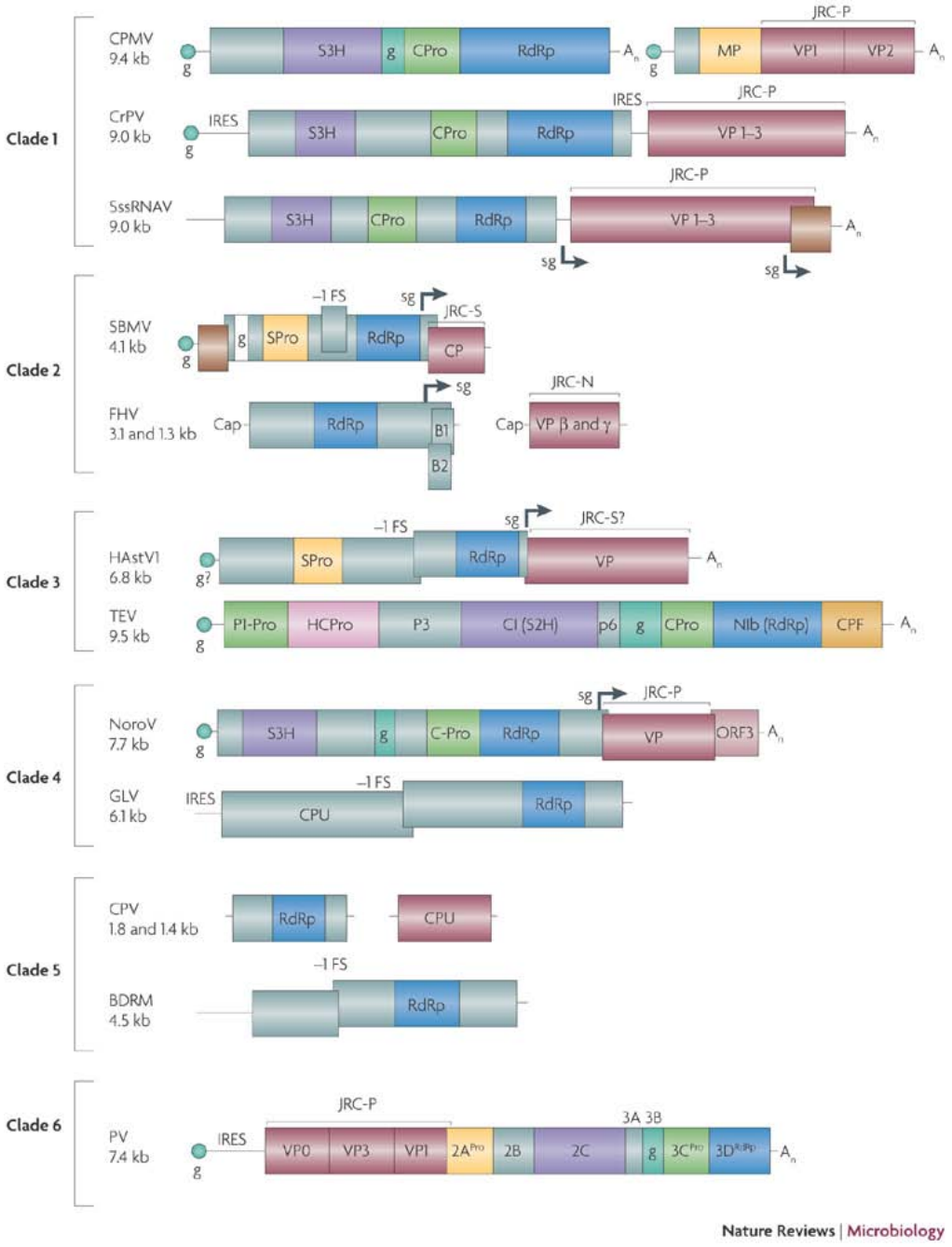
Nagasaki *et al*. (2005) *Appl. Env. Microbiol.* **71**:8888

Culley *et al*. (2006) *Science* **312**:1795

The amended Picornavirus-like superfamily includes 14 recognized viral families, 4 floating genera and 15 unclassified positive-strand and double-strand RNA viruses that infect hosts from **4 of the 5** eukaryotic supergroups

-6 distinct clades from RdRp phylogeny

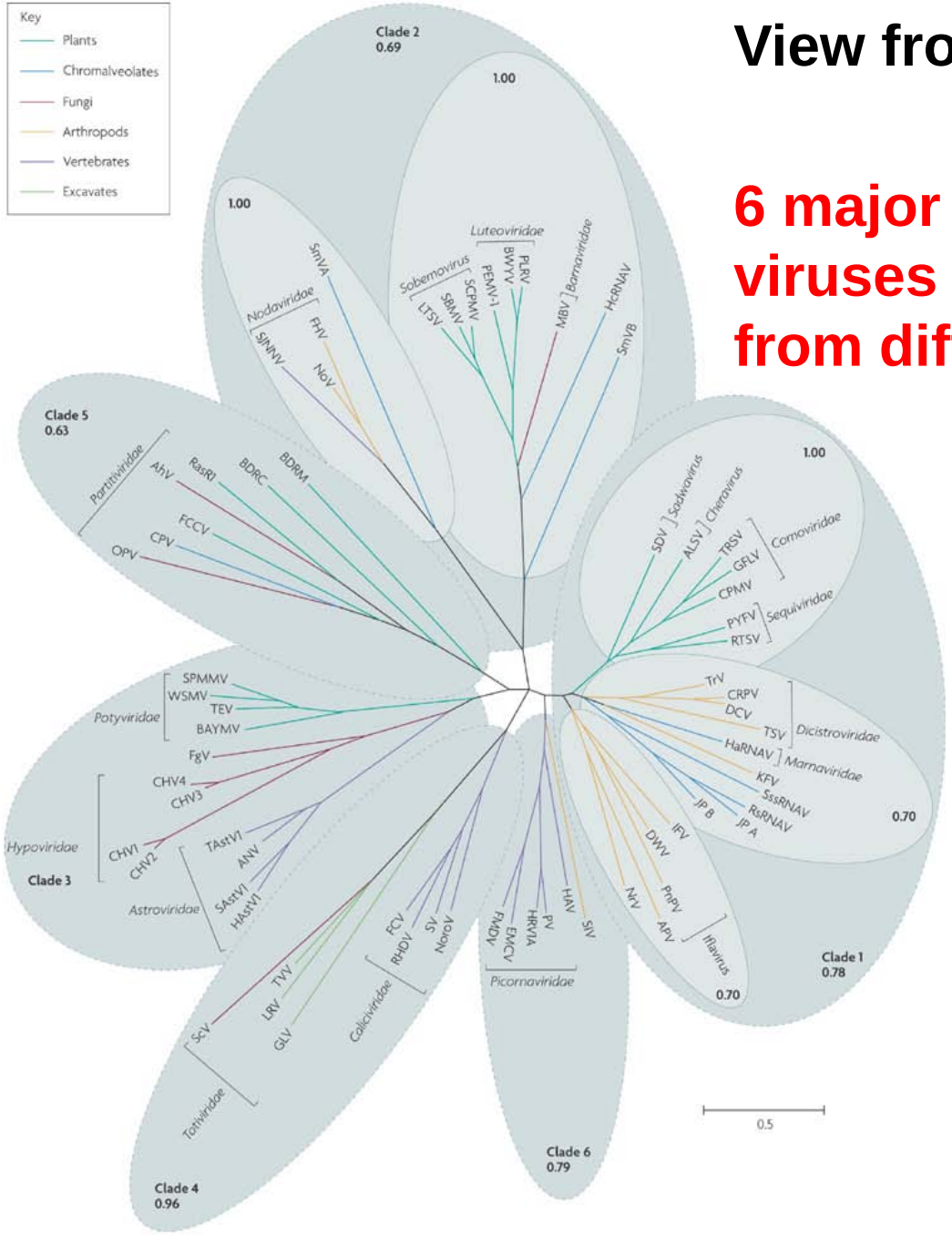
-diverse genome layouts



Nature Reviews | Microbiology

View from the viral side

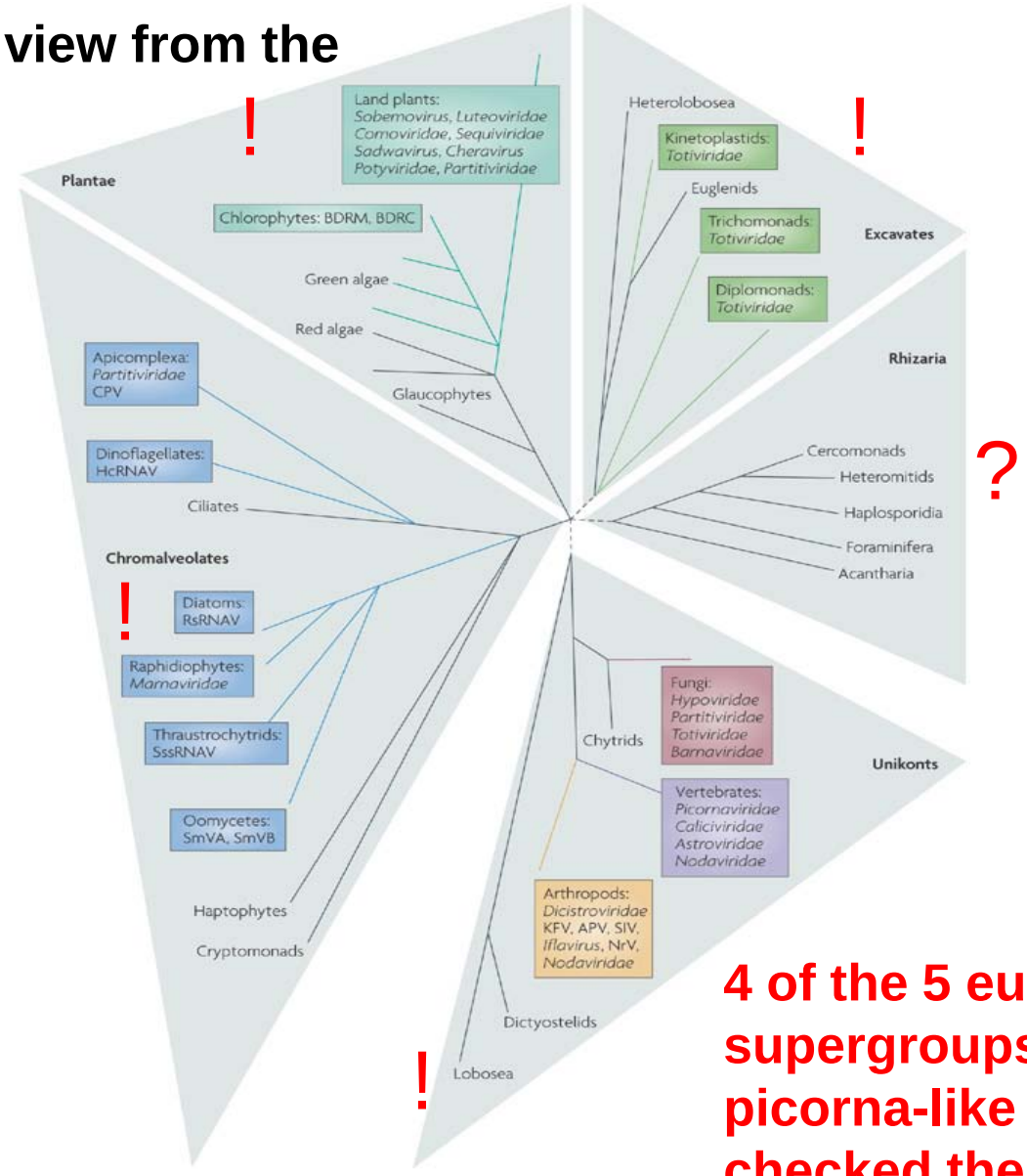
6 major clades of picorna-like viruses - 5 infect eukaryotes from different supergroups



Nature Reviews | Microbiology

The 5 supergroups of eukaryotes and their picorna-like viruses

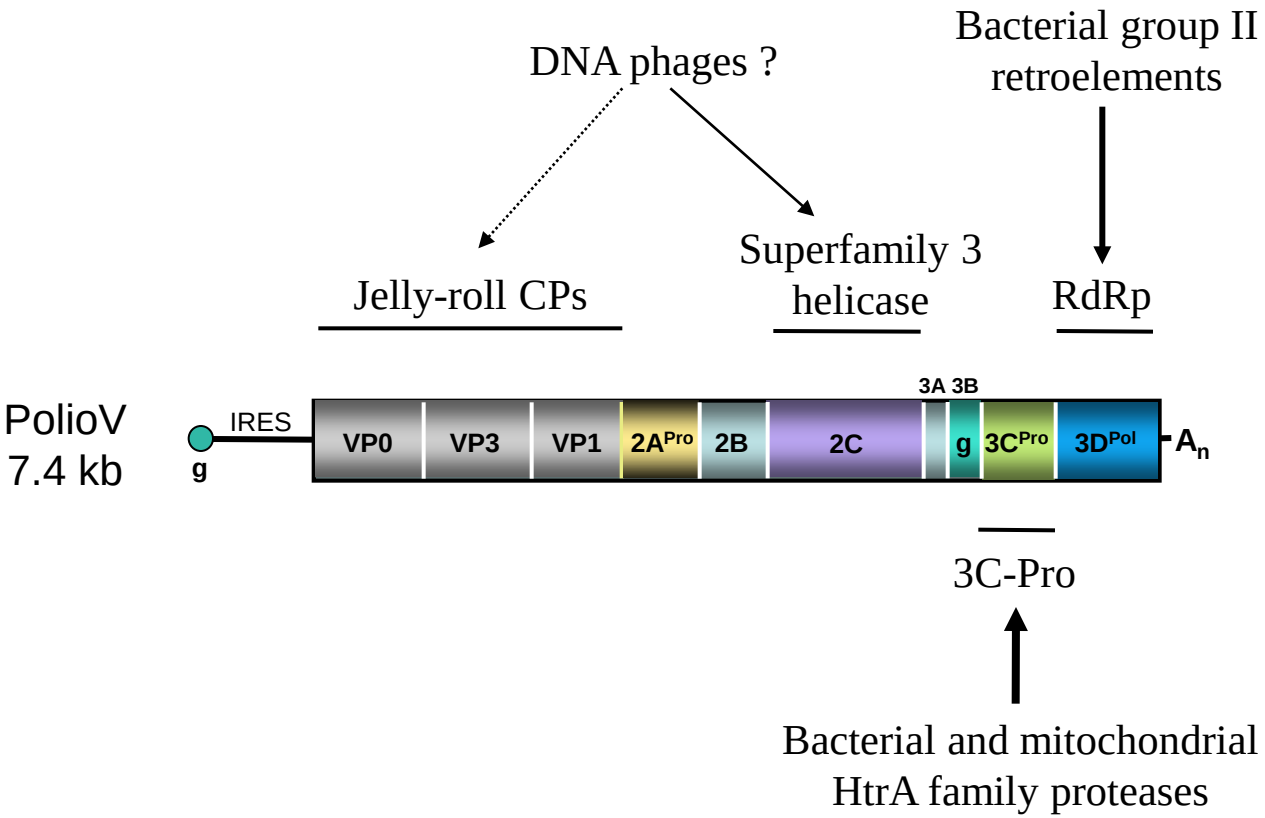
Complementary view from the host side

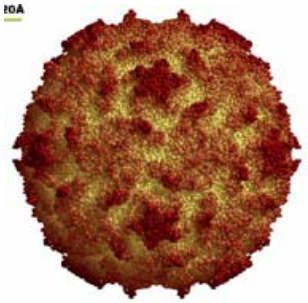


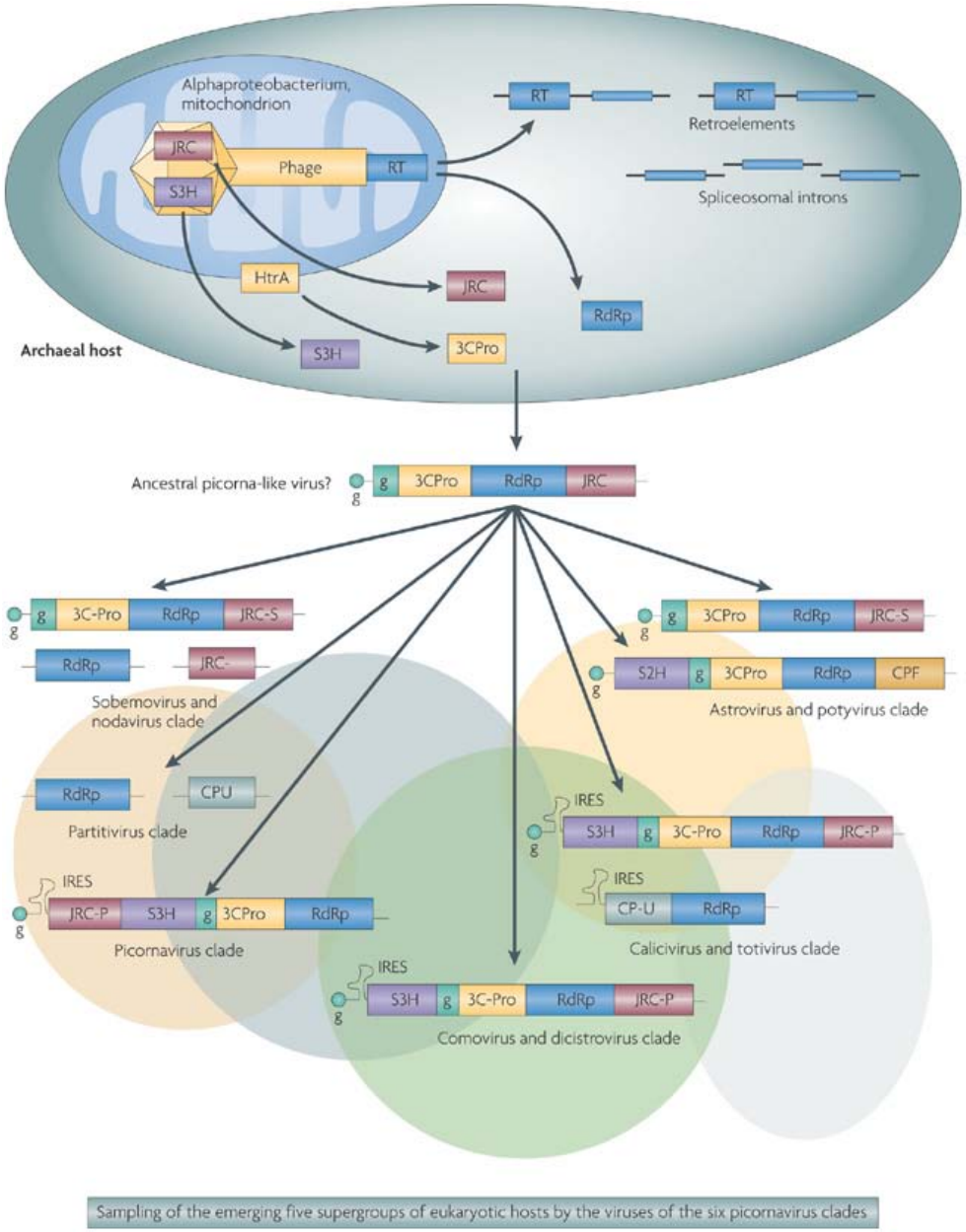
4 of the 5 eukaryotic supergroups host diverse picorna-like viruses (no one checked the 5th)

