Genetic variation and recombination in Aichi virus

Alexander N. Lukashev,1,2† Jan Felix Drexler,1† Ilya S. Belalov,2
Monika Eschbach-Bludau,1 Sigrid Baumgarte3 and Christian Drosten1

1Institute of Virology, University of Bonn Medical Centre, Bonn, Germany
2Chumakov Institute of Poliomyelitis and Viral Encephalitides, Moscow, Russia
3Institute of Hygiene and the Environment, Hamburg, Germany

Aichi virus (AiV), a member of the genus Kobuvirus in the family Picornaviridae, causes gastroenteritis in humans. It was noted that AiV differs from other picornaviruses in its unusually high C content and a very high degree of genome-ordered RNA secondary structures. However, the genetic variability and mutational restrictions on a full-genome scale have not been studied. In addition to the available five complete AiV genomes, we determined here another five complete coding sequences of AiV sampled in Germany, 2004. Distinctive AiV genetic features included a low incidence of recombination along the genome without obvious hotspots or spared regions and very low rates of synonymous and non-synonymous variation, supporting an absence of AiV serotypes. In addition, the absence of recombination between AiV genotypes A and B suggested the existence of reproductive isolation between taxonomic units below the species level. In contrast to most other picornaviruses, AiV genomes strongly avoided the UpA dinucleotide, while there was no obvious selection against the CpG dinucleotide. AiV genomes also appeared to contain a codon usage bias (CUB) apparent as an effective number of codons of 39.5, which was amongst the most extreme among RNA viruses. A set of sequence scrambling algorithms was developed to determine the origin of CUB in AiV. While in most picornaviruses the genomic dinucleotide content contributed significantly to CUB, in AiV its extreme nucleotide content, i.e. 57 % third codon position C, was the main driving force behind the apparent CUB.

INTRODUCTION

Aichi virus (AiV), a member of the family Picornaviridae, was first isolated in 1989 and is the only known member of the genus Kobuvirus infecting humans (Yamashita et al., 1991). AiV has been identified worldwide in patients with gastroenteritis, with detection rates varying from 0.5 to 0.9 % in Europe and up to 4 % in Asia and Africa (Ambert-Balay et al., 2008; Oh et al., 2006; Pham et al., 2007; Yamashita et al., 1993; Yang et al., 2009). Seroprevalence of AiV is close to 60 % in infants below 10 years of age and reaches 90 % later in life (Oh et al., 2006; Ribes et al., 2010; Sdiri-Loulizi et al., 2010).

The AiV genome comprises approximately 8250 nt with a single ORF that encodes three structural proteins (VP0, VP3 and VP1) and seven non-structural proteins, 2A–2C and 3A–3D. The ORF is preceded by an approximately 1200 nt 5′-NTR that contains an internal ribosomal entry site, and ends with a short 3′-NTR. Compared with most other picornaviruses, AiV differs in the absence of an L protease, absence of VP0 cleavage into VP2 and VP4 upon capsid assembly, and a divergent sequence of its 2A protein. AiV has a G+C content of 59 %, considerably higher than other picornaviruses (Yamashita et al., 1998) and contains pure G+C stretches of up to 30 nt. Also, kobuviruses were reported to possess the highest degree of genome-order secondary RNA structure among all picornaviruses (Davis et al., 2008; Simmonds et al., 2004). Primary and secondary RNA structure complicates sequencing of the virus, which could explain why only five full genome sequences have been reported to date, in contrast to over 100 partial sequences. Owing to this limitation, it has not been studied if recombination in AiV is as frequent as reported for most other picornaviruses (reviewed by Lukashev, 2010). In addition, it was known that AiV isolates fall into two phylogenetically distinct subtypes, A and B, but not whether these subtypes correspond to biologically functional taxonomic units.

In this study, we have determined the complete polyprotein sequences of five AiV and characterized genetic variation and recombination on a full-genome scale.

†These authors contributed equally to this work.

