K-mer approaches provide valuable insight into mobilome evolution in the domain Archaea

Ariane Bize, Violette Da Cunha, Cédric Midoux, Sophie Schbath, Patrick Forterre

K-mer approaches provide valuable insight into mobilome evolution in the domain Archaea

Ariane Bize, Violette Da Cunha, Cédric Midoux, Sophie Schbath, Patrick Forterre

12th July 2018

K-mer signatures

k-mers = all possible subsequences of length k from a DNA sequence

4 possible bases: A, T, G, C → 4^k different k-mers

e.g. k = 4 → 4-mers or tetramers (e.g: ATAA, CGAG, GTTC, …)

→ 4^4 = 64 different tetramers

k-mer profile

ATTAGGCCGCAAGGGCCTTCATAGTTTTAGCGATTTGGGA

64 words

AGGC	1
ATTA	1
GGCC	2
TAGG	1
TTAG	2
…	

Counts

Normalization

AGGC	0.270
ATTA	0.270
GGCC	0.054
TAGG	0.270
TTAG	0.054
…	

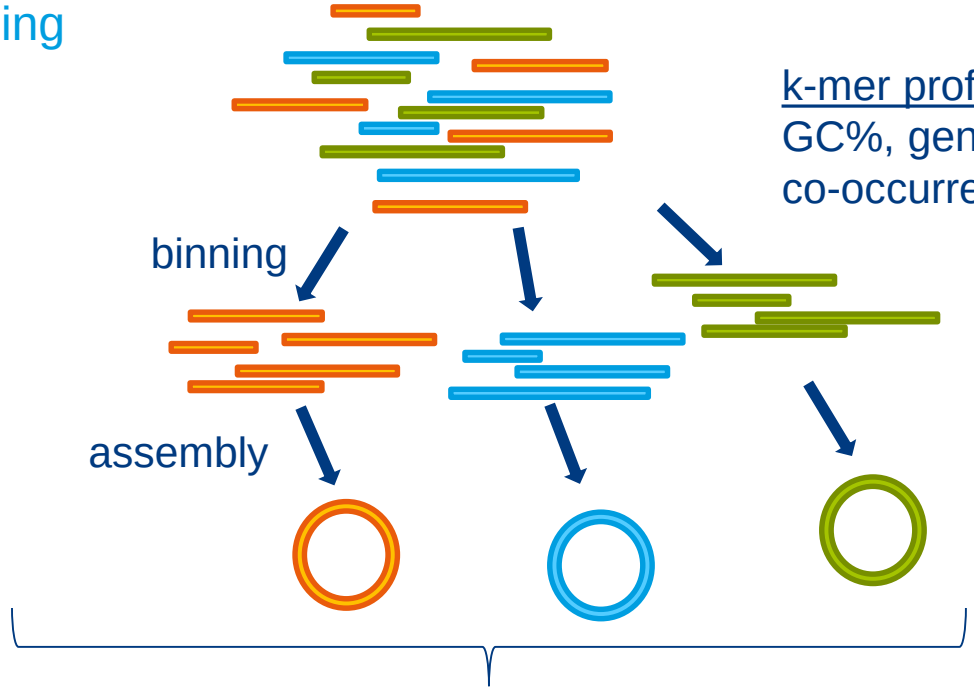
Frequencies

Tetramer profile

- Annotation independent
- Faster computation

K-mer approaches: metagenomics

Contig binning



k-mer profiles, coverage, GC%, gene content, co-occurrence…

"metagenome assembled genomes" = MAGs
or "population genomes"

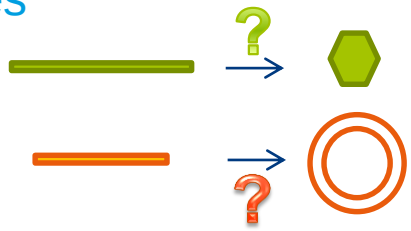
Alneberg et al, **Binning metagenomic contigs by coverage and composition**. *Nature Methods* 2014
Kang et al, **MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities**. *PeerJ* 2015

K-mer approaches: mobilome

Detection of viral or plasmid sequences

VirFinder (Ren et al, Microbiome, 2017)

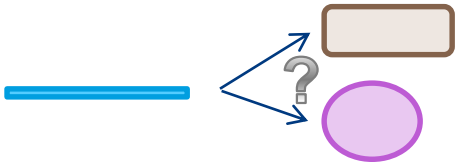
PlasFlow (Krawczyk et al, NAR, 2018)



Host prediction of viruses or plasmids

WiSH (Galiez et al, Bioinformatics, 2017)

PlasFlow (Krawczyk et al, NAR, 2018) [phylum level]



Evolutionnary biology? → Also being explored

-Detection and characterization of horizontal transfers in prokaryotes using genomic signature, Dufraigne et al, 2005, NAR