Nature Reviews | Microbiology

Likely origins of the signature genes of the picorna-like superfamily were inferred using signature PSSMs and PSI-BLAST searches







Sampling of the emerging five supergroups of eukaryotic hosts by the viruses of the six picornavirus clades

Hallmark genes of picorna-like viruses have distinct prokaryotic origins

Radiation of major viral clades occurred in a "Big Bang" during eukaryogenesis and antedates the divergence of eukaryotic supergroups -the viruses then "sampled" the hosts

Conclusions on the picorna and NCLDV stories:

Big Bangs of virus evolution

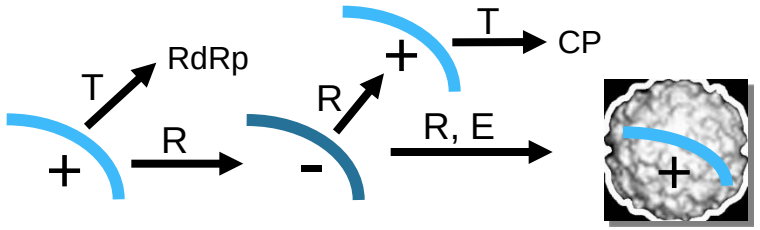
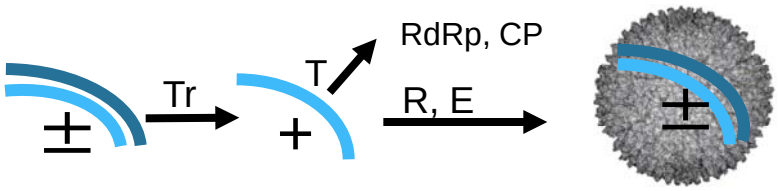
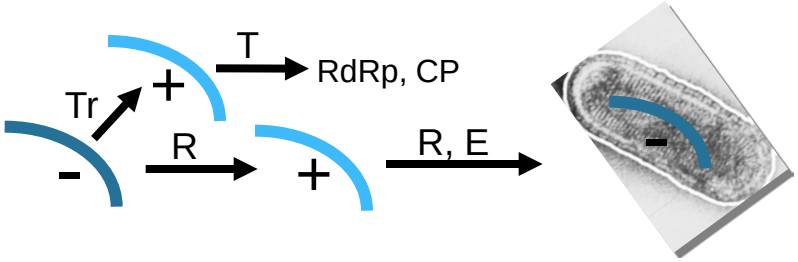
- Phylogenomic analysis of the picorna-like superfamily of eukaryotic RNA viruses indicates that the major lineages within this superfamily diverged prior to the divergence of the eukaryotic supergroups
- **Big Bang** – an explosive early phase of viral evolution
- Most likely, the same pattern holds for other major groups of viruses as illustrated by the evolutionary study of NCLDV, a completely different group of viruses
- The **Big Bangs** of eukaryotic virus evolution occurred concomitantly with a similar rapid phase of host evolution and could be a manifestation of a general model of major evolutionary transitions

Koonin, Wolf, Nagasaki, Dolja, Nat. Rev. Microbiol. 2008 Dec;6(12):925-39

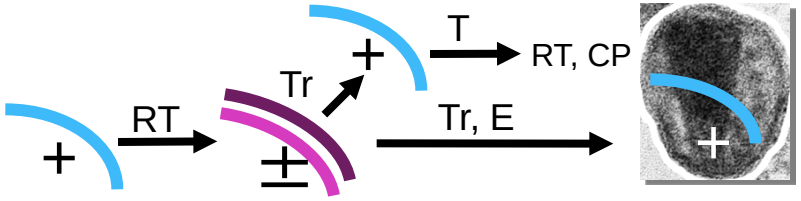
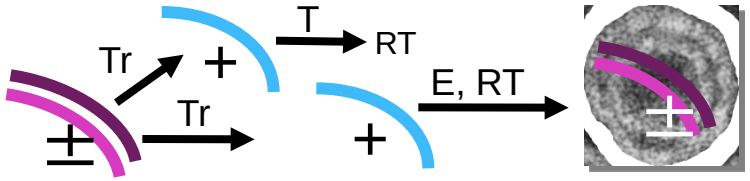
Koonin. The Biological Big Bang model for the major transitions in evolution. *Biol Direct*. 2007 Aug 20;2:21

Diversity of viral genetic cycles versus the uniform genetic cycle of all cellular organisms: Viruses are the biosphere's laboratory of genomic strategies

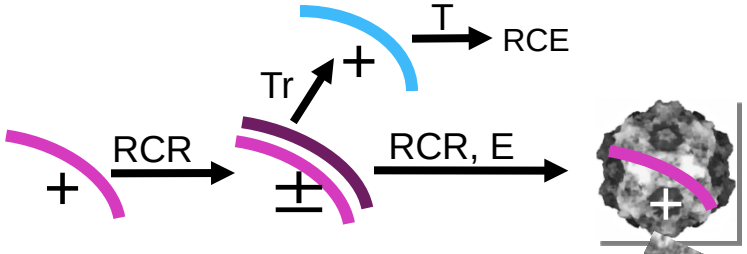
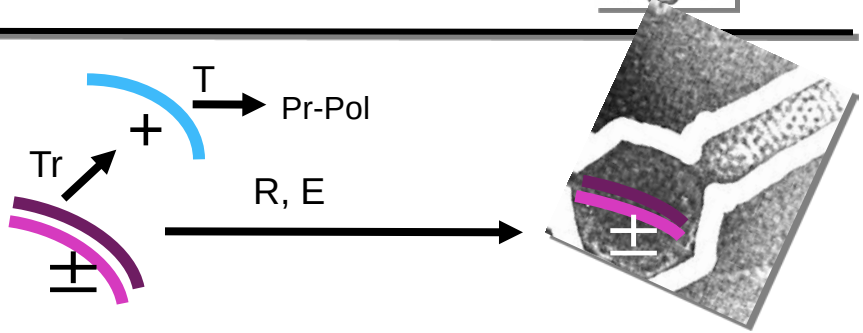
Genetic cycles of RNA viruses

Class	Replication cycle
1. Positive-strand RNA 3-30 kb	
2. Double-strand RNA 4-25 kb	
3. Negative-strand RNA 11-20 kb	

Genetic cycles of retroid viruses and retroelements

Class	Replication cycle
4. Retroid RNA viruses 7-12 kb	
5. Retroid DNA viruses, elements 2-10 kb	

Genetic cycles of DNA viruses and plasmids

6. ssDNA viruses, plasmids 2-11 kb	
7. dsDNA viruses, plasmids 5-1,200 kb	

YOU ARE HERE

ALL CELLULAR ORGANISMS

Natural history of viral genes: a one-page summary of *viral comparative genomics*

I. Genes with readily detectable homologs from cellular life forms:

1. Genes with closely related homologs from cellular organisms (typically, the host of the given virus) present in a narrow group of viruses
2. Genes that are conserved within a virus lineage or even several lineages and have moderately close cellular homologs

Origin: relatively recent (1) or ancient (2) acquisition from host

II. Virus-specific genes

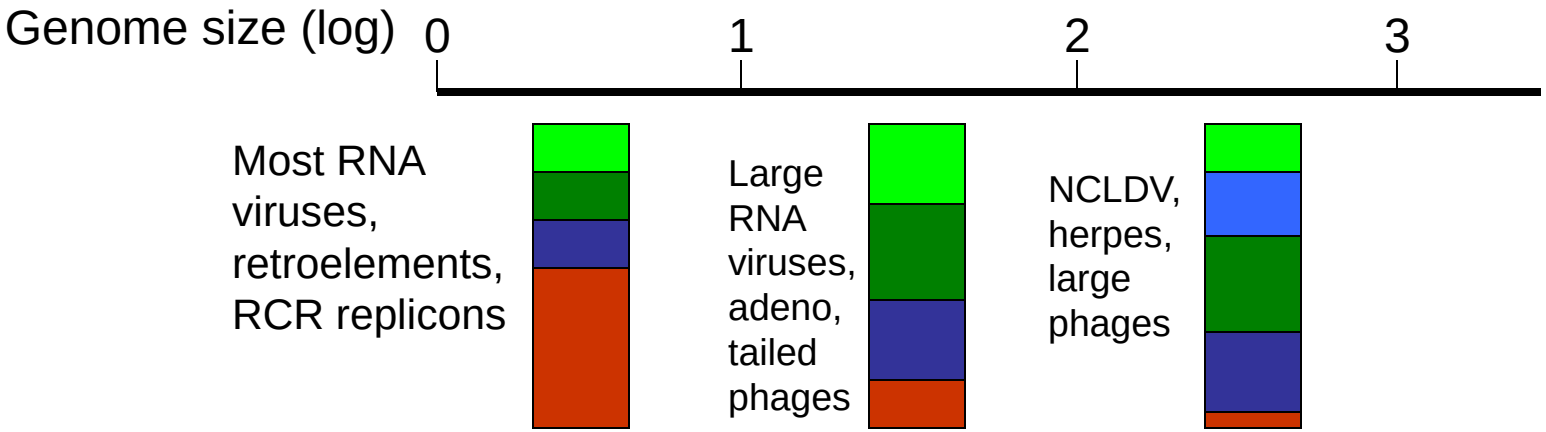
3. ORFans, i.e., genes without detectable homologs except, possibly, in closely related viruses
4. Virus-specific genes that are conserved within a virus lineage

Acquisition from host but with rapid divergence from ancestor once within viral genomes?

III. Viral hallmark genes

5. Genes shared by many diverse virus lineages, with only very distant homologs in cellular organisms

Contributions of different classes of viral genes to the genomes of different classes of viruses: strong dependence on genome size



Recent acquisitions

Old acquisitions

Virus-specific ORFans

Virus-specific conserved

Virus hallmark

Natural history of viral genes: Viral Hallmark Genes

Shared by many diverse groups of viruses: **from the smallest RNA viruses to the giant DNA viruses**

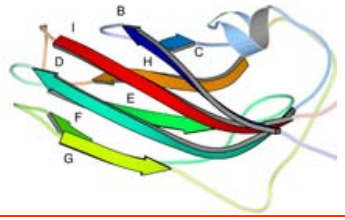
Strong support for monophyly of all viral members of the respective gene families

Only distant homologs in cellular organisms

Play major roles in genome replication, packaging and assembly

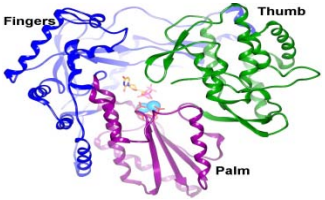
Can be viewed as signatures of the 'virus state'

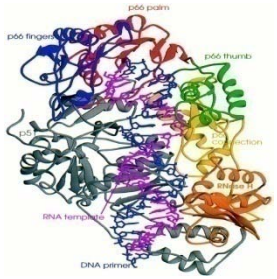
Protein products of viral hallmark genes



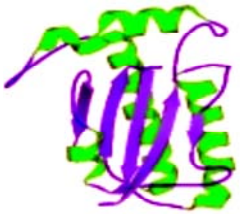
1. Jelly-roll capsid protein

2. Superfamily 3 helicase





3. RNA-dependent RNA polymerase and Reverse transcriptase



4. Rolling circle replication initiation endonuclease

5. Viral archaeo-eukaryotic DNA primase

6. UL9-like superfamily 2 helicase

7. Packaging ATPase of the FtsK family

8. ATPase subunit of terminase

Viral hallmark genes are:

- ***present in a huge diversity of viruses and other selfish elements***
- ***represented only by remote homologs in cellular life forms***

The primordial gene pool hypothesis
(*extremely counterintuitive! – Santiago Elena, Feb 16, 2011*)

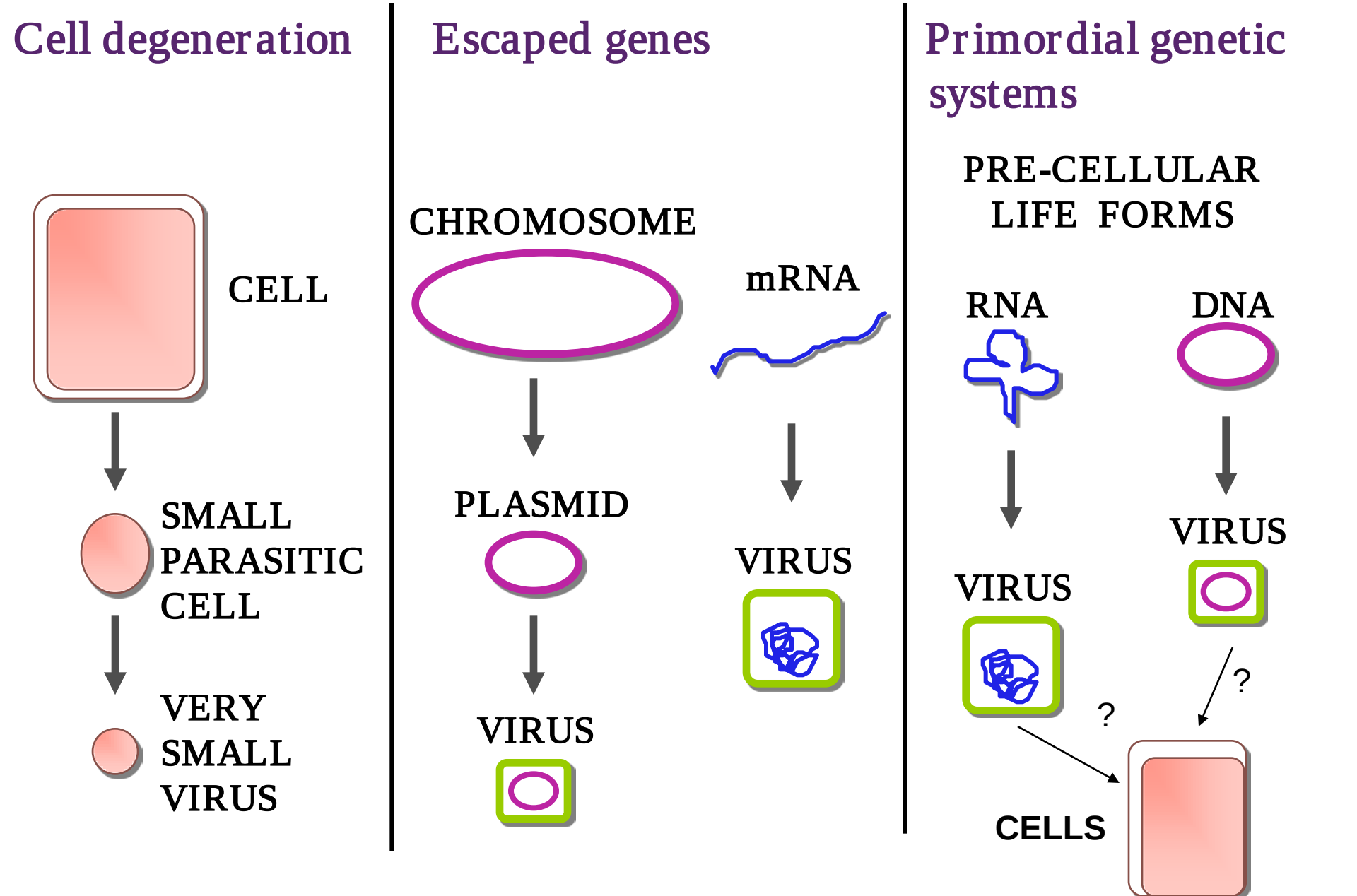
The hallmark genes AND, by implication, the major lineages of modern viruses (at least, viruses of prokaryotes) descend directly from a primordial gene pool

