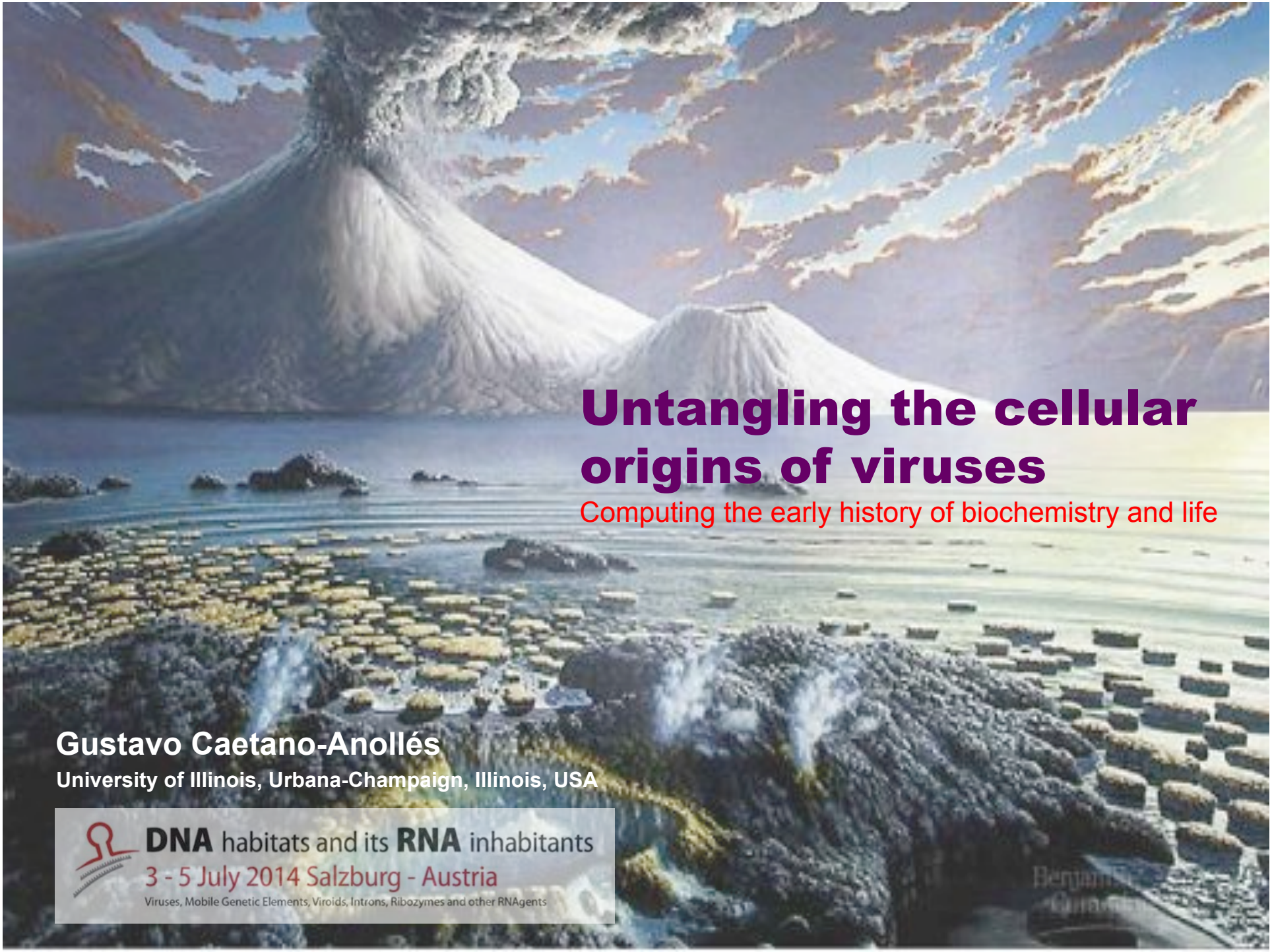


Untangling the cellular origins of viruses

Computing the early history of biochemistry and life

Gustavo Caetano-Anollés

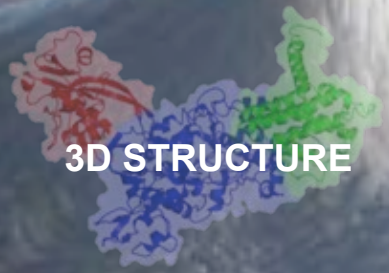
University of Illinois, Urbana-Champaign, Illinois, USA

DNA habitats and its **RNA** inhabitants

3 - 5 July 2014 Salzburg - Austria

Viruses, Mobile Genetic Elements, Viroids, Introns, Ribozymes and other RNAgents

Retrodictive exploration



QUESTIONS:

What is the role of viruses in cellular evolution?

Is there a truly universal tree of life?

Are viruses monophyletic or polyphyletic?

How are viral groups related to each other?

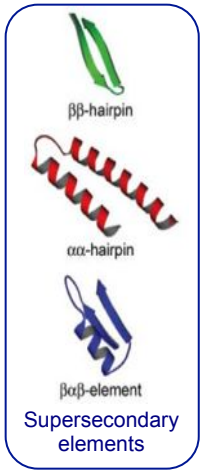
Is there a preferential direction of gene transfer?

Ideographic framework:
historical and retrodictive

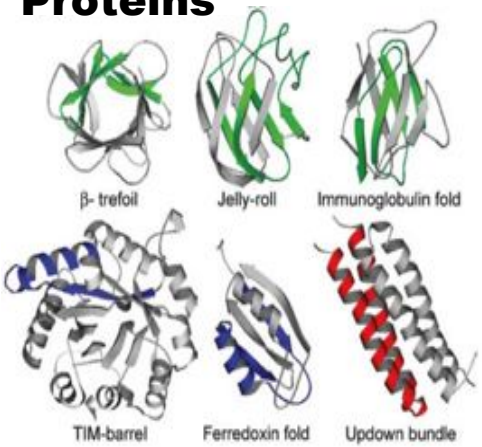
Structure provides a window into function and evolution

The macromolecular world is modular

Successful structural designs are recurrent

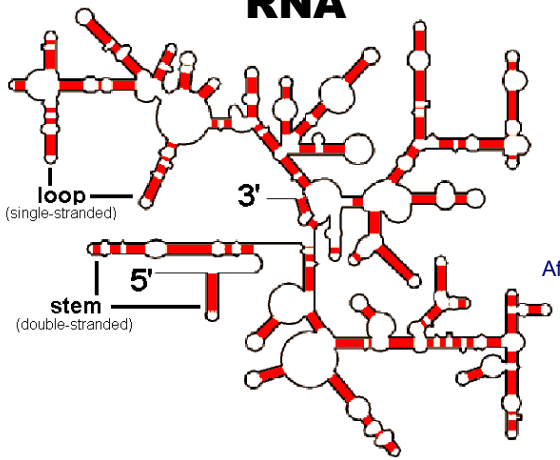


Proteins



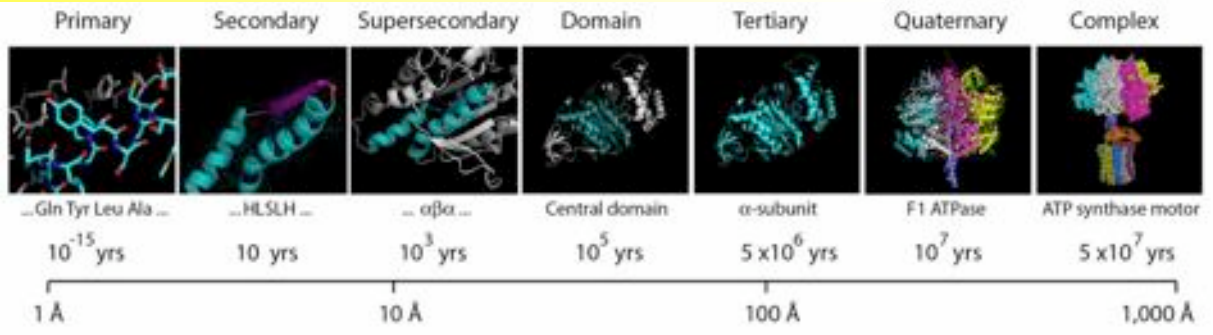
After Söding & Lupas (2003)

RNA



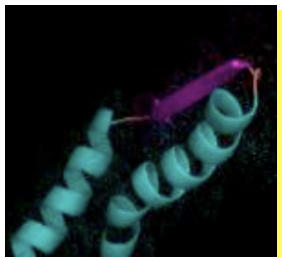
After Karp (1997)

Modules carry deep evolutionary signal

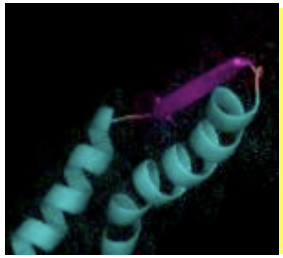


Evolutionary *conservation* increases with hierarchical complexity

Domains exhibit levels of structural abstraction



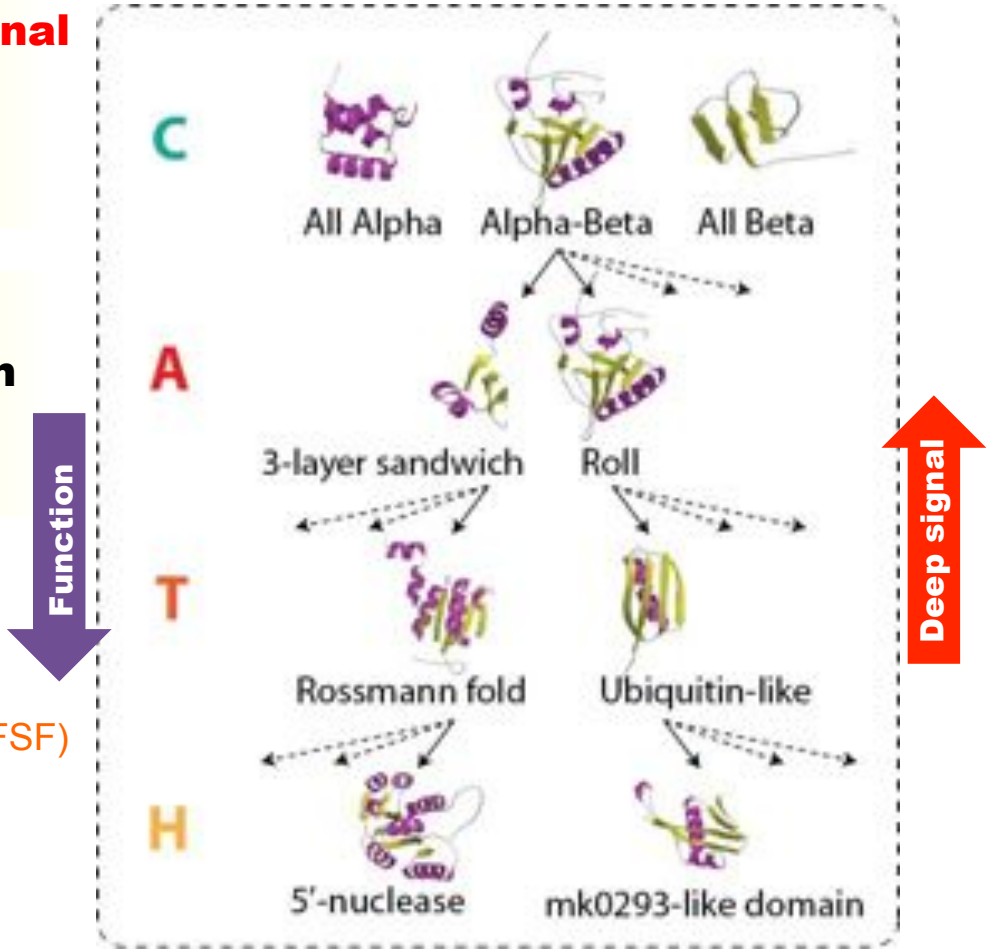
Deep phylogenetic signal increases with higher levels of structural abstraction



