Journal of Molecular Evolution (2019) 87:7–15
https://doi.org/10.1007/s00239-018-9877-1

ORIGINAL ARTICLE

Genomic Signatures Among *Acanthamoeba polyphaga* Entoorganisms Unveil Evidence of Coevolution

Víctor Serrano-Solís[1] · Paulo Eduardo Toscano Soares[1] · Sávio T. de Farías[1]

Received: 6 August 2018 / Accepted: 9 November 2018 / Published online: 20 November 2018

Abstract

The definition of a genomic signature (GS) is "the total net response to selective pressure". Recent isolation and sequencing of naturally occurring organisms, hereby named entoorganisms, within *Acanthamoeba polyphaga*, raised the hypothesis of a common genomic signature despite their diverse and unrelated evolutionary origin. Widely accepted and implemented tests for GS detection are oligonucleotide relative frequencies (OnRF) and relative codon usage (RCU) surveys. A common pattern and strong correlations were unveiled from OnRFs among *A. polyphaga*'s Mimivirus and virophage Sputnik. RCU showed a common A-T bias at third codon position. We expanded tests to the amoebal mitochondrial genome and amoeba-resistant bacteria, achieving strikingly coherent results to the aforementioned viral analyses. The GSs in these entoorganisms of diverse evolutionary origin are coevolutionarily conserved within an intracellular environment that provides sanctuary for species of ecological and biomedical relevance.

Keywords Entoorganism · Genomic Signature · Virophage · Mimivrus · *Acanthamoeba polyphaga* · Coevolution

Introduction

Several organisms have been isolated within the protist *Acanthamoeba polyphaga*. The giant virus Acanthamoeba polyphaga Mimivirus (APMV) of the genus *Mimivirus*, family *Mimiviridae* (La Scola et al. 2003), is a virus that has shaken the ideas about the tree of life (Raoult and Forterre 2008; Forterre 2010). It has a huge size for a virus, a giant genome encoding tRNAs, repair proteins, translational-related proteins and, for the first time in a virus, aminoacyl-tRNA synthetases (Legendre et al. 2011), which by definition were thought to exist only in cellular organisms and not in viral genomes (Raoult et al. 2004). APMV is in turn infected by the virophage mimivirus-dependent virus Sputnik (La Scola et al. 2008), genus *Sputnikvirus*, family *Lavidaviridae*. Other giant viruses able to infect *A. polyphaga* are Megavirus chilensis (Arslan et al. 2011), Megavirus courdo7 (Desnues et al. 2012), Mimivirus terra2 (Filée 2015) and Acanthamoeba polyphaga Moumouvirus (Yoosuf et al. 2012), members of the three lineages belonging to the *Mimivirus* genus. Another non-mimivirus virus isolated in this amoeba is the Marseillevirus marseillevirus (Boyer et al. 2009) of the *Marseillevirus* genus, *Marseilleviridae* family. Regarding cellular entoorganisms, the amoeba-resistant bacteria (ARB) category has several cases isolated from Acantamoebas (Greub and Raoult 2004), such as the lytic ARB *Francisella tularensis* (Berdal et al. 1996). Recently, an intracellular bacterium, Candidatus *Babela massiliensis*, was isolated in *A. polyphaga* (Pagnier et al. 2015).

Here we propose the prefix *ento* (from the Greek ἐντός, inside), denoting an organism, whether cellular or viral, inhabiting and coevolving within the membrane boundaries of a cellular host. The coexistence of these very organisms within a single cell makes the inner parts of this protist an unexpected new environment for studying evolutionary processes (Moliner et al. 2010). This entoorganism (Table 1; see Methods) raised the hypothesis of a common feature, a feature lying at the genomic level.

Classifications of contemporary methods for comparative genomics form two groups: parametric and phylogenetic methodologies (Ravenhall et al. 2015). Since there is no homology between all the entoorganisms of this study, phylogenetic methodologies may not be effective. Parametric

✉ Víctor Serrano-Solís
vicmsersol@gmail.com

[1] Laboratório de Genética Evolutiva Paulo Leminsk, Departamento de Biologia Molecular, Centro de Ciencias Exatas e da Natureza, Universidade Federal da Paraíba, João Pessoa, Brazil

Table 1 List of genomes and general characteristics of the amoeba, amoebal mitochondrion, amoebal entoorganisms and negative controls studied in this work

Organism	GC (%)	Length (bp)	CDS	Acession ID
Acantamoeba polyphaga	57.17	**	**	PRJNA307312 ***
A. polyphaga str. Linc Ap-1 mitochondrion	28.96	39,215	35	KP054475.2
Acanthamoeba polyphaga Mimivirus	27.95	1,181,549	979	NC_014649.1
Acanthamoeba polyphaga Moumovirus	24.61	345,413	349	NC_020104.1
Megavirus chilensis	25.23	1,259,197	1120	NC 016072.1
Megavirus courdo 7	25.20	529,672	535	J N885991.1
*Mimivirus terra2**	27.94	1,168,989	N/A	NC 023639.1
Mimivirus-dependent virus Sputnik 2	27.05	18,338	20	J N603369.1
Candidatus Babela massiliensis strain BABL 1	27.38	1,118,422	983	NC 023003.1
Francisella tularensis	32.26	1,892,775	1556	NC_006570.2
Marseillevirus marseillevirus	44.73	368,454	428	NC_013756.1
Deinococcus radiodurans	67.01	2,648,638	2629	NC 001263.1
Human immunodeficiency virus 1	42.12	9,181	10	NC 001802.1

*Partial genome, **no data available, genome at the scaffold level, ***BioProject, genome's draft

methods for sequence analyses search for characteristic patterns of a particular clade and can unveil a genomic signature (GS) that reflects a "total net response to selective pressure" (Karlin and Burge 1995; Abe et al. 2003). Oligonucleotide distributions and codon usage (CU) profiles are well-known and accepted GS methodologies (Burge et al. 1992; van Passel et al. 2006). Regardless of the sequence length and region selected, composition biases are detected. This phenomenon is called pervasivity (Deschavanne et al. 2000; Jernigan and Baran 2002). This pervasivity is constant in a species genome and differs between related species (Gentles and Karlin 2001; Lerat et al. 2002).

Several GS analyses have been performed on plasmids, phages and viruses (Blaisdell et al. 1996; Campbell et al. 1999; Robins et al. 2005; Pride et al. 2006; Mrázek and Karlin 2007; Suzuki et al. 2008). An advantage of using GS instead of traditional phylogenetic methods is that results will not vary regarding the set of sequences utilized (Campbell et al. 1999). Another advantage is that the use of GS allows comparisons regarding a lack of common ancestor, independence of base composition, coding or noncoding regions, making comparison of viral and cellular organisms possible.

In this article, we will put forward the hypothesis of the entoorganisms and inner organisms—amoebal in this case—coevolving and sharing a common genomic pattern and possible explanations for such evolutionary phenomena.