-***synergy with the diversity of genomic strategies***

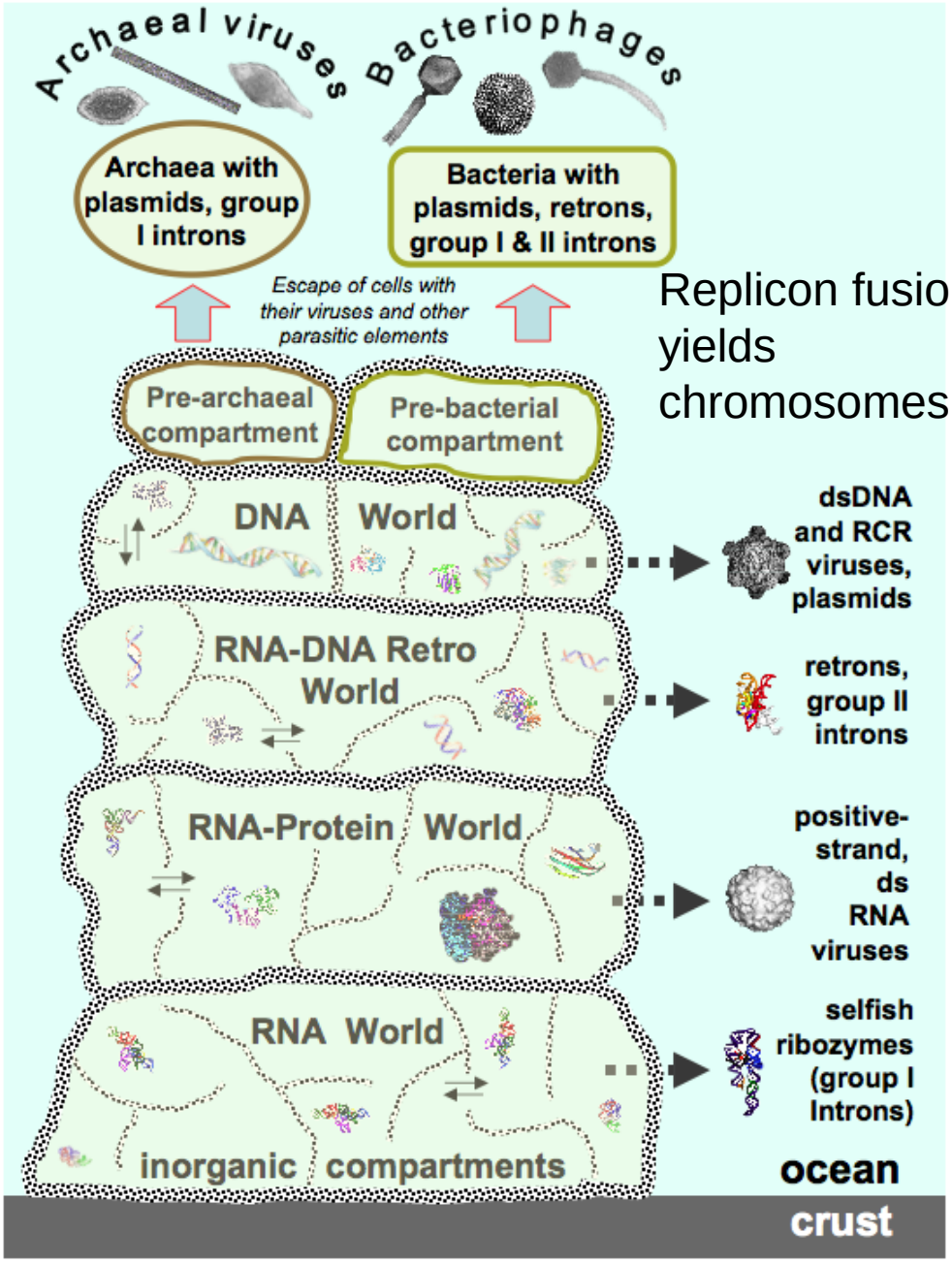
A crucial corollary: If viruses come directly from a primordial gene pool, then, origin of viruses is *inextricably linked to the origin of cells*

Koonin, Senkevich, Dolja. Biol. Direct., 2006, 1:29

Competing concepts of the origin of viruses



Origin and evolution of virus-like genetic elements in the pre-cellular era



Replicon fusion yields chromosomes

Koonin, Martin, TIG, 2005
Koonin, Senkevich, Dolja.
Biol. Direct. 2006, 1:29
Koonin, Ann NY Acad Sci, 2009

The ancient Virus World (VW)

- **V**iruses and virus-like genetic elements are not "just" pathogens: they are ***dominant entities*** in the biosphere
- **E**mergence of virus-like parasites is inevitable in any replicating system
- **I**n the pre-cellular epoch, the genetic elements that later became viral and cellular genomes comprised a ***single pool*** in which they mixed, matched, and evolved new, increasingly complex gene ensembles
- **D**ifferent replication strategies including RNA replication, reverse transcription, and DNA replication evolved already in the primordial genetic pool
- **W**ith the emergence of prokaryotic cells, a ***distinct pool*** of viral genes formed that retained its identity ever since as evidenced by the extant distribution of viral hallmark genes: ***"virus world" or the virosphere***
- **T**he emergence of the eukaryotic cell was a second melting pot of virus evolution, from which viruses of eukaryotes originated via recombination of genes from prokaryote viruses, retroelements, and the evolving eukaryotic host
- **V**iruses make essential contributions to the evolution of the genomes of cellular life forms, in particular, as vehicles of HGT: GTAs, transducing phages

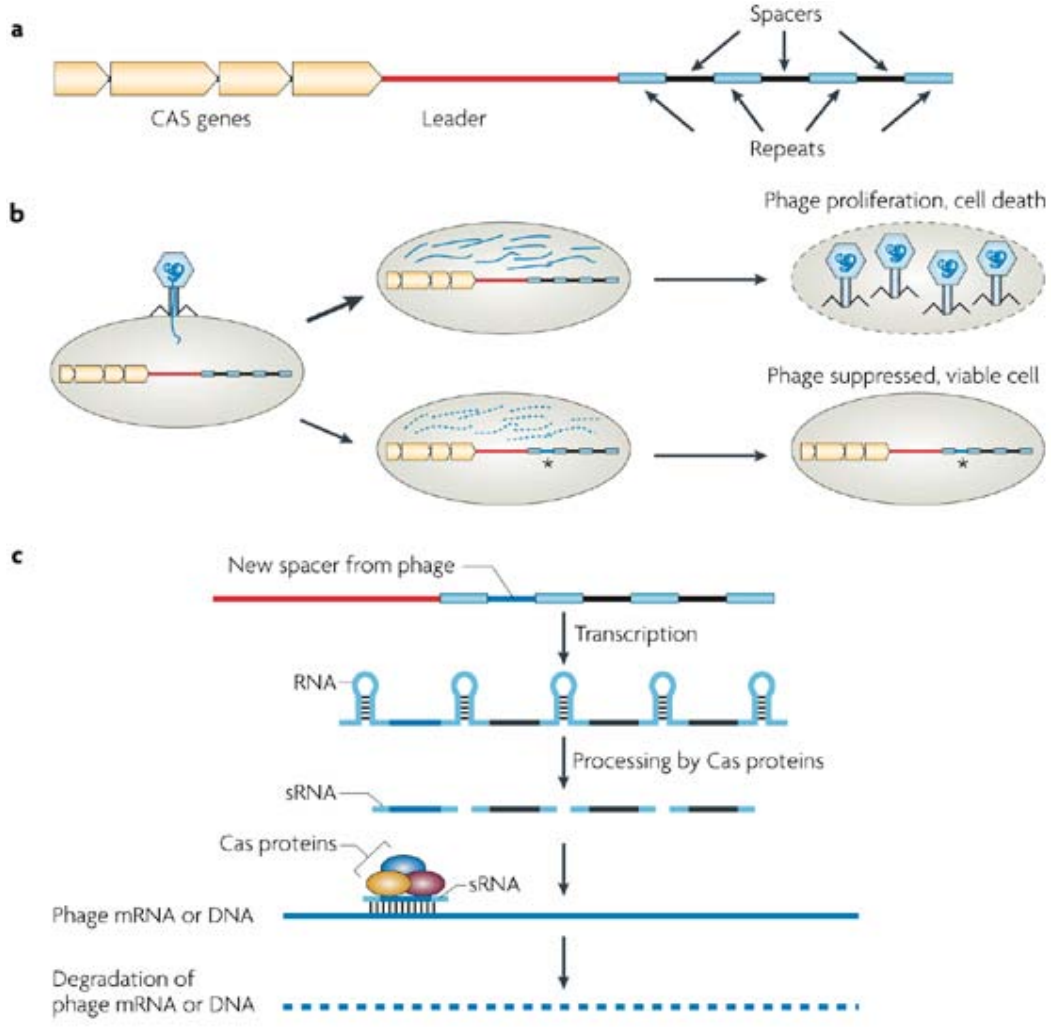
Koonin EV, Senkevich TG, Dolja VV. **The ancient Virus World and evolution of cells.** *Biol Direct*. 2006
Koonin EV, Wolf YI, Nagasaki K, Dolja VV. **The complexity of the virus world.** *Nat Rev Microbiol*. 2009

Evolution of antivirus defense systems

- CRISPR/Cas system of adaptive immunity in prokaryotes
- A case for Lamarckian evolution
- The perennial virus-host arms race

CRISPR repeats and Cas genes

CRISPR: Clustered, Regularly interspaced short palindromic repeats
Cas: CRISPR-associated (genes)



Sorek et al. Nature Rev Microbiol 2008

Nature Reviews | Microbiology

Makarova KS Aravind L Grishin NV Rogozin IB Koonin EV

A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis.

During a systematic analysis of conserved gene context in prokaryotic genomes, a previously undetected, complex, partially conserved neighborhood consisting of more than 20 genes was discovered in most Archaea and some bacteria, including the hyperthermophiles Thermotoga maritima and Aquifex aeolicus. The gene composition and gene order in this neighborhood vary greatly between species, but all versions have a stable, conserved core that consists of five genes. One of the core genes encodes a predicted DNA helicase, often fused to a predicted HD-superfamily hydrolase, and another encodes a RecB family exonuclease; three core genes remain uncharacterized, but one of these might encode a nuclease of a new family.

……

The functional features of the proteins encoded in this neighborhood suggest that they comprise **a previously undetected DNA repair system**, which, to our knowledge, is the first repair system largely specific for thermophiles to be identified.

Protein components of the system: an update with unification of many diverse families ~25 families altogether

	Family	Subfamily[A]	Phyletic distribution[B]	Comments
1	COG1518	**COG1518 (cas1)**	All	Putative novel nuclease/integrase; Mostly α-helical protein
2	COG1343	**COG1343 (cas2)**, COG3512, ygbF-like; MTH324-like; y1723_N-like;	All	Small protein related to VapD, fused to helicase (COG1203) in y1723-like proteins
3	COG1203	**COG1203 (cas3)**	All	DNA helicase; Most proteins have fusion to HD nuclease
4	RecB-like nuclease	**COG1468 (cas4)**, COG4343	All	RecB-like nuclease; Contains three-cysteine C-terminal cluster
5	RAMP: Repair-associated mysterious protein	**COG1688 (cas5)**,COG1769, COG1583,COG1567, COG1336,COG1367, COG1604,COG1337, COG1332,COG5551, BH0337-like, MJ0978-like, YgcH-like, y1726-like, y1727-like	All	Belong to "RAMP" family, possibly RNA-binding protein, structurally related to duplicated ferredoxin fold (PDB: 1wj9)
6	COG1857	COG1857, COG3649, YgcJ-like, y1725-like	All	α/β protein; predicted nuclease or integrase
7	HD-like nuclease	COG1203 (N-terminus), COG2254	All	HD-like nuclease
8	BH0338	BH0338-like MTH1090-like	All, mostly archaea and FIRM	Large Zn-finger containing proteins, probably nucleases (nuclease activity was shown for MTH1090 (ref.)
9	COG1353, predicted polymerase	COG1353	Most archaea, some bacteria	Predicted palm-domain polymerase distantly related to viral RdRp and RT

…and ~15 other, less common proteins

Makarova et al. BD 2006

CRISPR: clustered regularly interspaced short palindromic repeats

1: Mol Microbiol. 2002 Mar;43(6):1565-75.

Related Articles, Links

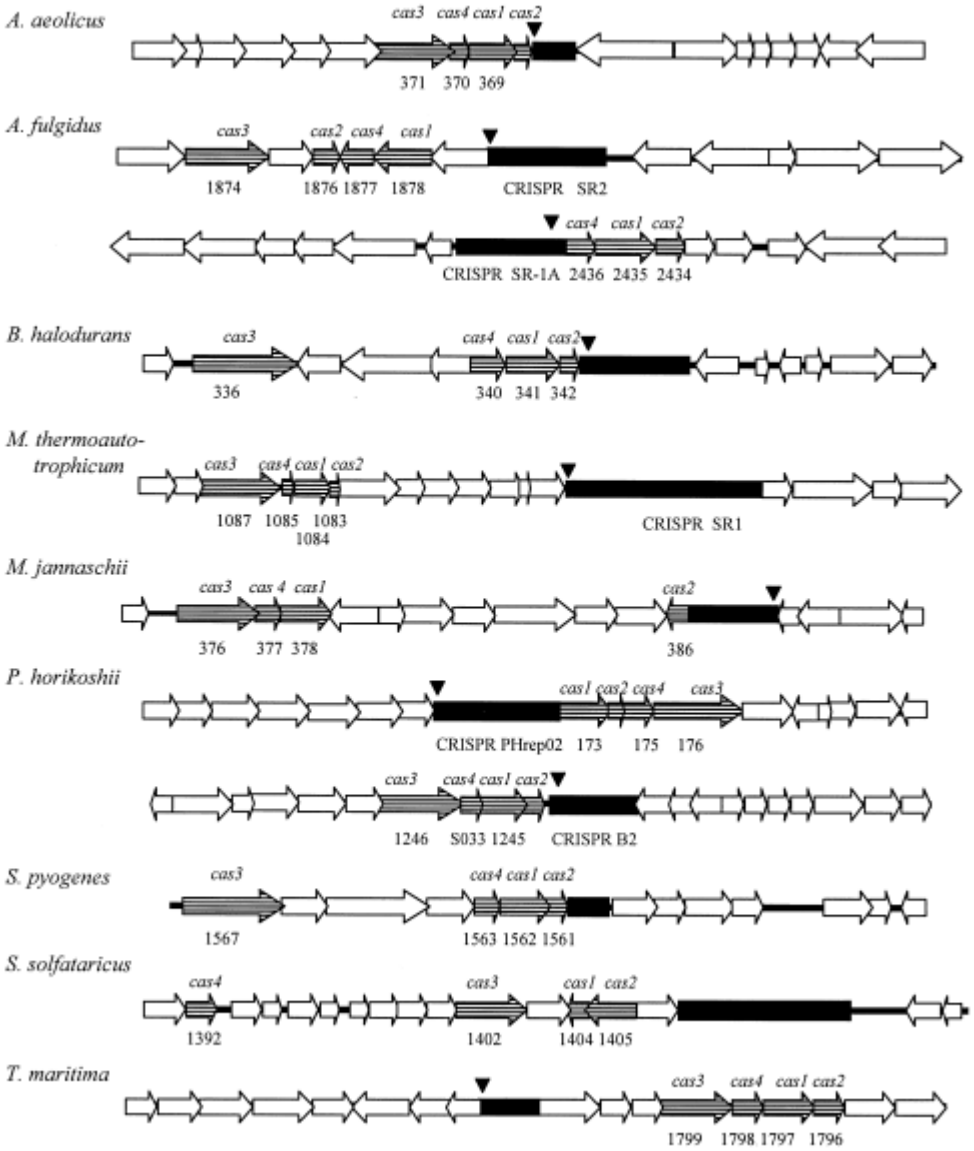
Identification of genes that are associated with DNA repeats in prokaryotes.

Jansen R, Embden JD, Gaastra W, Schouls LM.