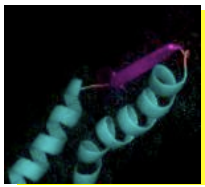
Modern functional definitions decrease in significance at higher abstraction levels

CATH:
Classes (C)
Architectures (A)
Topologies (T)
Superfamilies (S)
Domains

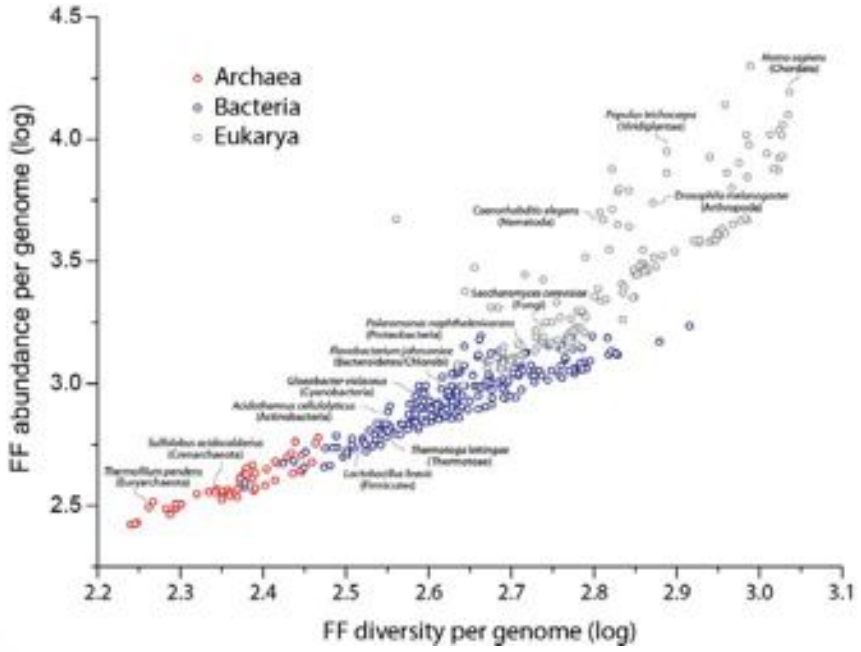
SCOP:
Classes (C)
Folds (F)
Fold superfamilies (FSF)
Fold families (FF)
Domains

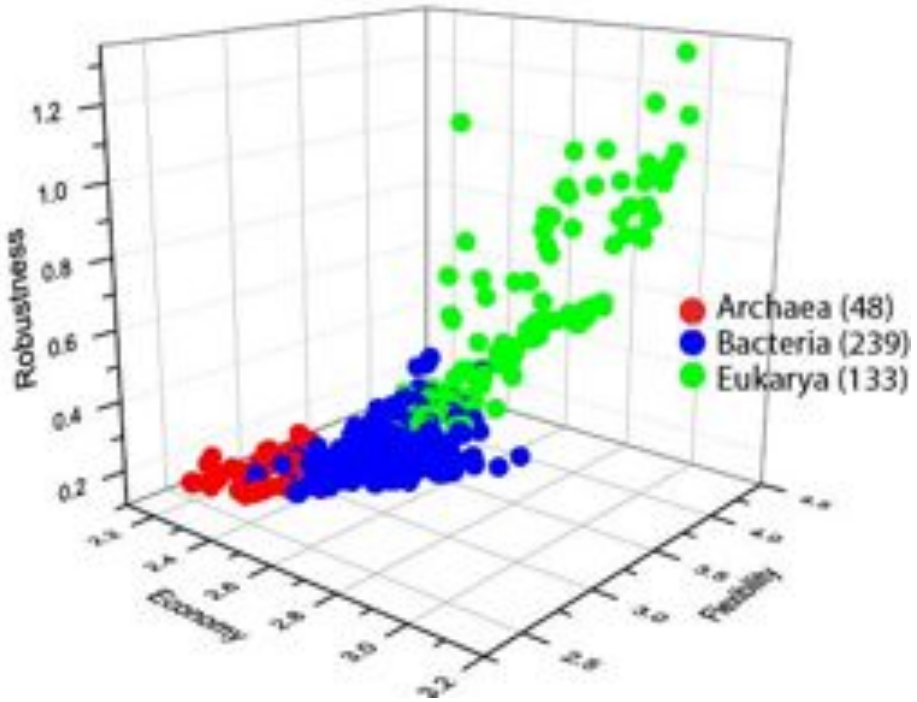


Modularity increases in evolution



Domain diversity (occurrence) and reuse (abundance) in proteomes carry strong and deep phylogenetic signal





Module diversity and reuse increase in evolution, even if domains are universal

Kim and Caetano-Anollés (2012) *BMC Evol Biol* 12: 13

Retrodictive exploration

Evolution of systems

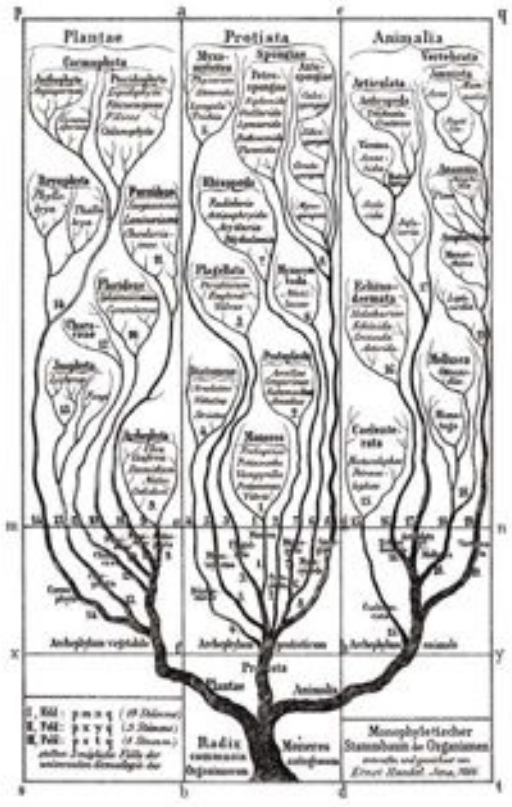
Systems

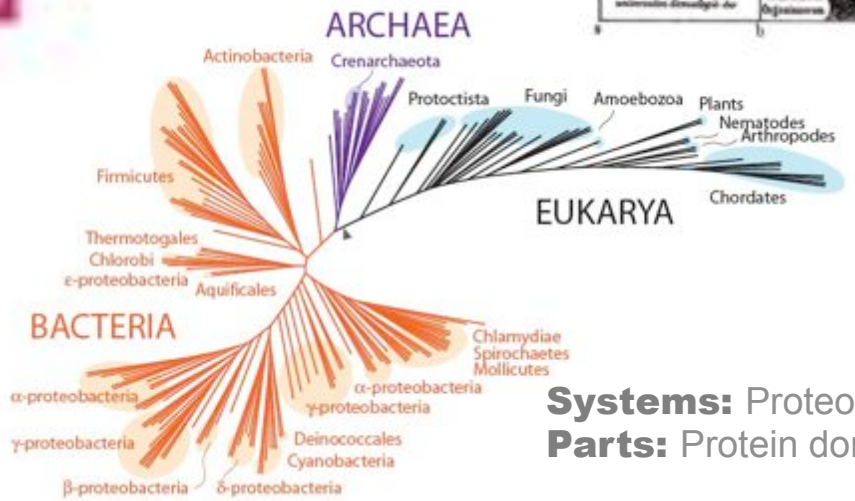
	s_1	s_2	s_3	s_4	s_5
p_1	4	3	3	5	7	9
p_2	5	A	6	8	9	B
p_3	2	2	8	C	D	6

Parts

Structure:

Model

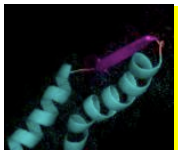




Ideographic framework:
historical and retrodictive

Systems: Proteomes
Parts: Protein domains

Retrodictive exploration



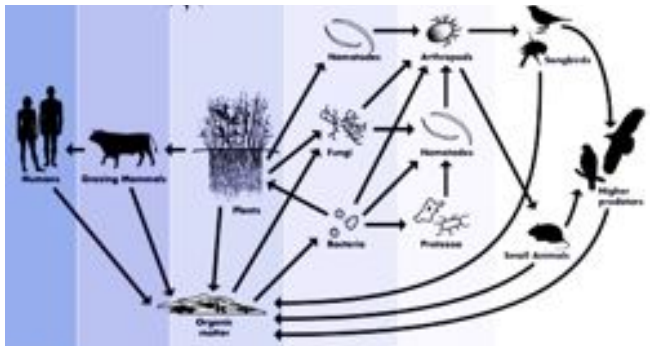
Evolution of parts

Parts

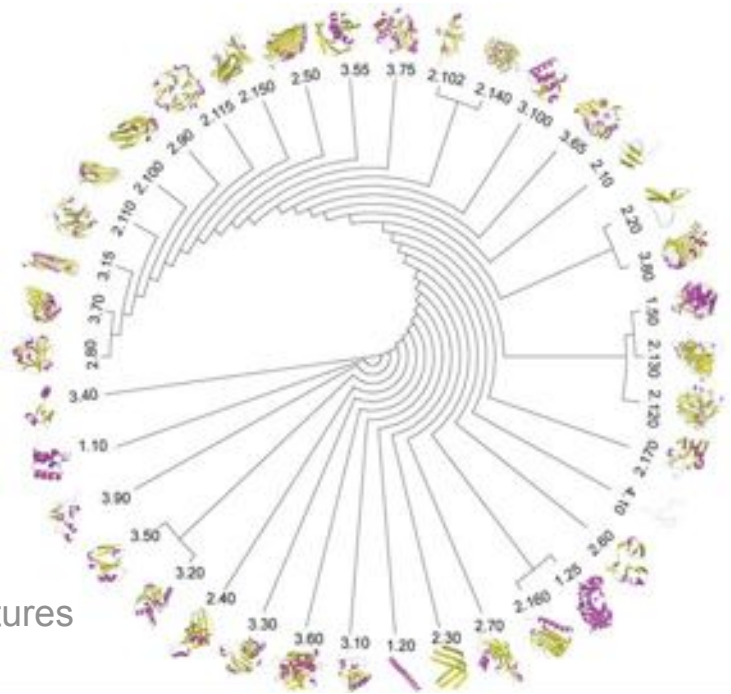
	p_1	p_2	p_3	p_4	p_5
s_1	B	B	5	9	2	2
s_2	4	A	6	8	3	B
s_3	4	9	8	5	D	6

Systems

Model



Structure: interaction of systems



NOVEL

Ideographic framework:
historical and retrodictive

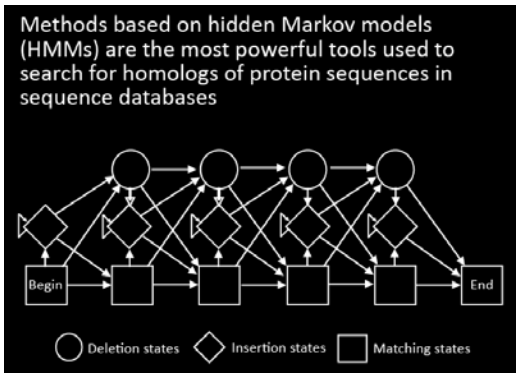
Systems: Proteomes
Parts: CATH architectures

