G+C content differs in conserved and variable amino acid residues of flaviviruses and other evolutionary groups.

Raphaëlle Klitting, Ernest Gould, Xavier De Lamballerie

Infection, Genetics and Evolution 45 (2016) 332–340

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

G + C content differs in conserved and variable amino acid residues of flaviviruses and other evolutionary groups

Raphaëlle Klitting [a,*], Ernest Andrew Gould [a], Xavier de Lamballerie [a,b]

[a] UMR "Emergence des Pathologies Virales" (EPV: Aix-Marseille University - IRD 190 - Inserm 1207 - EHESP), 27 bd Jean Moulin, 13385 Marseille, France
[b] Institut Hospitalo-Universitaire Méditerranée-Infection, 27 bd Jean Moulin, 13385 Marseille, France

ARTICLE INFO

Article history:
Received 8 July 2016
Received in revised form 1 September 2016
Accepted 19 September 2016
Available online 20 September 2016

Keywords:
Flavivirus
Evolution
G + C content

ABSTRACT

Flaviviruses are small RNA viruses that exhibit genetic and ecological diversity and a wide range of G + C content (GC%). We discovered that, amongst flaviviruses, the GC% of nucleotides encoding conserved amino acid (AA) residues was consistently higher than that of nucleotides encoding variable AAs. This intriguing phenomenon was also identified for a wide range of other viruses, and some non-viral evolutionary groups. Here, we analyse the possible mechanisms underlying this imbalanced nucleotide content (in particular the role of the specific G content and the AA composition in flaviviral genomes) and discuss its evolutionary implications. Our findings suggest that one of the most simple characteristics of the genetic code (i.e., the G or G + C content of codons) is linked with the evolutionary behavior of the corresponding encoded AAs.

1. Introduction

Flaviviruses constitute a genetically and ecologically diverse group of viruses some of which are arthropod-borne human pathogens (arboviruses) including yellow fever virus (YFV), dengue virus (DENV), West Nile virus (WNV), Japanese encephalitis virus (JEV), tick-borne encephalitis virus (TBEV) and recently emerging flaviviruses such as Zika virus. Other flaviviruses are transmitted by different and not necessarily defined routes, some of which exclusively infect arthropods. Flaviviruses are intimately linked with the history of virology, because YFV and DENV were amongst the first human pathogenic viruses of vertebrates to be isolated (Calisher and Gould, 2003). Moreover, the YFV live-attenuated 17D vaccine which was first used as early as 1938 in Brazil (Smith et al., 1938), and modified versions of this vaccine, are now some of the safest, most efficacious and successful viral vaccines ever produced.

Thus, in view of their importance as human pathogens and their genetic diversity their evolution has been extensively studied, taking advantage of the availability of a large corpus of genomic sequences. A wide range of genomic G + C content (GC%) has been reported within the genus *Flavivirus* (Kuno et al., 1998; Gaunt et al., 2001; Jenkins et al., 2001; de Lamballerie et al., 2002; Jenkins and Holmes, 2003; Cook and Holmes, 2006; Lobo et al., 2009; Schubert and Putonti, 2010; Kitchen et al., 2011; Belalov and Lukashev, 2013; Moureau et al., 2015). At an early stage of investigation, a possible association between nucleotide base composition and vector/host specificity was identified (Jenkins et al., 2001; de Lamballerie et al., 2002), and lineage-specific co-evolutionary processes between the viral and host/vector groups were proposed (Lobo et al., 2009). However, little is known about the mechanisms that drive or constrain flaviviral genetic evolution or even the factors that determine their GC% at the molecular level.

Here, in an initial molecular analysis of dengue 2 virus (DENV2), we detected higher GC% in sylvatic than non-sylvatic DENV2 isolates. By looking in detail at the differences between sylvatic and non-sylvatic sequences, we made a remarkable observation: when compared with variable amino acid residues, G + C content was strikingly higher in sequence encoding conserved amino acid residues in both sylvatic and non-sylvatic isolates.

This was our starting point for genetic analyses dedicated to comparison of nucleotide content in conserved and variable amino acid

Abbreviations: AA, Amino Acid; APOIV, *Apoi* virus; CA-Cons, Complete Alignment Conserved; CA-Var, Complete Alignment Variable; CDS, complete coding sequences; CXFV, Culex-borne flavivirus; DENV, dengue virus; GC%, G + C content; ISFV, insect specific flaviviruses; JEV, Japanese encephalitis virus; MBFV, mosquito-borne flaviviruses; MODV, Modoc virus; MMLV, Montana Myotis Leukoencephalitis virus; NKV, viruses with no known vector; ORF, open reading frame; PW, pairwise; RBV, Rio Bravo virus; TBEV, tick-borne encephalitis virus; TBFV, tick-borne flaviviruses; WNV, West Nile virus; YFV, yellow fever virus.

* Corresponding author.
E-mail addresses: raphaelle.klitting@posteo.de (R. Klitting), eag@ceh.ac.uk (E.A. Gould), xavier.de-lamballerie@univ-amu.fr (X. de Lamballerie).

http://dx.doi.org/10.1016/j.meegid.2016.09.017

residues of dengue virus, other flaviviruses and subsequently viruses from other taxonomic groups and eukaryotic and prokaryotic cells.

2. Results

2.1. GC % in genomic coding sequences of representative flaviviruses

It was previously observed (Jenkins et al., 2001; de Lamballerie et al., 2002) that GC% in flaviviruses, within a species, has a limited range. However, this range varied significantly (p-value = 0.01529, Wilcoxon-rank sum test, R software (Team, 2013)) between flavivirus species, from, e.g. 38.4% for Tamana bat virus to 54.3% for TBEV. This is illustrated in Fig. 1, which shows the GC% distribution of representative viruses in the different evolutionary groups of the genus *Flavivirus*. Thus, the range of GC% within a given species is narrow: WNV (median 51.2; min 50.5-max 52.7), YFV (49.7; 48.9–50.1), Culex-borne flavivirus (CXFV: 53.0; 52.8–53.3), TBEV (54.3; 53.8–55.2), *Apoi* virus (APOIV: 48.4; 48.4–48.4), Modoc virus (MODV: 45.2; 45.2–45.2), Montana Myotis Leukoencephalitis virus (MMLV: 43.9; 43.9–43.9) and Rio Bravo virus (RBV: 43.1; 43.1–43.2). However, for the major evolutionary groups there are wide differences in GC%: in the case of viruses with no known vector (NKV) the GC% ranges from 43.1 to 48.4. Similarly, the observed GC% ranges within the tick-borne flaviviruses (TBFV: 54.2; 48.6–55.2), mosquito-borne flaviviruses (MBFV: 46.3; 46.6–53.3) and insect specific flaviviruses (ISFV: 52.8; 46.6–53.3) are much wider than for individual species.

The wide range of GC% for the DENV can be explained by the clinical and historical criteria that prevailed for the delineation of the taxon (Sabin, 1952; Hammon et al., 1960; Calisher et al., 1989). Antigenic and genetic diversity of each serotype is mostly comparable with that of other flavivirus species, and this is compatible with the different GC% ranges observed for serotype 1 (DENV1: 46.3; 46.1–46.6), serotype 2 (DENV2: 45.7; 45.2–47.2), serotype 3 (DENV3: 46.2; 46.0–46.6) and serotype 4 (DENV4: 46.9; 46.7–47.2).