The GenBank/EMBL/DDBJ accession numbers for the sequences reported in this paper are GQ927704–GQ927706, GQ927711 and GQ927712

DOI 10.1099/vir.0.040311-0

Journal of General Virology (2012), 93, 1226–1235

Received 5 December 2011
Accepted 24 February 2012

© 2012 SGM
Printed in Great Britain
RESULTS

We determined the complete coding sequence of five AiV field samples (Table 1), four of which belonged to subtype B and one to subtype A. In line with the high cytosine content of AiV genomes described previously (Yamashita et al., 1998), the sequences contained on average 20.1 % A, 21.0 % U, 20.5 % G and 38.7 % C nucleotides. Importantly, most of this bias was contributed by the 54.5 % C content at the third synonymous codon position. This extreme nucleotide content could restrict possible mutational sequence space and limit genetic variation in AiV.

Besides nucleotide preference (base composition bias) as such, restrictions in mutational space could result from the preference or avoidance of certain dinucleotides. RNA viruses generally avoid specific dinucleotides, specifically CpG and UpA (Karlin et al., 1994; Rima & McFerran, 1997). We therefore compared the frequencies of these dinucleotides in the coding genome regions of AiV and other picornaviruses. To account for dinucleotide biases actually caused by the base composition bias, we calculated the ratio of observed dinucleotide frequencies to expected frequencies (e.g. \( f \text{C}_2 \times f \text{G}_3 \)) for CpG23 dinucleotide), i.e. the relative dinucleotide bias (RDB). Any deviation of the RDB from 1 would indicate specific avoidance or preference of a dinucleotide. The first dinucleotide of a codon depends mainly on the encoded amino acid, so we relied on dinucleotide frequencies at positions 2–3 of codons and at positions 3–1 between codons (the latter not shown).

As expected from previously published reports (Karlin et al., 1994; Rima & McFerran, 1997), the CpG dinucleotide was generally avoided in picornaviruses (Fig. 1a).

However, in contrast to the low CpG content observed in hepatitis A virus (HAV), parechoviruses, rhinovirus A and sapelovirus (SapV) (RDB below 0.57, Fig. 1a), AiV and foot-and-mouth disease virus (FMDV) exhibited strikingly higher RDB (0.78 and 0.81, respectively), indicating a much weaker bias against the CpG dinucleotide. In order to confirm this result, the arginine index (AI) was analysed as an additional measure of a CpG bias (Greenbaum et al., 2008). Arginine is encoded by six codons, CGN and AGR. In case of a bias against CpG, the index calculated as a ratio of CGN to all arginine codons should fall below 0.67 due to a preference towards AGR codons. The AI among picornaviruses generally corresponded to the CpG frequency (Fig. 1a). AiV had the highest AI of 0.92, indicating a preference towards CGN arginine codons, while FMDV had almost a neutral AI of 0.60, and other picornaviruses exhibited much lower values. An especially low AI of 0.08 was found in HAV, indicating a strong avoidance of the CpG dinucleotide.

Another dinucleotide commonly evaded to various degree by many viruses is UpA (Rima & McFerran, 1997). In contrast to the high CpG frequency in AiV and FMDV, these viruses had the lowest UpA content (RDB of 0.21 and 0.27, respectively), implying a strong avoidance of this dinucleotide. For comparison, other picornaviruses, the RDB for the UpA dinucleotide ranged from 0.49 to 0.88.

Nucleotide and dinucleotide usage has been shown to affect the preference of synonymous codons for a number of different viruses (D’Andrea et al., 2011; Jenkins & Holmes, 2003). We therefore analysed if these factors had any influence on codon usage in AiV. Codon usage bias (CUB) is commonly expressed as an effective number of codons (ENC), which can theoretically range from 61 (all codons used equally) to 20 (only one codon used per amino acid). Most RNA viruses have a low to moderate apparent CUB, evident as an ENC ranging from 39 to 58. Only two RNA viruses were shown previously to have an ENC below 40, HAV and rubella virus (Jenkins & Holmes, 2003). Similar to these viruses, AiV had an ENC of 39.5, one of the lowest among RNA viruses and picornaviruses in particular (Fig. 1b).