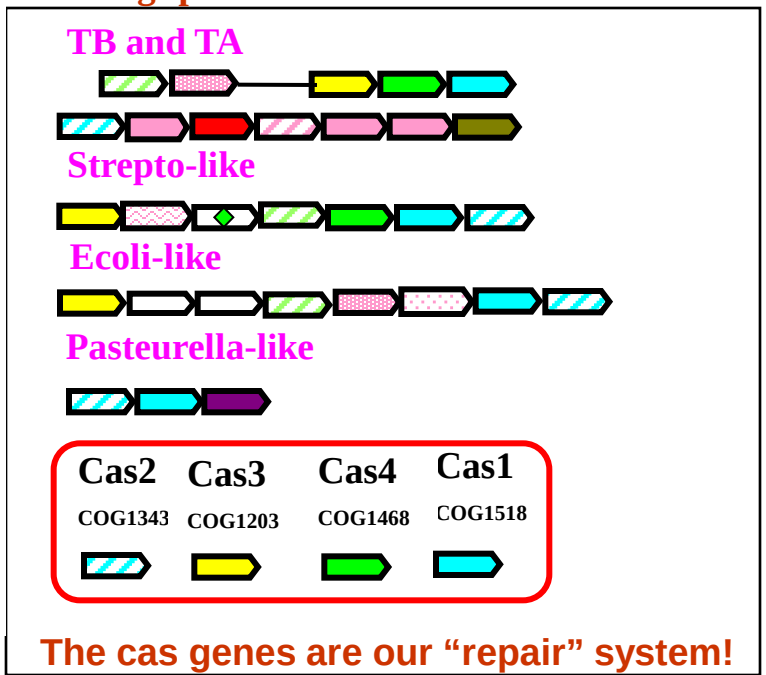
Department of Infectious Diseases and Immunology, Bacteriology Division, Veterinary Faculty, Utrecht University, Yalelaan 1, 3584 CL Utrecht, The Netherlands. R.jansen@vet.uu.nl

Using in silico analysis we studied a novel family of repetitive DNA sequences that is present among both domains of the prokaryotes (Archaea and Bacteria), but absent from eukaryotes or viruses. This family is characterized by direct repeats, varying in size from 21 to 37 bp, interspaced by similarly sized non-repetitive sequences. To appreciate their characteri-stic structure, we will refer to this family as the clustered regularly interspaced short palindromic repeats (CRISPR). In most species with two or more CRISPR loci, these loci were flanked on one side by a common leader sequence of 300-500 b. The direct repeats and the leader sequences were conserved within a species, but dissimilar between species. The presence of multiple chromosomal CRISPR loci suggests that CRISPRs are mobile elements. Four CRISPR-associated (cas) genes were identified in CRISPR-containing prokaryotes that were absent from CRISPR-negative prokaryotes. The cas genes were invariably located adjacent to a CRISPR locus, indicating that the cas genes and CRISPR loci have a functional relationship. The cas3 gene showed motifs characteristic for helicases of the superfamily 2, and the cas4 gene showed motifs of the RecB family of exonucleases, suggesting that these genes are involved in DNA metabolism or gene expression. The spatial coherence of CRISPR and cas genes may stimulate new research on the genesis and biological role of these repeats and genes.

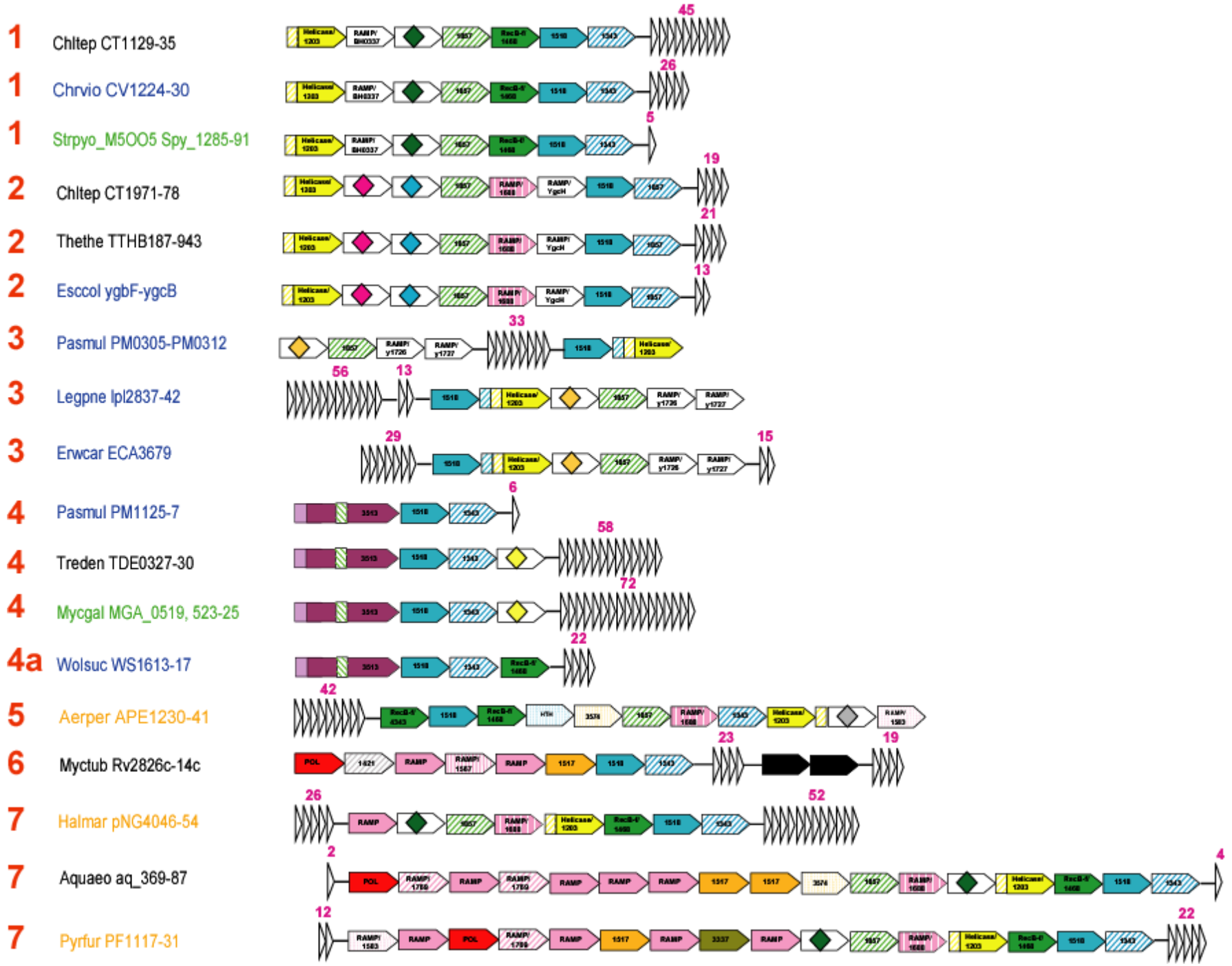
CRISPR repeats



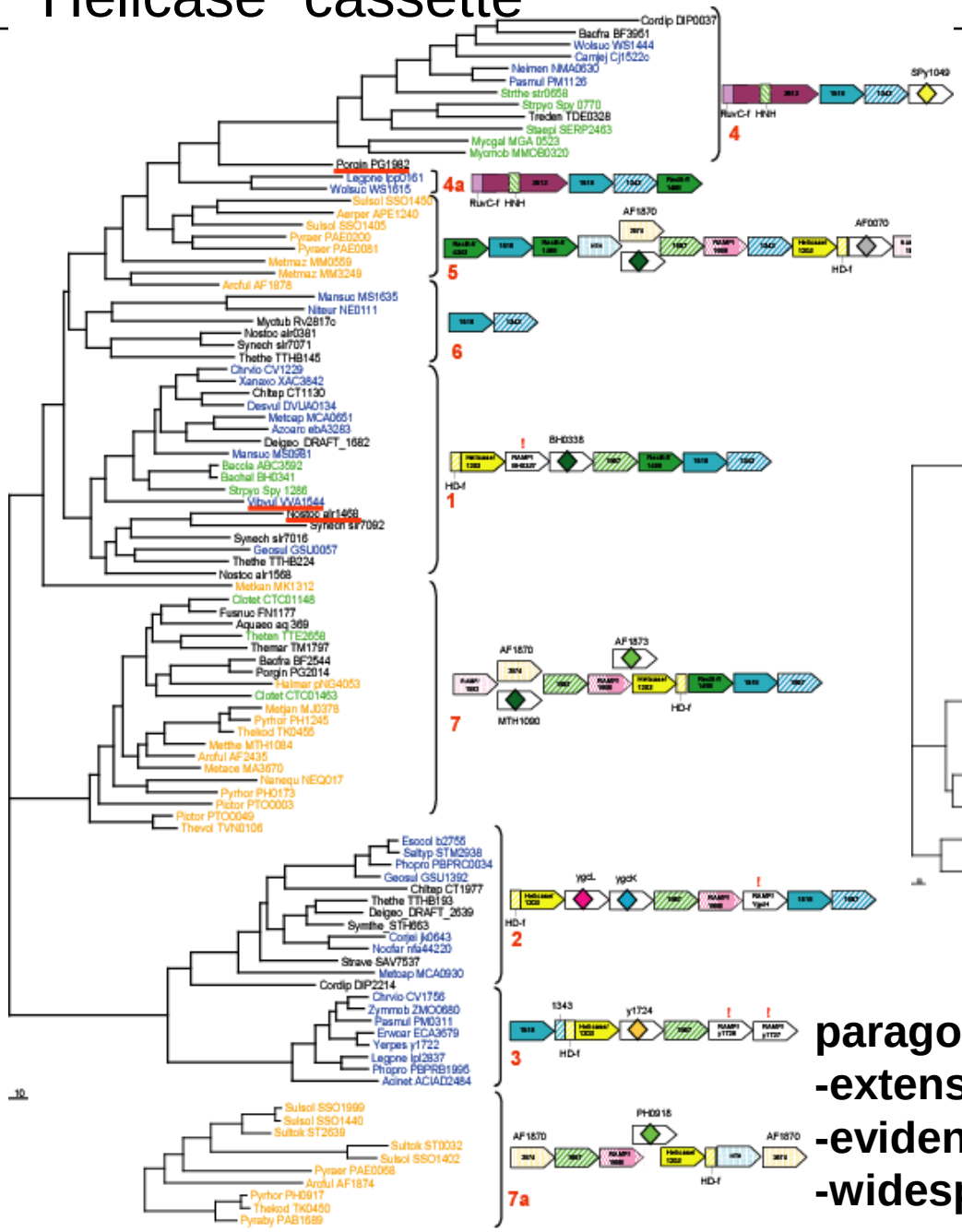
"The common structural characteristics of CRISPR loci are: (i) the presence of multiple short direct repeats, which show no or very little sequence variation within a given locus; (ii) the presence of non-repetitive spacer sequences between the repeats of similar size; (iii) the presence of a common leader sequence of a few hundred basepairs in most species harbouring multiple CRISPR loci; (iv) the absence of long open reading frames within the locus; and (v) the presence of the cas1 gene accompanied by the cas2, cas3 or cas4 genes in CRISPR-containing species."



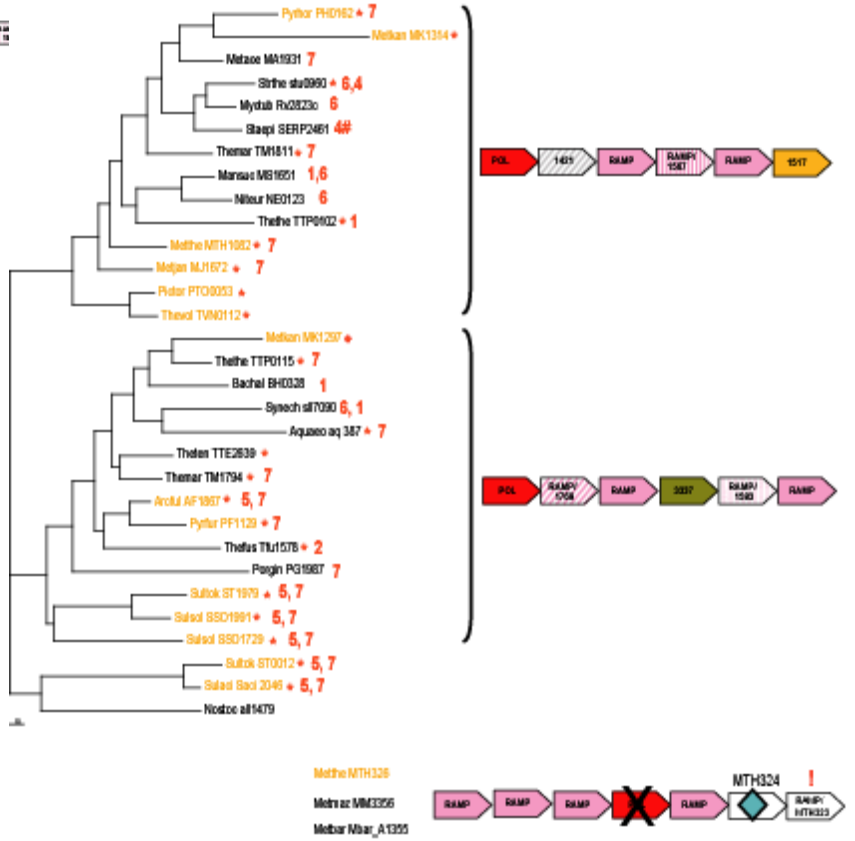
The cas genes are our "repair" system!


1
Chltep CT1129-35
45
1
Chrvio CV1224-30
26
1
Strpyo_M5OO5 Spy_1285-91
5
2
Chltep CT1971-78
19
2
Thethe TTHB187-943
21
2
Esccol ygbF-ygcB
13
3
Pasmul PM0305-PM0312
33
3
Legpne lpl2837-42
56
13
3
Erwcar ECA3679
29
15
4
Pasmul PM1125-7
6
4
Treden TDE0327-30
58
4
Mycgal MGA_0519, 523-25
72
4a
Wolsuc WS1613-17
22
5
Aerper APE1230-41
42
6
Myctub Rv2826c-14c
23
19
7
Halmar pNG4046-54
26
52
7
Aquaeo aq_369-87
2
4
7
Pyrfur PF1117-31
12
22

“Helicase” cassette



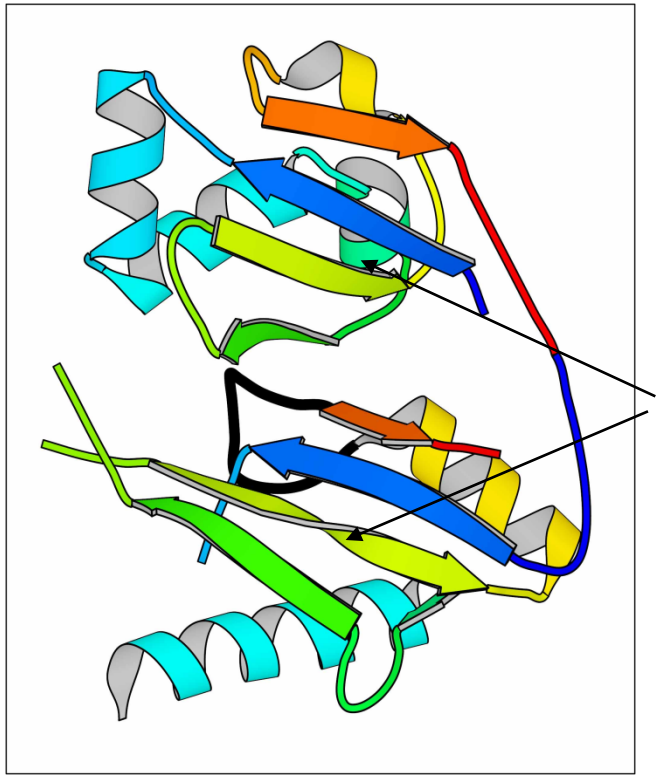
“Polymerase” cassette



paragon of prokaryotic genome plasticity:
-extensive gene shuffling
-evidence of multiple horizontal transfers
-widespread gene loss

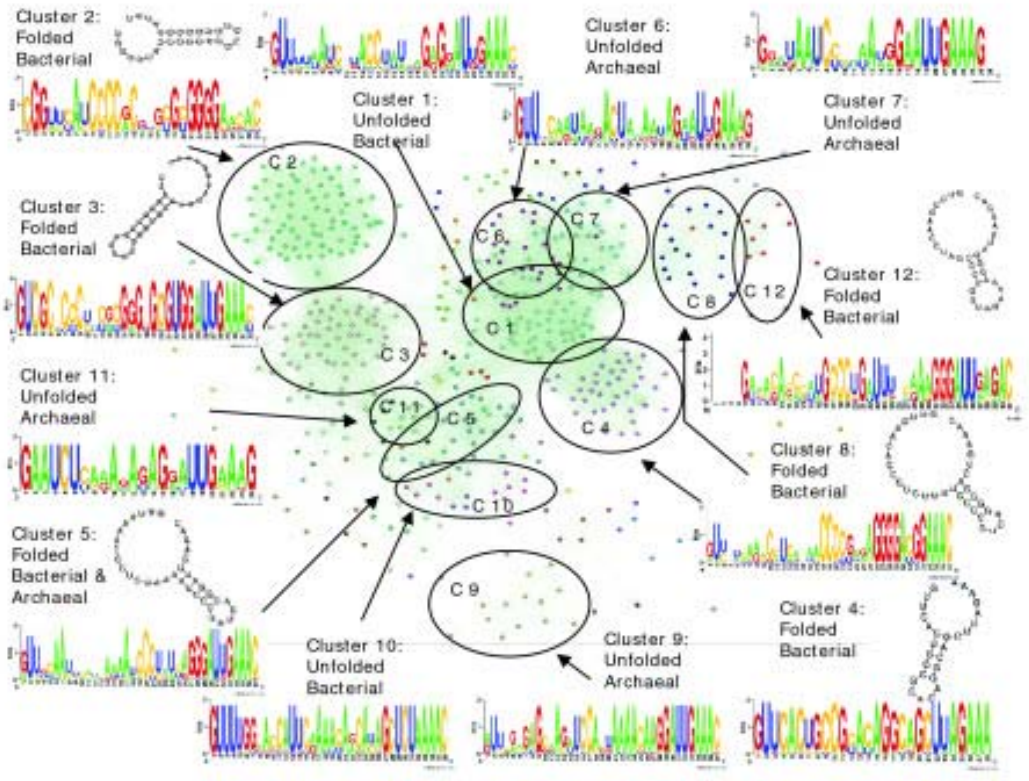
RAMP (Repeat-Associated Mysterious Proteins) superfamily: numerous families of Cas proteins, extreme sequence diversity