Amongst these dengue serotypes, DENV2 shows the most notable GC% heterogeneity. The DENV2 Asian/American genotype includes strains with the lowest GC% (45.2–46.3%) whereas the sylvatic genotype includes those with the highest GC% (46.4–47.3%).

2.2. Differences in mean GC% between conserved and variable amino acid residues in DENV

In order to investigate the underlying basis for the higher GC% observed in DENV2 sylvatic strains, pairwise (PW) comparison of two complete coding sequences (CDS), one from the sylvatic genotype (AN: EF105385; GC% = 46.5) and one from the Asian/American genotype (AN: EU482750; GC% = 45.5), was performed. Amino acid (AA) sequences were aligned and residues classified as "conserved", *i.e.* identical in both sequences, otherwise as "variable". GC% was estimated in both strains for the nucleotide triplets encoding either conserved or variable AA (Fig. 1 in supplemental data). It was strikingly higher in triplets corresponding to conserved amino acids for both the sylvatic (mean GC%: 46.9 in conserved residues vs 41.3 in variable residues) and the Asian/American strain (46.0 vs 38.2). This observation was repeated ten times using strains from different genotypes with similar results (data not shown).

An alignment of 250 DENV2 CDS was performed using Clustal W implemented in MEGA6 (Tamura et al., 2013) according to AA sequences. Alignment building methods and a comprehensive listing of all the sequences included in the analysis are provided as supplemental data.

We analysed GC% in AA residues conserved in the complete alignment (Complete Alignment Conserved (CA-Cons) residues), and in those not fully conserved (Complete Alignment Variable (CA-Var) residues). Altogether, for DENV2 the GC% of all CA-Cons residues was 53% and that of all CA-Var residues was 43% (See Fig. 2A and Fig. 2 in supplemental data). We defined a "HiGC-Cons" score as the proportion of cases in which the GC% of CA-Cons residues was higher than that of CA-Var residues when all sequences of an alignment were compared two by two (each pair generates two values: one for A vs B and another for B vs A). Importantly, we observed that the GC% of CA-Cons residues was higher than that of CA-Var residues for 100% of pairs compared in our DENV2 dataset (CA-HiGC-Cons score of 100%, Table 1). In Fig. 2A, CA-Cons and CA-Var GC% values were plotted for each of the 250 coding sequences. The difference between the two sets was found to be statistically significant (Wilcoxon-rank sum test, p-value = 2.2 e–16).

Similar analyses were achieved for dengue serotypes 1, 3 and 4 using alignments of 250, 246 and 122 sequences, respectively. For each serotype, the GC% of all CA-Cons residues (47%, 47% and 48%, for serotypes 1, 3 and 4, respectively) was higher than that of all CA-Var residues (43%, 40% and 40%, respectively) and 100% CA-HiGC-Cons scores were observed (Table 1).

Systematic pairwise comparison of all DENV2 coding sequences was then performed, using an in-house software programme. Here, the PW-Cons residues correspond to those residues conserved between the two individual sequences compared. The PW-Var residues correspond to those residues that are not conserved in the two individual sequences compared. Therefore, the numbers of DENV2 PW-Cons residues are globally higher than the numbers of DENV2 CA-Cons residues, and in very closely related sequences the number of DENV2 PW-Cons residues is close to 100%. For 96% of pairwise comparisons, GC% of PW-Cons residues was found to be higher than that of PW-Var residues (PW-HiGC-Cons score of 96%, Table 1). When similar analyses were achieved for dengue serotypes 1, 3 and 4, PW-HiGC-Cons scores of 90, 91 and 80% were obtained, respectively (Table 1).

2.3. Extension of the observation to the genus Flavivirus

A new dataset, extending analysis to the genus *Flavivirus*, was created. It comprised 131 CDS and involved 47 species, 11 tentative species and 10 unclassified species (see supplemental data for a comprehensive listing of these sequences), all recognized as members of the genus *Flavivirus* (ninth ICTV report (2014)). The amino acid sequences of these viruses were aligned as previously described (Moureau et al., 2015) and GC% was estimated for CA-Cons/Var residues and PW-Cons/Var residues, with a distinction between 1st/2nd and 3rd codon positions (Table 1). For the complete open reading frame (ORF) sequences, we observed 100% CA-HiGC-Cons scores (Table 1). The mean GC% of all CA-Cons residues (56%) was significantly higher (p-value = 2.2 e–16, Wilcoxon-rank sum test) than that of all CA-Var residues (49%)(Supplemental data Fig. 2). For convenience, the AA residues corresponding to CA-Cons residues in the flavivirus alignment will be called the "Core AA residues" from now on. Pairwise comparison also revealed 100% PW-HiGC-Cons scores (Table 1). The same analyses performed using an NS5 gene dataset identified CA-HiGC-Cons and PW-HiGC-Cons scores of 99% and 97% respectively.

Within the genus *Flavivirus*, seven additional alignments including sequences selected at various taxonomic levels (species, serotype, genotype) were prepared. At the species level, we observed 100% CA-HiGC-Cons scores and PW-HiGC-Cons scores of 98% and 96% for DENV and TBEV, respectively (Table 1).

One hundred percent CA-HiGC-Cons scores were observed for all DENV serotypes and PW-HiGC-Cons scores of 90%, 96%, 91% and 80% were obtained for serotypes 1, 2, 3 and 4, respectively (Table 1). For DENV2 genotype Asian/American and DENV3, genotype III, CA-HiGC-Cons scores reached 100% for both genotypes and PW-HiGC-Cons scores were 89% and 69% respectively (Table 1).

In summary, within the genus *Flavivirus*, CA-HiGC-Cons scores were systematically 100% regardless of the taxonomic level and PW-HiGC-Cons scores increased with the taxonomic level and genetic diversity.