**Table 1. AiV sequences used for analysis**

<table>
<thead>
<tr>
<th>GenBank accession no.</th>
<th>Genotype</th>
<th>Year of isolation</th>
<th>Country of isolation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>DQ028632</td>
<td>B</td>
<td>2001</td>
<td>Brazil</td>
<td>Oh et al. (2006)</td>
</tr>
<tr>
<td>AY747174</td>
<td>A</td>
<td>2003</td>
<td>Germany</td>
<td>Oh et al. (2006)</td>
</tr>
<tr>
<td>AB040749*</td>
<td>A</td>
<td>1988</td>
<td>Japan</td>
<td>Sasaki et al. (2001)</td>
</tr>
<tr>
<td>FJ890523</td>
<td>B</td>
<td>2008</td>
<td>China</td>
<td>Yang et al. (2009)</td>
</tr>
<tr>
<td>DJ/V12169/2004†</td>
<td>B</td>
<td>2004</td>
<td>Germany</td>
<td>Drexler et al. (2011)</td>
</tr>
<tr>
<td>DJ/V12359/2004†</td>
<td>B</td>
<td>2004</td>
<td>Germany</td>
<td>Drexler et al. (2011)</td>
</tr>
<tr>
<td>DJ/V12321/2004†</td>
<td>B</td>
<td>2004</td>
<td>Germany</td>
<td>Drexler et al. (2011)</td>
</tr>
<tr>
<td>DJ/V12287/2004</td>
<td>A</td>
<td>2004</td>
<td>Germany</td>
<td>Drexler et al. (2011)</td>
</tr>
<tr>
<td>DJ/V12244/2004†</td>
<td>B</td>
<td>2004</td>
<td>Germany</td>
<td>Drexler et al. (2011)</td>
</tr>
</tbody>
</table>

*These sequences are two versions of the sequence of the same strain A846/88; however, they differ by 17 substitutions in the coding genome region.

†These samples differed by only 16–21 nt substitutions in the coding genome region.

Nature of CUB

The two major forces influencing viral synonymous variation are (i) the mutation pressure, which includes factors acting on the RNA sequence independent of...
translation (e.g. nucleotide and dinucleotide content), and (ii) the translational pressure, which reflects a preference towards certain synonymous codons to match the cellular tRNA pool and thus optimize translation (Bulmer, 1987; Plotkin & Kudla, 2011). Mathematical correction for total G+C content termed Nc’ has been suggested (Novembre, 2002); however, it cannot account for distinct base frequencies in RNA viruses and has theoretical drawbacks (Fuglsang, 2006). Therefore, we relied on a universal way to distinguish between the sources of pressure by analysing sequences scrambled to preserve the overall amino acid sequence and (di)nucleotide composition, while shuffling synonymous codons (Table 2).

Randomization based on the combined G+C content at the third codon position produced ENC values notably higher than the original ENC (Fig. 1b); therefore, overall G+C content was not sufficient to explain all apparent CUB in picornaviruses. In AiV and HRV89, a shuffling of synonymous codons with preserved third position nucleotide frequencies (N3) produced ENC values very close to ENC of the original sequence, and dinucleotide-based correction could not further reduce this corrected ENC. Therefore, in these viruses the pressure to maintain the nucleotide content was nearly the sole factor that defined the CUB. In other picornaviruses, ENC values most similar to those of the original genomes could be obtained only in sequences shuffled by the NN23/31 algorithm (Table 2). Therefore, in these viruses dinucleotide content was a significant component of the mutational pressure. In most cases, with the exception of poliovirus, ENC corrected for dinucleotide content differed from the original ENC by less
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Explanation</th>
<th>Algorithm and examples</th>
</tr>
</thead>
</table>
| ENC     | The effective number of codons in the original sequence (Wright, 1990). ENC of 20 corresponds to extreme bias (only one synonymous codon per aa), ENC of 61 indicates absence of bias | 1. Number of synonymous codons for each degenerate amino acid is counted 2. Codon usage uniformity is calculated for each amino acid from squared codon frequencies using equation originally used to evaluate homozygocity of an allele (Wright, 1990)  
   \[ \bar{F} = \frac{n \sum_{i=1}^{k} p_i^2 - 1}{n - 1} \]  
   \( \bar{F} \) is ‘homozygocity’ of a given amino acid, \( n \) is count of this amino acid in the sequence, \( p \) is frequency of a codon and \( k \) is the possible number of codons for this amino acid 3. Inverse of codon usage uniformities for each amino acid are added up to produce ENC:  
   \[ \bar{N}_c = 2 + \left( \frac{9}{\bar{F}_2} \right) + \left( \frac{1}{\bar{F}_3} \right) + \left( \frac{5}{\bar{F}_4} \right) + \left( \frac{3}{\bar{F}_6} \right) \]  
   where \( \bar{N}_c \) is ENC and \( \bar{F}_i \) is mean homozygocity for i-class of degeneracy |
| (G+C)$_3$ | Randomization of synonymous codons (3rd positions only) based on combined (G+C) nucleotide frequencies at the third codon position | 1. Calculate combined frequency of (G+C) nucleotides at the 3rd codon position 2. Randomize syn. 3rd codon positions based on (G+C) usage in the original sequence |
| N$_3$    | Shuffling of synonymous codons (3rd positions only), while preserving amino acid sequence; sixfold degenerate codons are treated as independent two- and fourfold degenerate. | Swap compatible dinucleotides among synonymous codons in the sequence |
| NN$_{23;31}$ | Shuffling trinucleotides between codon positions 2–3–1, while preserving amino acid sequence; sixfold degenerate codons are treated as independent two- and fourfold degenerate. As a result, distinct dinucleotide content at codon positions 2–3 and 3–1 is identical to the original sequence. | Swap compatible triplets in codon positions 2–3–1 among synonymous codons in the sequence |
than one, implying very low, if any role of translational pressure in these viruses.