Materials and Methods

Sequences

For this study, we used scaffolds of the host *Acanthamoeba polyphaga*'s genome (Apss), *A. polyphaga's* mitochondrial genome (Apm), genomes of the viruses Megavirus chilensis (Mch), Megavirus courdo7 (Mc7), Mimivirus terra2 (Mt2), APMV, APMoV, the virophage mimivirus-dependent virus Sputnik (Spu) and Marseillevirus marseillevirus (Mma) and genomes of the cellular organisms Candidatus *Babela massiliensis* (Bab) and *Francisella tularensis* (Ftu). As negative controls, we used the genomes of the bacteria *Deinococcus radiodurans* (Dra) and the human immunodeficiency virus 1 (HIV) (Table 1). All sequences were downloaded from the NCBI GeneBank (Benson et al. 2017) and the Viral Genome Resource (Brister et al. 2015).

Oligonucleotide Frequencies

Oligonucleotide relative frequencies (OnRF), namely dinucleotide relative frequencies (DiRF), trinucleotide relative frequencies (TriRF) and tetranucleotide relative frequencies (TetRF), were obtained by an algorithm written at our group using Python (Rossum et al. 2010), which counted the frequency of each oligonucleotide and returned its relative frequency.

Relative Codon Usage (RCU)

The relative codon usage test was described by Sharp and Li (1987) to examine codon usage without the confounding influence of amino acid composition of different gene products. Here we implemented it with the modification of including methionine, tryptophan and stop codons, originally not considered by Sharp because of the lack of synonym codons. Calculations were done using the same script written in Python 3.5 to obtain the codon count of each CDS and then calculate the relative frequency among the synonymous codons. Heatmap was plotted using the PAST software v3.15 (Hammer et al. 2001).

Genomic Landscape at the Codon Usage (GLCU)

The genomic landscape at the codon usage (GLCU) was obtained by calculating the average RCU frequencies of each codon on every CDS of a genome. The codon count was obtained in the same step as the RCU, and the calculations were performed in the same script written in Python 3.5. As in the RCU, stop codons were retrieved from the CDS as well.

Correlation Analyses

Pearson's correlation analyses of oligonucleotide frequencies were calculated between the frequencies of each genome's dinucleotides, trinucleotides and tetranucleotides. It was also performed between the GLCU of each genome, including stop codons, obtained using the PAST v3.15 software using the 'Linear r (Pearson)' for the correlation statistic parameter and 'Statistic\p(uncorr)' for the table format parameter.

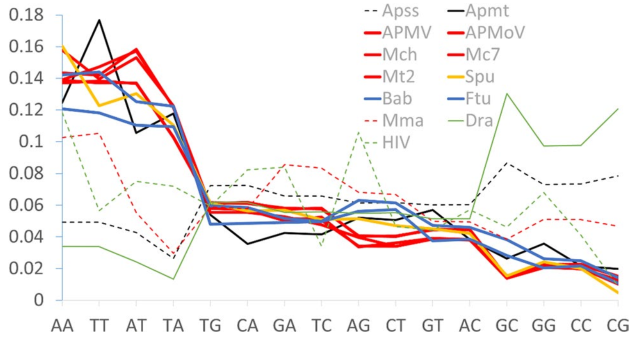

Fig. 1 Genomic signatures from dinucleotide relative frequencies. Dinucleotide distribution values are sorted in descending order. Black lines correspond to *Acanthamoeba polyphaga*'s genomic scaffolds (Ap) and mitochondrial genome (Apmt). Red lines correspond to viral genomes Acanthamoeba polyphaga Mimivirus (APMV), Acanthamoeba polyphaga Moumovirus (APMoV), Megavirus chilensis (Mch), Megavirus courdo7 (Mc7), Mimivirus terra2 (Mt2) and Marseillevirus marseillevirus (Mma). Orange line is the Sputnik (Spu) virophage. Blue lines correspond to bacteria isolated from *A. polyphaga: Candidatus Babela massillensis* (Bab) and *Francisella tularensis* (Ftu). Green lines correspond to negative controls: as cellular, *Deinococcus radiodurans* (Dra); as viral negative control, human immunodeficiency virus 1 (HIV)

Results

Dinucleotide Relative Frequencies

Since oligonucleotide frequencies are influenced by GC content, and GC content is usually related to the environment (Karlin 1998; Foerstner et al. 2005), a common genomic pattern may be elucidated from *A. polyphaga*'s entoorganisms.

Dinucleotide relative frequencies (DiRF) of the *Acanthamoeba polyphaga* mitochondrial genome, all Mimivirus and Sputnik genomes—i.e. the entovirals, endobacteria Candidatus *B. massiliensis* and ARB *Francisella tularensis*, showed strikingly similar DiRF profiles, as depicted in Fig. 1. These entoorganism profiles behave very similarly given any dinucleotide combination. Other ARBs (e.g., *Parachlamydia acanthamoebae*) were used as well, rendering very similar results to Ftu in every test; these are not shown because of redundancy. On the other hand, the *Acanthamoeba polyphaga* scaffolds, *Marseillevirus marseillevirus* and the negative controls *Deinococcus radiodurans* and HIV show clearly different patterns in their dinucleotide profiles.

Broadly, there are three groups of DiRF. First is the higher group comprised of AA, TT, AT and TA with a range of 12–18% DiRF each. Second is the medium group comprised of TG, CA, GA, TC, AG, CT and AC with a range of 4–6% DiRF observed. Third is a low group made up of GC, CC, CC and CG with DiRF values < 4% each. The profiles of Apmt, APMV, APMoV, Mch, Mc7, Mt2, Spu, Bab and Ftu show very similar distributions within ranges of 18–11% at the high value group, as described above. The medium and lower groups show even less variance.

In closer analyses, the mitochondrial genome of *A. polyphaga* showed the unique highest DiRF on TT with almost 18%; the rest of this organelle profile is highly correlated with the rest of the entoorganisms. The DiRF of the *A. polyphaga* scaffolds shows a different pattern regarding its mitochondria and the entoorganisms.

Candidatus *B. massiliensis* and *Francisella tularensis* show distributions resembling the entoviral and mitochondrial DiRF patterns. Unexpectedly, with Mma—a non-Mimivirus—we observed a different GS profile, similar to the negative controls. As negative controls, *Deinococcus radiodurans* and human immunodeficiency virus 1 were selected. Their genomic lengths, GC content and biology render completely different profiles, as expected.

Overall Pearson correlation tests r were performed on DiRF, TriRF and TetRF, namely r^{di}, r^{tri} and r^{tet}, respectively. Regardless of the OnRF, high correlation values $r \geq 0.89$ were detected in pairwise viral comparisons as well as with Candidatus *B. massiliensis* $r \geq 0.95$ and *Francisella tularensis* $r \geq 0.92$. Similar correlations $r \geq 0.89$ were detected between the mitochondria and entoamoebal organisms. A subtle decay in correlation values was detected with increasing OnRF complexity, TriRF and TetRF, as shown in Tables 2 and 3.