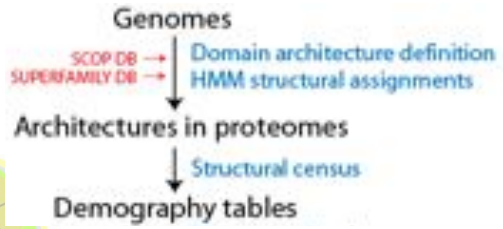
Structural phylogenomics

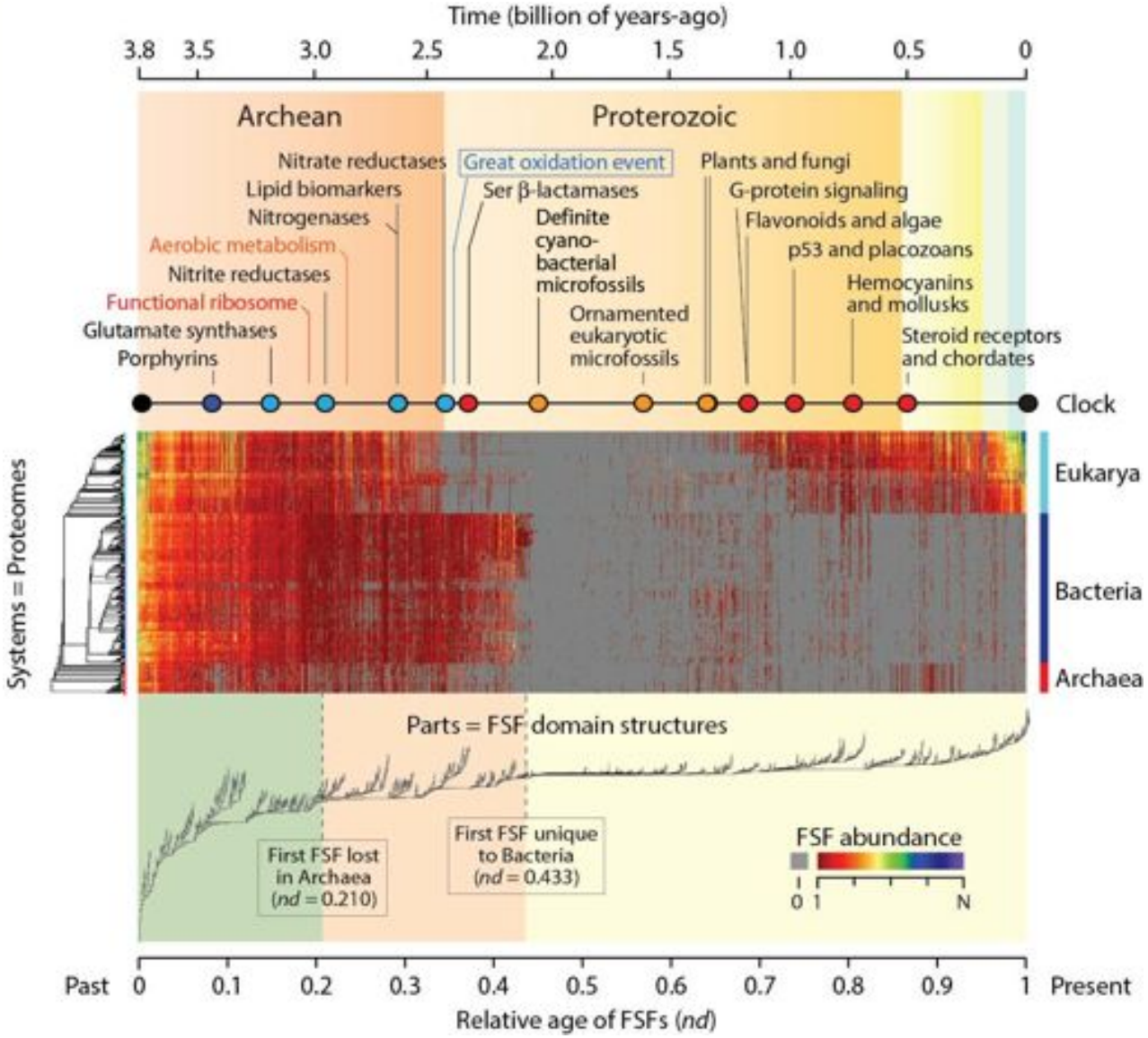
COMPUTING THE HISTORY OF THE PROTEIN WORLD:
Genomic demography and phylogeny reconstruction

Hidden Markov Models (HMMs)

Methods based on hidden Markov models (HMMs) are the most powerful tools used to search for homologs of protein sequences in sequence databases





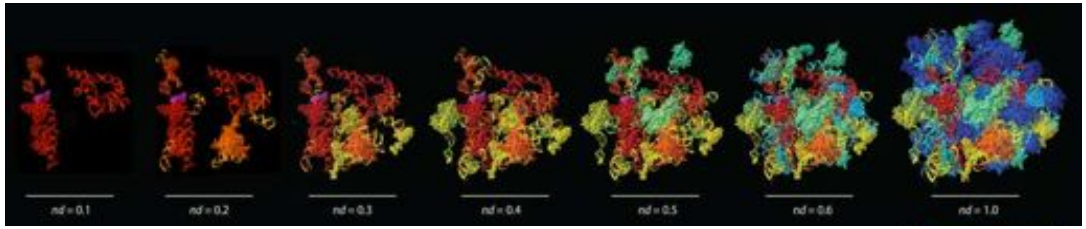


Structural phylogenomics

COMPUTING THE HISTORY OF THE PROTEIN WORLD:
Recent applications

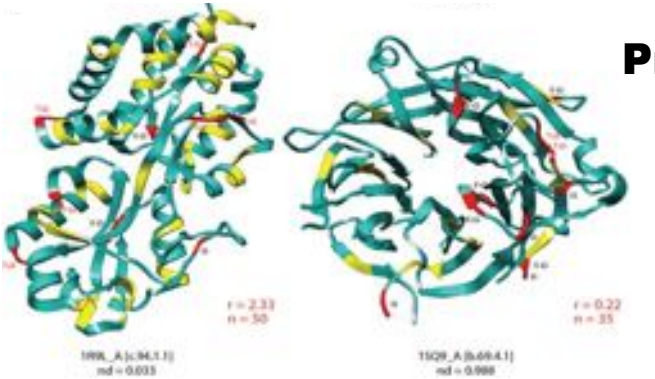
Coevolutionary history of the ribosome

Harish and Caetano-Anollés (2012) *PLoS ONE* 7: e32776
Caetano-Anollés and Sun (2014) *Front Genet* 5: 127.



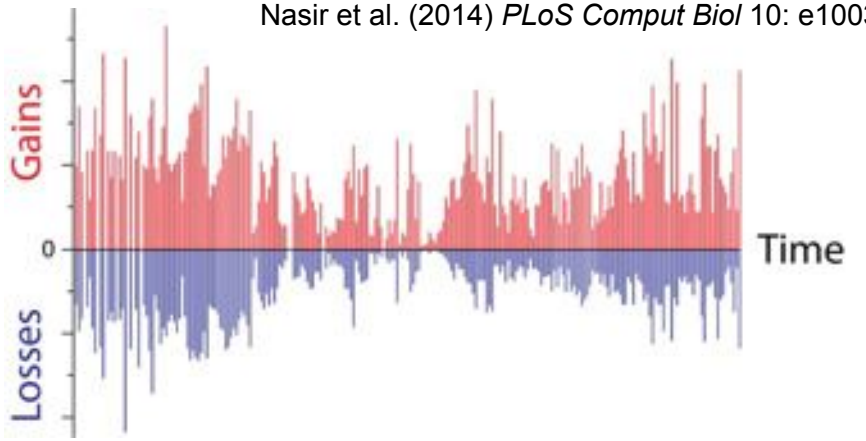
Origin of the genetic code in protein flexibility

Caetano-Anollés et al. (2013) *PLoS ONE* 8: e72225
Debes et al. (2013) *PLoS Comput Biol* 9: e1002861



Protein domain gain and loss in superkingdoms

Nasir et al. (2014) *PLoS Comput Biol* 10: e1003452



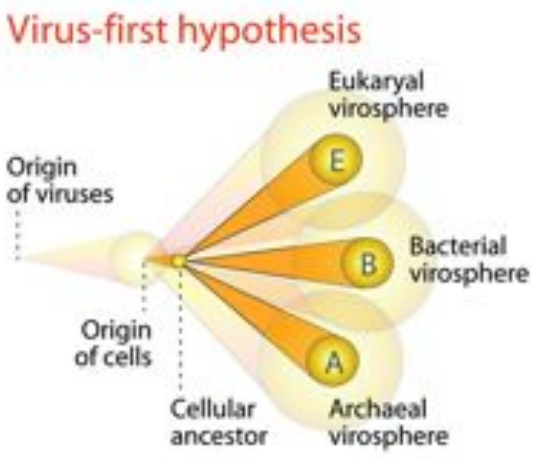
Origin of viruses

Three main lines of thought about three main events of emergence

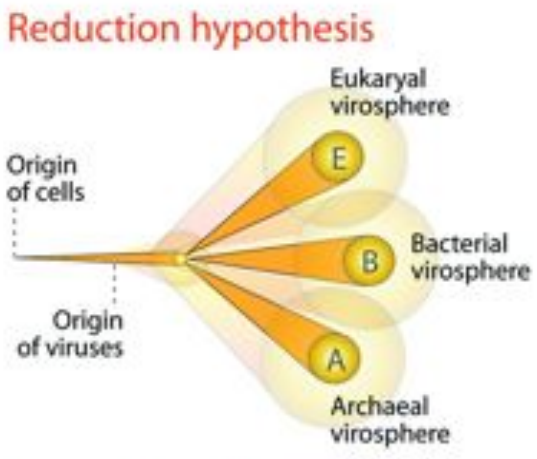
Nasir et al. (2012) *Mobile Genet Elements* 12:156