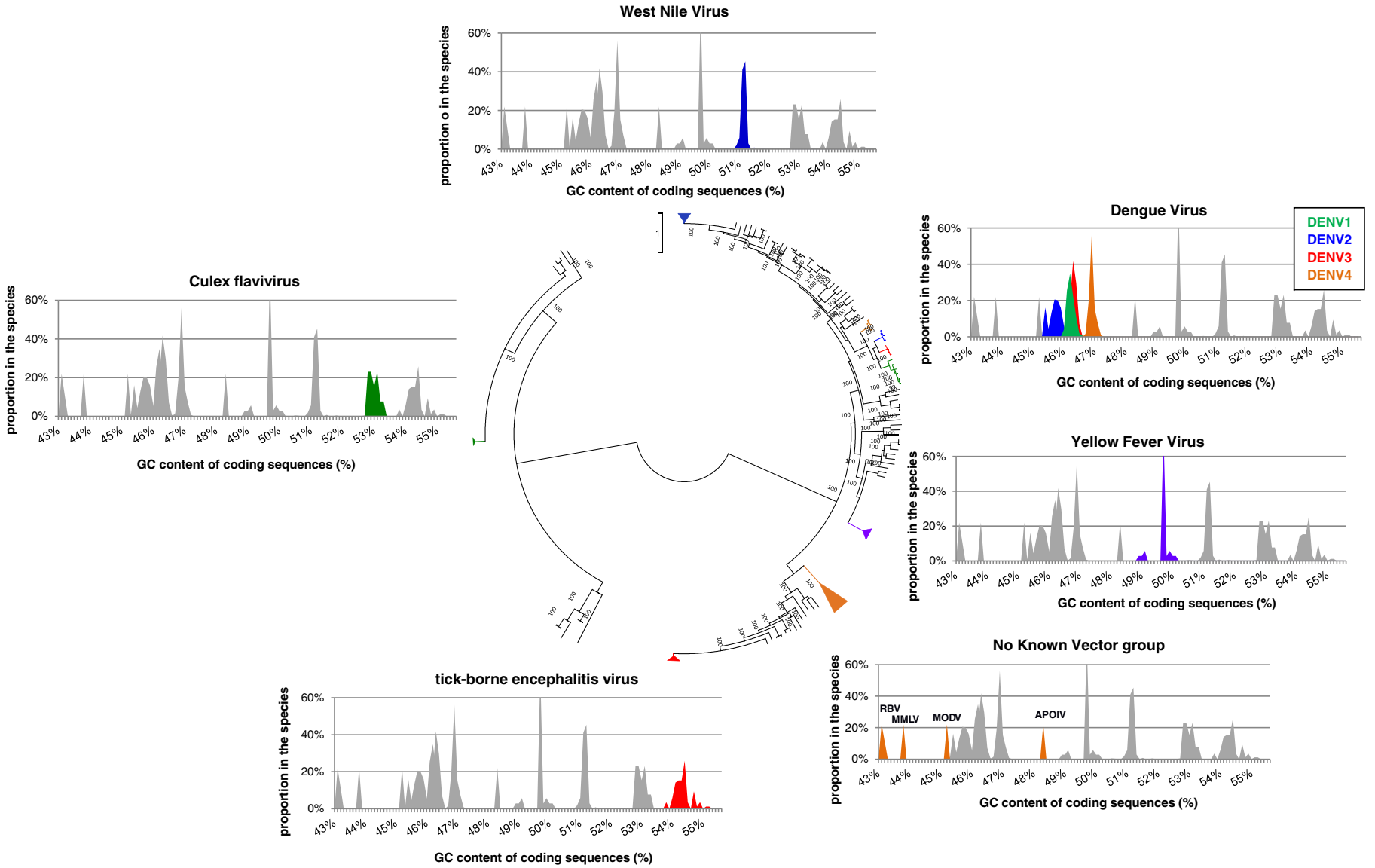

Fig. 1. GC% distribution within the *Flavivirus* genus. The number of sequences corresponding to each genomic GC content value is reported for a set of viral species (WNV, TBEV, DENV1/2/3/4, YFV, CXFV, MODV, MMLV, RBV, APOIV) from the genus *Flavivirus*. The proportion of sequences from the species corresponding to a given GC content value is presented on a separate graph for each species (or vector group in the case of the NKV group). The phylogenetic relationships between all species are illustrated in a phylogenetic tree of the genus *Flavivirus*. The tree includes 131 sequences corresponding to 47 different recognized species in the genus *Flavivirus*. A Maximum Likelihood (1000 bootstrap replicates) tree was prepared using a General Time Reversible Model with invariant sites and a gamma distribution of rates across site.

Table 1
CA- and PW-HiGC-Cons scores in a panel of viral evolutionary groups.

Taxonomy	Group tested	Number of sequences tested	Number of AA positions tested	% of conserved AA positions amongst all sequences tested	% of pairs tested in which nucleotide positions corresponding to conserved amino-acid residues have higher G + C content than those corresponding to variable AA					
					AA conserved amongst all sequences tested			AA conserved in pairwise comparison		
					Codon positions 123	Codon positions 12	Codon position 3	Codon positions 123	Codon positions 12	Codon position 3
Single stranded RNA, positive polarity										
Flaviviridae, *Hepacivirus*	*Hepatitis C virus* (complete ORF)	134	3165	32	**100**	100	70	**100**	100	71
Flaviviridae, *Flavivirus*	*Flavivirus* (ORF)	131	3645	7	**100**	100	76	**99**	100	62
Flaviviridae, *Flavivirus*	*Flavivirus* (polymerase NS5 gene)	131	920	17	**98**	100	83	**97**	95	79
Flaviviridae, *Flavivirus*	*Dengue virus* serotype 1 (ORF)	250	3392	78	**100**	100	44	**90**	90	64
Flaviviridae, *Flavivirus*	*Dengue virus* serotype 2 (ORF)	250	3391	75	**100**	100	87	**96**	96	58
Flaviviridae, *Flavivirus*	*Dengue virus* serotype 2, Asian American genotype (ORF)	250	3391	88	**100**	100	100	**89**	88	57
Flaviviridae, *Flavivirus*	*Dengue virus* serotype 3 (ORF)	246	3390	84	**100**	100	100	**91**	91	78
Flaviviridae, *Flavivirus*	*Dengue virus* serotype 3, genotype III (ORF)	249	3390	90	**100**	100	100	**69**	69	59
Flaviviridae,*Flavivirus*	*Dengue virus* serotype 4 (ORF)	122	3387	85	**100**	100	100	**80**	74	82
Flaviviridae, *Flavivirus*	*Tick-borne encephalitis virus* (ORF)	98	3414	73	**100**	100	97	**96**	97	62
Flaviviridae, *Pestivirus*	*Classical swine fever virus* (ORF)	78	3898	65	**100**	100	100	**90**	82	82
Togaviridae, *Alphavirus*	*Chikungunya virus* (concatened NS-S ORFS)	250	3726	79	**100**	100	58	**76**	74	63
Picornaviridae, *Enterovirus*	*Enterovirus C* (ORF)	246	2231	40	**100**	100	90	**86**	89	53
Leviviridae, *Levivirus*	*Levivirus* (concatenated ORFs)	120	604	38	**100**	100	100	**40**	44	28
Leviviridae, *Levivirus*	*Levivirus* (replicase ORF)	69	548	89	**88**	93	10	**38**	34	28
Retroviridae, *Orthoretrovirinae*, *Lentivirus*	*Human immunodeficiency virus 2* (concatenated gag-pol-env ORFs)	28	2149	54	**86**	100	54	**73**	86	41
Single stranded RNA, negative polarity										
Filoviridae, *Ebolavirus*	*Ebolavirus* (polymerase L gene)	214	2220	66	**100**	100	14	**75**	72	50
Filoviridae, *Ebolavirus*	*Ebolavirus* (concatenated ORFs)	237	4412	59	**100**	100	12	**60**	63	49
Filoviridae, *Ebolavirus*	*Zaire ebolavirus* (polymerase L gene)	183	2220	96	**31**	85	17	**61**	56	49
Filoviridae, *Ebolavirus*	*Zaire ebolavirus* (concatenated ORFs)	206	4839	93	**0**	2	0	**39**	40	36
Orthomyxoviridae, *Influenzavirus A*	*Influenzavirus A*, H1N1 (concatenated ORFs)	250	3899	87	**100**	100	100	**66**	72	45
Orthomyxoviridae, *Influenzavirus B*	*Influenzavirus B* (concatenated ORFs)	250	4271	83	**100**	100	100	**93**	95	73
Arenaviridae, *Mammarenavirus*	*Lassa* mammarenavirus (polymerase L gene)	212	2250	38	**100**	100	62	**60**	36	73
Double stranded RNA										
Reoviridae, *Sedoreovirinae*, *Orbivirus*	*Bluetongue virus* (polymerase gene)	241	1301	69	**100**	100	97	**82**	70	78
Reoviridae, *Rotavirus*	*Rotavirus A* (polymerase gene)	250	1091	42	**99**	100	58	**39**	25	72
Single stranded DNA										
Parvoviridae, *Parvovirinae*, *Erythroparvovirus*	*Human parvovirus B19* (NS1)	77	671	80	**96**	99	6	**63**	46	77
Parvoviridae, *Parvovirinae*, *Erythroparvovirus*	*Human parvovirus B19* (concatenated NS1-VP ORFs)	86	1453	77	**77**	46	74	**70**	57	71
Double stranded DNA										
Adenoviridae, *Mastadenovirus*	*Mastadenovirus* (polymerase gene)	246	1202	49	**7**	3	64	**48**	46	80
Adenoviridae, *Mastadenovirus*	*Human Mastadenovirus D* (concatenated E2B - L3 ORFs)	67	2060	84	**100**	16	100	**98**	90	98