High correlation values were detected among entoviruses at $r^{di} \geq 0.99$, $r^{tri} \geq 0.98$ and $r^{tet} \geq 0.85$. *APMV* pairwise correlations were the highest detected, for example, *APMV* and

Table 2 Pearson correlation analyses of the DiRF, TriRF and TetRF of entoorganisms

		APss	Apmt	APMV	APMoV	Mch	Mc7	Mt2	Spu	Bab	Ftu	Mma	Dra	HIV
DiNT	APss		3.0E-04	2.8E-04	1.2E-04	1.5E-04	1.3E-04	2.8E-04	1.6E-04	5.9E-05	5.6E-05	5.8E-01	3.3E-06	1.1E-01
	Apmt	-0.787		6.7E-07	2.0E-07	1.0E-06	4.7E-07	6.0E-07	2.8E-06	1.4E-08	2.1E-07	5.3E-02	1.7E-03	1.4E-01
	APMV	-0.789	0.915		1.3E-14	6.1E-14	8.7E-14	3.7E-26	1.9E-11	8.9E-10	5.5E-09	4.0E-02	1.9E-04	4.5E-02
	APMoV	-0.815	0.929	0.994		1.4E-18	1.1E-18	1.5E-14	9.3E-11	6.2E-11	2.1E-09	8.2E-02	3.8E-04	5.5E-02
	Mch	-0.809	0.910	0.992	0.998		1.3E-20	8.1E-14	3.3E-10	9.7E-10	8.5E-09	1.0E-01	4.1E-04	5.8E-02
	Mc7	-0.812	0.920	0.991	0.998	0.999		6.9E-14	1.1E-09	5.5E-10	7.2E-09	1.0E-01	3.9E-04	7.4E-02
	Mt2	-0.789	0.917	1.000	0.993	0.992	0.992		4.5E-11	8.9E-10	6.0E-09	4.0E-02	1.9E-04	4.9E-02
	Spu	-0.807	0.895	0.981	0.977	0.972	0.967	0.979		3.2E-10	3.4E-10	4.4E-02	1.2E-04	1.1E-02
	Bab	-0.834	0.952	0.968	0.978	0.967	0.970	0.968	0.972		3.8E-13	5.9E-02	3.1E-04	3.8E-02
	Ftu	-0.835	0.929	0.958	0.964	0.955	0.956	0.958	0.972	0.989		8.3E-02	9.0E-05	2.0E-02
	Mma	-0.149	0.492	0.517	0.447	0.421	0.425	0.518	0.510	0.482	0.447		1.7E-01	1.6E-01
	Dra	0.893	-0.719	-0.802	-0.779	-0.776	-0.778	-0.802	-0.814	-0.785	-0.823	-0.363		3.8E-02
	HIV	-0.411	0.389	0.508	0.489	0.482	0.459	0.500	0.614	0.523	0.575	0.372	-0.523	
		Apss	Apmt	APMV	APMoV	Mch	Mc7	Mt2	Spu	Bab	Ftu	Mma	Dra	HIV
TriNT	Apss		2.8E-09	9.8E-10	3.0E-10	3.9E-10	2.9E-10	1.0E-09	1.2E-10	1.0E-11	6.1E-12	9.7E-01	9.8E-17	8.4E-03
	Apmt	-0.661		7.9E-22	6.0E-25	7.2E-22	4.0E-23	5.5E-22	6.1E-20	7.4E-30	9.2E-25	5.5E-04	1.9E-06	3.7E-03
	APMV	-0.675	0.881		6.3E-49	3.5E-50	1.2E-49	1.8E-103	2.2E-39	2.6E-32	2.2E-28	6.6E-04	3.5E-09	9.1E-05
	APMoV	-0.689	0.907	0.985		2.4E-69	2.8E-70	1.4E-48	5.1E-36	8.4E-39	1.9E-29	4.3E-03	5.5E-08	2.1E-04
	Mch	-0.686	0.881	0.986	0.997		3.9E-81	1.4E-49	1.7E-34	8.2E-34	8.6E-28	1.0E-02	5.4E-08	3.2E-04
	Mc7	-0.690	0.893	0.986	0.997	0.999		5.7E-50	1.8E-32	7.4E-35	3.9E-28	8.5E-03	5.0E-08	6.4E-04
	Mt2	-0.674	0.883	1.000	0.984	0.986	0.986		8.9E-38	3.3E-32	3.0E-28	6.4E-04	3.5E-09	1.3E-04
	Spu	-0.700	0.862	0.969	0.960	0.955	0.948	0.965		3.6E-33	2.4E-31	5.9E-04	1.0E-09	6.0E-07
	Bab	-0.727	0.936	0.947	0.968	0.953	0.956	0.947	0.951		6.6E-43	1.6E-03	1.2E-08	5.2E-05
	Ftu	-0.732	0.906	0.929	0.934	0.925	0.927	0.928	0.943	0.976		3.1E-03	1.6E-10	5.9E-06
	Mma	0.006	0.420	0.415	0.353	0.318	0.326	0.415	0.418	0.387	0.365		3.9E-02	3.1E-03
	Dra	0.821	-0.555	-0.658	-0.617	-0.618	-0.619	-0.658	-0.674	-0.640	-0.697	-0.259		1.5E-04
	HIV	-0.327	0.358	0.469	0.447	0.436	0.415	0.461	0.577	0.483	0.532	0.364	-0.456	
		Apss	Apmt	APMV	APMoV	Mch	Mc7	Mt2	Spu	Bab	Ftu	Mma	Dra	HIV
TetraNT	Apss		3.7E-16	4.2E-17	4.1E-16	1.6E-16	1.5E-16	5.0E-17	2.2E-17	5.2E-20	2.8E-22	1.7E-01	6.9E-54	5.5E-04
	Apmt	-0.480		1.4E-73	1.7E-84	7.1E-74	4.4E-78	4.8E-74	3.1E-66	5.9E-103	1.5E-83	1.8E-10	1.5E-11	3.9E-08
	APMV	-0.493	0.853		6.4E-170	1.1E-180	1.2E-178	0.0E+00	3.2E-133	1.2E-110	2.6E-98	1.8E-10	3.4E-17	1.1E-13
	APMoV	-0.479	0.881	0.976		2.4E-254	4.6E-258	1.4E-168	1.7E-119	5.4E-136	3.8E-98	1.3E-07	1.0E-13	4.2E-12
	Mch	-0.485	0.854	0.980	0.995		1.2E-307	1.9E-178	2.7E-116	5.7E-118	2.4E-94	2.0E-06	4.3E-14	1.8E-11
	Mc7	-0.485	0.865	0.979	0.995	0.998		1.5E-179	3.4E-110	1.7E-121	2.3E-95	8.9E-07	4.9E-14	1.6E-10
	Mt2	-0.492	0.854	1.000	0.975	0.979	0.980		5.9E-128	4.2E-110	1.1E-97	1.6E-10	3.6E-17	3.7E-13
	Spu	-0.497	0.830	0.953	0.939	0.935	0.927	0.948		1.6E-108	2.2E-103	7.1E-11	7.1E-17	4.8E-21
	Bab	-0.531	0.916	0.928	0.955	0.937	0.941	0.927	0.925		3.6E-143	7.6E-09	1.1E-15	1.2E-14
	Ftu	-0.557	0.879	0.909	0.908	0.901	0.903	0.908	0.917	0.961		1.5E-08	1.2E-20	9.2E-18
	Mma	0.087	0.385	0.385	0.323	0.292	0.302	0.386	0.393	0.351	0.345		2.9E-03	3.5E-09
	Dra	0.781	-0.405	-0.494	-0.443	-0.449	-0.448	-0.494	-0.490	-0.473	-0.539	-0.186		8.0E-09
	HIV	-0.214	0.335	0.442	0.415	0.404	0.386	0.434	0.543	0.458	0.502	0.359	-0.351	