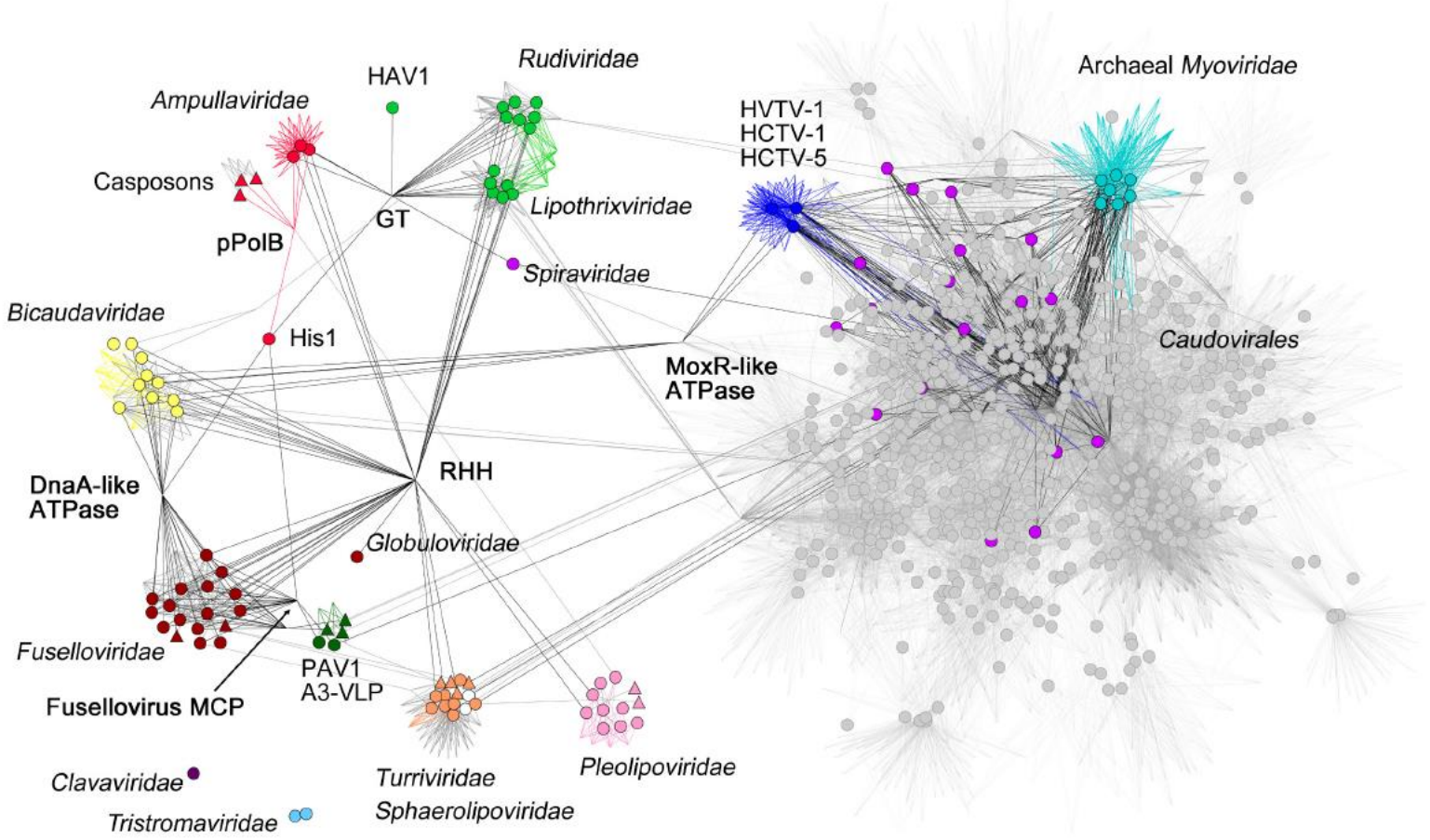
-K-mer natural vector and its application to the phylogenetic analysis of genetic sequences, Wen J, et al, 2014, Gene

-Phenetic comparison of prokaryotic genomes using k-mer, Désrape et al, 2017, MBE

Present study: plasmids, viruses and hosts from the domain Archaea

- For the hosts, topology based on k-mer signatures consistent with the phylogeny of archaea?
- Regarding archaeal viruses and plasmids, topology similar to that of the hosts?
- Specific signature of archaeal extrachromosomal elements?
- Factors underlying these distributions?