V : Origin of viruses ? **C :** Origin of cells

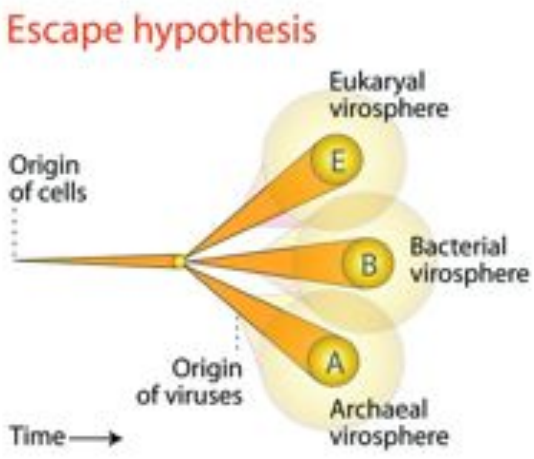
L : Origin of diversified cells*



V > C > L



C > V > L



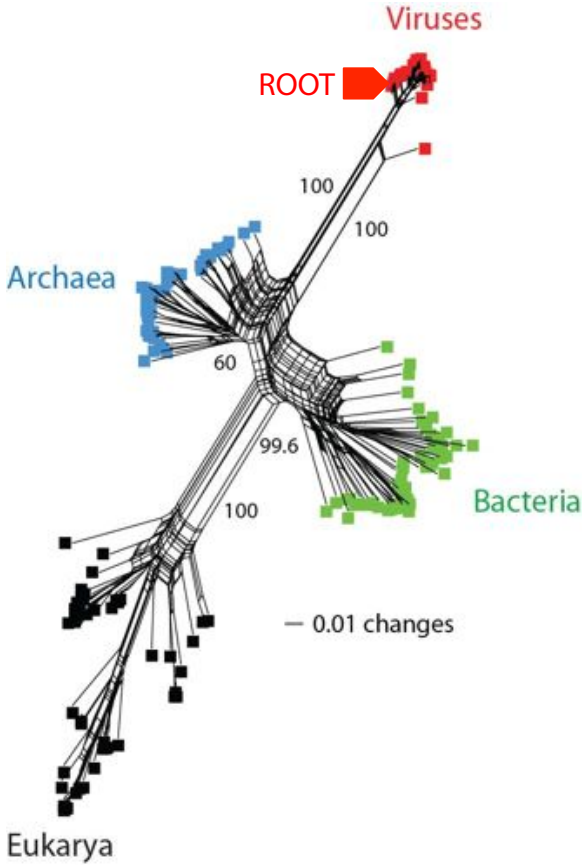
C > L > V

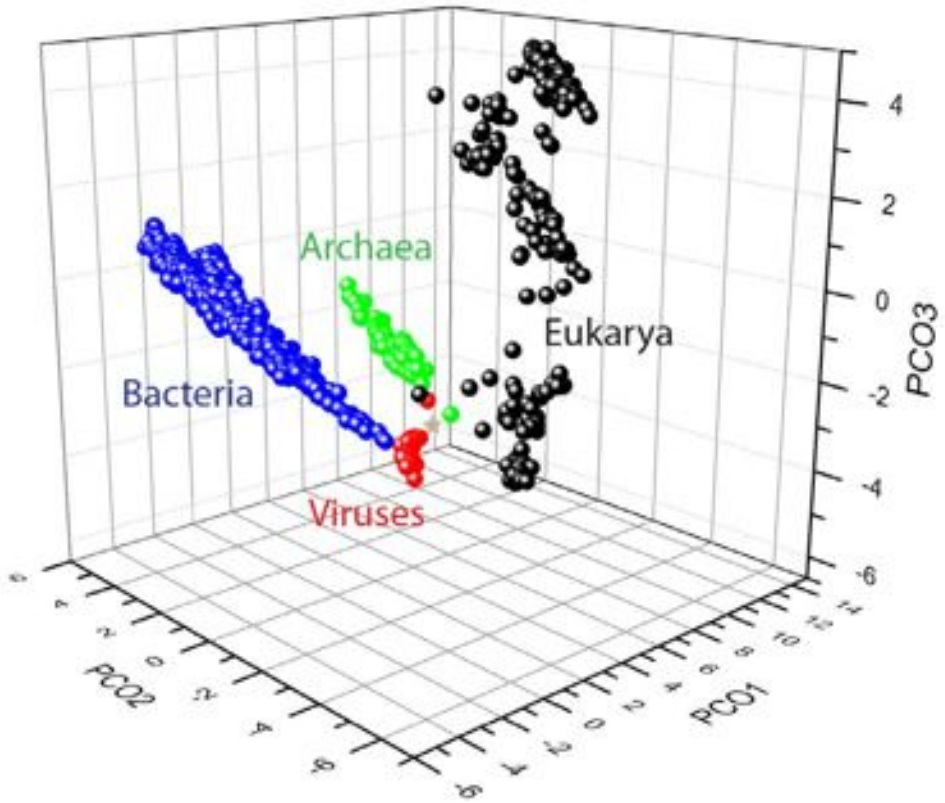
Last universal cellular ancestor [**LUCELLA**]

The root of the 'universal' tree of life

Trees of proteomes and evolutionary PCA analyses derived from *fold superfamilies* (FSFs) place the root of the tree in VIRUSES with large-to-medium size genomes

Nasir et al. (2012) *BMC Evol Biol* 12:156

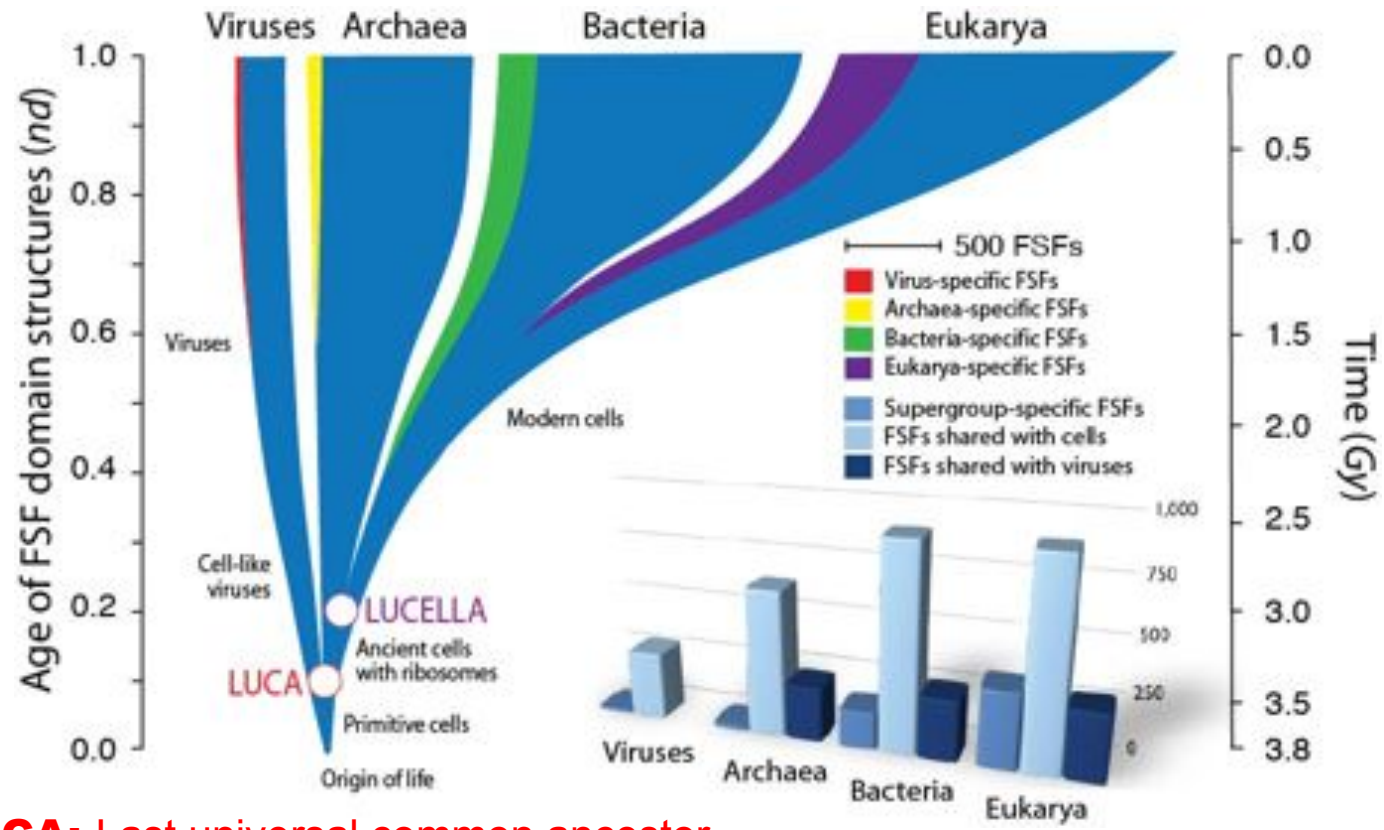




The root of the 'universal' tree of life

The ancient cellular origin of viruses supports the 'reductive' evolution hypothesis and redefines urancestors of life

Nasir et al. (2012) *Mobile Genet Elements* 12:156

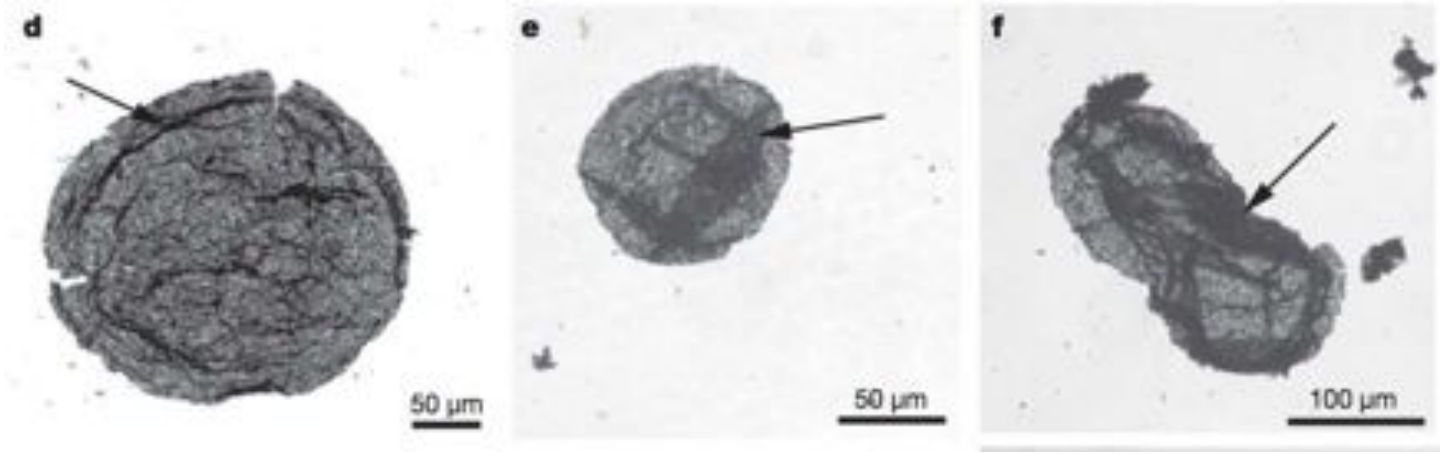


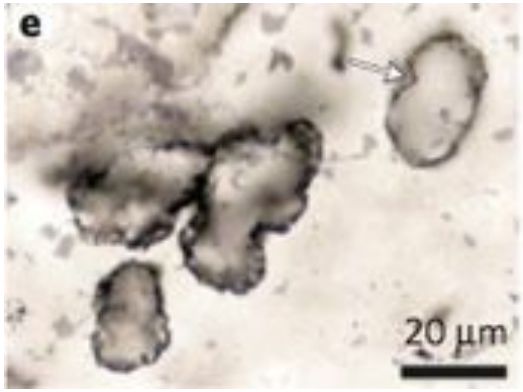
LUCA: Last universal common ancestor

LUCELLA: Last universal cellular ancestor

The root of the 'universal' tree of life

Phylogenomics support the biogenicity of 2.9-3.5 Gy-old microfossils of cellular microbes





Two organic-walled microfossil size ranges:
5-90 µm (Pilbara Craton, Western Australia)
50-300 µm (Moodies Group, South Africa)

Sugitani et al. (2007) *Precamb Res* 158: 228-226
Sugitani et al. (2009) *Astrobiology* 9:603-615
Javaux et al. (2010) *Nature* 463:934-938
Wacey et al. (2011) *Nature Geosci* 4:698-702

Expanded proteome dataset

[March 2013]