Over the diagonal, are the statistical significance (*p*) values. Under the diagonal, correspond to Correlation (*r*) values. Strong correlations are observed among the entoorganisms studied in this work. (Color table online)

Table 3 Pearson correlation analyses of GLCU between genomes

	Apmt	APMV	APMoV	Mch	Mc7	Spu	Bab	Ftu	Mma	Dra	HIV
Apmt		8.7E-27	3.2E-27	2.5E-27	3.0E-27	5.4E-25	1.9E-30	2.4E-32	1.2E-07	4.9E-01	3.0E-11
APMV	0.919		2.1E-47	6.9E-53	1.5E-47	5.8E-42	5.8E-35	3.7E-33	2.1E-09	2.9E-01	3.8E-11
APMoV	0.922	0.983		3.4E-66	1.4E-66	3.1E-42	3.9E-42	2.3E-34	1.9E-07	9.9E-02	1.3E-10
Mch	0.923	0.989	0.996		4.3E-74	3.3E-40	5.7E-40	6.1E-34	1.4E-07	1.4E-01	2.7E-10
Mc7	0.922	0.983	0.996	0.998		3.2E-39	2.6E-40	2.8E-33	2.6E-07	1.3E-01	2.7E-10
Spu	0.907	0.975	0.975	0.971	0.969		5.1E-36	1.5E-32	8.3E-07	5.1E-02	1.6E-10
Bab	0.939	0.957	0.975	0.970	0.971	0.960		3.4E-40	1.0E-07	1.5E-01	6.3E-12
Ftu	0.947	0.950	0.955	0.953	0.951	0.948	0.971		3.5E-08	2.8E-01	6.8E-14
Mma	0.605	0.665	0.597	0.603	0.592	0.571	0.607	0.625		1.2E-04	1.8E-09
Dra	-0.088	-0.133	-0.208	-0.184	-0.192	-0.245	-0.180	-0.138	0.461		3.3E-01
HIV	0.716	0.713	0.699	0.691	0.691	0.697	0.732	0.773	0.667	0.125	

Values under the diagonal show the Correlation (*r*) values. Red cells highlight overall lower correlation values, green cells highlight overall higher correlation values. Over the diagonal values, correspond to statistical significance (*p*). (Color table online)

Mt2 values $r^{di} = 1$, $r^{tri} = 1$ and $r^{tet} = 1$. Also, high correlations were detected between *APMV* and Spu $r^{di} = 0.98$, $r^{tri} = 0.96$ and $r^{tet} = 0.95$.

The virophage Sputnik showed lower correlation with the mitochondria $r^{di} = 0.89$ $r^{tri} = 0.86$ and the highest with APMoV $r^{di} = 0.98$, $r^{tri} = 0.96$ and $r^{tet} = 0.95$. In the case of Bab, r^{di} are always close to 0.97 with every entoorganism, except for Apmt and APMV. For the ARB Ftu, the highest correlations are with Bab $r^{di} = 0.98$, $r^{tri} = 0.97$ and $r^{tet} = 0.96$ and lower with Apmt $r^{di} = 0.92$, $r^{tri} = 0.9$ and

APMV r^{tet} = 0.87. For the mitochondria, the highest values are with Bab r^{di} = 0.95and Babr^{tri} = 0.93.

For the negative cellular control Dra, regardless of OnRF combination, every pairwise comparison resulted in negative values except at pairing with Apss. For the viral negative control HIV, values are near $r \leq 0.5$ and negatives at pairing with Apss.

An interesting case is the Mma, showing values of $r^{di} \leq 0.52$, $r^{tri} \leq 0.42$ and $r^{tet} \leq 0.39$. Its higher values are constant with Spu.

Relative Codon Usage (RCU)

The RCU test on all CDSs unveiled the GSs at the codonic level on every genome. The RCU of all entoorganisms showed a high preference for codons ending in A or T (darker halves on Fig. 2 profiles) and low or no preference for codons ending in C or G (clearer halves on Fig. 2 profiles). A bias was expected because of the low GC content (see Table 1); what was not expected was a common expression at the third position of the codons. Non-entoorganisms, namely Mma and the negative controls Dra and HIV, do not have this RCU pattern. Organisms with > 1000 CDSs (Table 1) and random samples of 1000 CDSs of the given genome were used.

Broadly, neat strips of preference are common for entoorganisms, namely the Apmt, entoviruses and entobacteria. Analyzing the A/T-ending high-frequency codons, AAA (lysine), CAA (glutamine), GAA (glutamic acid), TAA (Ochre), AAT (asparagine), CAT (histidine), GAT (aspartic acid), TAT (tyrosine), TGT (cysteine) and TTT (phenylalanine) are the most frequent common entootganism codons.