Note: GC contents were calculated on all parts of the sequences according to the conservation of the encoded amino acids either across the whole alignment (CA-Cons/CA-Var residues) or in pairwise comparisons (PW-Cons/PW-Var residues). For each comparison (CA or PW) the proportion of sequences for which the GC% of conserved residues is higher than that of variable residues (HiGC-Cons score) is reported.
Numbers highlighted in bold correspond to those that are cited in the main text. They indicate the % of pairs tested in which nucleotide positions 1, 2 and 3 corresponding to conserved amino-acid residues have higher G + C content than those corresponding to variable AA.

2.4. Extension of the study to other virus evolutionary groups

Amino acid alignments were produced for representatives of species *Hepatitis C virus* (genus *Hepacivirus*), *Classical swine fever vi*rus (genus *Pestivirus*), *Chikungunya virus* (genus *Alphavirus*) and *Enterovirus C* (genus *Picornavirus*). Results were comparable with those reported above for the genus *Flavivirus* (CA-HiGC-Cons scores 100% and PW-HiGC-Cons scores above 95%). However, this was not the case for levivirus bacteriophages: CA-HiGC-Cons scores were commonly high, but PW-HiGC-Cons scores were below 50% when an alignment of concatenated ORFs and the highly conserved replicase ORF were analysed. In the case of retroviruses, a CA-HiGC-Cons score of 86% was obtained and the PW-HiGC-Cons score was 73%.

In a selection of negative-stranded RNA viruses, (genus *Ebolavirus*, species *Influenzavirus A* and *B*, and species *Lassa mammavirus*), 100% CA-HiGC-Cons scores were recorded and PW-HiGC-Cons scores ranged

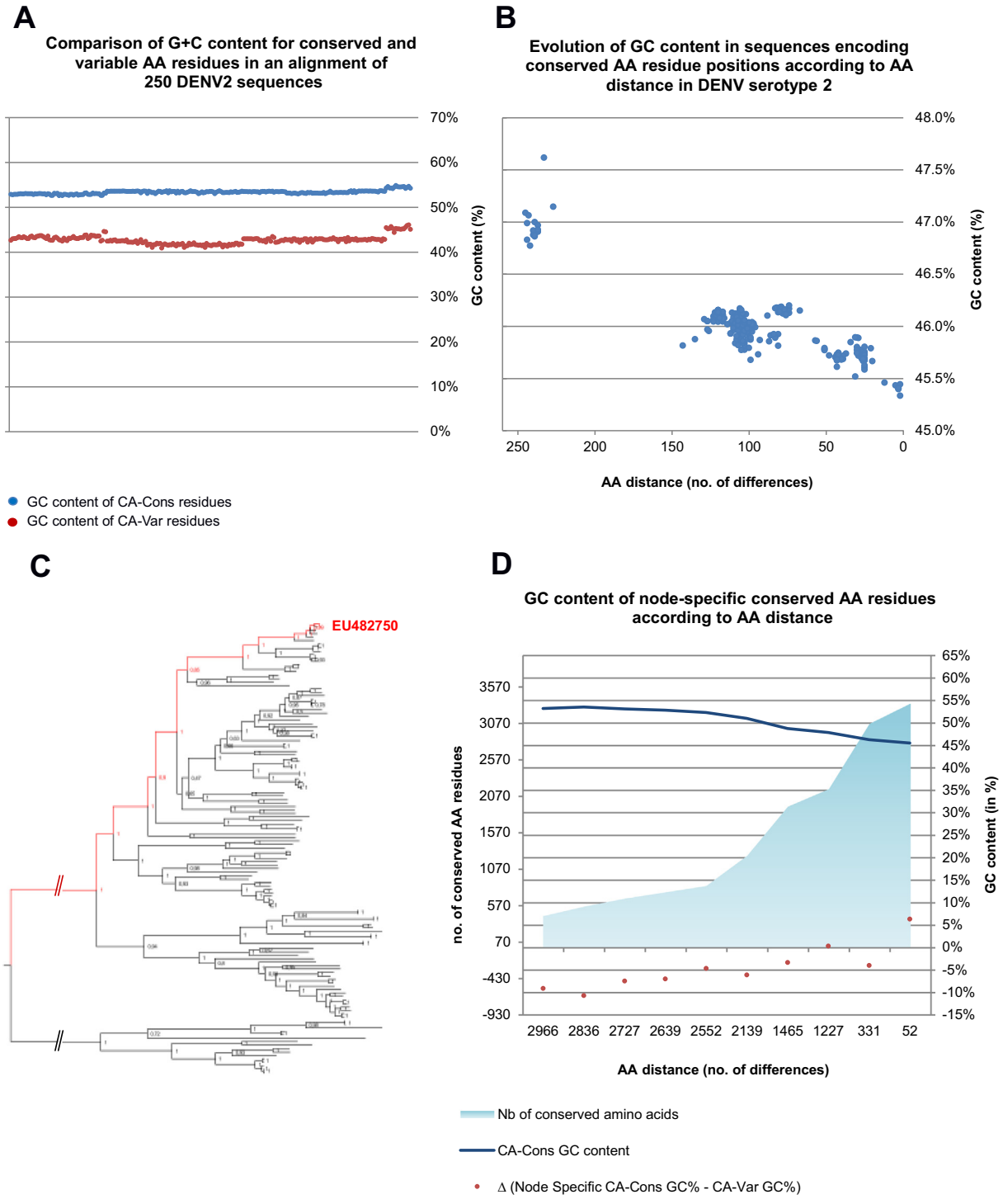

Fig. 2. G + C content of conserved and variable residues and its evolution according to AA distance in dengue virus and the genus *Flavivirus*. A. 250 DENV2 sequences were aligned. CA-Cons (in blue) and CA-Var (in red) GC% values were plotted for each of the 250 CDS. B. GC content of PW-Cons AA residues in sequence EU482750 (in %, on Y axis) is plotted as a function of the AA distance (in no. of differences on X axis) between the two compared sequences. This analysis was carried out on a 250 DENV2 sequences alignment. C. Phylogenetic tree including 131 species from the genus *Flavivirus*. Maximum Likelihood tree (1000 bootstrap replicates) was built using a General Time Reversible model (with gamma distribution and invariant rates amongst sites). The evolutionary pathway beginning on the EU482750 DENV2 sequence is highlighted (in red). D. Values of GC content (in % on Z axis) correspond either to CA-cons AA residues (blue line) or to the difference between the node-specific CA-Cons GC% and the CA-Var GC% for the same node. They are both plotted as a function of amino acid distances (in no. of differences, on X axis). Evolution of the number of CA-Cons AA residues (in no. of AA) is indicated by blue area.

from 60 to 93%. In contrast, species *Zaire Ebolavirus* recorded a strikingly low CA-HiGC-Cons score when alignments of concatenated ORFs and the highly conserved polymerase ORF were analysed (Table 1).