MOTIFs	specific	I	II	III	IV	V
COGs 1336,1367, 1604,1337,1332		h.h...s.h.hG.s	ust-lKGhh+.hh	hhGtt	h.D	lGht.t.s.G.h
y1726-like			slhlpEKuVRGT		lRTIDs	YGuVTs.Ghuh
COG1851			hGphpG.psaFh			hGFGRh
BH0337-like			T.pA-h+GIh-uIh		hhLpDV	LGsREh.u.ht
COG1567		.hhhp.p	up.s-lhtAh...h			lusc.o.GhG.h
COG1769		hhhh+Ph-.hhh	.s.s-hhGhlhs.h	h.G.h		hGtcp+hsthchp
COG1688 (Cas5)		h..h...hh.ht.s	ss.sshhGhl..sh			lGttp..h.h
COGs 1583,5551	hh.hhOPhhl					hGtppshGFG.l
YgcH-like	hHphlh					hG.u+uhGhGhh
y1727-like	LHphLh					.G.FsthGLStss
MJ0978-like	hHNH					lG+tsuhGhGol



Ferredoxin fold/ RNA-recognition motif (RRM)

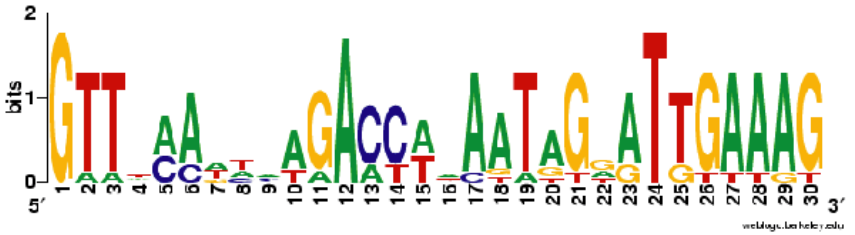
CRISPR show extreme diversity and complex clustering

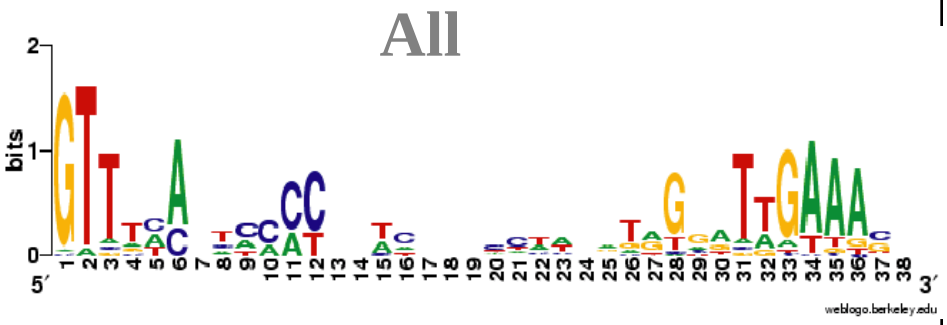


The sequence similarity space of CRISPR repeats visualized with the BioLayout program [26]. Dots denote individual repeat sequences; connecting lines represent Smith-Waterman similarities, such that closer dots represent more similar sequences. Dot colors denote cluster association as derived from MCL clustering. The 12 largest clusters are indicated by circles together with their sequence logos, coarse phylogenetic composition, and sample secondary structures where applicable.
Kunin *et al. Genome Biology* 2007 **8**:R61 doi:10.1186/gb-2007-8-4-r61

A large subset of CRISPRs is conserved among diverse species, even between archaea and bacteria, suggesting that CRISPRs are horizontally transferred together with *cas* genes

```
Arcful    GTTGAAATC-------AGACCAAAATGGGATTGAAAG-
Metthe    GTTAAAATC-------AGACCAAAATGGGATTGAAAT-
Pyraby    GTTCCAATA-------AGACTAGAATAGAATTGAAAG-
Pyrhor    GTTTAATAA--------GACTAAAATAGAATTGAAAG-
Sulsol    GATTAATCC------------CAAAAGAATTGAAAG-
Pyrhor    GTTTCCGTA--------GAACTAAATAGTGTGGAAAG-
Pyraby    GTTTCCGTA--------GAACTTAGTAGTGTGGAAAG-
Pyraer    GTTTCAACT------------ATCTTTTGATTTTTGG-
Pictor    GTTAAATAA-------TAACCTAAATAGGATTGAAAG-
Theaer    GTAAAATAG---------ACCTTAATAGGATTGAAAG-
Metace    GTTTCAATC-------CCTCAAAGGTCTGATTTTAAC-
Sultok    GATGAATCC------------CAAAAGGAATTGAAAG-
Sulaci    GTTTTAGTT------------TCTTGTCGTTATTAC-
Pictor    GTTTAAGAA--------TTACTAGATAGTATGGAGT--
Halmar    GTTTCAGAC-------GGACCCTTGTGGGGTTGAAGC-
Metjan    ATTAAAATC-------AGACCGTTTCGGAATGGAAAT-
Metkan    GTTTCATTA-CCCGTATTATTACGGGTTAATTGCGAG-
Nanequ    CTTTCAATA---------TTTCTAATATATTAGAAAC-
Themar    GTTTCAATA-------CTTCCTTAGAGGTATGGAAAC-
Theten    GTTTCAATC-------CCTTTTAGGTAGGCTAAAAAC
Thethe    GTTGCAAAC-CTCGTTAGCCTCGTAGAGGATTGAAAC-
Aquaeo    GTTTTAACT--------CCACACGGTACATTAGAAAC-
Bachal    GTCGCACTC----TATATGGGT-GCGTGGATTGAAAT-
Azoarcus  GTGTTCCCC-------GCGCATCGCGGGGGTTGAAG--
Chltep    GTCTTCCCC--------ACGCC-CGTGGGGGTGTTTC-
Chrvio    GTGCTCCCC-------CACGCA-CGTGGGGATGAACCG
Clotet    GTATTAGTA-------GCACCA-TATTGGAATGTAAAT
Anabaena  GTTTTAATTAACAAAAATCCCTATCAGGGATTGAAAC-
Myctub    GTTTCCGTC--CCCTCTCGGGGTTTTGGGTCTGACGAC
Yerpes    GTTCACTGC--------CGCACAGGCAGCTTAGAAA--
Metcap    GTTTCAATCCACTCCCGGCTATTTAGCCGGGAGATAC-
Mycgal    GTTTTAGCA-CTGTACAATACTTGTGTAAGCAATAAC-
Pasmul    GTTAACTGC--------CGTATAGGCAGCTTAGAAA--
Pasmul    GTTGTAGTTCCCTCTCTCATTTCGCAGTGCTACAAT--
Pasmul    GTTCACCAT--------CGTGTAGATGGCTTAGAAA--
```





Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Soria E.

Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements.
J Mol Evol. 2005 Feb;60(2):174-82

Here we show that CRISPR spacers derive from preexisting sequences, either chromosomal or within transmissible **genetic elements such as bacteriophages and conjugative plasmids.**

……
The transcription of the CRISPR loci (*Tang et al. 2002*) suggests that such activity could be executed by CRISPR-RNA molecules, acting as regulatory RNA that specifically recognizes the target through the homologous RNA-spacer sequence, **similarly to the eukaryotic interference RNA**.

Biology Direct

BioMed Central

Research

Open Access

A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action

Kira S Makarova[1], Nick V Grishin[2], Svetlana A Shabalina[1], Yuri I Wolf[1] and Eugene V Koonin*[1]

Address: [1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA and [2]Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9050, USA

Email: Kira S Makarova - makarova@ncbi.nlm.nih.gov; Nick V Grishin - grishin@chop.swmed.edu; Svetlana A Shabalina - shabalin@ncbi.nlm.nih.gov; Yuri I Wolf - wolf@ncbi.nlm.nih.gov; Eugene V Koonin* - koonin@ncbi.nlm.nih.gov

* Corresponding author

Published: 16 March 2006

Biology Direct 2006, 1:7 doi:10.1186/1745-6150-1-7

Received: 08 February 2006
Accepted: 16 March 2006

This article is available from: http://www.biology-direct.com/content/1/1/7

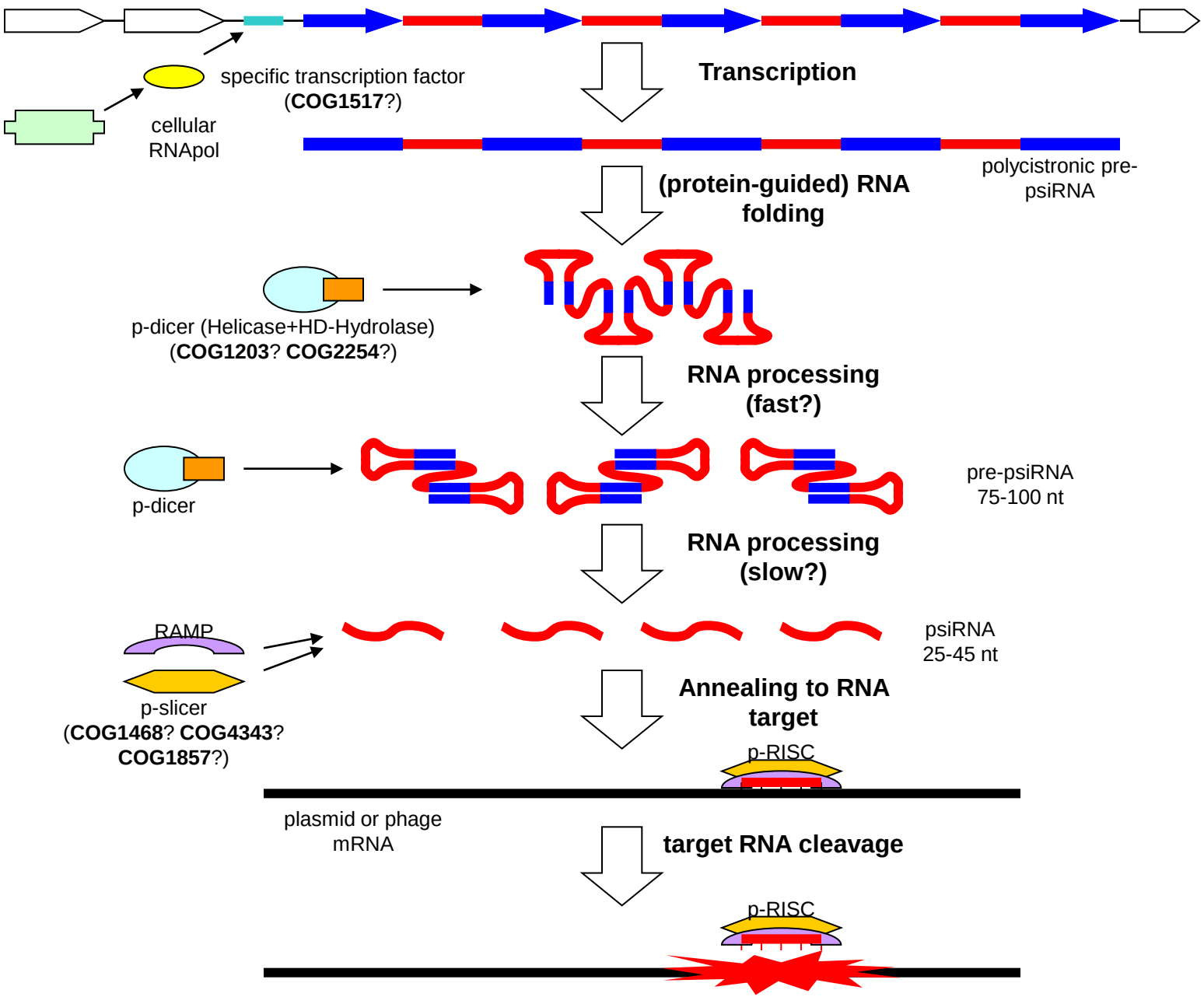
Abstract

Hypothesis

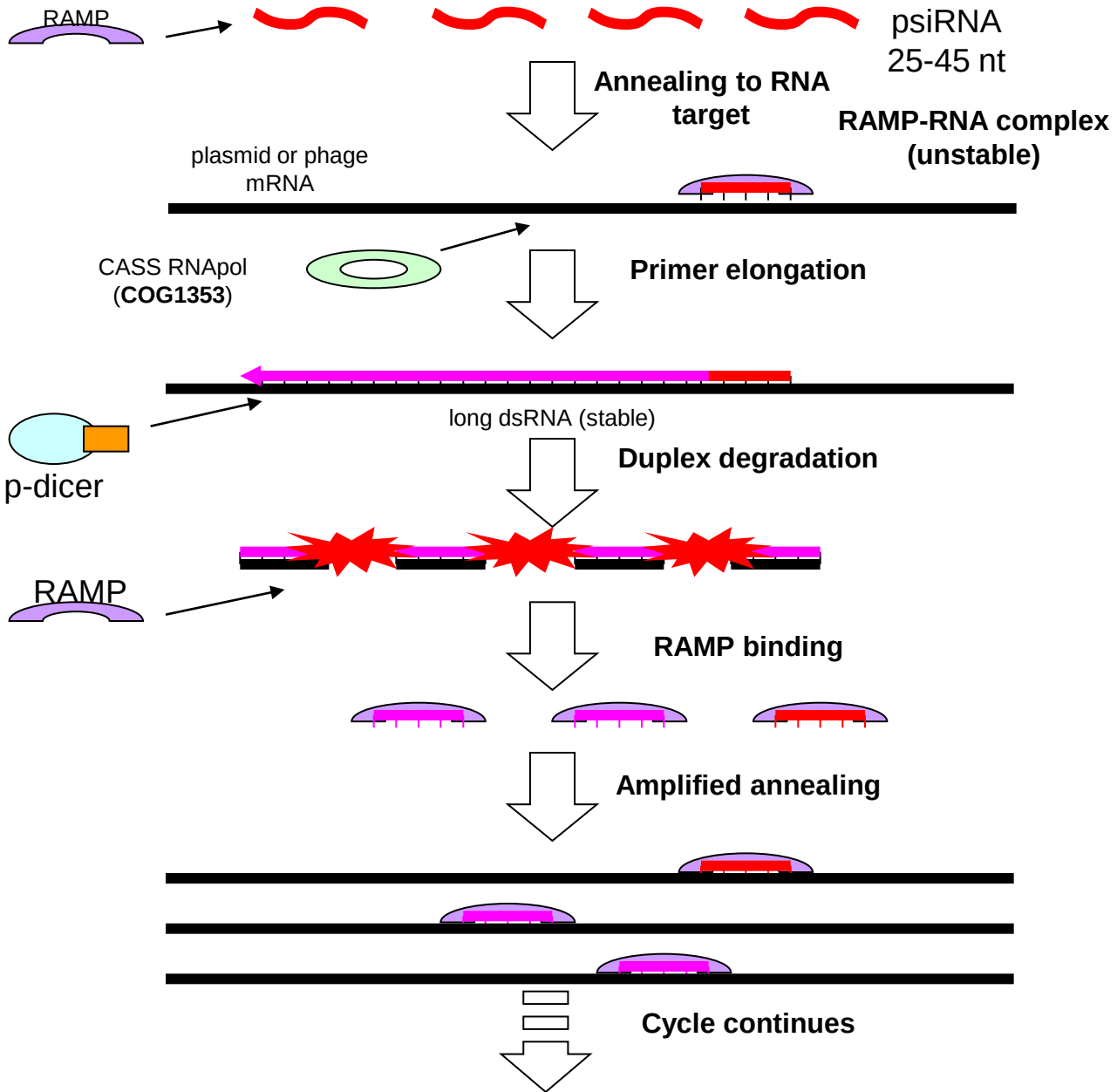
CRISPR/Cas

- is a prokaryotic immunity system that functions on the RNAi principle
- integrates short fragments of essential phage/plasmid genes into CRISPRs
- When expressed, these fragments (psiRNA – after prokaryotic siRNA) silence the target gene and make the organism immune to the respective agent
- contains all or most of the protein activities involved in these processes
- Some of the Cas proteins are functional analogs of the eukaryotic proteins involved in RNAi, in particular, components of RISC (RNA-Induced Silencing Complex), and form prokaryotic analogs of RISC (pRISCs)