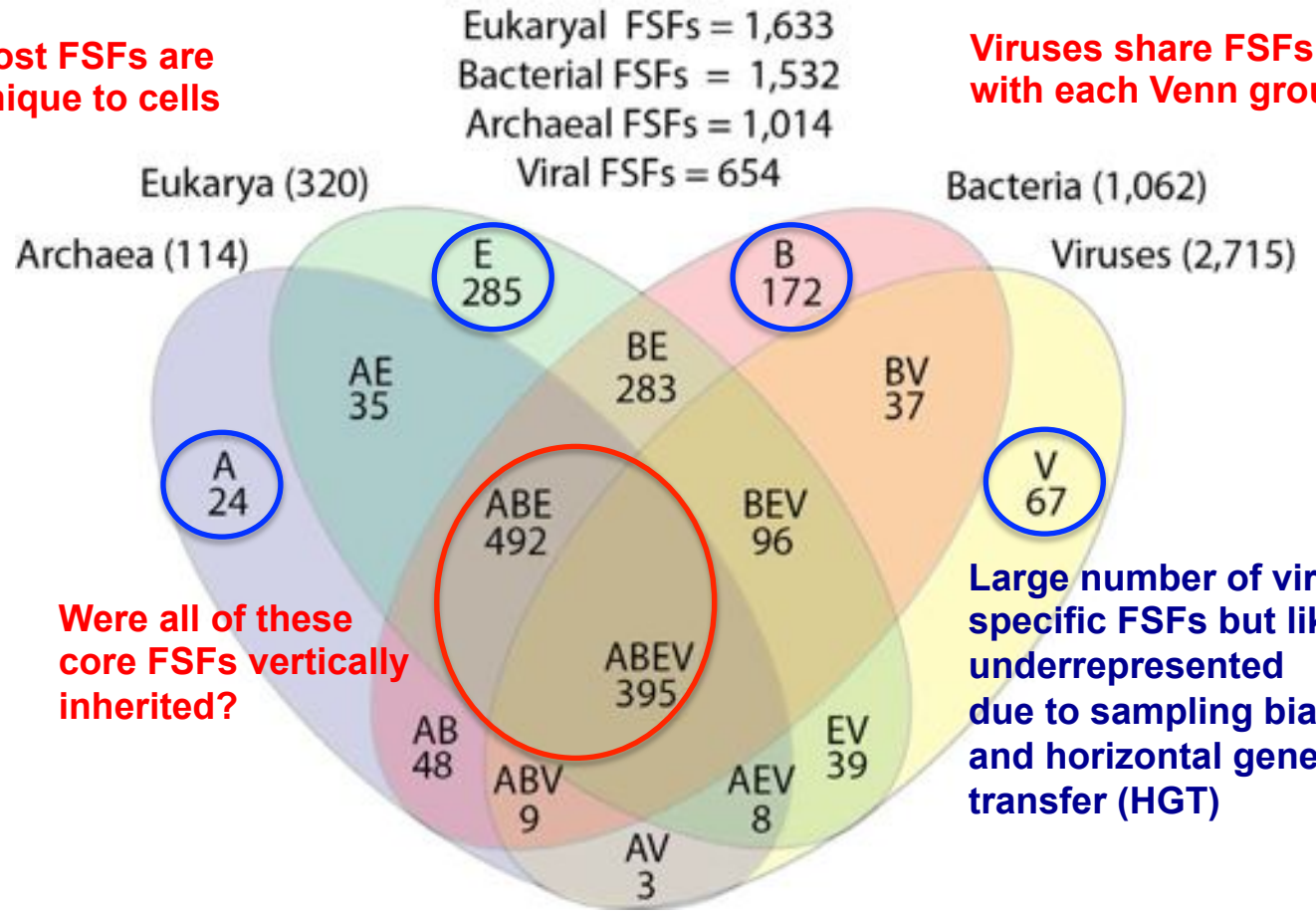
- **2,715 proteomes from all 6 viral groups** [NCBI RefSeq Viral Resource]
  - 1,125 dsDNA
  - 453 ssRNA
  - 122 dsRNA
  - 806 ssRNA(+)
  - 95 ssRNA(-)
  - 114 Retrotranscribing viruses
- **1,496 proteomes from cellular organisms** [SUPERFAMILY database]
  - 114 Archaea, 1,062 Bacteria, and 320 Eukarya

Hierarchical level: FSF

- Includes protein domains with low sequence identity (< 15%)
- Common conserved 3D cores and biochemical properties
- Homologous as defined by the SCOP evolutionary theory
- More conserved and better suited to study remote relationships

Sharing patterns

Most FSFs are unique to cells



Viruses share FSFs with each Venn group

Were all of these core FSFs vertically inherited?

Large number of virus-specific FSFs but likely underrepresented due to sampling biases and horizontal gene transfer (HGT)

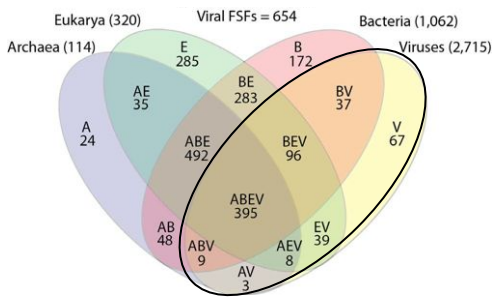
4,211 proteomes from cells and viruses
1,993 FSF domains ($E < 0.0001$)

Spread of viral FSFs in cellular proteomes

***f* index ***

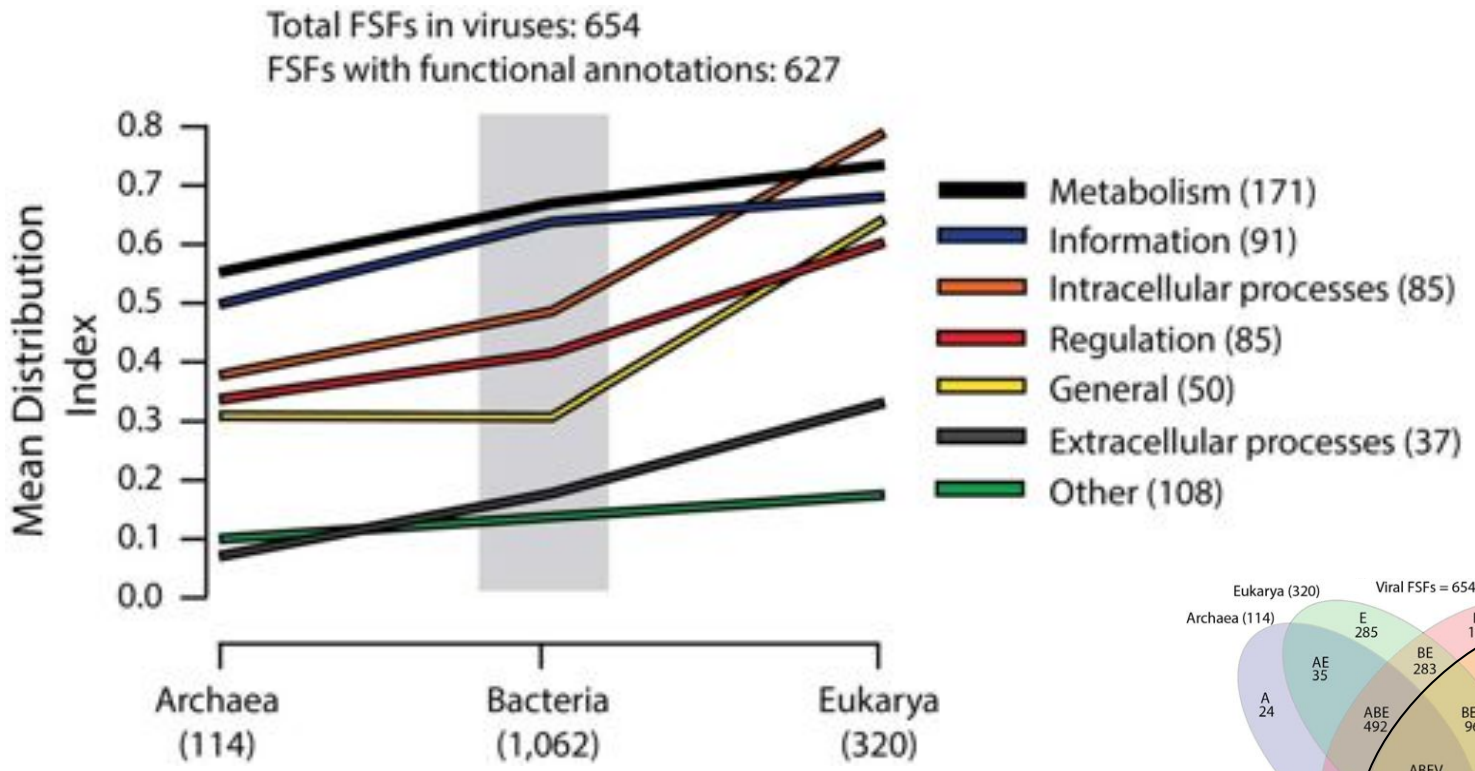
FSF Description	Molecular Function	Detailed Function	Distribution in Viruses	Archaea (*f*)	Bacteria (*f*)	Eukarya (*f*)
Inhibitor of apoptosis (IAP) repeat	Intracellular processes	Cell cycle, Apoptosis	dsDNA, ssRNA(+)	0.00	0.00	0.73
ATP-dependent DNA ligase DNA-binding domain	Regulation	DNA-binding	dsDNA	0.98	0.12	0.99
Ribonuclease H-like	Metabolism	Nucleotide m/tr	dsDNA, ssDNA, ssRNA(-), retrotranscribing	1.00	1.00	1.00
DNA/RNA polymerases	Information	DNA replication/repair	dsDNA, dsRNA, ssRNA(+), ssRNA(-), retrotranscribing	1.00	0.98	1.00
Translation initiation factor 2 beta, aIF2beta, N-t	Information	Translation	dsDNA	1.00	0.00	1.00
RING/U-box	Information	DNA replication/repair	dsDNA, ssRNA(-)	0.04	0.11	1.00
dUTPase-like	Metabolism	Nucleotide m/tr	dsDNA, retrotranscribing	0.98	0.89	0.91
Nucleic acid-binding proteins	Information	DNA replication/repair	dsDNA, ssDNA	1.00	1.00	1.00
R1 subunit of ribonucleotide reductase, N-termir	Metabolism	Nucleotide m/tr	dsDNA	0.61	0.78	0.96
DNA ligase/mRNA capping enzyme, catalytic dor	Information	DNA replication/repair	dsDNA	1.00	0.99	1.00
Zinc-binding domain of translation initiation fact	General	Ion binding	dsDNA	1.00	0.00	0.99
P-loop containing nucleoside triphosphate hydro	General	Small molecule binding	dsDNA, ssDNA, dsRNA, ssRNA(+)	1.00	1.00	1.00
Ferritin-like	Intracellular processes	Ion m/tr	dsDNA	0.99	0.99	1.00
PFL-like glycyl radical enzymes	Metabolism	Other enzymes	dsDNA	1.00	0.98	0.98
Protein kinase-like (PK-like)	Regulation	Kinases/phosphatases	dsDNA, dsRNA, retrotranscribing	1.00	0.96	1.00
Cytidine deaminase-like	Metabolism	Other enzymes	dsDNA	0.96	0.99	1.00
Ribosomal protein S5 domain 2-like	Information	Translation	dsDNA	1.00	1.00	1.00
Cryptochrome/photolyase FAD-binding domain	General	General	dsDNA	0.32	0.53	0.83
FAD-linked reductases, C-terminal domain	Metabolism	Redox	dsDNA	0.61	0.85	1.00
ADP-ribosylation	Metabolism	Secondary metabolism	dsDNA, ssDNA	0.47	0.33	0.98
RPB5-like RNA polymerase subunit	Information	Transcription	dsDNA	0.98	0.00	0.93
TNF-like	Extracellular processes	Immune response	dsDNA	0.00	0.05	0.35
GroES-like	Intracellular processes	Protein modification	dsDNA	0.80	0.98	1.00
Translation proteins	Information	Translation	dsDNA	1.00	1.00	1.00

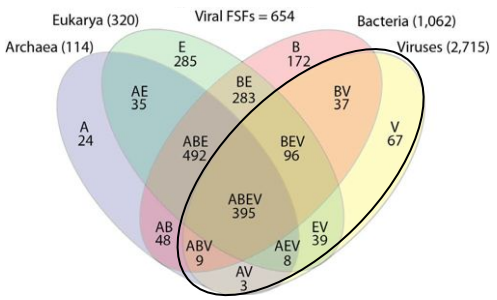
*Percentage of proteomes encoding an FSF divided by the total number of proteomes (in a relative 0-1 scale)