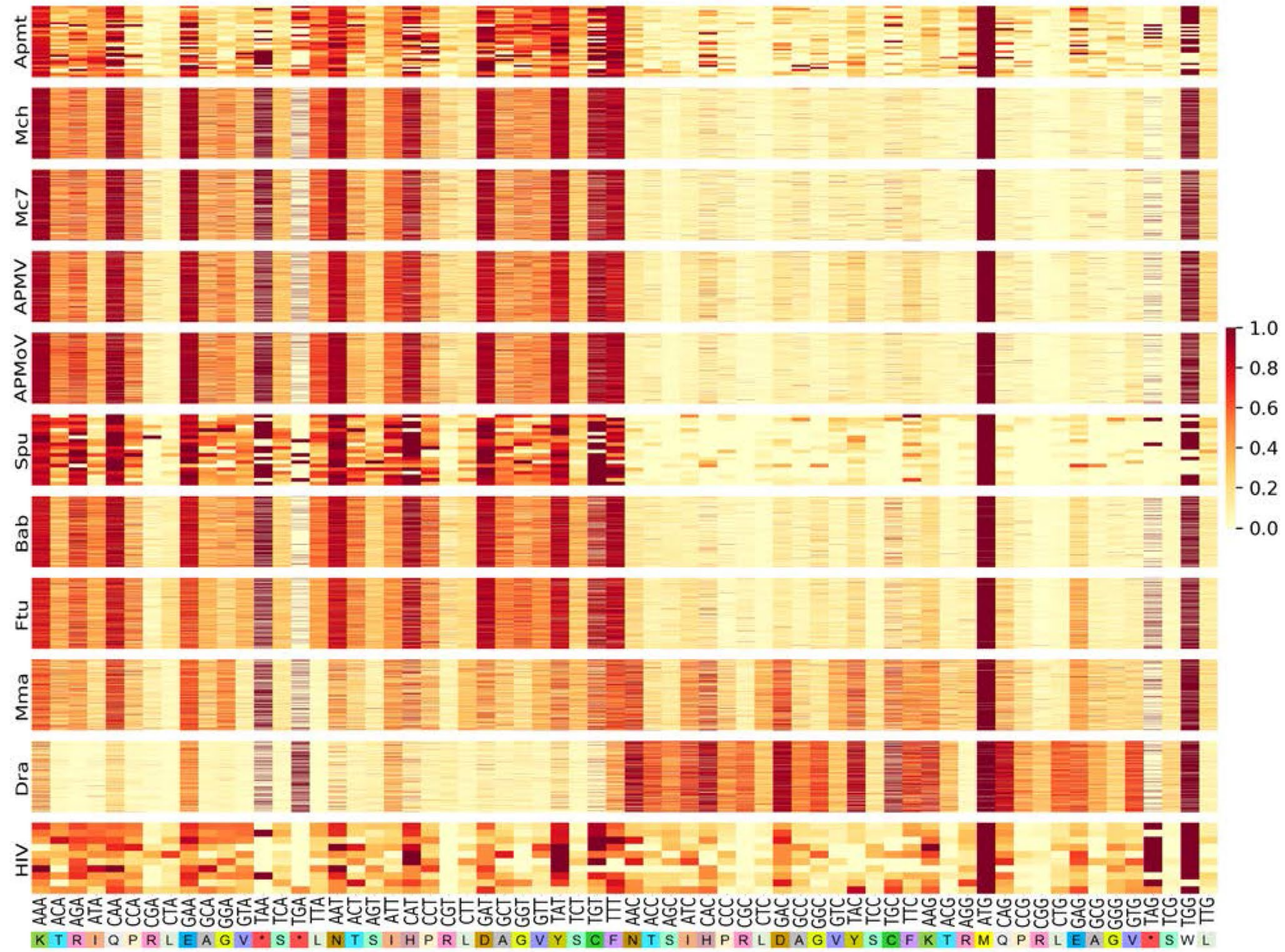

Fig. 2 Genomic signature at relative codon usage (RCU). Each CDS of the given genome is displayed as a row. Codons correspond to columns and are sorted by the third position. Mt2 is not included because of its partial genome. Apss is not included because of the lack of available CDS information. RCUs were calculated according the standard codon table. Stop codons (*) are included. Methionine (M) and tryptophan (W) codons are depicted as well. The legend for the amino acid translation is based on the standard genetic code and serves as a global reference

The A/T-ending low-frequency codons also form a neat common pattern for entoorganisms, e.g., CGA (arginine), CTA (leucine), TGA (stop), AGT (serine), CGT (arginine), CTT (leucine) and TCT (serine).

Among the C/G-ending frequency codons of entoorganisms, all are close to zero, but ATG (methionine), TAG (Amber) and TGG (tryptophan) show a presence-absence behavior because of the lack of synonym codons.

Ftu presents a slight reduced preference in GGA (glycine) with respect to all entoorganisms. Ftu and Bab show higher preferences for GGC (glycine), TGC (cysteine), AAG (lysine) and GAG (glutamic acid) than the entovirus and the mitochondrion.

Mma also shows a preference for the most frequent codons used by entoorganisms, but with lower values. The viral negative control, HIV, among its more homogeneous RCUs, shows some common preferences to entoorganisms, e.g., CAA (glutamine), TAT (tyrosine) and TGT (cysteine). Interestingly, CGA is rather low in all. On the other hand, AAA (lysine) and GAA (glutamic acid) are universally preferred.

Genomic Landscape at Codon Usage

Calculating the RCU bias per each CDS of every genome leads us to construct a new picture of codonic genomic values for a faster and condensed overall visualization and comparison of entoorganisms, the genomic landscape at codon usage (GLCU). Codons were sorted by the third position as well, and the common codon preference pattern is shown in Fig. 3, confirming Fig. 2's results.

This test shows again a general preference for A/T-ending synonymous codons in all entoorganisms. There are shared preferences of codon usage such as AAA (Lys), CAA (Gln), GAA (Glu), AAT (Asn), GAT (Asp) and TTT (Phe) and slightly lower ones auch as TAA (Ochre), GAT (Asp) and TGT (Cys). The C/G-ending preferred codons are with ATG (Met) and TGG (Trp) because of their uniqueness.

A second group of prevalence codons comprises those with a frequency of $<40\%$, namely GTT and GTA (both Val); TCT, TCA and AGT (all Ser); CCA and CCT (Pro); ACT and ACA (both Thr); GCT and GCA (both Ala); GGA and GGT (both Gly).

There are overall differences at stop codon frequencies. Despite TAA (Ochre) being the most frequent among entoorganisms, TGA (Opal) is the second most preferred, with TAG (Amber) almost avoided.

Mma shows a homogeneous distribution in preferences regarding the entoorganisms. The negative cellular control Dra is prone to G/C-ending codons. The negative viral control HIV shows a bias for A-ending codons.

Pearson correlation analyses were performed for the GLCU to compare the patterns found for each genome. Very high correlation values are detected among the entoviral genomes closely followed by Sputnik and the entobacteria. Bab had the highest correlation with Moumovirus $r = 0.975$, but correlated strongly with the entoviruses as well as with Ftu $r = 0.97$ and Apmt $r = 0{,}93$. The mitochondria have the least correlated codon usage with viruses (ranging from $r = 0.907$ to $r = 0.92$). Ftu correlation values range from $r = 0.94$ pairing with Apmt to $r = 0.97$ pairing with Bab.

As expected, correlation values of the cellular negative control Dra were nonsignificant compared with every other genome. Values were always $r \leq 0.46$. Interesting results are the pairwise comparisons of Mma and HIV with every entoorganism. As depicted in Figs. 2 and 3, HIV and Mma showed a different pattern regarding entoorganisms. In this analysis they do as well, but HIV shows higher correlation values than Mma, though not significantly regarding entoorganisms.