Archaeal viruses



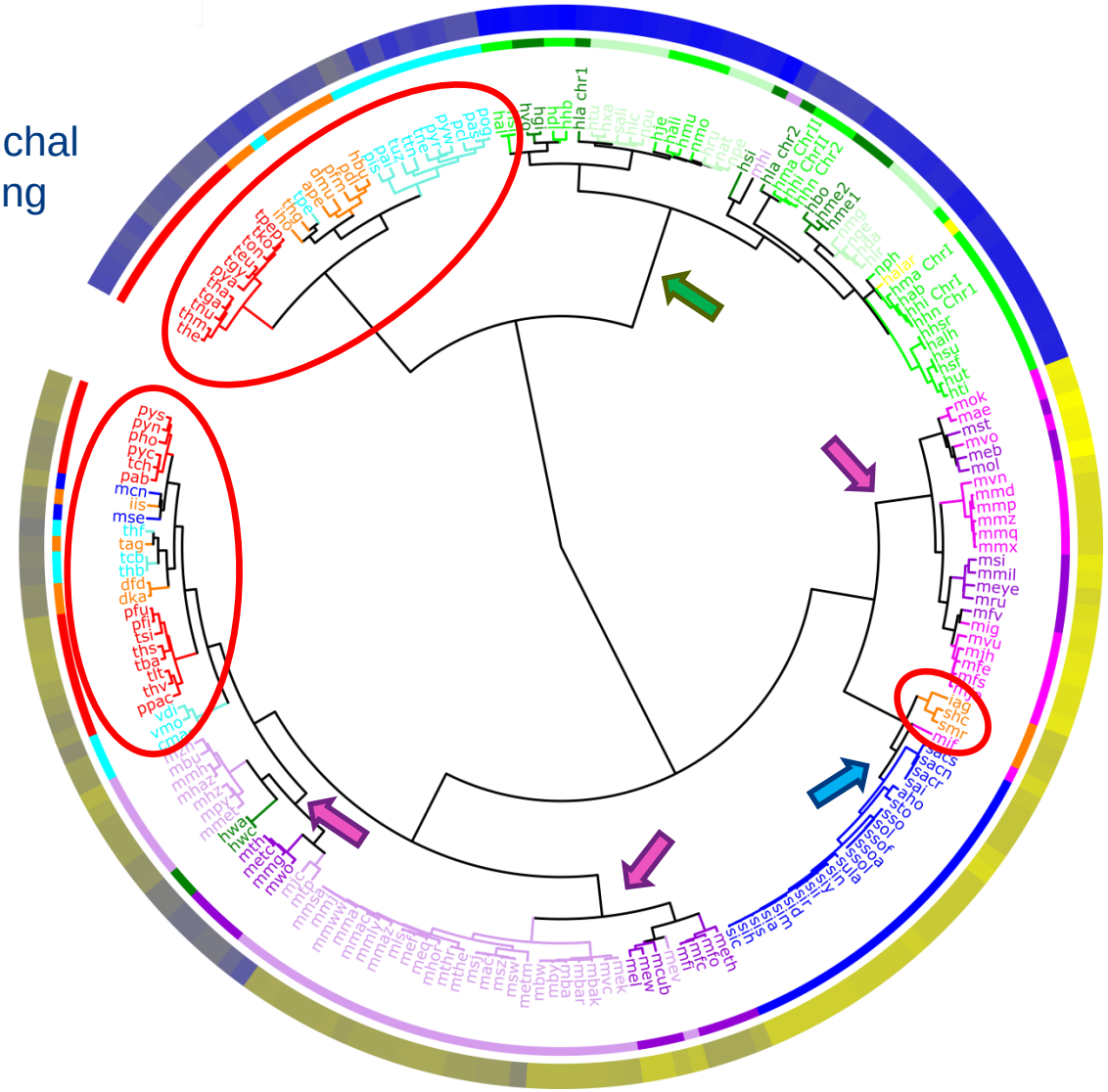
J. Iranzo, M. Krupovic, E.V. Koonin. "The double-stranded DNA virosphere as a modular hierarchical network of gene sharing." MBio 7.4 (2016): e00978-16.

Dataset

	Cells	Conjugative plasmids	Other plasmids	Viruses	Total
Crenarchaeota	***55***	***13***	***13***	***51***	***132***
Desulfurococcales	14			4	**18**
Sulfolobales	24	13	11	41	**89**
Thermoproteales	17		2	4	**23**
undet. hyperthermophilic archaea				2	**2**
Euryarchaeota	***141***	***12***	***149***	***39***	***341***
Halobacteriales	22	10	37	15	**84**
Haloferacales	10	1	29	15	**55**
Natrialbales	12		29	2	**43**
undet. haloarchaea	1	1	1	1	**4**
Gp I **Methanosarcinales**	36		12		**48**
Gp II **Methanococcales**	16		9	2	**27**
Methanobacteriales	19		7	2	**28**
Thermococcales	25		25	2	**52**
Total	**196**	**25**	**162**	**90**	**473**