Biases in functional preferences of viral FSFs

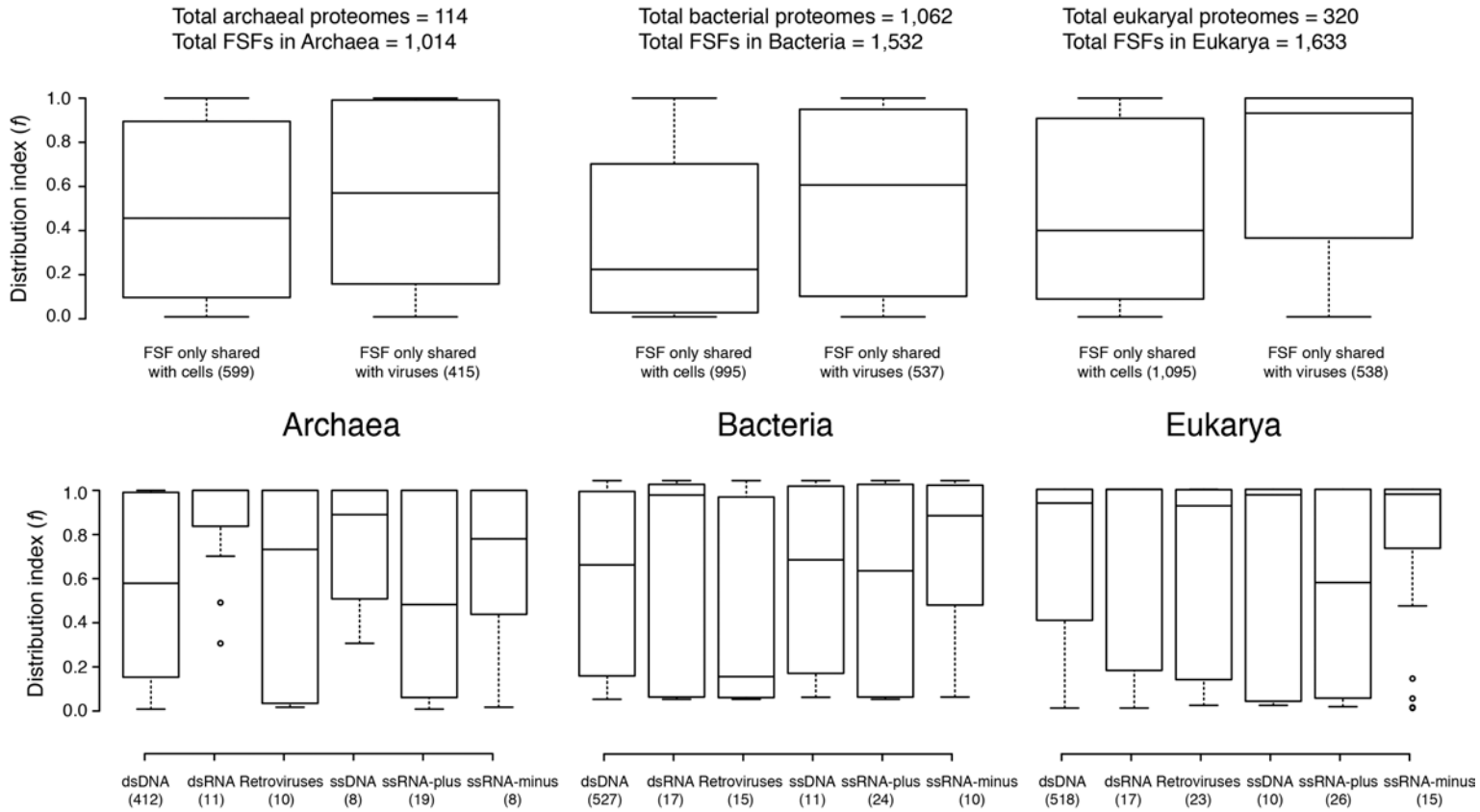
Extracellular processes underrepresented in Archaea
Intracellular processes and general metabolism overrepresented in Eukarya





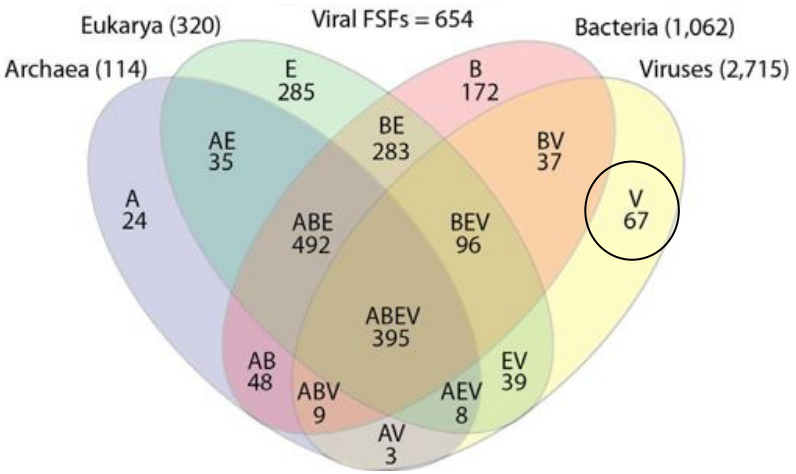
Viruses enhance molecular biodiversity

FSFs shared with viruses are widely distributed in the world of proteomes, especially with ssDNA, dsRNA and ssRNA (-) viruses



67 virus-specific FSFs

- Three times more abundant than Archaea-specific FSFs
- Under-represented (HGT and sampling biases)
- Harbor pathogenic and immunological roles
- Include major viral capsid proteins
- Interesting drug targets
- Present in all six viral subgroups
- Could not have originated in cells!



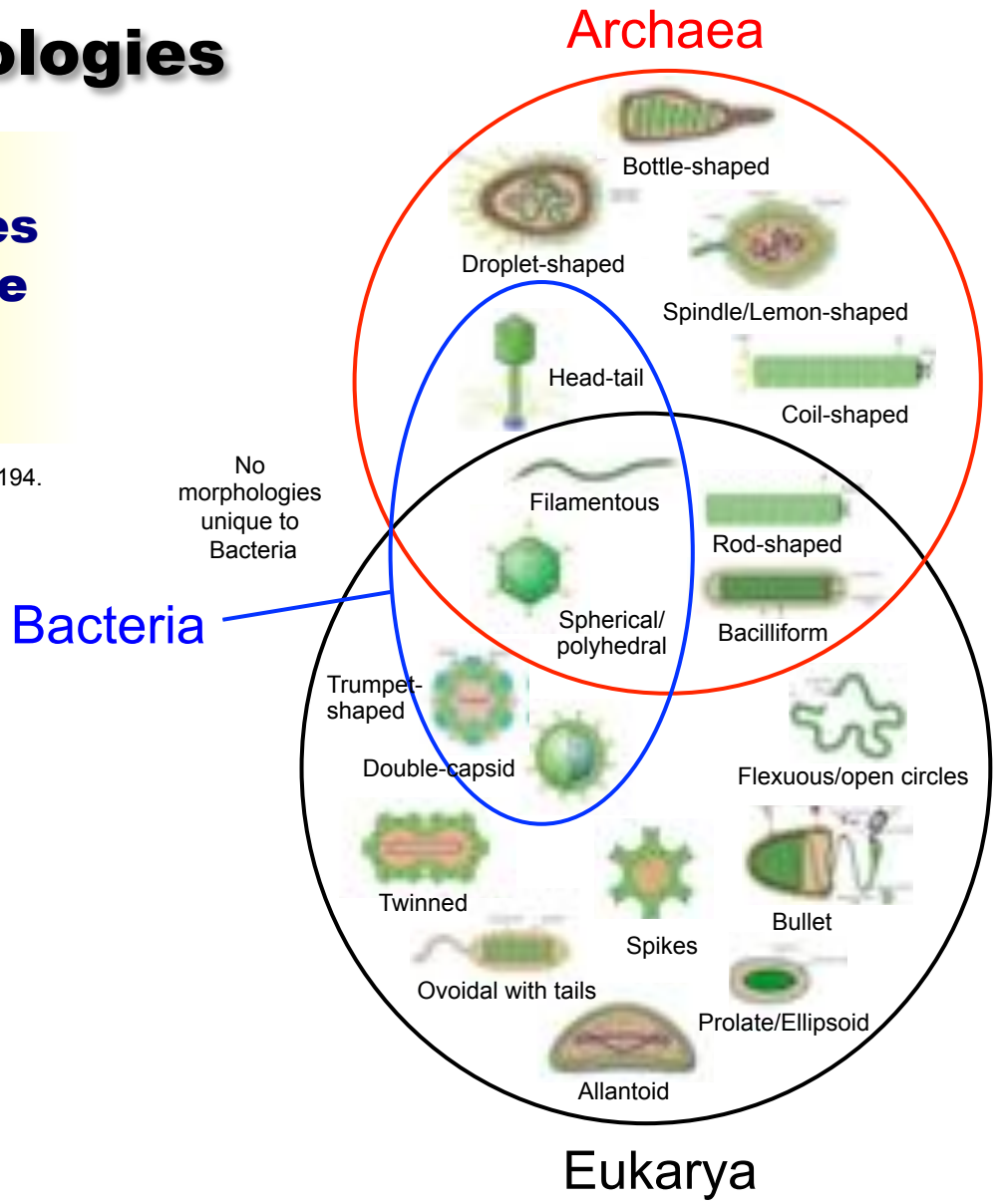
a.115.1.1
Orbivirus capsid

a.115.1.3
Phitoreovirus capsid

Virion morphologies

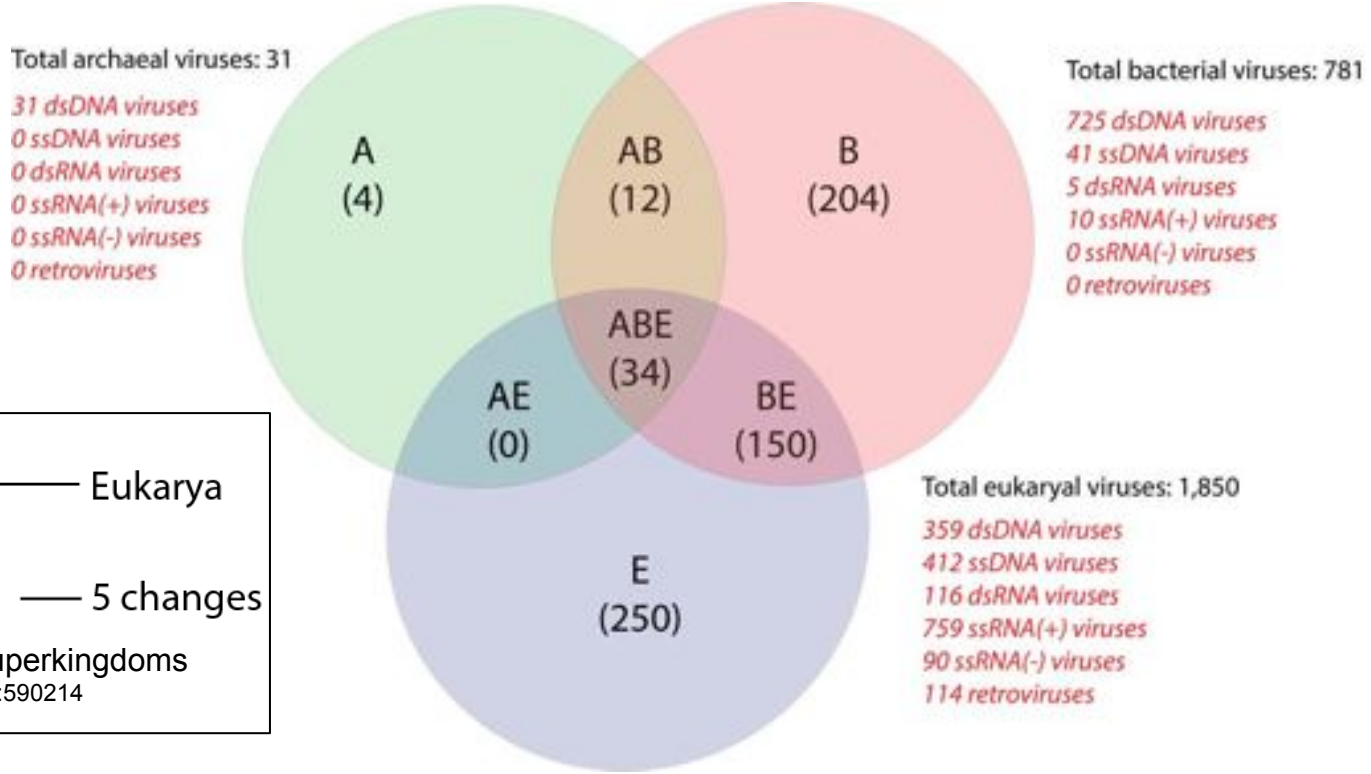
A common set of simple forms gives rise to a multitude of complex structures

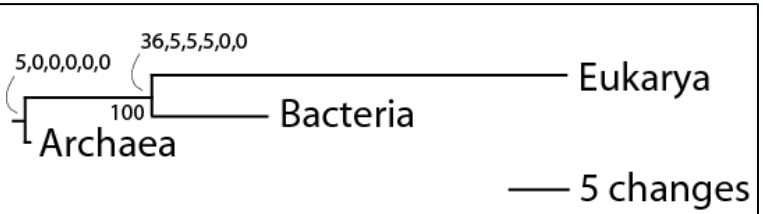
Nasir et al. (2014) *Frontiers Microbiology* 5: 194.