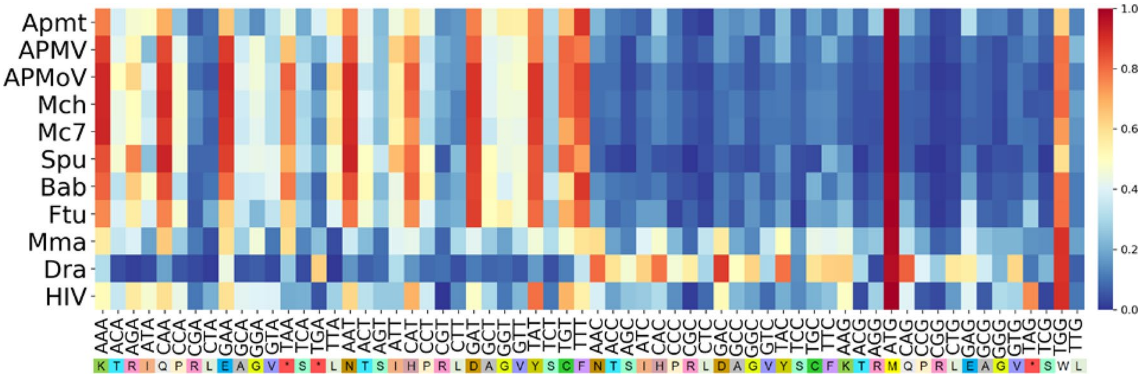

Fig. 3 Genomic signature at the genomic landscape at codon usage (GLCU). Rows are the average of codon frequencies of the given genome; columns represent codons. They are sorted in the same way as RCU for easy comparison. Stop codons (*), methionine and tryptophan codons are also included. The legend for the amino acid translation is based on the standard genetic code and serves as a global reference

Discussion

Karlin and Burge (1995) defined the genomic signature as the "total net response to selective pressure." Several factors interact to maintain the constant and coherent uniqueness of a GS: restriction avoidance (McDowall et al. 1994), core processes such as replication, recombination and reparation of DNA (Moran 2002), physical constraints such as the DNA structure stacking energy (Sinden 1994) and DNA curvature (Kozobay-Avraham et al. 2006; Mrázek 2009). Mutational processes include methylation, short oligonucleotide modifications and context-dependent mutation biases (Karlin 1996). There are also environmental factors such as energy sources and temperature (Kirzhner et al. 2007), g-radiation damage and osmolarity gradients (Prabha and Singh 2014). Even habitats and lifestyles exert selective pressure on maintaining a GS (Foerstner et al. 2005; Xia et al. 2002).

Phylogenetic-oriented genomic comparisons between the *A. polyphaga* and *A. polyphaga* mitochondria and entoorganisms presented here would not be effective because of their different evolutionary origins (Blaisdell et al. 1996; Serrano-Solís et al. 2016). Analyses for GS detection have advantages: they do not depend on homolog aligning tests, use the whole genome, present small variances and are unaffected by mutations such as rearrangements (Karlin 1998). In this study, we demonstrate an ecologic GS that transcends the species specifity hallmark of a GS between the *A. polyphaga* entoorganisms.

Detection of the GS is based on a pattern such as oligonucleotide frequencies being maintained pervasively on a given species genome and accumulating variations as the phylogenetic distance increases. Dinucleotide relative abundance is a demonstrated GS with pervasive characteristics (Karlin and Cardon 1994; Karlin et al. 1994; Prasaot and Vemuri 2007; Prabha and Singh 2014).

The dinucleotide genomic patterns studied here are clearly discernable by the dinucleotide relative abundance as depicted in Fig. 1. For entoorganisms, the cellular and viral profiles show striking similarity, while the negative controls *D. radiodurans* and HIV are clearly differentiated.

A special case is the Marseillevirus, whose genome is hypothesized to derive from several sources because of HGT, nonetheless grouping phylogenetically with APMV (Boyer et al. 2009). While for the Mimivirus linage—the irophage included—a common pattern is clear: Mma shows a homogeneous codon usage, in terms of the GS, clearly different from entoorganisms. An explanation for this may be a recent host range addition into *A. polyphaga* and that not enough time has passed for the entoorganism line of adaptation—the GS—to be adopted. It is pertinent to remember that the entoamoebales studied here also replicate in *Acantamoeba castellanii* as Mma.

The A-T dinucleotide combinations comprise the high-frequency group, an intuitive bias given the high A-T content of these genomes. This implies a grade of compromise because the high A-T content may increase improper binding of regulatory factors such as TATA boxes and polyadenylation signals (Nussinov 1987; Burge et al. 1992).

Correlation analyses statistically support dinucleotide profiles in the sense that all entoorganisms, whether cellular or viral, are very similar generally and particularly given any oligonucleotide combination. The viruses are the most correlated organisms, followed by entobacteria and mitochondria. This higher viral correlation value may mean a longer coevolutive process, even for the virophage case.

It is worth mentioning that the entobacteria Bab is more correlated to viruses than to the *mt*DNA in DiRF, TriRF or TetRF comparisons. Candidatus *Babela massiliensis* is an obligatory intracellular bacterium that interestingly shows common adaptations to NCLDV such as the ankyrin repeats implicated at virus-host interactions (Pagnier et al. 2015). Therefore, the comparison of the entoviral results with the mitochondria and this bacterium is necessary for evaluating the common adaptation of all these entoorganisms to the amoebal host.

It has been reported that similar dinucleotide relative abundance profiles could reflect the similarity of the enzymes engaging in a replication process (Frick and Richardson 2001). As speculation, replication processes of this entoorganism might be performed with the same replication machinery, either that of the amoeba or APMV, or a mixture of both. For APMVs, this would suggest a case where complex viral replication machinery (Raoult et al. 2004) might be recruited by bacteria for their normal genome replication without suffering a viral infection process, overtaking the hypothesis of the melting pot (Moliner et al. 2010). Our hypothesis would inevitably expand the host range definition into a new notion of the viral "accessory-host" range and a "core-host" range, both adding to a phenotypic complementation of a PAN-host range.

The codon usage patterns were clearly discernable as well. The results of CU preferences of entoorganisms were compared for detecting a GS. Codon bias is a direct consequence of dinucleotide bias (Kunec and Osterrieder 2016). CU is related to an efficiency increase in the translation speed (Plotkin and Kudla 2011; Kumar et al. 2016) and to a correlation of the tRNA repertoire (Sharp et al. 1986; Kumar et al. 2016; Duan and Antezana 2003).

Coincidental convergence would be the current scenario for the entoorganisms studied here because of crowding of the GS space (Mrázek 2009). However, this phenomenon is detected only at low order oligonucleotides such as dinucleotides and ruled out at the higher order ones (Mrázek 2009) such as trinucleotides and tetranucleotides.