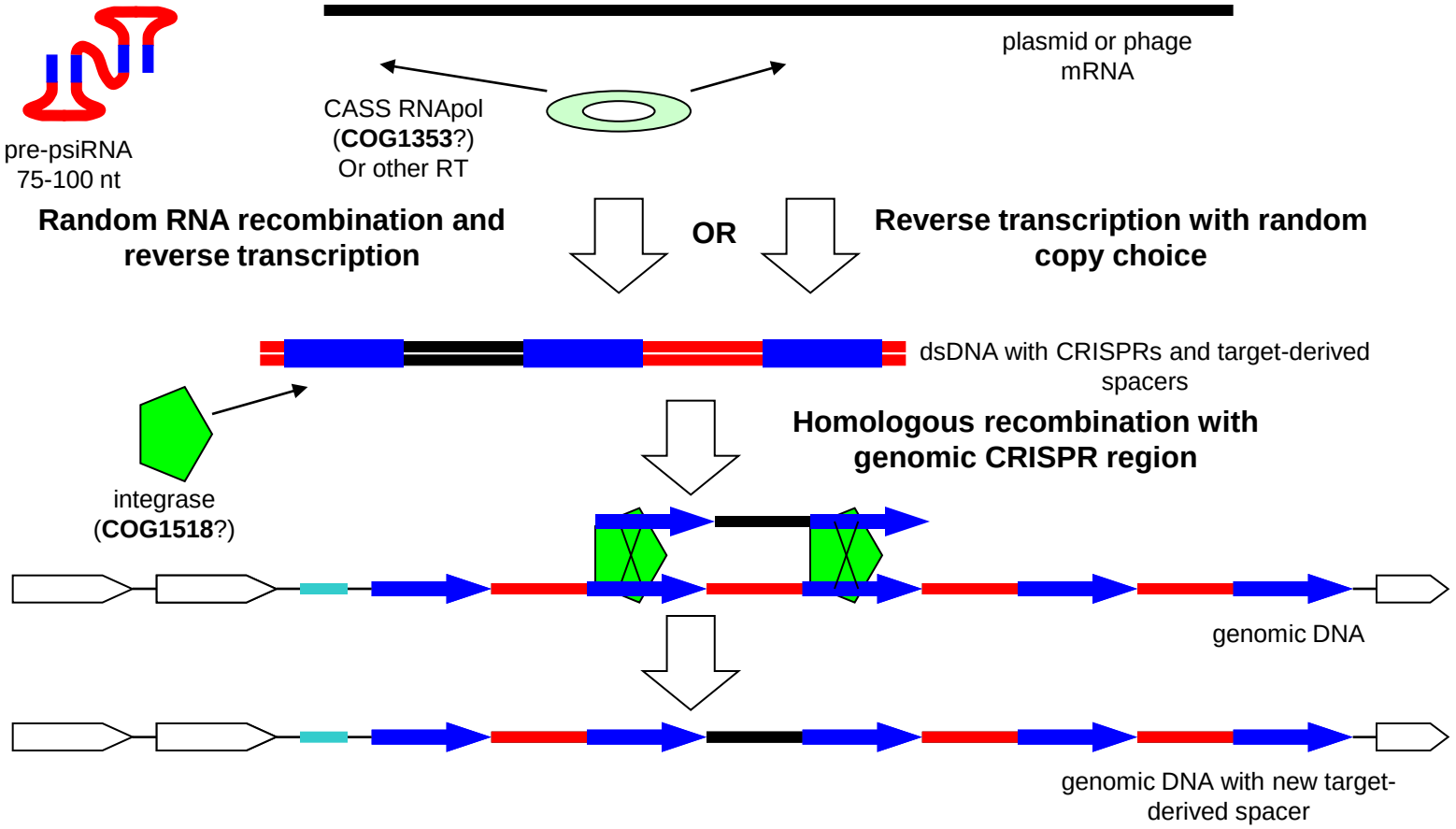
The basic scheme of CASS functioning



Variant of CASS functioning with polymerase/psi-RNA amplification (mostly, in thermophiles, but also in Mycobacteria)



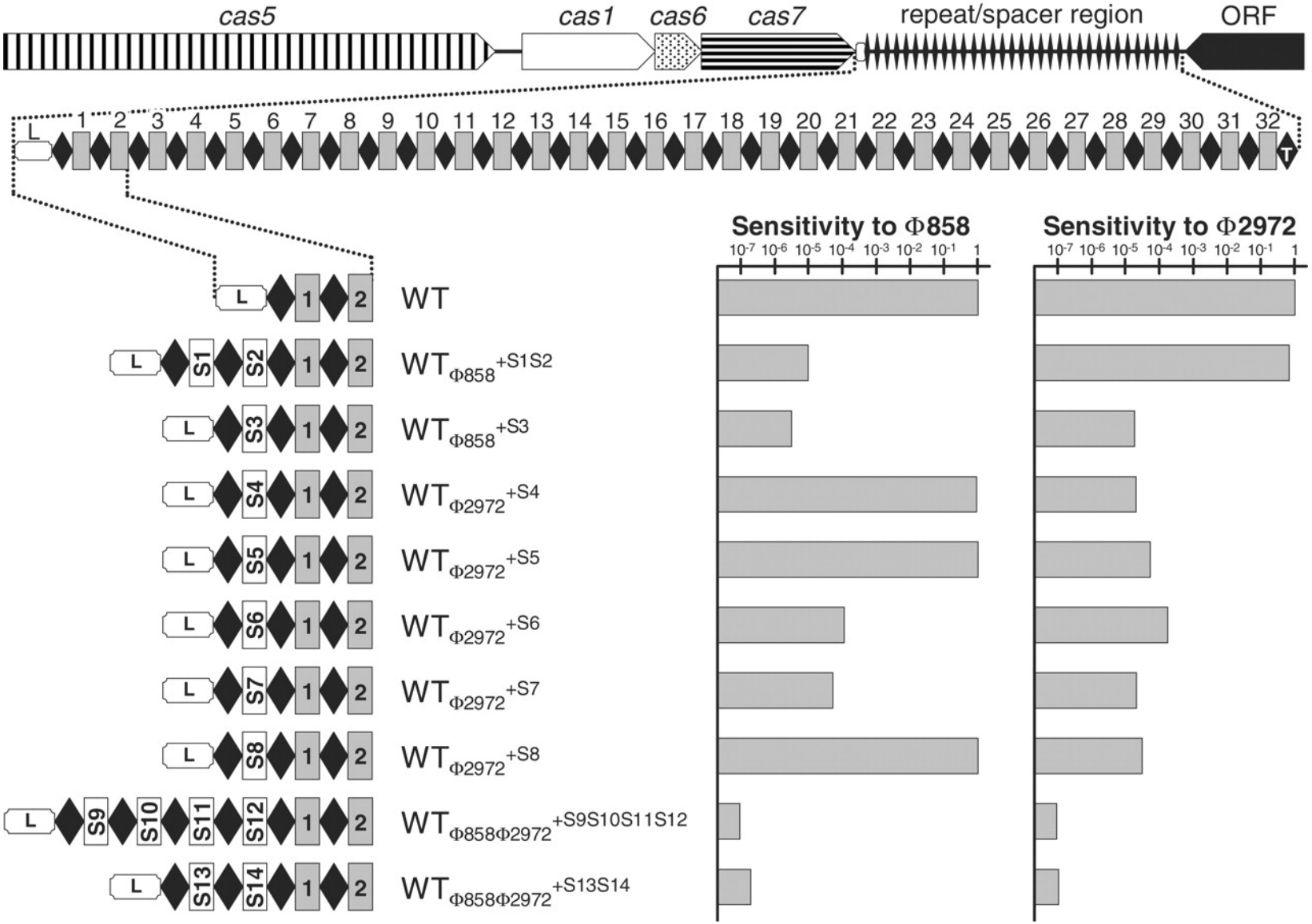
New psiRNA generation



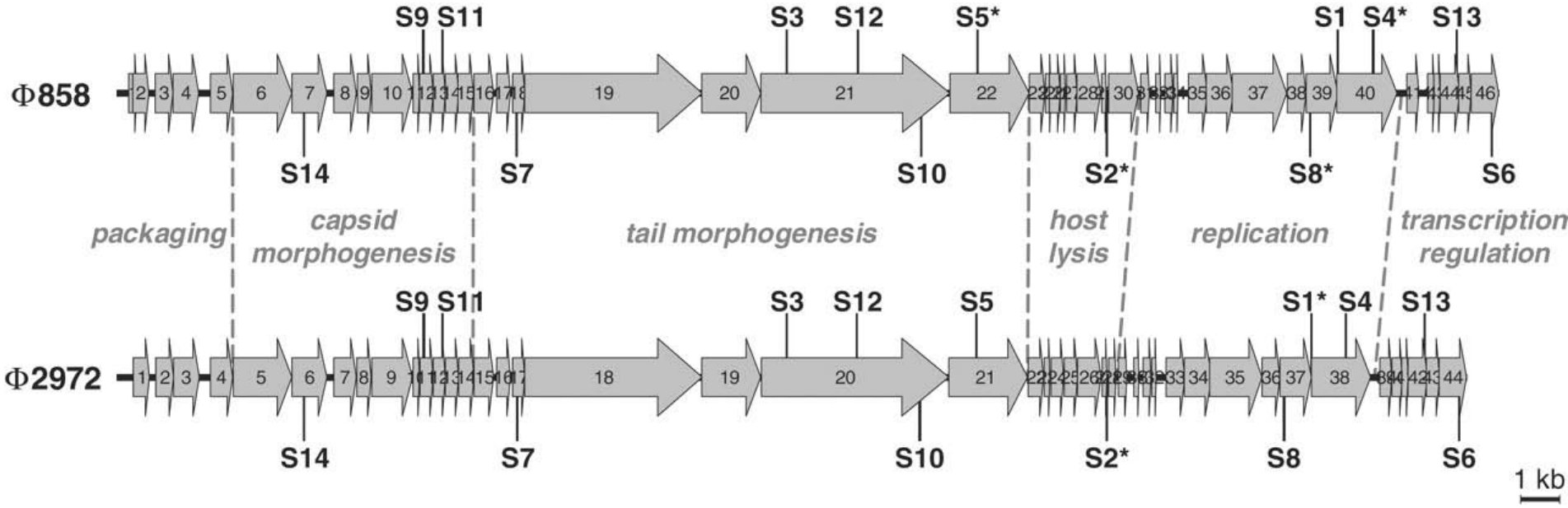
Key validation:

Barrangou R, **Fremaux C**, **Deveau H**, **Richards M**, **Boyaval P**, **Moineau S**, **Romero DA**, **Horvath P**.
CRISPR provides acquired resistance against viruses in prokaryotes. Science. 2007 Mar 23;315(5819):1709-12

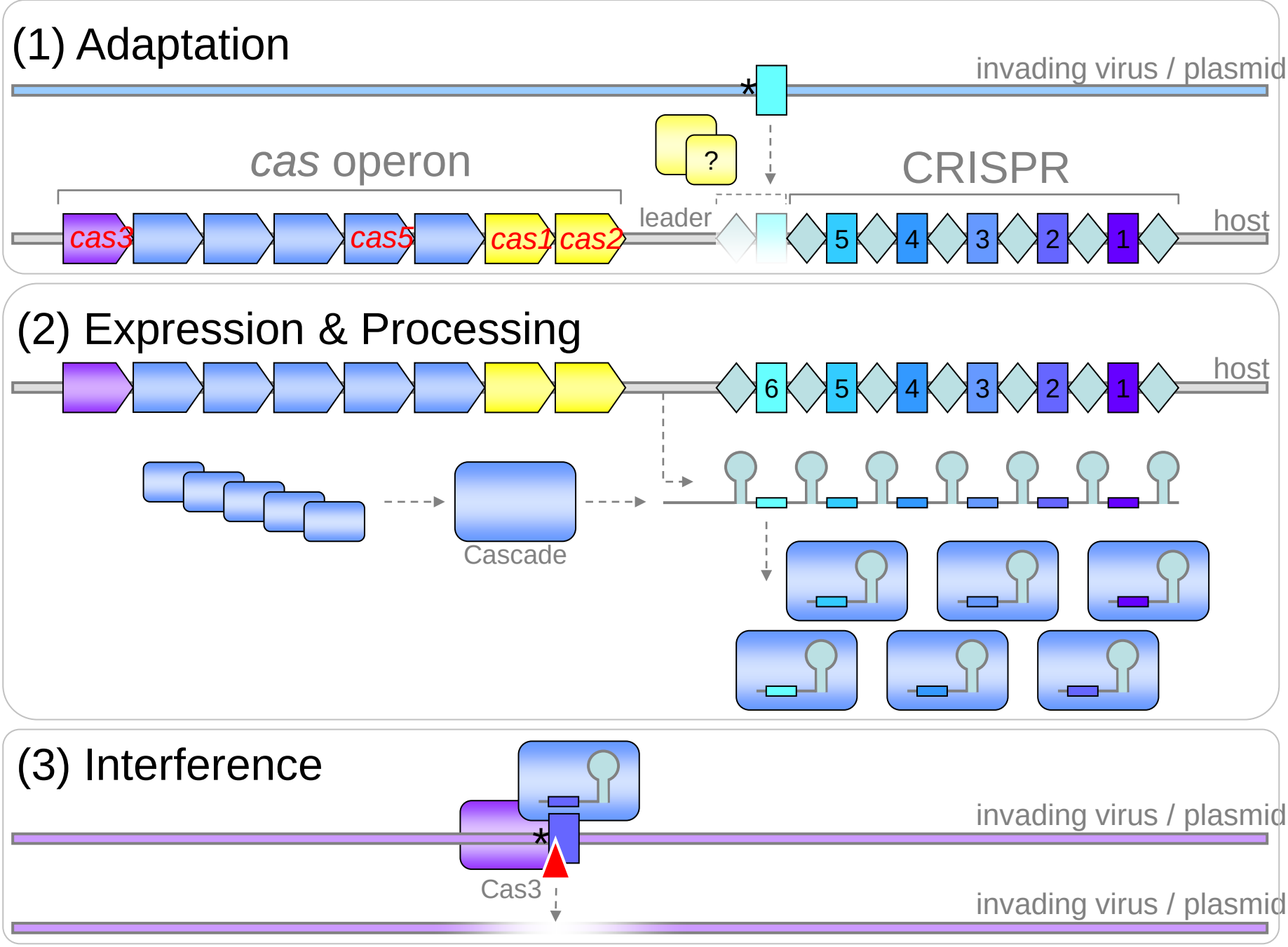
Clustered regularly interspaced short palindromic repeats (CRISPR) are a distinctive feature of the genomes of most Bacteria and Archaea and are thought to be involved in resistance to bacteriophages. We found that, after viral challenge, bacteria integrated new spacers derived from phage genomic sequences. Removal or addition of particular spacers modified the phage-resistance phenotype of the cell. Thus, CRISPR, together with associated cas genes, provided resistance against phages, and resistance Specificity is determined by spacer-phage sequence similarity.



- **Phage-specific inserts confer resistance that is highly sequence-specific: a single substitution (SNP) reverts to sensitivity**
- **The spacers worked only when inserted between CRISPR**
- **Resistance required COG3513 (cas5), a predicted nuclease**

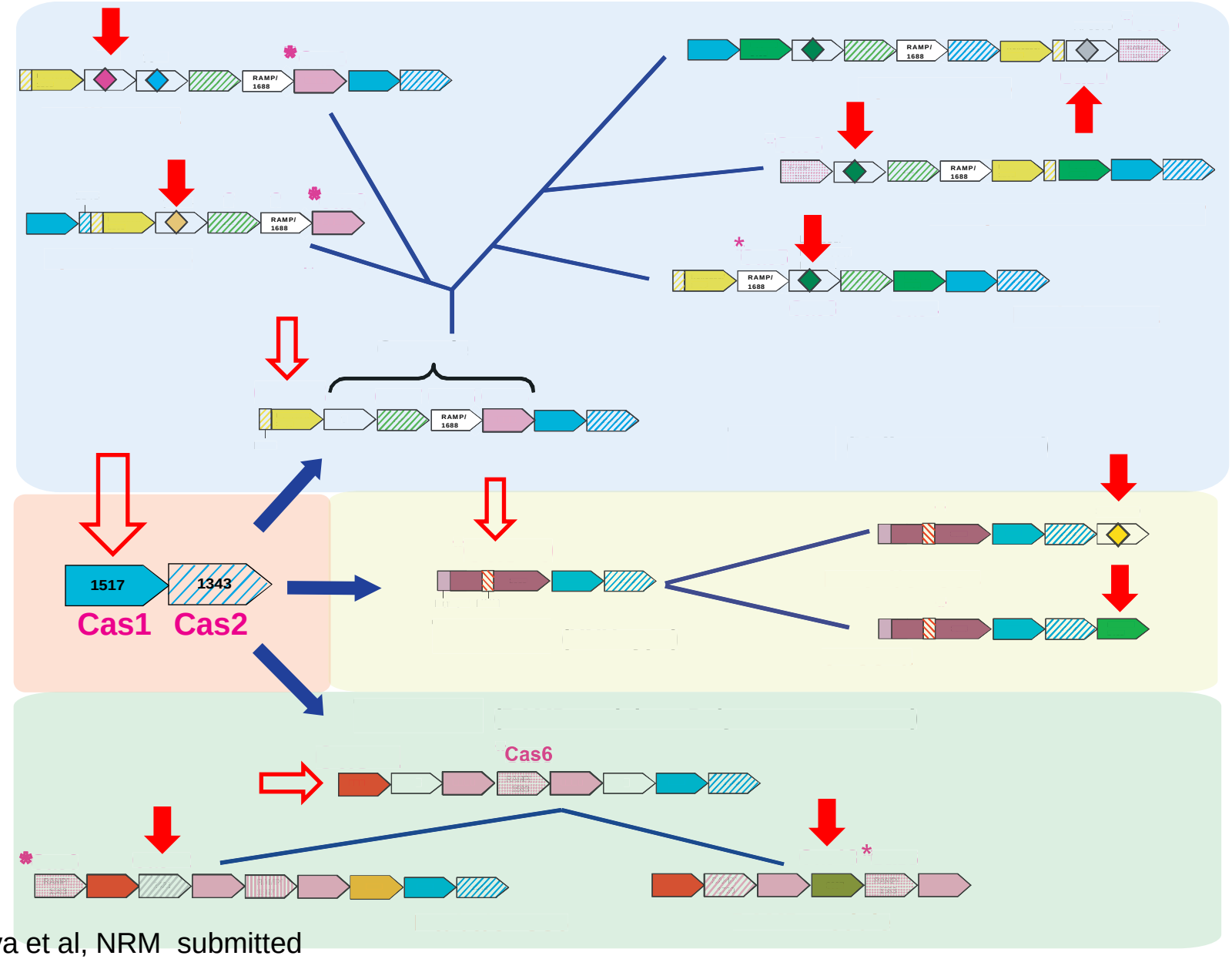


Inserts from phage-resistant mutants were homologous to regions scattered over the phage genomes