Virus-host relationships and FSF sharing

While viruses do not cross superkingdom barriers and have specific host preferences, they harbor a common structural core





Distribution of viral replicons in superkingdoms
Caetano-Anollés et al. (2014) *Archaea* 2014:590214

- **Loss of viral lineages in Archaea and Bacteria?**
- **Late appearance of RNA viruses?**

34 FSFs common to viruses associated with the three superkingdoms

	FSF Description	Molecular Function	Detailed Function	Distribution	A (J)	B (J')	E (J')
2	Ribonuclease H-like	Metabolism	Nucleotide m/tr	dsDNA, ssDNA, ssRNA(-), retrotranscribing	1.00	1.00	1.00
3	DNA/RNA polymerases	Information	DNA replication/repair	dsDNA, dsRNA, ssRNA(+), ssRNA(-), retrotranscribing	1.00	0.98	1.00
4	dUTPase-like	Metabolism	Nucleotide m/tr	dsDNA, retrotranscribing	0.98	0.89	0.91
5	Nucleic acid-binding proteins	Information	DNA replication/repair	dsDNA, ssDNA	1.00	1.00	1.00
6	P-loop containing nucleoside triphosphate h	General	Small molecule binding	dsDNA, ssDNA, dsRNA, ssRNA(+)	1.00	1.00	1.00
7	(Phosphotyrosine protein) phosphatases II	Regulation	Kinases/phosphatases	dsDNA	0.51	0.51	1.00
8	PIN domain-like	Other	Unknown function	dsDNA	1.00	0.99	1.00
9	N-terminal nucleophile aminohydrolases (Ntn	Metabolism	Other enzymes	dsDNA	1.0		00
10	Chaperone J-domain	Intracellular processes	Protein modification	dsDNA	0.4		00
11	Putative DNA-binding domain	Regulation	DNA-binding	dsDNA	1.0		99
12	Winged helix DNA-binding domain	Regulation	DNA-binding	dsDNA	1.0		00
13	Uracil-DNA glycosylase-like	Information	DNA replication/repair	dsDNA	1.0		99
14	Nucleotide-diphospho-sugar transferases	Metabolism	Transferases	dsDNA, dsRNA	1.0		00
15	vWA-like	Extracellular processes	Cell adhesion	dsDNA	0.9		00
16	S-adenosyl-L-methionine-dependent methylt	Metabolism	Transferases	dsDNA, dsRNA, ssRNA(+)	1.0		00
17	Thioredoxin-like	Metabolism	Redox	dsDNA	0.9		00
18	DNA clamp	Information	DNA replication/repair	dsDNA	1.0		00
19	Metallo-dependent phosphatases	Intracellular processes	Proteases	dsDNA	1.0		00
20	DNA breaking-rejoining enzymes	Information	DNA replication/repair	dsDNA	0.8		99
21	ATPase domain of HSP90 chaperone/DNA to	Intracellular processes	Protein modification	dsDNA	1.0		00
22	Rad51 N-terminal domain-like	Information	DNA replication/repair	dsDNA	1.0		81
23	NAD(P)-binding Rossmann-fold domains	General	Small molecule binding	dsDNA	1.0		00
24	Radical SAM enzymes			dsDNA	1.0		00
25	Restriction endonuclease-like			dsDNA	1.0		00
26	TPR-like			dsDNA	0.8		00
27	UDP-Glycosyltransferase/glycogen phosphor			dsDNA, dsRNA	1.0		00
28	Thymidylate synthase-complementing protei			dsDNA	0.5		91
29	Concanavalin A-like lectins/glucanases			dsDNA, dsRNA	0.4		00
30	ARM repeat			dsDNA, ssRNA(+)	0.8		00
31	beta-beta-alpha zinc fingers	Regulation	DNA-binding	dsDNA	0.3		00
32	Bacterial hemolysins	Metabolism	Other enzymes	dsDNA, ssDNA, dsRNA	0.3		43
33	Trypsin-like serine proteases	Intracellular processes	Proteases	dsDNA, ssRNA(+)	0.5		97
34	Fibronectin type III	Extracellular processes	Cell adhesion	dsDNA	0.25	0.51	0.93
35	Pectin lyase-like	Metabolism	Polysaccharide m/tr	dsDNA, ssRNA(+)	0.68	0.72	0.89

All viral groups

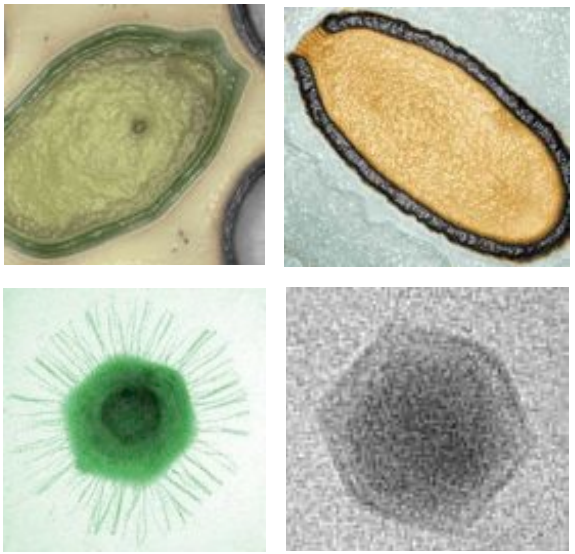
dsDNA viruses: Common denominator!

Many FSFs are superfolds with central functions

Widespread in cells!

A possible early origin of dsDNA viruses

dsDNA viruses can be as large as cells

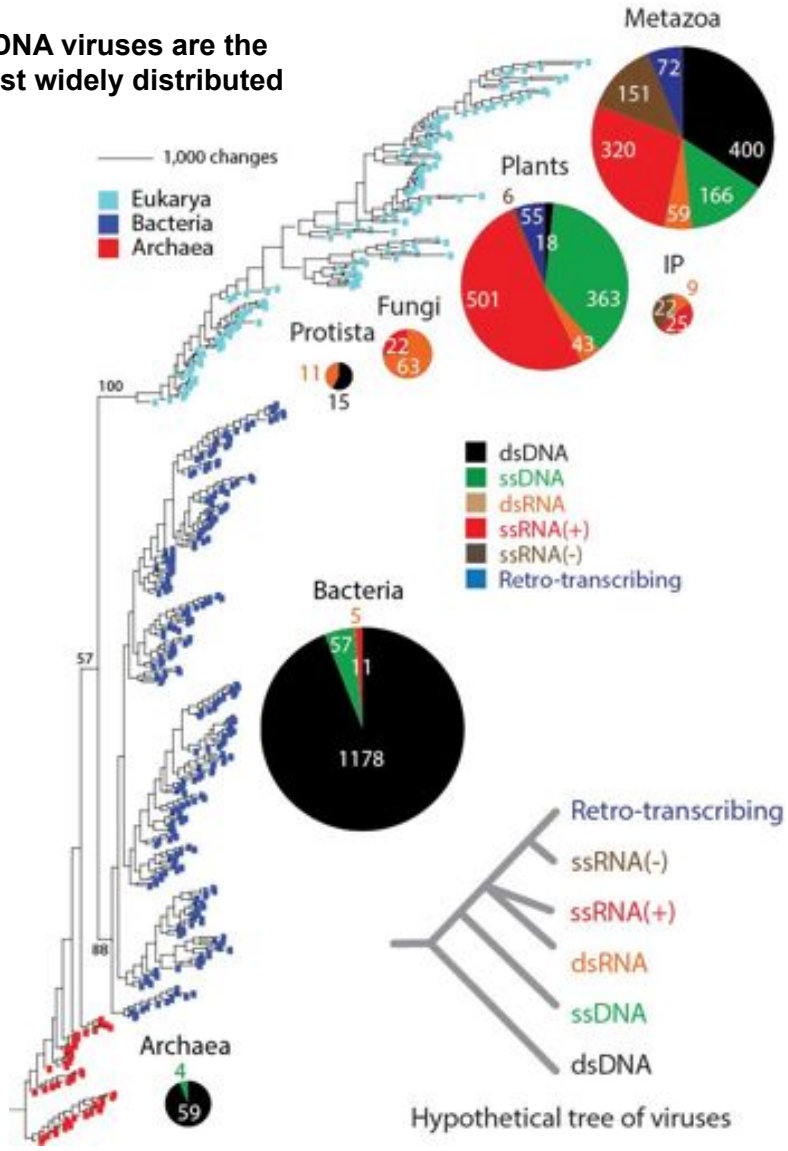


dsDNA viruses harbor the largest number and the most shared FSFs

Subgroup	Total	Unique	ssDNA	retroviruses	dsRNA	ssRNA
dsDNA	587	530	7	11	9	22
ssDNA	14	4	-	0	0	0
retroviruses	27	12	-	-	0	1
dsRNA	28	11	-	-	-	1
ssRNA	68	38	-	-	-	-

Nasir et al. (2014) *Frontiers Microbiology* 5:194

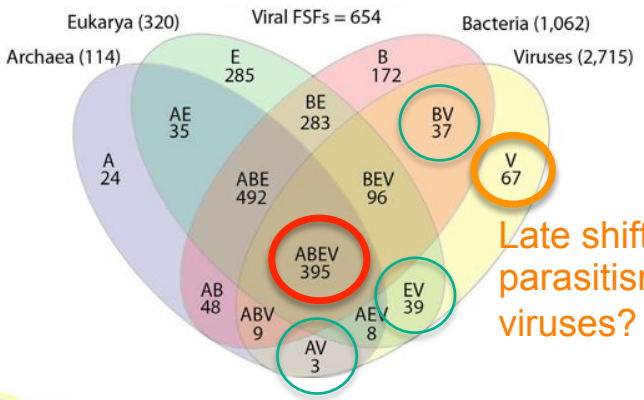
dsDNA viruses are the most widely distributed



Rooted tree of proteomes built from an FSF domain census
Kim and Caetano-Anollés (2011) *BMC Evol Biol* 11: 140.