Therefore, a possible explanation is through low DNA recombination and repairing activity, since reduced genomes have lost sensitive genes related to these pathways (Moran 2002; Bentley and Parkhill 2004). Their absence in virophage genomes and decreased function on APMV (Abergel et al. 2007; Silva et al. 2015) allows for mutations to accumulate. Experiments demonstrate that the most frequent random mutation occurring in cells is C to T (or G to A) because of the deamination of cytosine to form uracil, which is subsequently replicated as thymidine (Glass et al. 2000). Thus, in the absence of DNA repair, genomes tend to become more AT-rich, leading to amelioration (Paz et al. 2006). Naturally, the low GC content of Candidatus *Babela massiliensis* might occur through another process because of replication proteins coded into the bacterial genome (Pagnier et al. 2015).

In conclusion, here we provide evidence of shared genomic signatures between *A. polyphaga* and its entoorganisms. It is not clear how all these organisms interact, but the presence of common GSs reveals a coevolutionary process with two probable scenarios: (1) multiple coincidental evolutionary convergences or (b) an adaptive process to selective pressures caused by the intracellular environment of the host. What seems clear is the current adaptation to the ecologic affinity and dynamics for this unique amoebal intracellular environment. Further work is needed to determine the actual mechanisms driving this coevolution.

The ability of *A. polyphaga* to resist harsh conditions, such as extreme temperatures, pH and osmolarity, suggests its usefulness as a safe harbor for pathogenic bacterial and viral vectors (Greub and Raoult 2004; Moliner et al. 2010; Khan and Siddiqui 2014) possibly facilitating lateral transfer events of virulence and resistance traits among concurrent entoorganisms. Understanding the amoebomics and entoecologies has the utmost animal and human biomedical importance. Further isolation and sequencing of new entoamoebal organisms, either transient or perennial, may reveal the broader hallmark of a probable wide genomic signature associated with each amoeba species.

Acknowledgements The authors thank Dr. Jonathas Abrahão for his valuable comments. This work was supported by a postdoctoral grant from PNPD/CAPES to VSS.

Author Contributions VSS and STF conceived the project. VSS designed the experiments. VSS and PETS performed the experiments. VSS, STF and PETS analyzed the data. VSS and PETS wrote the paper.

Data Availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Compliance with Ethical Standards

Competing interests The authors declare no competing interests.

References

Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T (2003) Informatics for unveiling hidden genome signatures. Genome Res 13(4):693–702

Abergel C, Rudinger-Thirion J, Giegé R, Claverie JM (2007) Virus-encoded aminoacyl-tRNA synthetases: structural and functional characterization of mimivirus TyrRS and MetRS. J Virol 81(22):12406–12417

Arslan D, Legendre M, Seltzer V, Abergel C, Claverie JM (2011) Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. Proc Natl Acad Sci USA 108:17486–17491. https://doi.org/10.1073/pnas.1110889108

Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2017) GenBank. Nucleic Acids Res 42:D32–D37. https://doi.org/10.1093/nar/gkt1030

Bentley SD, Parkhill J (2004) Comparative genomic structure of prokaryotes. Annu Rev Genet 38:771–792

Berdal BP, Mehl R, Meidell NK, Lorentzen-Styr A-M, Scheel O (1996) Field investigations of tularemia in Norway. FEMS Immunol Med Microbiol 13:191–195

Blaisdell BE, Campbell AM, Karlin S (1996) Similarities and dissimilarities of phage genomes. Proc Natl Acad Sci USA 93(12):5854–5859

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 (2009) Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. Proc Natl Acad Sci USA 106(51):21848–21853. https://doi.org/10.1073/pnas.0911354106

Brister JR, Ako-Adjei D, Bao Y, Blinkova O (2015) NCBI viral genomes resource. Nucleic Acids Res 43:D571–577. https://doi.org/10.1093/nar/gku1207

Burge C, Campbell AM, Karlin S (1992) Over- and under-representation of short oligonucleotides in DNA sequences. Proc Natl Acad Sci USA 89(4):1358–1362

Campbell A, Mrázek J, Karlin S (1999) Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proc Natl Acad Sci USA 96(16):9184–9189

Deschavanne P, Giron A, Vilain J, Dufraigne C, Fertil B (2000) Genomic signature is preserved in short DNA fragments. In: Bio-Informatics and Biomedical Engineering, Proceedings. IEEE International Symposium on. IEEE, 2000. pp 161–167. https://doi.org/10.1109/BIBE.2000.889603

Desnues C, La Scola B, Yutin N, Fournous G, Robert C, Azza S, Jardot P, Monteil S, Campocasso A, Koonin EV, Raoult D (2012) Provirophages and transpovirons as the diverse mobilome of giant viruses. Proc Natl Acad Sci USA 109(44):18078–18083. https://doi.org/10.1073/pnas.1208835109

Duan J, Antezana MA (2003) Mammalian mutation pressure, synonymous codon choice, and mRNA degradation. J Mol Evol 57(6):694–701

Filée J (2015) Genomic comparison of closely related Giant Viruses supports an accordion-like model of evolution. Front Microbiol 6:593. https://doi.org/10.3389/fmicb.2015.00593

Foerstner KU, von Mering C, Hooper SD, Bork P (2005) Environments shape the nucleotide composition of genomes. EMBO Rep 6(12):1208–1213

Forterre P (2010) Defining life: the virus viewpoint. Orig Life Evol Biosph 40(2):151–60. https://doi.org/10.1007/s11084-010-9194-1

Frick DN, Richardson CC (2001) DNA primases. Annu Rev Biochem 70:39–80

Gentles AJ, Karlin S (2001 Apr) Genome-scale compositional comparisons in eukaryotes. Genome Res 11(4):540–546

Glass JI, Lefkowitz EJ, Glass JS, Heiner CR, Chen EY, Cassell GH (2000) The complete sequence of the mucosal pathogen Ureaplasma urealyticum. Nature 407(6805):757–762

Greub G, Raoult D (2004 Apr) Microorganisms resistant to free-living amoebae. Clin Microbiol Rev 17(2):413–433

Hammer Ø, Harper DAT, Ryan PD (2001) PAST: Paleontological statistics software package for education and data analysis. Palaeontol Electron 4(1):1–9

Jernigan RW, Baran RH (2002) Pervasive properties of the genomic signature. BMC Genom 3(1):23

Karlin S (1996) Genomic Signature and Bacterial Phylogeny. In: Bacterial genomes: physical structures and analysis, De Bruijn FJ, Lupski JR, Weinstock G (eds.). Chapman & Hall, New York

Karlin S (1998 Oct) Global dinucleotide signatures and analysis of genomic heterogeneity. Curr Opin Microbiol 1(5):598–610

Karlin S, Burge C (1995 Jul) Dinucleotide relative abundance extremes: a genomic signature. Trends Genet 11(7):283–290