Archaeal cells

k=4
Hierarchichal clustering



GC_percent

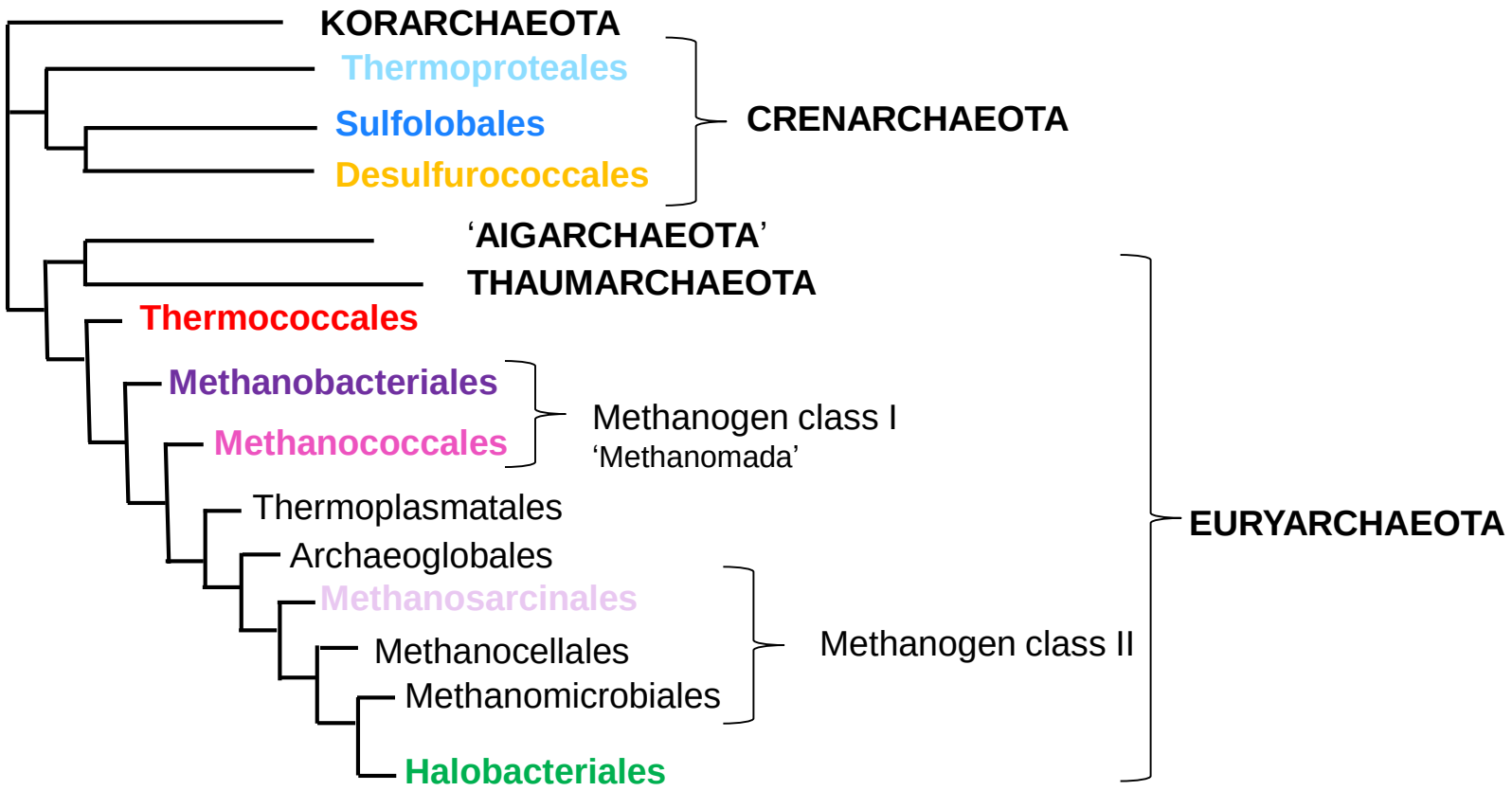
- Minimum
- Maximum

Taxonomic Order

- Haloferacales
- Halobacteriales
- Natrialbales
- haloarchaea
- Methanobacteriales
- Methanococcales
- Methanosarcinales
- Thermococcales
- Desulfurococcales
- Thermoproteales
- Sulfolobales

Archaeal cells & phylogeny

- The topology of the dendogram based on tetramer profiles is **not** consistent with the phylogeny of archaea



Based on Petitjean C et al (2015). Extending the conserved phylogenetic core of archaea disentangles the evolution of the third domain of life. MBE.

Archaeal cells & underlying factors

- Influential inter-related factors (MANOVA):

Factor	% of variance	Statistical significance
Taxonomy (order)	78%	+++
GC% content	74%	+++
“Niche”	67%	+++
Taxonomy (phylum)	6%	+++
Genome length	2%	+