In double-stranded RNA viruses, CA-HiGC-Cons scores were almost always 100%. However, different results were recorded in pairwise comparisons of *Bluetongue virus* (PW-HiGC-Cons score = 82%) and *Rotavirus A* (PW-HiGC-Cons score = 39%).

We extended our analysis to DNA viruses. For the single-stranded *Human Parvovirus B19* species, CA-HiGC-Cons scores were higher than 75% and PW-HiGC-Cons scores were above 60%. For the double-

stranded mastadenoviruses, high HiGC-Cons scores were observed with sequences of species *Human Mastadenovirus D*, but not when analysis was extended to sequences from the complete genus (Table 1). Overall, these results suggest that, within a given evolutionary group, high HiGC-Cons scores can be observed for a large range of genetic distances (e.g. in the case of flaviviruses) or in a more limited range (e.g., at the genus level for *Filoviridae*, at the species level for the species *Human Mastadenovirus D*).

When we investigated alignments within the RpoB bacterial gene, 100% CA-HiGC-Cons scores were observed but PW-HiGC-Cons scores varied with the proportion of conserved AA residues, i.e., from 59% in an alignment of *Mycobacterium tuberculosis* genes (98% AA identity) to 96% in an alignment of different eubacterial genes (32% AA identity).

In alignments from primate genes (Apobec 3G and concatenated ORFs in 54 nuclear gene regions (Perelman et al., 2011), high CA-HiGC-Cons scores were observed, but PW-HiGC-Cons scores were close to equilibrium (50%).

2.5. Relationship between G + C content of conserved AA residues and AA distance

We performed pairwise comparisons between DENV2 sequence EU482750 and the other 249 sequences included in our DENV2 dataset, and plotted the GC% of PW-Cons AA residues as a function of the AA distance. The Spearman's rank-sum correlation test (R software) identified a positive correlation (p-value = 2.2 e–16), with minimum and maximum values of GC% corresponding to minimum and maximum AA distances respectively (Fig. 2B).

A similar analysis was performed using the DENV2 sequence EU482750 and the other 130 sequences included in our flavivirus dataset (supplemental data Fig. 4). Again, a positive correlation (p-value = 7.6e–7) was identified between the GC% of PW-Cons AA residues and AA distance.

When comparing DENV2 sequence EU482750 with increasingly distant flaviviruses, the number of conserved AA residues decreases whilst their corresponding GC% increases. Therefore, the observed parallel decrease of GC% and AA distance may be due to a simple dilution of the high GC% "core" conserved residues along the nodes of the evolutionary tree (from root to terminal branches). To investigate this issue, we generated a flavivirus phylogenetic tree (Fig. 2C) and created an alignment using all the sequences from each node between the root and the DENV2 EU482750 leaf. The AA distance at the node was set as the number of AA residues of the DENV2 sequence EU482750 which were not conserved in the complete alignment generated at the node. The corresponding GC% of CA-Cons AA was calculated from the CA-Cons AA residues of DENV2 sequence EU482750 in this alignment.

During the evolutionary progression from the root of the tree to its terminal branches, at each node, the number of conserved residues increased as the result of new "node-specific" conserved AA residues. If the GC% of these "node-specific" conserved AA residues is higher than that of the variable residues at the node, then a simple dilution of the high GC% "core" conserved residues cannot alone explain the decrease of GC% in conserved residues observed with the decrease of genetic diversity. If this is not the case, the dilution hypothesis would be valid. To test this hypothesis, the GC% of "node-specific" CA-Cons AA residues was calculated from the DENV2 sequence EU482750 as follows:

$$(GC_n * L_n - GC_{dn} * L_{dn})/(L_n - L_{dn})$$

where GC_n and L_n represent GC% and number of conserved AA residues at the considered node, respectively, and GC_{dn} and L_{dn} represent GC% and the number of conserved AA residues at the first deeper node. Fig. 2D shows that, for the vast majority of nodes, the GC% of "node-specific" CA-Cons AA residues is higher than that of CA-Var residues. Therefore, a pure dilution effect cannot account for the observed higher GC% of CA-Cons versus CA-Var residues observed at the most peripheral nodes of the tree.

Table 2
Nucleotide content of conserved amino acids within the DENV2 sequence.

	DENV2 EU482750		
	All residues	S2 CA-Cons residues	Core CA-Cons residues
GC%	45	47	53
A%	33	31	29
T%	21	21	18
G%	25	26	35
C%	20	21	18

Note: Nucleotide content within DENV2 sequence EU482750 were measured: for the complete sequence (All residues), for AA residues fully conserved in the DENV2 alignment (S2 CA-Cons residues) and for AA residues fully conserved in the flavivirus alignment (Core CA-Cons residues).

2.6. HiGC-cons score: which nucleotide frequency accounts for different G + C content in conserved and variable AA residues?

Nucleotide content within DENV2 sequence EU482750 was measured as follows (i) for the complete sequence (all residues), (ii) for AA residues fully conserved in the DENV2 alignment (S2 CA-Cons AA residues), (iii) for AA residues fully conserved in the flavivirus alignment (Core CA-Cons AA residues). For both S2 CA-Cons and Core CA-Cons AA residues, increased G% and, to a lesser extent, decreased A% accounted for the observed increase of GC% in conserved AA residues (Table 2).

We repeated pairwise comparison for the 250 sequences of the DENV2 dataset and considered the four nucleotides separately. The most important variations between PW-Cons and PW-Var AA residues were observed for the G and A content (Table 3). Results were even more clear-cut when the flavivirus dataset was analysed, with respectively 100% and 0% of pairs for which the G% (respectively A%) of PW-Cons AA residues was higher than that of PW-Var residues.

2.7. Extended analysis of G + C content of conserved residues in relation to AA distances

The AA composition of a protein sequence accounts for ca. two thirds of the nucleotide composition of the sequence by which it is encoded. Accordingly, for S2 CA-Cons/Var AA residues and for Core CA-Cons/Var AA residues, (i) nucleotide content was analysed for each of the 20 AA, separately for conserved and variable residues; (ii) the proportion of each of the 20 AA was calculated separately for conserved and variable residues.

Analysis was performed for all four nucleotides (supplemental data). To increase legibility, results below are provided for G content only (Table 4). Each AA type whose G% is higher in Core CA-Cons or S2 CA-Cons parts of the sequence can potentially contribute to the increase in G%. Similarly, any AA type whose G% is intrinsically "high" (>25%) and which is found in higher proportion in Core CA-Cons or S2 CA-

Table 3
PW-Hi[A/T/G/C]-Cons scores in the DENV2 and the flavivirus datasets.