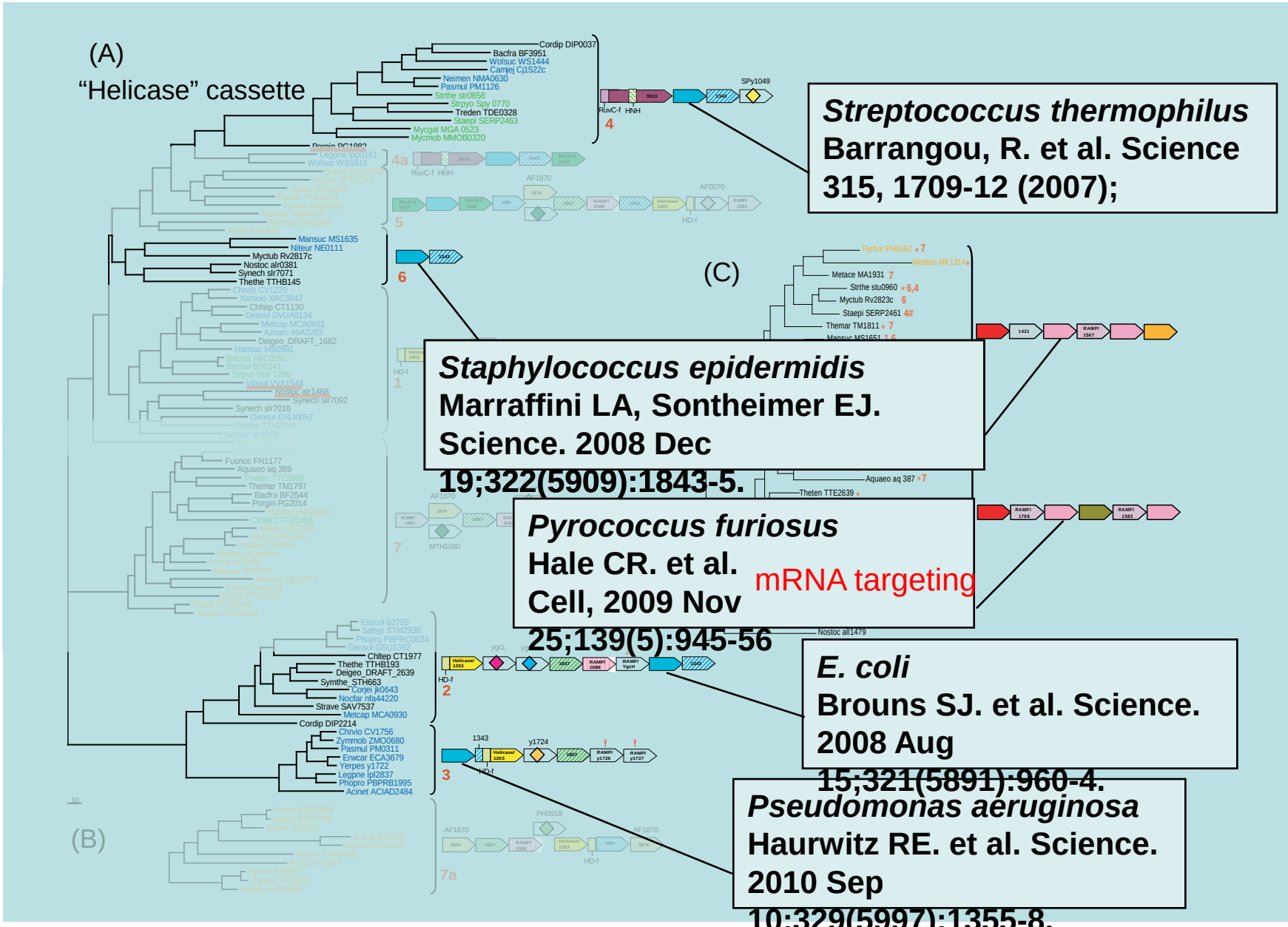
Van der Oost et al. TIBS 2009

The three types of CRISPR/Cas systems and their signature genes

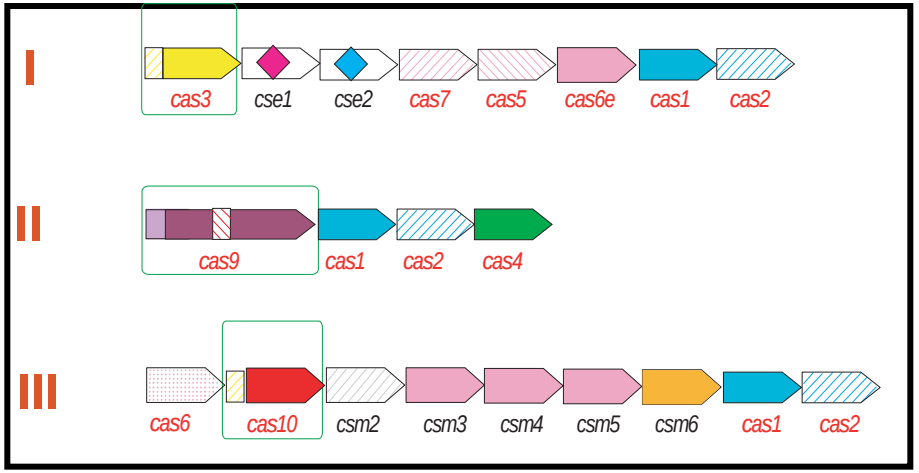


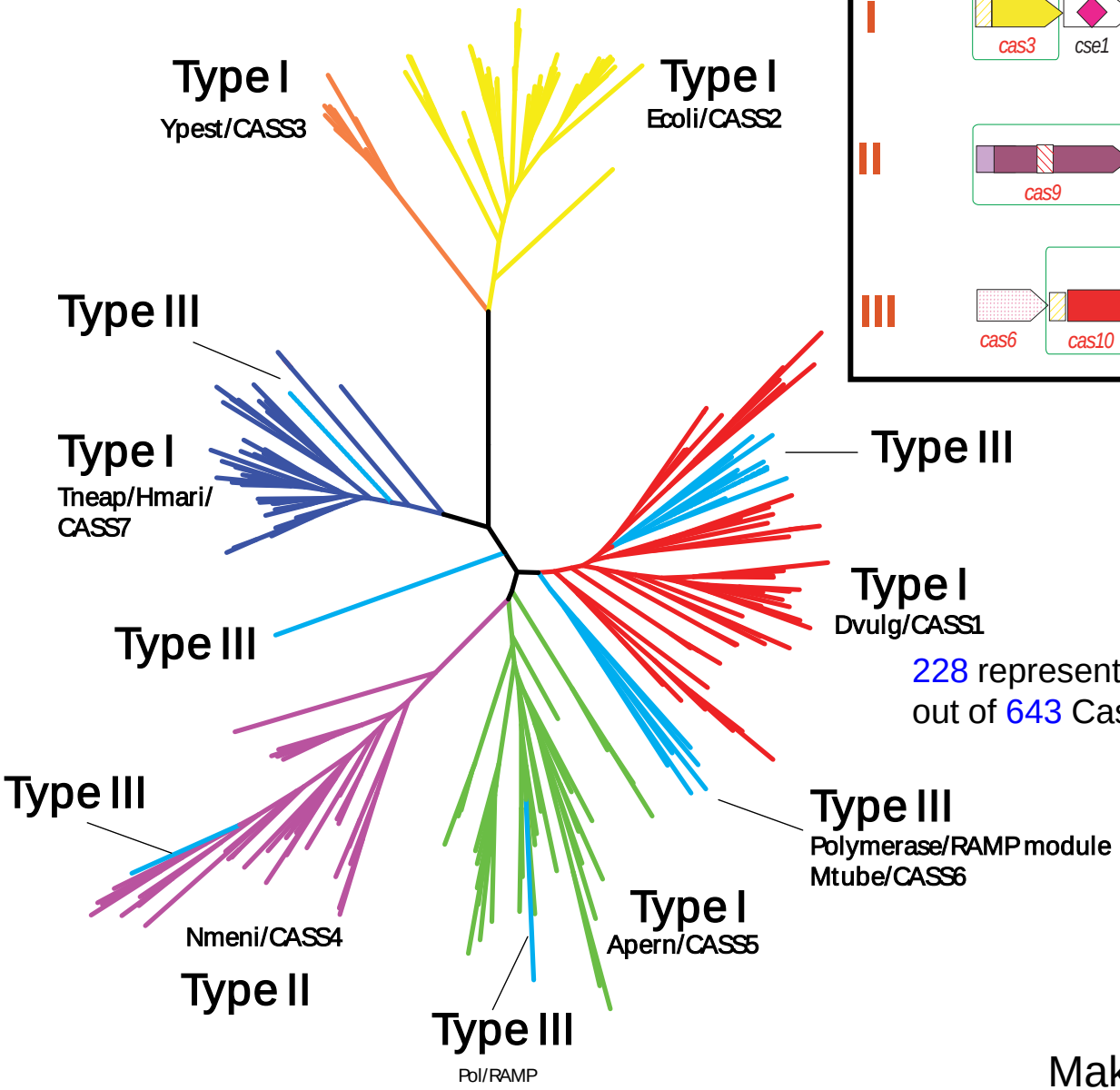
Makarova et al, NRM submitted

Experimental data on CRISPR/Cas systems



Phylogeny of Cas1 and the 3 types of CRISPR/Cas systems

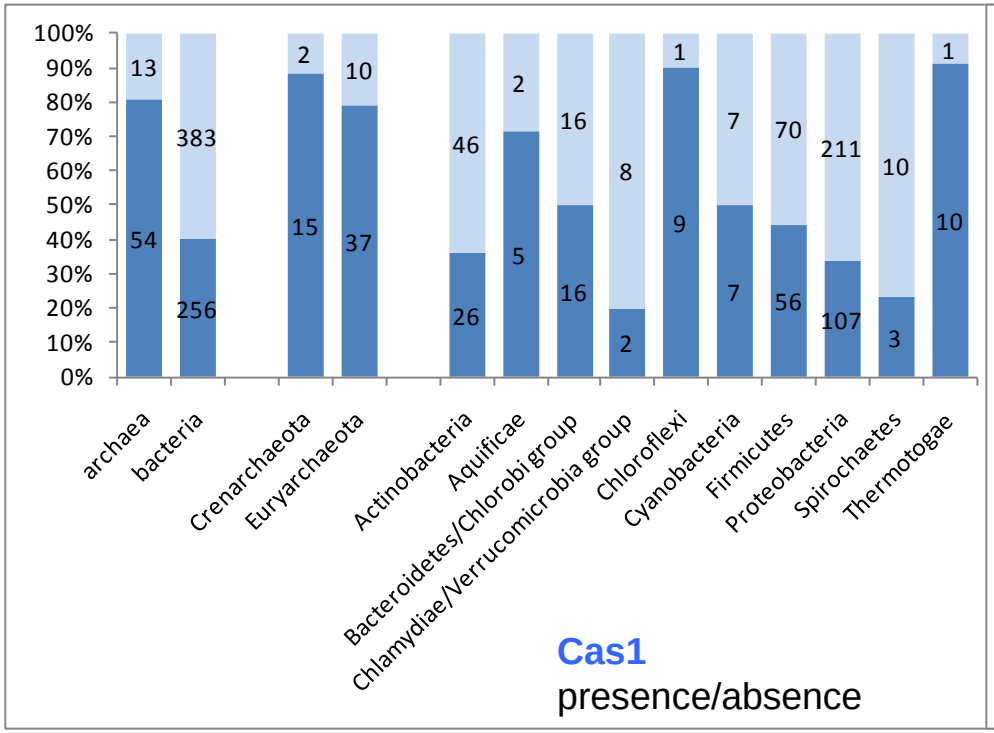


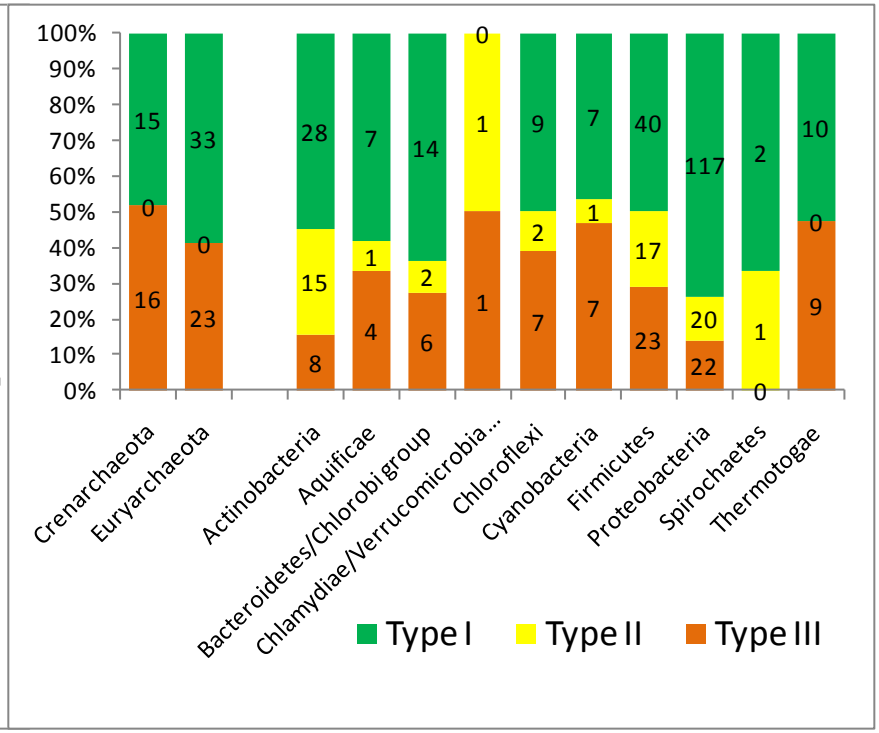


228 representative Cas1 sequences, selected out of 643 Cas1 proteins in 442 genomes

Makarova et al., NRM submitted

CRISPR/CAS systems in 703 selected complete genomes of archaea and bacteria

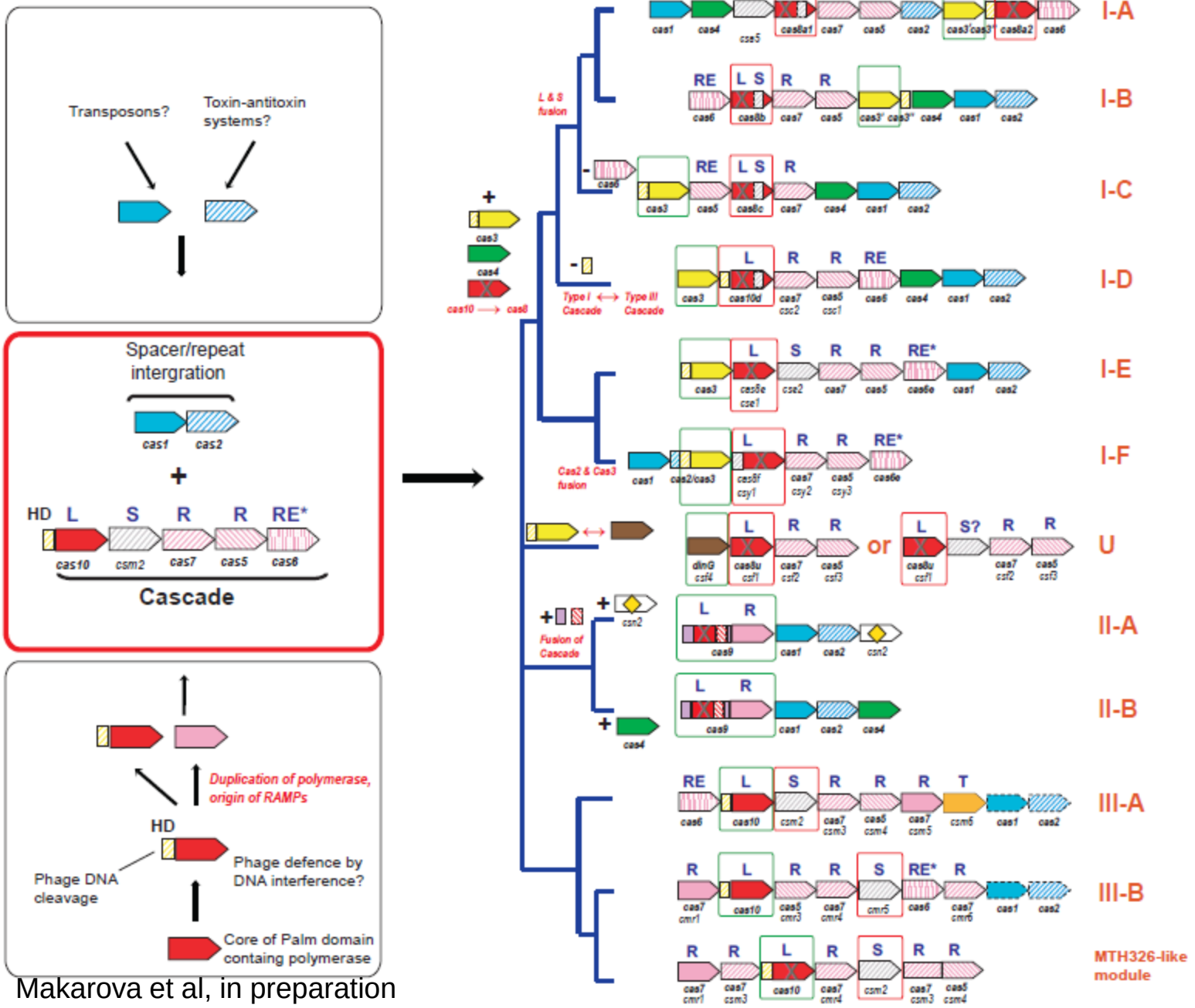




- Cas1 is present in 310 (44%) genomes
- ~90% archaea but only ~35% of bacteria
- Type I is present in 42% genomes; Type II – 9%; Type III – 20%;
- Two or three systems of different types are present in 128 (20%) genomes

Makarova et al, in preparation

Modular evolution of the 3 types of CRISPR/Cas



Makarova et al, in preparation

Back to repair functions?