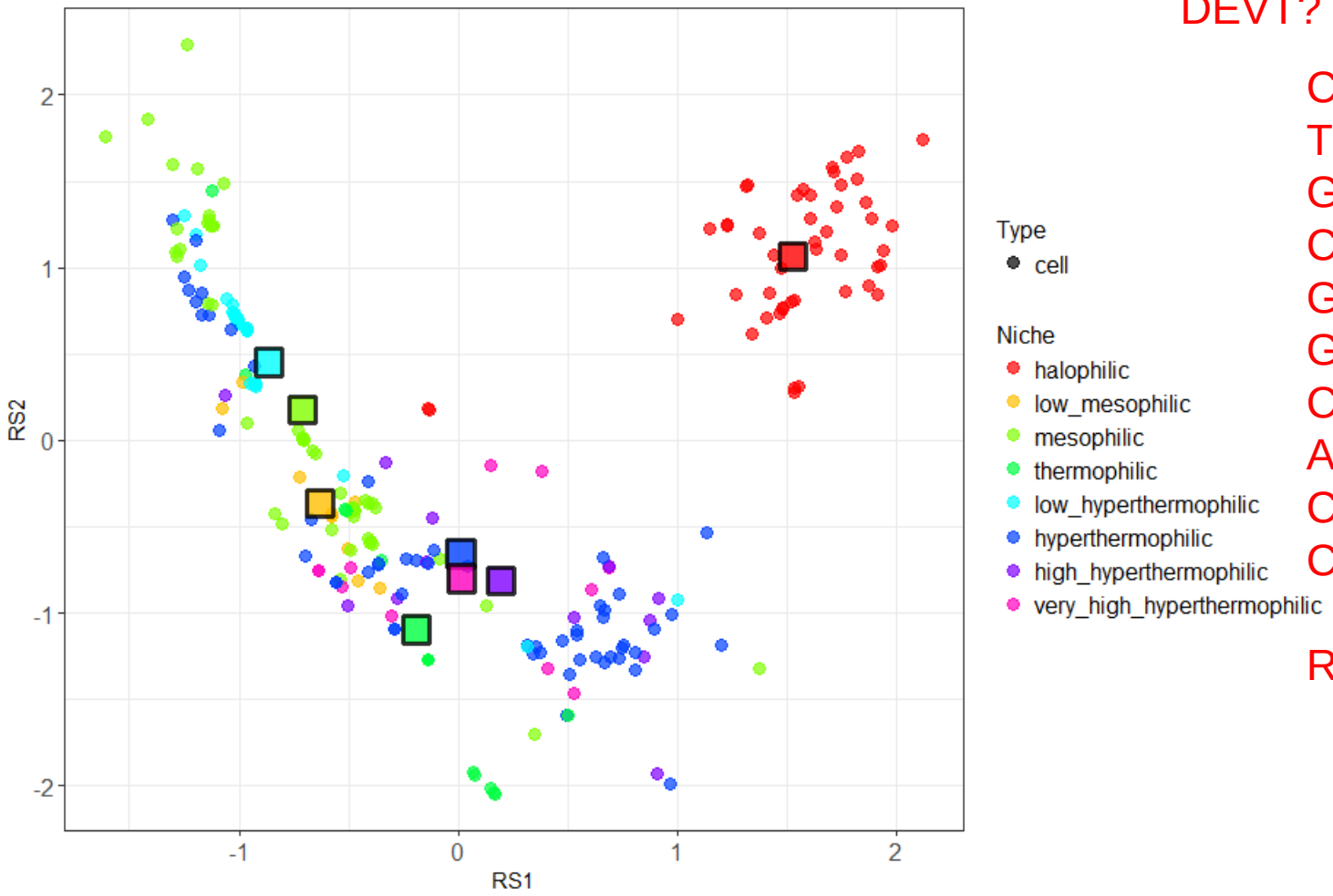
(Link between life style and codon usage: Botzman and Margalit Genome Biology 2011, 12:R109, http://genomebiology.com/2011/12/10/R109)

→ Ancient evolutionary links are not detected

- k-mer composition could evolve rapidly (e.g. Thermococcales)

Proteome adaptation to extremophilic conditions?

Principal component analysis, k=3



DEVT?