	Frequency	
	DENV2 dataset	flavivirus dataset
A% PW-Cons > A% PW-Var	4%	0%
T% PW-Cons > T% PW-Var	65%	44%
G% PW-Cons > G% PW-Var	96%	100%
C% PW-Cons > C% PW-Var	85%	27%

Note: Indicated scores (in %) represent the proportion of pairs in the DENV2 and flavivirus datasets for which the A, T, G or C% of PW-Cons AA residues is higher than that of PW-Var residues.

Table 4
DENV2 conserved and variable AA residues.

AA	DENV2 EU482750											
	S2 CA-Cons/Var residues		Core CA-Cons/Var residues		Complete sequence	S2 CA-Cons/Var residues		Core CA-Cons/Var residues		S2 CA-Cons/Var residues		Core CA-Cons/Var residues	
	AA proportion in conserved residues	AA proportion in variable residues	AA proportion in conserved residues	AA proportion in variable residues	AA proportion	G%				Participation in G content			
						In conserved residues	In variable residues	In conserved residues	In variable residues	In conserved residues	In variable residues	In conserved residues	In variable residues
A	0,06	0,08	0,03	0,07	0,07	35%	36%	33%	35%	2%	3%	1%	3%
C	0,02	0,00	0,09	0,01	0,02	33%	33%	33%	33%	**1%**	0%	**3%**	0%
D	0,04	0,03	0,07	0,04	0,04	33%	33%	33%	33%	1%	1%	2%	1%
E	0,08	0,06	0,06	0,07	0,07	43%	45%	45%	43%	3%	3%	3%	3%
F	0,03	0,02	0,02	0,03	0,03	0%	0%	0%	0%	0%	0%	0%	0%
G	0,10	0,03	0,16	0,07	0,08	73%	72%	70%	73%	**7%**	2%	**12%**	5%
H	0,02	0,02	0,02	0,02	0,02	0%	0%	0%	0%	0%	0%	0%	0%
I	0,05	0,12	0,01	0,07	0,07	0%	0%	0%	0%	0%	0%	0%	0%
K	0,05	0,10	0,03	0,07	0,06	12%	11%	13%	11%	1%	1%	0%	1%
L	0,11	0,05	0,03	0,10	0,09	15%	13%	11%	15%	2%	1%	0%	1%
M	0,03	0,05	0,02	0,04	0,04	33%	33%	33%	33%	1%	2%	1%	1%
N	0,03	0,05	0,04	0,04	0,04	0%	0%	0%	0%	0%	0%	0%	0%
P	0,05	0,01	0,06	0,04	0,04	3%	3%	4%	2%	0%	0%	0%	0%
Q	0,03	0,03	0,01	0,03	0,03	10%	20%	8%	12%	0%	1%	0%	0%
R	0,06	0,06	0,12	0,05	0,06	39%	43%	36%	41%	2%	3%	4%	2%
S	0,05	0,06	0,04	0,06	0,06	12%	15%	10%	13%	1%	1%	0%	1%
T	0,07	0,09	0,04	0,08	0,08	4%	5%	3%	4%	0%	0%	0%	0%
V	0,05	0,10	0,02	0,07	0,06	48%	43%	50%	46%	2%	4%	1%	3%
W	0,03	0,00	0,09	0,02	0,03	67%	67%	67%	67%	**2%**	0%	**6%**	1%
Y	0,02	0,01	0,04	0,02	0,02	0%	0%	0%	0%	0%	0%	0%	0%

Note: For each AA type, nucleotides were counted in conserved and variable AA residues of the DENV2 EU482750 sequence according to alignments of 131 flavivirus sequences (Core alignment) and also of the 250 DENV2 sequences (S2 alignment). Mean G% of variable AA residues in DENV2 EU482750 according to Core and S2 alignments has been calculated as follows: ([(mean G% in D2 seq) * (no. of AA in D2 seq)] − [(G% in conserved AAs) * (no. of conserved AAs)]) / (no. of variable AAs). The participation to the G% in each AA category (conserved/variable residues in Core/S2 alignment) has been calculated as follows: (mean G% of the considered AA in the category) * (no. of the considered AA in the category) / (total no. of AA in the category).

Cons residues can also contribute to increase their G%. We calculated a participation score that accounts for these two phenomena using the following formula:

$$\frac{(\text{mean G\% in sequence}) * (\text{AA proportion in sequence})}{\text{total no.of AA in sequence}}.$$

In the above formula, "sequence" refers either to Core CA-Cons or S2 CA-Cons residues in DENV2 EU482750, according to whether it refers to genus *Flavivirus* or DENV2 alignments. This score enables evaluation of the participation of each AA type to the G content.

Overall, the AAs that contribute the most to the high G content of conserved residues are Glycine, Tryptophan, and Cysteine (Table 4).

Therefore, the difference in G content between conserved and variable residues is due to AA composition rather than to a "high G" encoding of the considered sequence. For example, Glycine residues participate in the high G content of conserved residues and yet are not necessarily encoded with a higher G% in these residues: the mean G% of Glycine residues is similar in S2 CA-Cons residues and S2 CA-Var residues (73% vs 72%), and even lower in Core CA-Cons residues than in Core CA-Var residues (70% vs 73%).

Overall, we conclude that the high G content in conserved residues of the DENV2 EU482750 sequence is due to the limited number of AA for which the G% is intrinsically high (Glycine, Tryptophan and Cysteine) and which are found in higher proportion amongst conserved residues of the sequence.

3. Discussion

Our study was stimulated by the original concept that flavivirus evolutionary groups are endowed with high genetic variability and specific G + C content (Jenkins et al., 2001; de Lamballerie et al., 2002). An extreme example is provided by the comparison of two sister evolutionary groups: viruses with no known vector are characterised by low G + C content (e.g., RBV, GC% = 43%) whereas tick-borne flaviviruses are characterised by high G + C content (e.g., TBEV, GC% = 54%). This is first explained by strong differences in the AA content of viral polyproteins, that have a higher proportion of AA residues with intrinsically high GC% (namely alanine, glycine and arginine) in TBEV sequence and a lower proportion of AA residues with intrinsically low GC% (isoleucine, lysine and asparagine). According to Sueoka (Sueoka, 1988) and Singer and Hickey (Singer and Hickey, 2000) "the most parsimonious explanation of the observed patterns of amino acid composition in these genomes (would be) an underlying mutational bias that varies between lineages". In addition, the encoding of AAs also contributes to the difference observed, since, as previously reported, the G + C content at the third position of codons faithfully follows that of the complete polyprotein (Bellgard and Gojobori, 1999a,b; de Lamballerie et al., 2002).

In previous studies, relationship between G + C content and AA sequence has been extensively analysed. It has been shown that AAs encoded by GC-rich (and GC-poor) codons are more frequent in sequences with high (and low) G + C content (Singer and Hickey, 2000; Banerjee et al., 2005; Li et al., 2015). Authors have investigated the relationship between G + C content and the frequency of hydrophobic, hydrophilic or amphipatic AAs, the function and the structure of the protein (e.g., membrane proteins, protein secondary structures)(Lobry, 1997; Gu et al., 1998; Banerjee et al., 2005). Studies were performed on a genome wide scale and on genes (Lobry, 1997; Gu et al., 1998; Singer and Hickey, 2000; Banerjee et al., 2005), or on fragments of coding regions (Li et al., 2015), but not at the scale of individual AA residues.