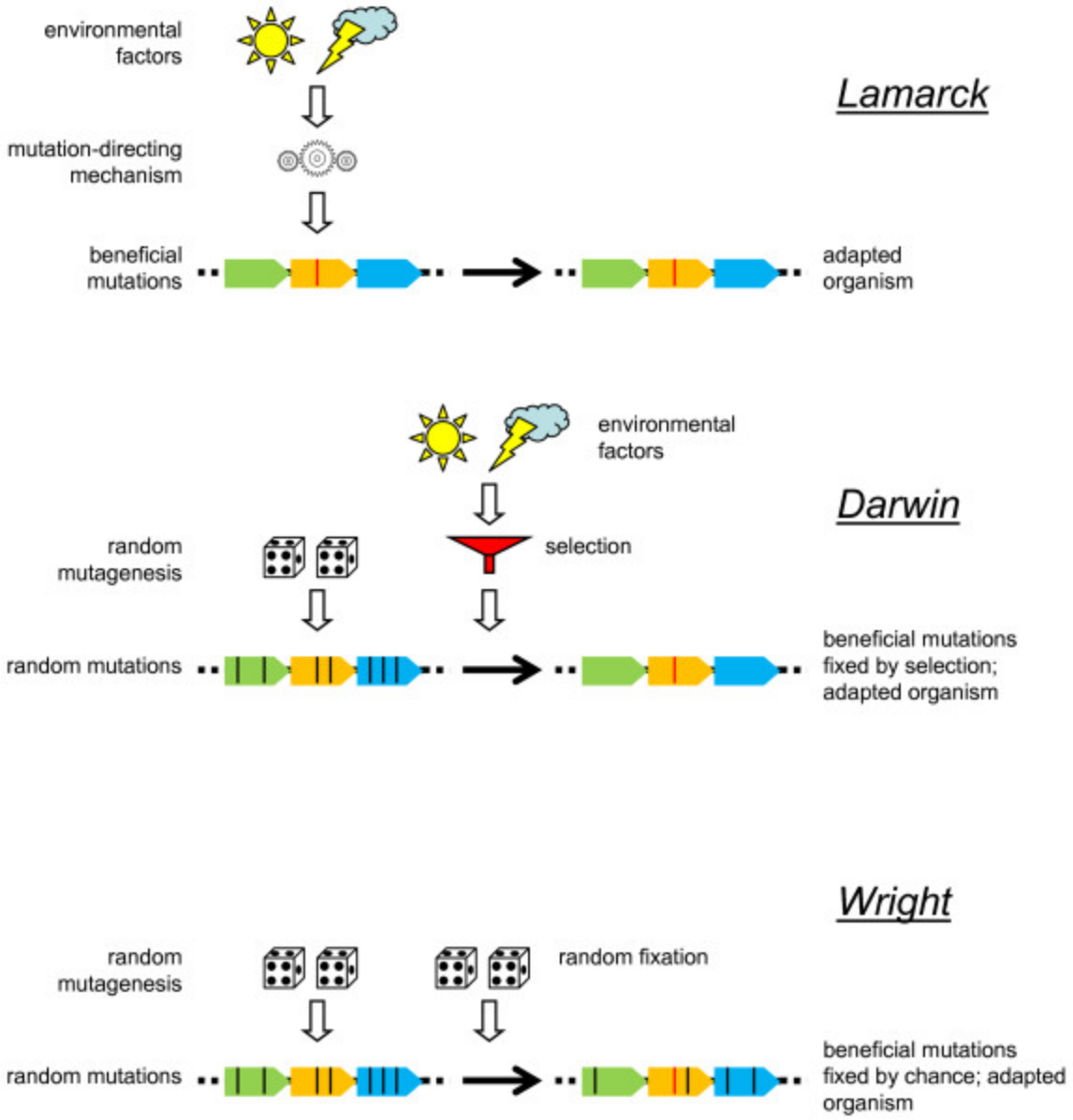
Mol Microbiol. 2011 Jan;79(2):484-502.

A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair.

Babu M, Beloglazova N, Flick R, Graham C, Skarina T, Nocek B, Gagarinova A, Pogoutse O, Brown G, Binkowski A, Phanse S, Joachimiak A, Koonin EV, Savchenko A, Emili A, Greenblatt J, Edwards AM, Yakunin AF.

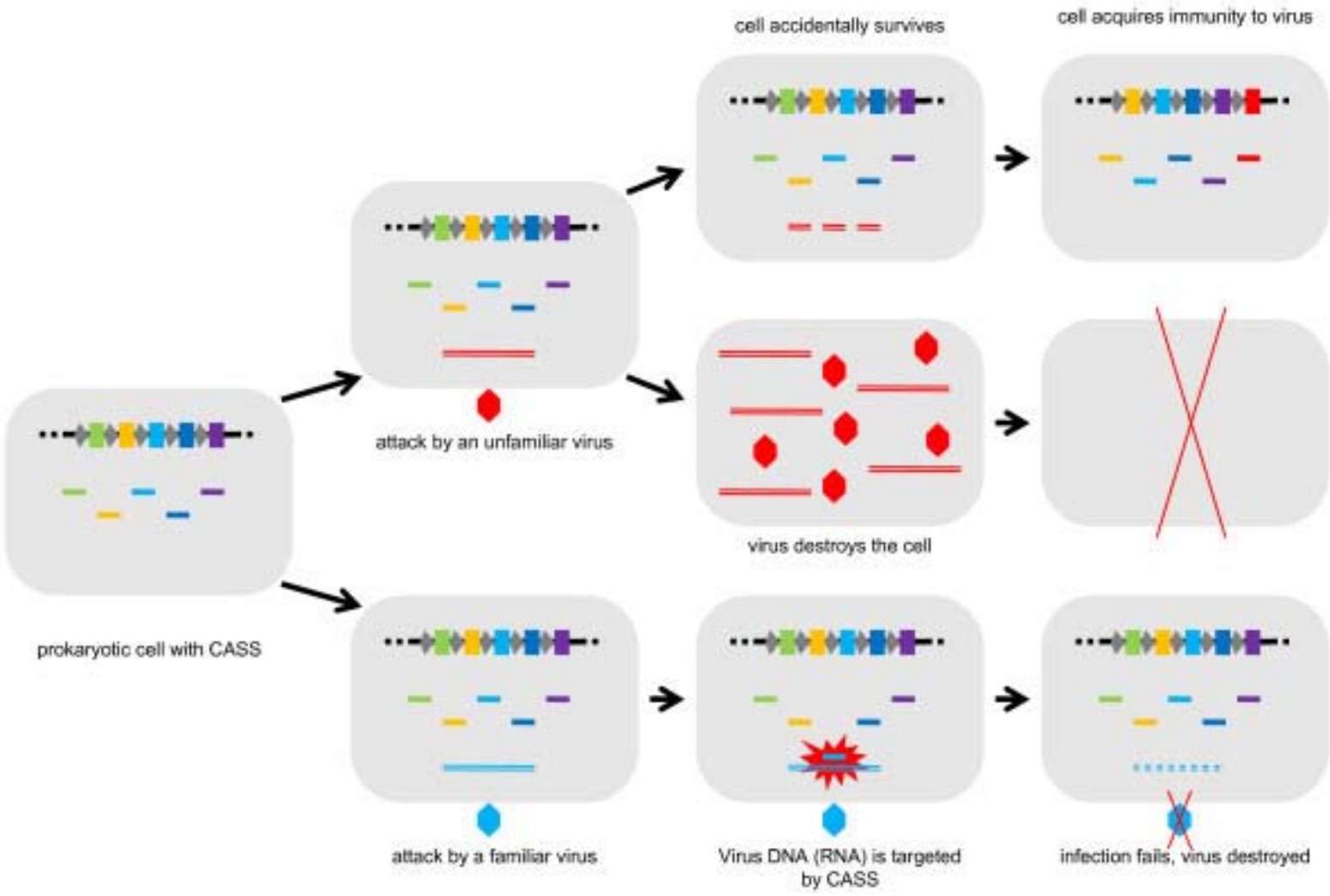
Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) and the associated proteins (Cas) comprise a system of adaptive immunity against viruses and plasmids in prokaryotes. Cas1 is a CRISPR-associated protein that is common to all CRISPR-containing prokaryotes but its function remains obscure. Here we show that the purified Cas1 protein of Escherichia coli (YgbT) exhibits nuclease activity against single-stranded and branched DNAs including Holliday junctions, replication forks and 5'-flaps. The crystal structure of YgbT and site-directed mutagenesis have revealed the potential active site. Genome-wide screens show that YgbT physically and genetically interacts with key components of DNA repair systems, including recB, recC and ruvB. Consistent with these findings, the ygbT deletion strain showed increased sensitivity to DNA damage and impaired chromosomal segregation. Similar phenotypes were observed in strains with deletion of CRISPR clusters, suggesting that the function of YgbT in repair involves interaction with the CRISPRs. **These results show that YgbT belongs to a novel, structurally distinct family of nucleases acting on branched DNAs and suggest that, in addition to antiviral immunity, at least some components of the CRISPR-Cas system have a function in DNA repair**.

The 3 major modalities of evolution



Koonin, Wolf, Biol. Direct 2009

CRISPR/Cas as a bona fide Lamarckian system



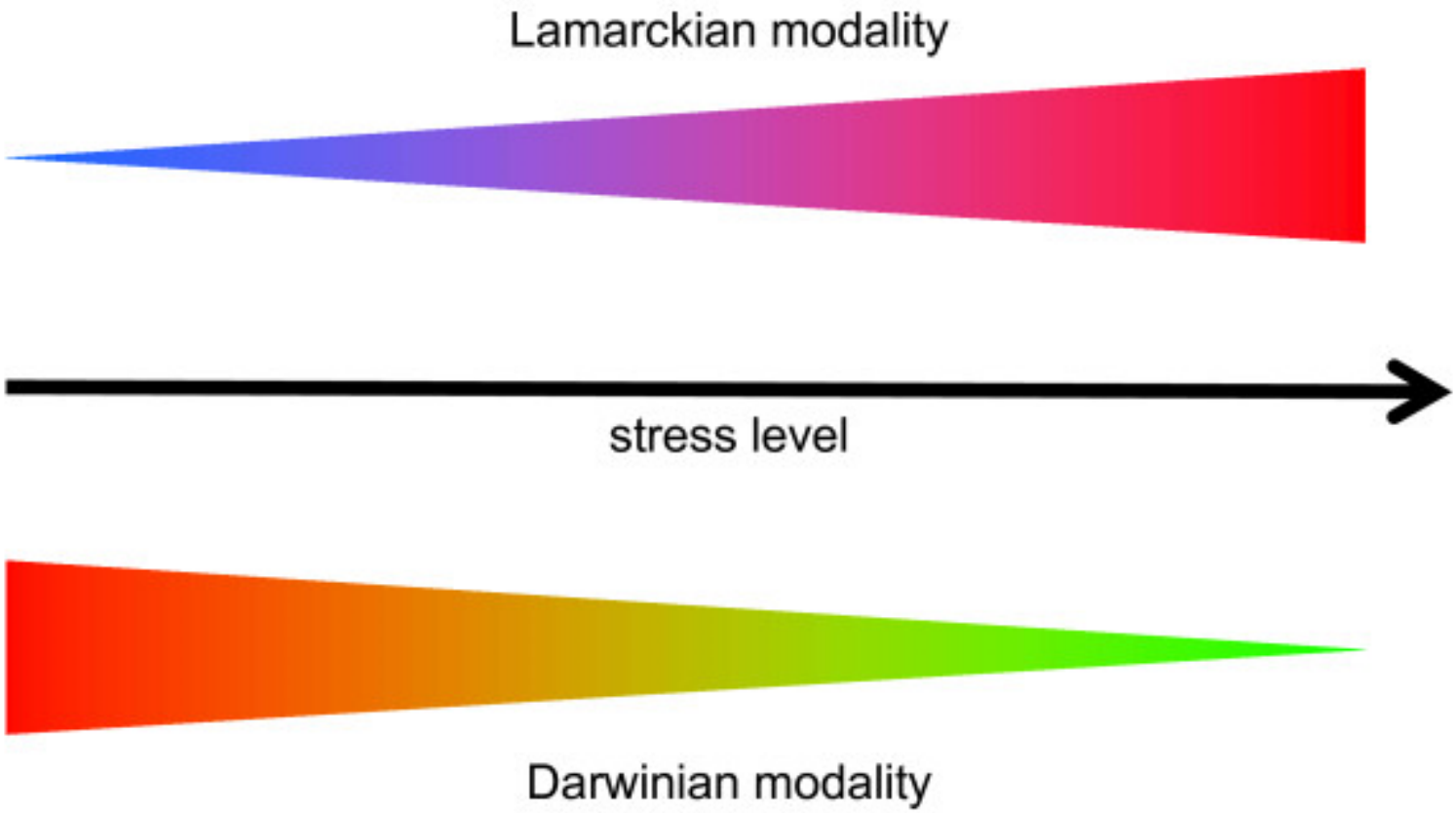
Koonin, Wolf, Biol. Direct 2009

Diverse Lamarckian and quasi-Lamarckian phenomena

Phenomenon	Biological role/function	Phyletic spread	Lamarckian criteria		
			Genomic changes caused by environmental factor	Changes are specific to relevant genomic loci	Changes provide adaptation to the causative factor
Bona fide Lamarckian					
CRISPR/Cas	Defense against viruses and other mobile elements	Most of the Archaea and many bacteria	Yes	Yes	Yes
piRNA	Defense against transposable elements in germline	Animals	Yes	Yes	Yes
HGT (specific cases)	Adaptation to new environment, stress response, resistance	Archaea, bacteria, unicellular eukaryotes	Yes	Yes	Yes
Quasi-Lamarckian					
HGT (general phenomenon)	Diverse innovations	Archaea, bacteria, unicellular eukaryotes	Yes	No	Yes/no
Stress-induced mutagenesis	Stress response/resistance/ adaptation to new conditions	Ubiquitous	Yes	No or partially	Yes (but general evolvability enhanced as well)

Koonin, Wolf, Biol. Direct 2009

Stress as a gauge of evolutionary modality



Koonin, Wolf, Biol. Direct 2009

•**E**volution of parasites is intrinsic to any replicator system

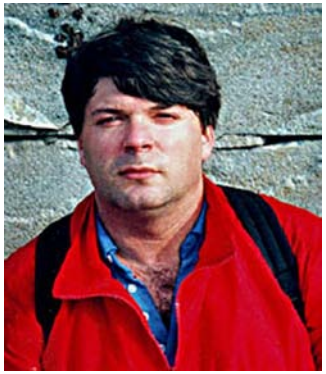
•**D**efense systems, in particular, those based on the RNAi principle, appeared concomitantly with cells and coevolved with cells and viruses ever since

•**D**efense systems occupy a substantial fraction of the genomes in all cellular life forms

•**P**erennial arms race between parasites and hosts is one of the principal factors of evolution

Valerian Dolja,
Oregon State University

Tatiana Senkevich,
NIAID, NIH

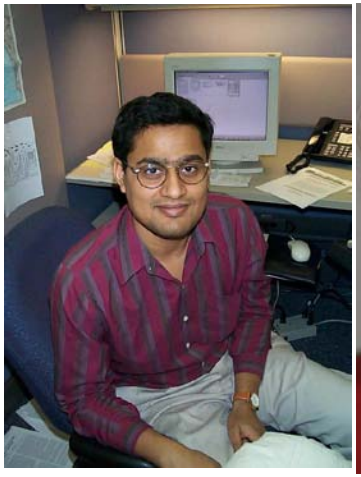


Yuri Wolf, NCBI

Kira Makarova, NCBI

Bill Martin,
Univ. Duesseldorf

L. Aravind, NCBI



Laks Iyer, NCBI

Natalia Yutin, NCBI

Didier Raoult **et al.**

Universite de la
Mediterranee-Marseille