CGA
TCG
GCG
CGC
GAC
GTC
CGT
ACG
CCG
CGG

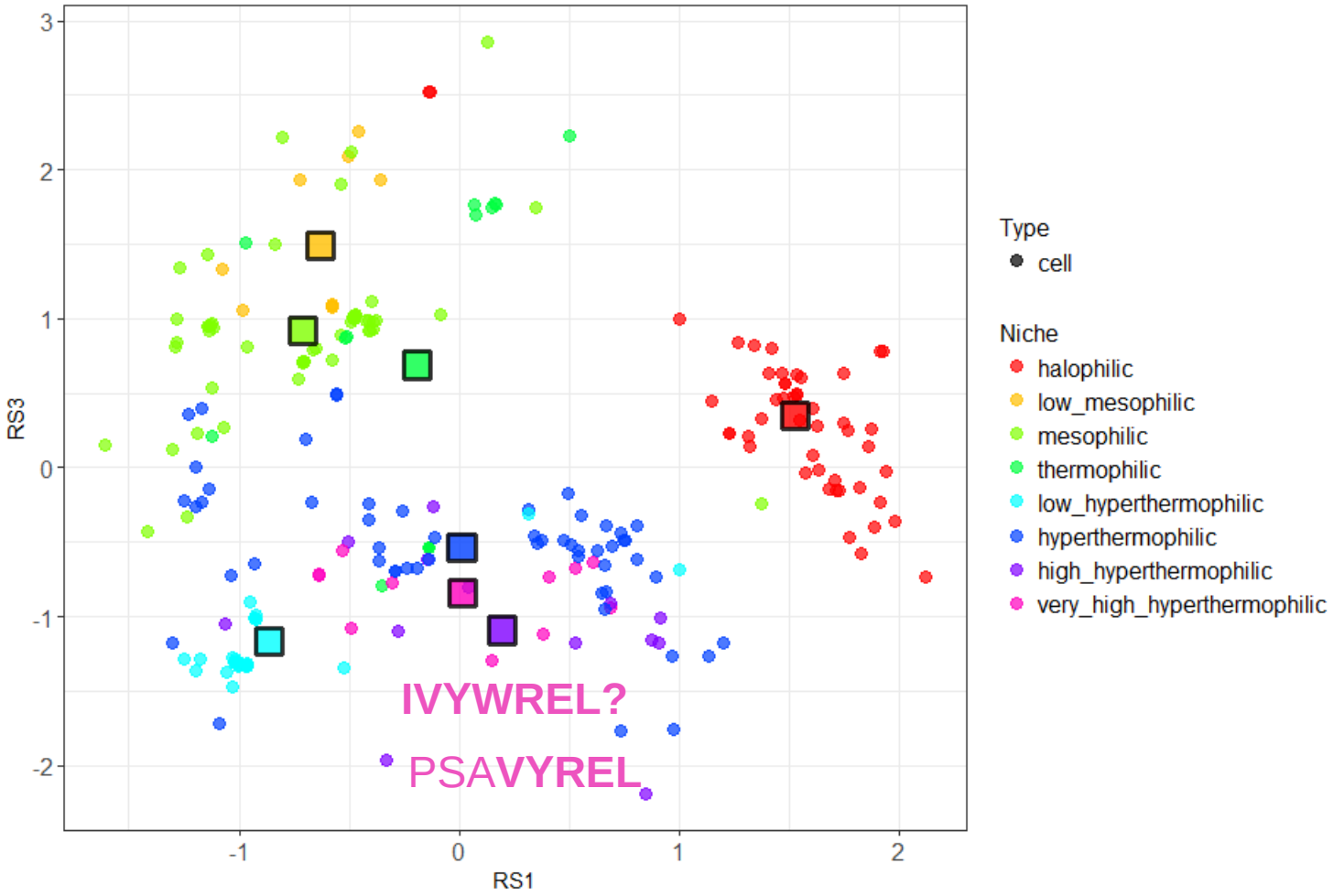
RSA**DVT**P

Link with proteome adaptation to extremophilic conditions

Principal component analysis, k=3

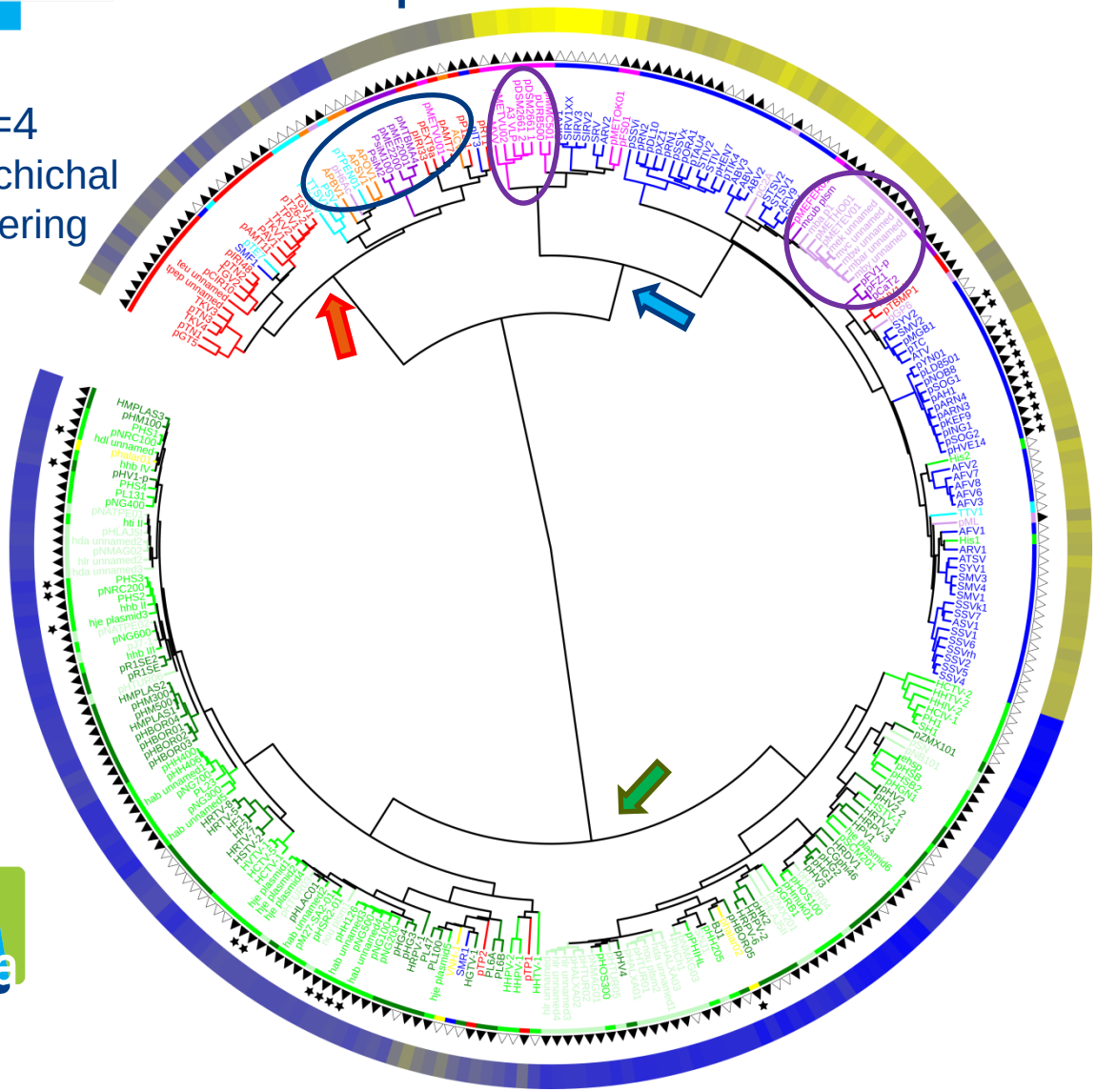
ACIDQLTS*C
TS*CQLIDAC

ACIDQTSCL



Archaeal plasmids and viruses

k=4
Hierarchichal clustering



GC_percent
- Minimum
- Maximum