Remarkably, whilst analysing the characteristics of G + C content amongst flaviviruses, we made a founding observation: in an alignment of flaviviral sequences, G + C content is strikingly different in conserved and variable AA residues: higher values are observed in conserved than in variable residues. Within the genus *Flavivirus*, this observation was valid for the complete range of observed genetic distances, and for both complete alignments- and pairwise- conserved residues. For

example, in an alignment of two closely related dengue virus sequences (i.e., including a large majority of conserved residues), or an alignment of two distantly related flaviviral species (i.e., including a large majority of variable residues), G + C content was consistently higher in the conserved residues than in the variable residues. Moreover, in the case of flaviviruses, this difference is directly attributable to the specific G content in the conserved and variable AA residues.

The association between G + C or G content and conservation of AA residues was remarkably strong. In a dataset of 250 DENV2 sequences, the conservation rate in AA residues with a 100% G + C content (~6% of all AAs) was ~85% and in residues with a 100% G + C content for codon positions 1 and 2 (~21% of all AAs), it was ~84%. In other words, without considering the nature of the AAs or their position in the viral polyprotein, the G + C content can predict with 84% certainty the conservation amongst all DENV2 sequences of ~21% of AA residues. Furthermore, when the G content alone is considered, a similar trend is observed in glycine residues, i.e. the only AA with a 100% G content (Supplemental Data).

Indirect support to our findings can be found in previous analyses of prokaryotic genomes. Singer and Hickey (Singer and Hickey, 2000) analysed the relationship between the degree of AA bias and sequence divergence in *Mycobaterium tuberculosis* and *Borrelia burgdorferi*. Interestingly, they reported higher frequency of GARP amino acids (encoded by GC-rich codons) in the conserved genes of *Borrelia burgdorferi*, and higher frequency of FYMINK amino acids (encoded by GC-poor codons) in the variable genes. Several studies have highlighted a relationship between hydropathy of AAs and G + C content (D'Onofrio et al., 1999; Jabbari et al., 2003; Banerjee et al., 2005). Our finding may therefore derive from a higher conservation of hydrophobic AA residues. However, AAs which are at least two times more frequent amongst flavivirus core-conserved residues (i.e., residues shared by all flaviviral species) than amongst variable residues include glycine, arginine, tryptophan and cystein. Glycine, arginine and tryptophan have an intrinsically high G content and are either amphipathic (glycine and tryptophan) or hydrophilic (arginine). Cystein is the only hydrophobic AA in the list and its G content is not intrinsically high. Hence, the link between nucleotide content and conservation of AA is not a consequence of the hydropathy of the latters.

The next issue will be to resolve whether AAs such as glycine, arginine, tryptophan and cysteine are conserved because of their function, their intrinsic properties or their high G content. If G content drives AA conservation, then a high G content should be fostered on the third codon position within conserved AA residues. This is not what is observed: in our alignment of 131 flavivirus sequences, the mean G% of the core conserved AA residues is higher (35%) than that of the variable AA residues (28%), but the mean G3% of the core conserved AA residues (28%) is similar to that of the variable AA residues (29%).

Here we have reported for the first time that the G + C content of conserved residues increases with genetic distance between the flavivirus sequences under comparison. This implies that the most conserved AA residues have the highest GC% values. It is therefore tempting to think that the most conserved residues offer a picture of what the ancestral sequence of the ancient flavivirus progenitor was, and therefore to deduce that this ancestral sequence had a high G + C content, which progressively decreased during evolution. This approach would be in coherence with the previously proposed concept that "each coding system may come to a unique optimal (DNA) base composition" (Sueoka, 1962). However, the fact that the most conserved AA residues have a specifically high G + C content does not necessarily imply that the complete ancestral sequence had a similarly high G + C content. Rather, we can imagine that a pool of residues with a lower G + C content existed in the ancestral sequence and that most non-synonymous substitutions occurred over time within this pool, in accordance with the concept that selective pressure maintains most AA frequencies close to an optimal value (Lobry, 1997). Or, put differently, the fact that the most conserved AA residues have a high G + C content does not imply that the evolutionary mechanisms tend to replace an AA by another with a lower G + C content. Previously, Jenkins and collaborators (Jenkins et al., 2001) used phylogenetic methods to propose estimates of the G + C content of progenitors for the genus *Flavivirus* or its main evolutionary branches. Their results provided no clear evidence that the G + C content at the deepest nodes of the phylogenetic reconstruction was higher than that of extant species. Using similar methods but a more recent and larger dataset of flavivirus sequences, we also failed to identify a trend for G + C content decrease over time; similarly, when we compared the sequences of the first isolates of the West-Nile virus made in the United States of America (1999) with late isolates in the 2010–2012 period, or when we tried to apply to flaviviruses the approach proposed by Gojobori for estimating the G + C content of ancestral mycobacterial sequences (Bellgard and Gojobori, 1999a,b), we found no evidence of such a trend (data not shown). Finally, we tested but did not confirm, the hypothesis that the evolution of the G + C content with respect to genetic distance may be explained by the progressive dilution of an ancestral core of AAs with a high G + C content during the course of virus evolution.

Our founding observation, that G + C content of AAs is higher in conserved than in variable residues applies to the whole genus *Flavivirus* but its scope is likely to be wider. Pairwise and complete alignments were assembled in 11 different virus families and a "HiGC-Cons" score was defined as the proportion of asymmetric pairs for which the G + C content was higher in conserved than in variable residues (Table 1). Altogether, these results indicate that association between G + C content and AA conservation is encountered in a large number of viral and possibly, non-viral- evolutionary groups. High HiGC-Cons scores were observed in RNA viruses whose genomic structure is close to that of flaviviruses, e.g. alphaviruses and enteroviruses (short single-stranded RNA genomes of positive polarity). For more distant viral genera including retroviruses, negative and double-stranded RNA viruses, or DNA viruses, the association between G + C content and AA conservation may be limited to a specific range of genetic distances: our observation is not valid for a number of RNA and DNA datasets that include either highly conserved (sp. *Zaire Ebolavirus*) or poorly conserved sequences (*Mastadenovirus* and *Levivirus*). Further investigations on non-viral evolutionary groups gave us reasonable grounds to believe that our observation may also apply to some prokaryotic sequences but we did not identify relevant datasets from which to draw a robust conclusion in the case of high eukaryotes. In sum, our investigations should be regarded as a first detailed description of a completely new phenomenon that should serve as a starting point for further studies of specific evolutionary groups, based on robust and validated genomic datasets, and associating nucleotide content of codons with the biological properties of the corresponding AAs by both analytical and experimental means.

Finally, our study sheds new light on the characteristics of the genetic code. It is clear that the distribution of codons assigned to AAs is not randomly determined (Taylor and Coates, 1989; Freeland and Hurst, 1998) since AAs with similar structures or properties commonly have adjacent codons or share the first base of their codon (Wong, 1980; Di Giulio, 1989). Our findings would additionally suggest a relationship between simple characteristics of the genetic code (i.e., the G or G + C content of codons) and the evolutionary behavior of the corresponding encoded AA residues. The mechanisms behind this association remain to be determined. Since the most obvious parameter associated with G + C content is the strength of molecular hybridization of polynucleotide duplexes, fundamental driving forces, such as thermodynamic constraints (England, 2013) may be operating.