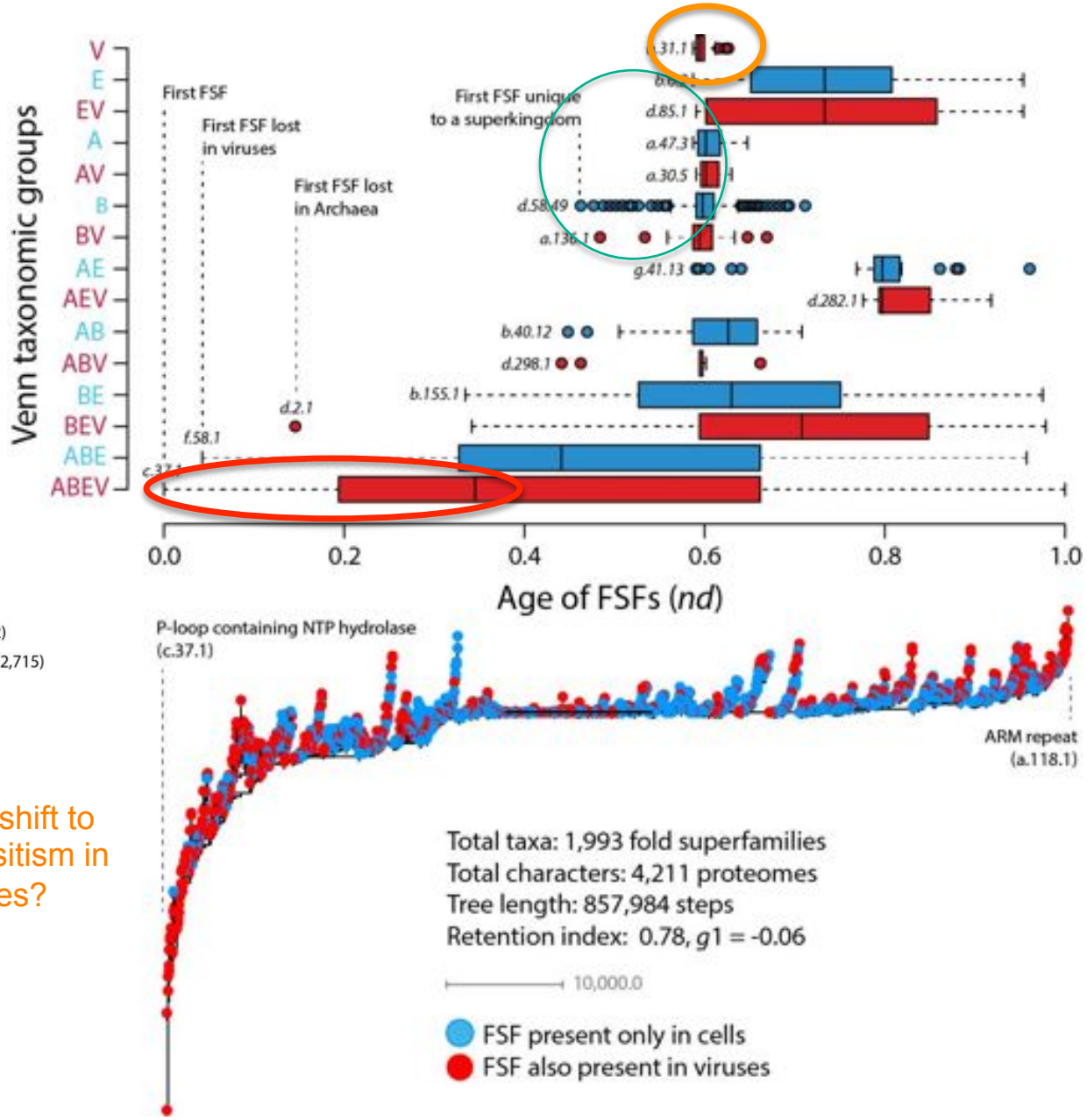
History of the viral and cellular FSF repertoire

Superkingdom-specific FSFs shared with viruses (AV, BV, and EV) appear after those that are not shared

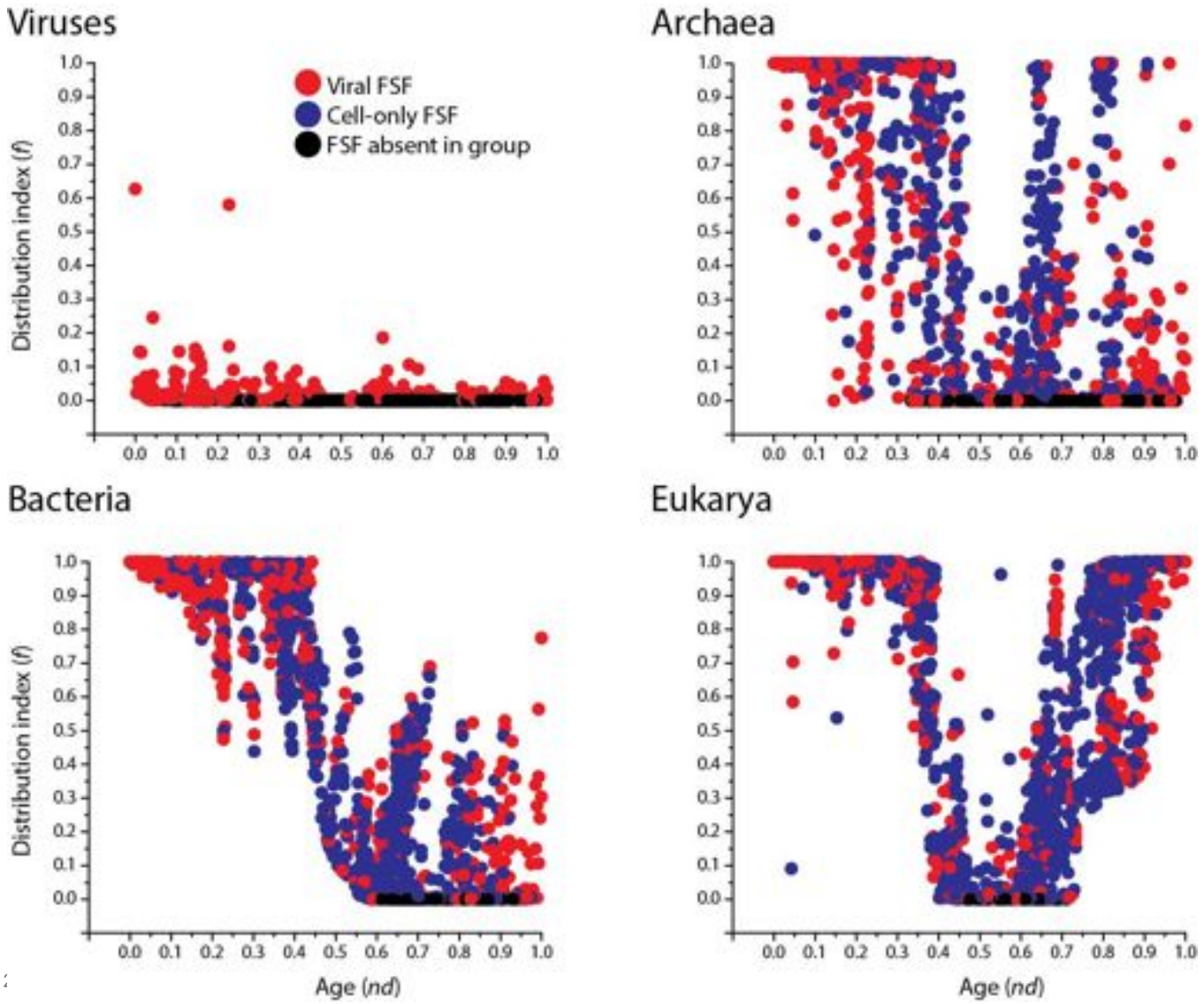


Late shift to parasitism in viruses?

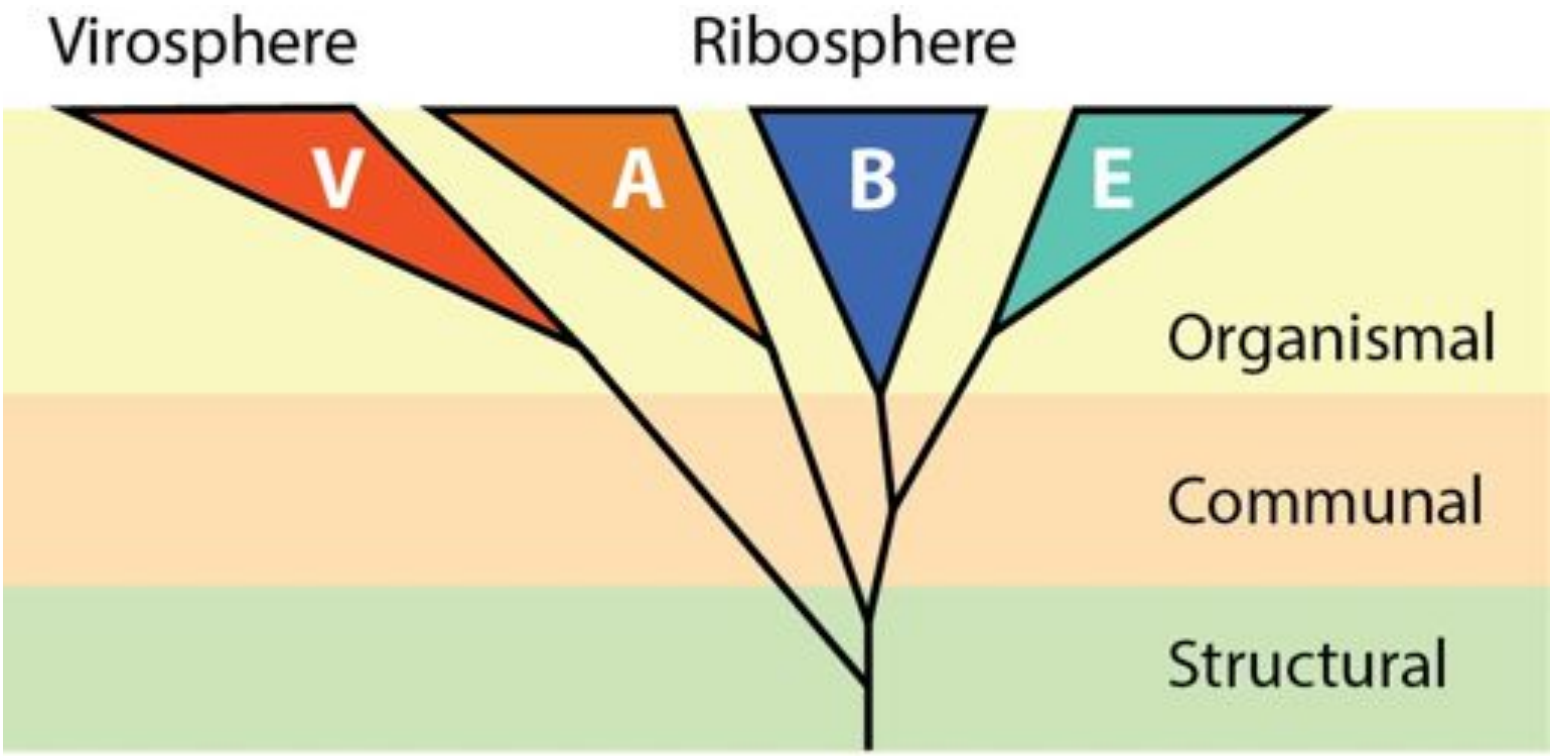
The majority of FSFs shared between cells and viruses appear very early!



Extreme reductive tendencies in viruses



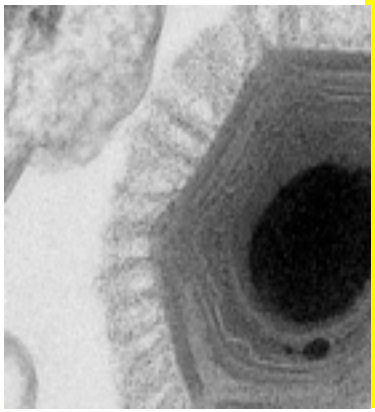
Diversification of life



Some conclusions

Early cellular origins of viruses

- Viruses originated from proto-cells that coexisted with cellular ancestors ~3.4 Gy ago
- Early origin of virocell lineages by reductive evolution
- Ancient virocells had DNA replicons
- DNA viruses are ancient and monophyletic



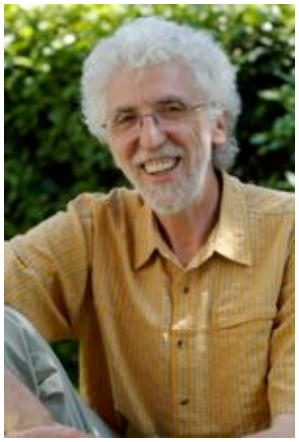
Late rise of viral novelties

- Late appearance of capsids and parasitic life cycles
- Viruses enhance cellular molecular diversity
- Early origins of archaeal DNA viruses and spread of novelties to RNA viruses in other superkingdoms

Acknowledgements

Arshan Nasir

Collaborators:



Kyung Mo Kim (Daejeon)
Patrick Forterre (Paris)

NSF

COMSATS

KRIBB
Korea Research Institute of Bioscience & Biotechnology

UNIVERSITÉ PARIS SUD

Thank you!

Questions