13

Type of element
- Virus
- Plasmid
- Conjugative plasmid

Taxonomic Order
- Haloferacales
- Halobacteriales
- Natrialbales
- haloarchaea
- Methanobacteriales
- Methanococcales
- Methanosarcinales
- Thermococcales
- Desulfurococcales
- Thermoproteales
- Sulfolobales

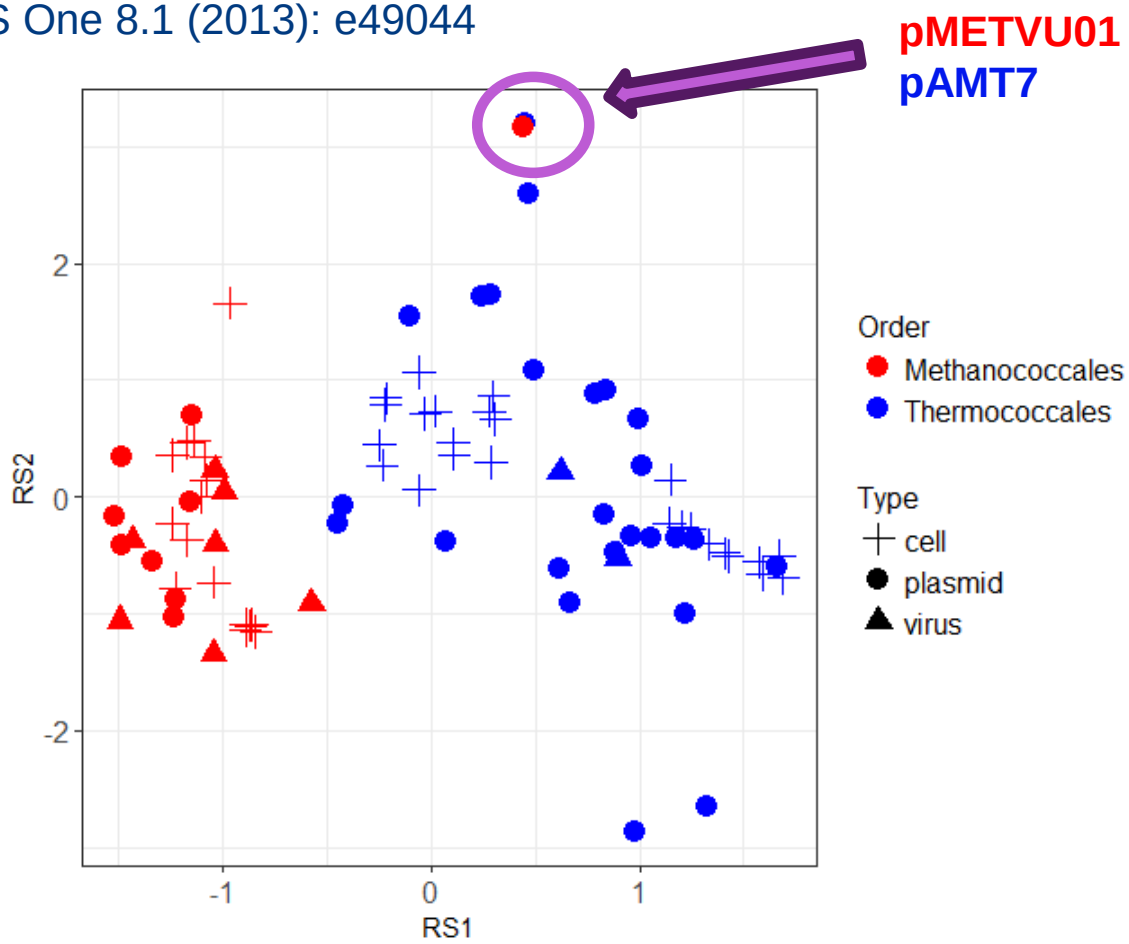
Archaeal plasmids and viruses: conclusions