4. Conclusion

Within the genus *Flavivirus*, G + C content is strikingly higher in conserved than in AA residues, and this observation is valid for a wide range of genetic distances, and for both complete alignments- and

pairwise- conserved residues. We observed that the G + C content of conserved residues increases with genetic distance, but could not provide any evidence of an evolutionary mechanism that would tend to replace an AA by another with a lower G + C content.

The association between G + C content and AA conservation is also encountered in a large number of viral and possibly, non-viral- evolutionary groups. It is specifically strong in RNA viruses whose genomic structure is close to that of flaviviruses (short single-stranded RNA genomes of positive polarity) and may be limited to specific ranges of genetic distances in more distantly related virus groups. In the case of flaviviruses, this association is directly attributable to the increased frequency of AAs with an intrinsically high G content (namely glycine, arginine and tryptophan) amongst conserved residues, and not to the biochemical characteristics of the AAs (e.g. their hydropathy).

Beyond the specific case of flaviviruses, the intriguing discovery of a widely encountered relationship between the nucleotide content and the conservation of AA residues obviously requires detailed analysis in an enlarged variety of evolutionary branches of the tree of life.

Competing interests

The authors have declared that no financial and non-financial competing interests exist.

Acknowledgements

We thank Morgan Seston for software development.

This work was supported by the European Virus Archive goes Global, http://global.europeanvirus-archive.com/ (European Union's Horizon 2020 research and innovation programme under grant agreement no. 653316); and the Agence Nationale de la Recherche, http://www.agence-nationalerecherche.fr/ (grant ANR-14-CE14-0001 RNA VacciCode). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.meegid.2016.09.017.

References

Banerjee, T., Gupta, S.K., et al., 2005. Role of mutational bias and natural selection on genome-wide nucleotide bias in prokaryotic organisms. Biosystems 81 (1), 11–18.
Belalov, I.S., Lukashev, A.N., 2013. Causes and implications of codon usage bias in RNA viruses. PLoS One 8 (2), e56642.
Bellgard, M.I., Gojobori, T., 1999a. Inferring the direction of evolutionary changes of genomic base composition. Trends Genet. 15 (7), 254–256.
Bellgard, M.I., Gojobori, T., 1999b. Significant differences between the G + C content of synonymous codons in orthologous genes and the genomic G + C content. Gene 238 (1), 33–37.
Calisher, C.H., Gould, E.A., 2003. Taxonomy of the virus family Flaviviridae. Adv. Virus Res. 59, 1–19.
Calisher, C.H., Karabatsos, N., et al., 1989. Antigenic relationships between flaviviruses as determined by cross-neutralization tests with polyclonal antisera. J Gen Virol 70 (Pt 1), 37–43.
Cook, S., Holmes, E.C., 2006. A multigene analysis of the phylogenetic relationships among the flaviviruses (family: *Flaviviridae*) and the evolution of vector transmission. Arch. Virol. 151 (2), 309–325.
de Lamballerie, X., Crochu, S., et al., 2002. Genome sequence analysis of Tamana bat virus and its relationship with the genus *Flavivirus*. J. Gen. Virol. 83 (Pt 10), 2443–2454.
Di Giulio, M., 1989. The extension reached by the minimization of the polarity distances during the evolution of the genetic code. J. Mol. Evol. 29 (4), 288–293.
D'Onofrio, G., Jabbari, K., et al., 1999. The correlation of protein hydropathy with the base composition of coding sequences. Gene 238 (1), 3–14.
England, J.L., 2013. Statistical physics of self-replication. J. Chem. Phys. 139 (12), 121923.
Freeland, S.J., Hurst, L.D., 1998. The genetic code is one in a million. J. Mol. Evol. 47 (3), 238–248.
Gaunt, M.W., Sall, A.A., et al., 2001. Phylogenetic relationships of flaviviruses correlate with their epidemiology, disease association and biogeography. J Gen Virol 82 (Pt 8), 1867–1876.
Gu, X., Hewett-Emmett, D., et al., 1998. Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. Genetica 102-103 (1–6), 383–391.
Hammon, W.M., Rudnick, A., et al., 1960. Viruses associated with epidemic hemorrhagic fevers of the Philippines and Thailand. Science 131 (3407), 1102–1103.
Jabbari, K., Cruveiller, S., et al., 2003. The correlation between GC3 and hydropathy in human genes. Gene 317 (1–2), 137–140.
Jenkins, G.M., Holmes, E.C., 2003. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res. 92 (1), 1–7.
Jenkins, G.M., Pagel, M., et al., 2001. Evolution of base composition and codon usage bias in the genus *Flavivirus*. J. Mol. Evol. 52 (4), 383–390.
Kitchen, A., Shackelton, L.A., et al., 2011. Family level phylogenies reveal modes of macroevolution in RNA viruses. Proc. Natl. Acad. Sci. U. S. A. 108 (1), 238–243.
Kuno, G., Chang, G.J., et al., 1998. Phylogeny of the genus *Flavivirus*. J. Virol. 72 (1), 73–83.
Li, J., Zhou, J., et al., 2015. GC-content of synonymous codons profoundly influences amino acid usage. G3 (Bethesda) 5 (10), 2027–2036.
Lobo, F.P., Mota, B.E., et al., 2009. Virus-host coevolution: common patterns of nucleotide motif usage in *Flaviviridae* and their hosts. PLoS One 4 (7), e6282.
Lobry, J.R., 1997. Influence of genomic G + C content on average amino-acid composition of proteins from 59 bacterial species. Gene 205 (1–2), 309–316.
Moureau, G., Cook, S., et al., 2015. New insights into flavivirus evolution, taxonomy and biogeographic history, extended by analysis of canonical and alternative coding sequences. PLoS One 10 (2), e0117849.
Perelman, P., Johnson, W.E., et al., 2011. A molecular phylogeny of living primates. PLoS Genet. 7 (3), e1001342.
Sabin, A.B., 1952. Research on dengue during World War II. J Trop Med Hyg]–>Am.J.Trop. Med. Hyg. 1 (1), 30–50.
Schubert, A.M., Putonti, C., 2010. Evolution of the sequence composition of *Flaviviruses*. Infect. Genet. Evol. 10 (1), 129–136.
Singer, G.A., Hickey, D.A., 2000. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol. Biol. Evol. 17 (11), 1581–1588.
Smith, H.H., Penna, H.A., Paoliello, A., 1938. Yellow fever vaccination with cultured virus (17D) without immune serum. American Journal of Tropical Medicine 18, 437–468.
Sueoka, N., 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. U. S. A. 48, 582–592.
Sueoka, N., 1988. Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. U. S. A. 85 (8), 2653–2657.
Tamura, K., Stecher, G., et al., 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30 (12), 2725–2729.
Taylor, F.J., Coates, D., 1989. The code within the codons. Biosystems 22 (3), 177–187.
Team, R.C., 2013. R: A Language and Environment for Statistical Computing.
Virus Taxonomy, 2014. Ninth Report of the International Committee on Taxonomy of Viruses.
Wong, J.T., 1980. Role of minimization of chemical distances between amino acids in the evolution of the genetic code. Proc. Natl. Acad. Sci. U. S. A. 77 (2), 1083–1086.