The strong nucleotide and dinucleotide bias in AiV could also restrict the degree of synonymous variation along the genome. Therefore, we studied synonymous variation rates per site in AiV compared to two picornaviruses known for their relatively high (human enterovirus species A; HEV-A) and low (HAV) genetic variability. In all three viruses, synonymous variation rates were almost constant over the genome (Fig. 2a). However, synonymous variation in AiV was about two times lower than in HAV and four times lower than in HEV-A. To confirm that this output did not result from a limited dataset for AiV, we repeated the analysis with only two to four randomly selected sequences for each virus and obtained similar results (data not shown). The differences in synonymous variation rates were not equally reflected in non-synonymous variation rates, which were much lower in both HAV and AiV compared with HEV-A (Fig. 2b). Importantly, in HAV and AiV, non-synonymous variation was relatively constant over the genome. This was in contrast to HEV-A, which showed about five times higher non-synonymous rate in the capsid-encoding genome region than in the remaining part of the genome encoding the non-structural proteins. HEV-A species includes multiple serotypes, which results in increased apparent variation at genome regions encoding capsid proteins. On the contrary, HAV has little antigenic variability and is considered a single serotype (Stapleton & Lemon, 1987). The profile of non-synonymous variation in AiV was therefore compatible with presumed absence of serotypes.

Recombination

The recombination profile of AiV was compared with that of HAV and HEV-A, with HAV showing lower rates of recombination and HEV-A representing typical, frequently recombining picornaviruses (Belalov et al., 2011; Lukashev, 2005; Simmonds & Welch, 2006). A phylogenetic compatibility matrix indicated a low rate of recombination throughout the AiV genome (Fig. 3a). This result was similar to HAV (Fig. 3b), but in clear contrast to HEV-A (Fig. 3c) and most other picornaviruses (data not shown). Picornaviruses generally exhibit a high apparent frequency of recombination over the non-structural part of the genome and almost no signs of recombination within regions encoding the major capsid proteins (Heath et al., 2006; Lukashev et al., 2005; Simmonds & Welch, 2006). A typical bootscan graph for AiV genotype B (Fig. 3d) indicated a mosaic genome structure relative to other AiV genotype B sequences. Analysis of AiV genotype A complete sequences showed no evidence of recombination, probably due to the low number of unique sequences available so far (data not shown). The overall recombination frequency as indicated by compatibility matrices, recombination breakpoints could be found in all genome regions of AiV, similar to HAV (Fig. 3e) (Belalov et al., 2011). This recombination pattern was distinct from other picornaviruses such as HEV-A (Fig. 3f), where breakpoints most frequently map to the borders of the capsid-encoding genome region, often to the non-structural genome region, and rarely (never between different serotypes) within the structural genome part (Lukashev, 2010; Simmonds, 2006). Importantly, no phylogenetic evidence of recombination was detected between AiV genotypes A and B in any part of the genome.