- **Compared to cells, a slightly redistributed pattern**

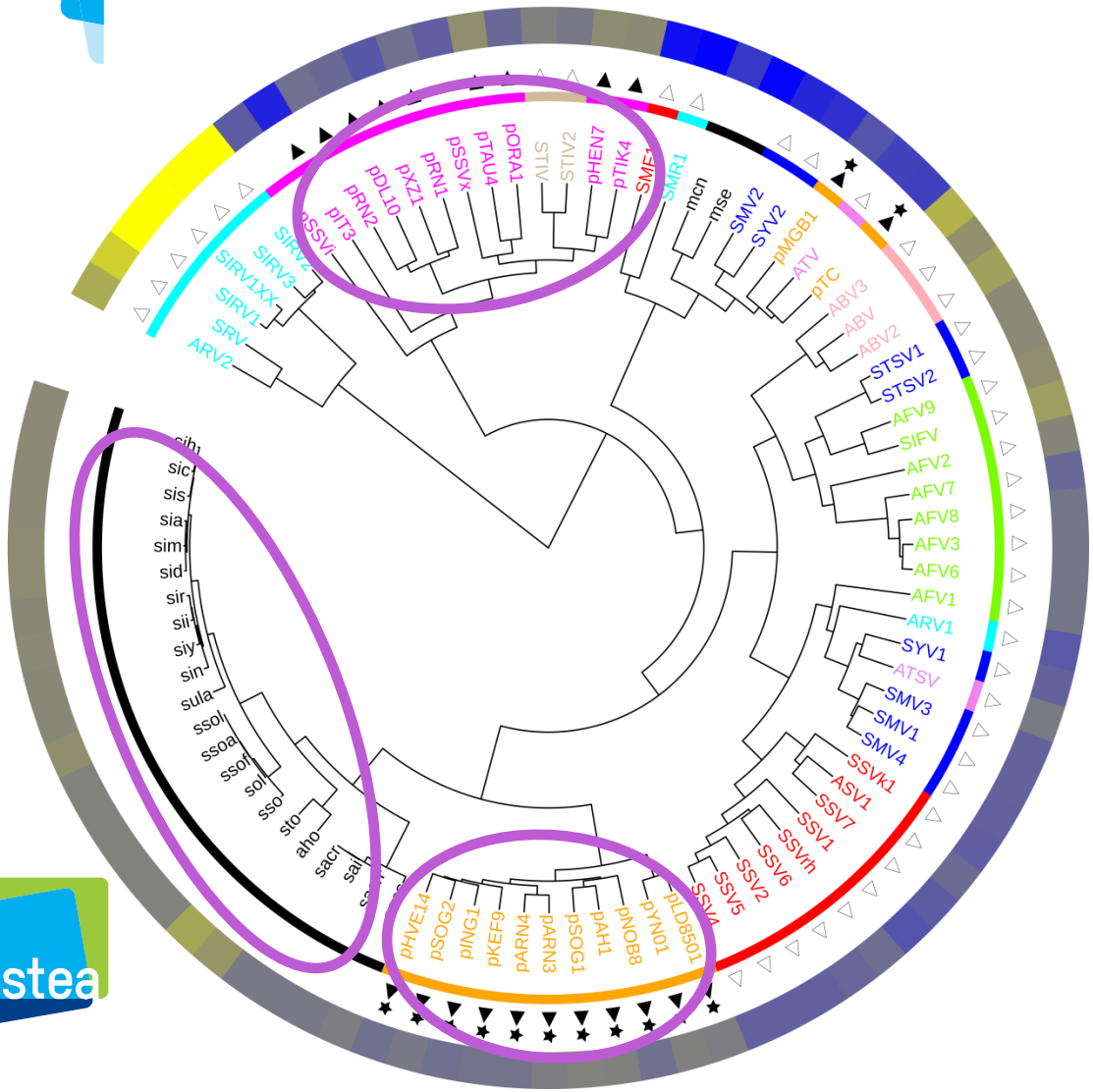
→ link with recent evolutionary events?

Krupovic, M, et al. PLoS One 8.1 (2013): e49044

Principal component analysis



Virus, plasmids, hosts: Sulfolobales (Crenarchaeota)



GC_percent

- Minimum
- Maximum

Family

- Fuselloviridae
- Lipothrixviridae
- Rudiviridae
- Ampullaviridae
- Bicaudaviridae
- Turriviridae
- Other viruses
- pNOB8-like
- pRN-like
- Sulfolobaceae

Type of element

- Virus
- Plasmid
- Conjugative plasmid

Conclusions, implications

Conclusions

- Phylogenetic position of the host (order), GC%, environment
- **Mobile genetic elements have their own signature**
- Rapidly evolving ? → recent evolutionnary events

Implications

- Host prediction: better results on archaea with WiSH than on the whole prokaryote dataset. WiSH also worked for plasmids.

- **Metagenomic binning based on k-mer**
  → possible loss of integrated mobile genetic elements?

Violette
Da Cunha

Cédric
Midoux

Sophie
Schbath

Patrick
Forterre

Thank you for your attention!

VIRAME ANR-17-CE05-0011-01

EVOLMOBIL-ERC