Karlin S, Cardon LR (1994) Computational DNA sequence analysis. Annu Rev Microbiol 48:619–654

Karlin S, Ladunga I, Blaisdell BE (1994) Heterogeneity of genomes: measures and values. Proc Natl Acad Sci USA 91(26):12837–12841

Khan NA, Siddiqui R (2014) Predator vs aliens: bacteria interactions with Acanthamoeba. Parasitology 141(7):869–874. https://doi.org/10.1017/S003118201300231X

Kirzhner V, Paz A, Volkovich Z, Nevo E, Korol A (2007) Different clustering of genomes across life using the A-T-C-G and degenerate R-Y alphabets: early and late signaling on genome evolution? J Mol Evol 64(4):448–456

Kozobay-Avraham L, Hosid S, Bolshoy A (2006 May) Involvement of DNA curvature in intergenic regions of prokaryotes. Nucleic Acids Res 5(8):2316–2327

Kumar S, Kumari R, Sharma V (2016) Coevolution mechanisms that adapt viruses to genetic code variations implemented in their hosts. J Genet 95(1):3–12

Kunec D, Osterrieder N (2016) Codon pair bias is a direct consequence of dinucleotide bias. Cell Rep 14(1):55–67. https://doi.org/10.1016/j.celrep.2015.12.011

La Scola B, Audic S, Robert C, Jungang L, de Lamballerie X, Drancourt M, Birtles R, Claverie JM, Raoult D (2003) A giant virus in amoebae. Science 299(5615):2033

La Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, Fournous G, Merchat M, Suzan-Monti M, Forterre P, Koonin E, Raoult D (2008) The virophage as a unique parasite of the giant mimivirus. Nature 455(7209):100–104. https://doi.org/10.1038/nature07218

Legendre M, Santini S, Rico A, Abergel C, Claverie JM (2011) Breaking the 1000-gene barrier for Mimivirus using ultra-deep genome and transcriptome sequencing. Virol J 8:99. https://doi.org/10.1186/1743-422X-8-99

Lerat E, Capy P, Biémont C (2002) The relative abundance of dinucleotides in transposable elements in five species. Mol Biol Evol 19(6):964–967

McDowall KJ, Lin-Chao S, Cohen SN (1994) A + U content rather than a particular nucleotide order determines the specificity of RNase E cleavage. J Biol Chem 269(14):10790–10796

Moliner C, Fournier PE, Raoult D (2010) Genome analysis of microorganisms living in amoebae reveals a melting pot of evolution. FEMS Microbiol Rev 34(3):281–294. https://doi.org/10.1111/j.1574-6976.2010.00209.x

Moran NA (2002) Microbial minimalism: genome reduction in bacterial pathogens. Cell 108(5):583–586

Mrázek J (2009) Phylogenetic signals in DNA composition: limitations and prospects. Mol Biol Evol 26(5):1163–1169. https://doi.org/10.1093/molbev/msp032.

Mrázek J, Karlin S (2007) Distinctive features of large complex virus genomes and proteomes. Proc Natl Acad Sci USA 104(12):5127–5132

Nussinov R (1987) Theoretical molecular biology: prospectives and perspectives. J Theor Biol 125(2):219–235

Pagnier I, Yutin N, Croce O, Makarova KS, Wolf YI, Benamar S, Raoult D, Koonin EV, La Scola B (2015) Babela massiliensis, a representative of a widespread bacterial phylum with unusual adaptations to parasitism in amoebae. Biol Direct 10:13. https://doi.org/10.1186/s13062-015-0043-z

Paz A, Kirzhner V, Nevo E, Korol A (2006 Jan) Coevolution of DNA-interacting proteins and genome “dialect”. Mol Biol Evol 23(1):56–64

Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 12(1):32–42 https://doi.org/10.1038/nrg2899

Prabha R, Singh DP (2014) Analysis of dinucleotide bias and genomic signatures across cyanobacterial genomes. J Adv Biotechnol 3(3):2348–6201

Prasaot BVLS, Vemuri MC (2007) Genome analysis for nucleotide interactions in fully sequenced genomes of selective prokaryotes. J Biosci 23(3):255–263

Pride DT, Wassenaar TM, Ghose C, Blaser MJ (2006) Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. BMC Genom 7:8

Raoult D, Forterre P (2008) Redefining viruses: lessons from Mimivirus. Nat Rev Microbiol 6(4):315–319. https://doi.org/10.1038/nrmicro1858

Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie JM (2004) The 1.2-megabase genome sequence of Mimivirus. Science 306(5700):1344–1350

Ravenhall M, Škunca N, Lassalle F, Dessimoz C (2015) Inferring horizontal gene transfer. PLoS Comput Biol 11(5):e1004095. https://doi.org/10.1371/journal.pcbi.1004095

Robins H, Krasnitz M, Barak H, Levine AJ (2005 Dec) A relative-entropy algorithm for genomic fingerprinting captures host-phage similarities. J Bacteriol 187(24):8370–8374.

Rossum van G, Drake JR, Fred L (2010) Python tutorial. History

Serrano-Solís V, Cocho G, José MV (2016) Genomic signatures in viral sequences by in-frame and out-frame mutual information. J Theor Biol 403:1–9. https://doi.org/10.1016/j.jtbi.2016.05.014

Sharp PM, Li WH (1987) The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15(3):1281–1295

Sharp PM, Tuohy TM, Mosurski KR (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res 14(13):5125–5143

Silva LC, Almeida GM, Assis FL, Albarnaz JD, Boratto PV, Dornas FP, Andrade KR, La Scola B, Kroon EG, da Fonseca FG, Abrahão JS (2015) Modulation of the expression of mimivirus-encoded translation-related genes in response to nutrient availability during Acanthamoeba castellanii infection. Front Microbiol 6:539. https://doi.org/10.3389/fmicb.2015.00539

Sinden RR (1994) DNA structure and function. Academic Press, San Diego

Suzuki H, Sota M, Brown CJ, Top EM (2008) Using Mahalanobis distance to compare genomic signatures between bacterial plasmids and chromosomes. Nucleic Acids Res 36(22):e147. https://doi.org/10.1093/nar/gkn753

van Passel MW, Kuramae EE, Luyf AC, Bart A, Boekhout T (2006) The reach of the genome signature in prokaryotes. BMC Evol Biol 6:84

Xia X, Wei T, Xie Z, Danchin A (2002) Genomic changes in nucleotide and dinucleotide frequencies in Pasteurella multocida cultured under high temperature. Genetics 161(4):1385–1394

Yoosuf N, Yutin N, Colson P et al (2012) Related giant viruses in distant locations and different habitats: acanthamoeba polyphaga moumouvirus represents a third lineage of the mimiviridae that is close to the megavirus lineage. Genome Biol Evolut 4(12):1324–1330. https://doi.org/10.1093/gbe/evs109