Absence of recombination is a feature of distinct picornavirus taxonomic units, such as enterovirus species (Lukashev, 2005) or HAV genotypes (Belalov et al., 2011). However, these recognized taxonomic units are primarily distinguished by sequence distance criteria. The sampling of AiV complete genomic sequences was too small to test such criterion, therefore all available partial sequences for VP1 (n=33) and 3CD (n=103) genome regions commonly used for typing were used. In VP1 genome region, viruses of the same subtype differed by no more than 8.1% of nucleotide sequence, while different subtypes had at least 11.6% different positions (Fig. 4), therefore the subtypes...
were perfectly distinguishable. In a more representative 3CD dataset, sequence distance within a subtype was up to 8.7%, while intertypic distances could be as low as 7%. This resulted in a ‘grey zone’ that did not allow unambiguous identification of subtypes. Use of amino acid sequence distances also did not provide reliable demarcation of AiV subtypes in the 3CD genome region (data not shown).

Fig. 3. Phylogenetic compatibility matrix of the coding genome region of (a) AiV, (b) HAV and (c) HEV-A. Axes indicate genome positions, colour shows frequencies of phylogeny violations per clade according to the colour scheme. Window size was 250 nt, step size 50 nt and bootstrap cut-off was 70%. Bootscan graphs of (d) AiV strain Goiania/GO/03/01/Brazil (DQ028632), (e) HAV isolate LY6 (AF485328) and (f) HEV-A enterovirus 90 (AY773285). The x-axis indicates genome position, y-axis shows per cent of permuted trees that supported grouping with the query sequence. Window size was 600 nt, step size was 500 nt. The dotted line shows the 70% bootstrap reliability limit.
DISCUSSION

Our analysis of complete AiV genomes exposed distinctive features of AiV in comparison to other picornaviruses. The first of these was related to the overall genomic nucleotide content of AiV, for which a high C nucleotide content of about 39% was described previously (Yamashita et al., 1998). We found that this was mostly provided by the third-position C content in codons (54.5% in our sequences). The second feature was related to the genomic dinucleotide usage pattern of AiV, which was also different from most other picornaviruses in the weak avoidance of CpG dinucleotides. The CpG dinucleotide is generally under-represented in both vertebrate genomic DNA (Lander et al., 2001) and in mammalian RNA viruses (Rima & McFerran, 1997). In vertebrate DNA, this can be explained by methylation of cytosine residues in a CpG sequence context (Lister et al., 2009) and subsequent deamination of C into T, which facilitates recognition of non-self DNA through Toll-like receptor 9 activation by CpG-rich DNA (Krieg et al., 1995). In RNA viruses, however, the reasons for CpG avoidance remain unclear. It has been suggested that CpG in RNA viruses may be under-represented due to editing by the human APOBEC3 cytosine deaminase (Harris & Lidament, 2004). Another hypothesis implied recognition of CpG in viral RNA by yet unidentified innate immunity sensors (Greenbaum et al., 2009). In our analyses, AiV and other picornaviruses demonstrated an extensive variation in their CpG frequencies, with AiV having the second highest CpG content after FMDV. The genetic similarity of AiV and FMDV in their low CpG bias was extended by a remarkable avoidance of the UpA dinucleotide, which exceeded that of other picornaviruses. Under-representation of UpA dinucleotides in RNA viruses could be explained by recognition of this sequence motif by RNase L, an antiviral host RNase activated by dsRNA (Bisbal & Silverman, 2007; Player & Torrence, 1998). Some viruses have mechanisms to inhibit RNase L, which, amongst picornaviruses, has so far only been demonstrated for poliovirus (Han et al., 2007). The low UpA content in AiV and FMDV might imply that these viruses lack the capacity to inhibit RNase L and therefore have a selection pressure against UpA acting on the RNA sequence. Picornaviruses have developed diverse ways to evade innate immunity via security proteins (Agol & Gmyl, 2010), and it can be speculated that AiV and FMDV have the capacity to suppress hypothetical pathways that recognize CpG motifs, but not RNase L.

The third distinctive feature of AiV was a strong CUB. It is commonly accepted that synonymous genetic variation is subject to multiple constraints that vary among organisms with different levels of complexity (Plotkin & Kudla, 2011). The reasons and implications of restrained synonymous variation in viruses remain poorly understood. In particular, (di)nucleotide composition has been implicated as the major driver of CUB in RNA viruses (Jenkins & Holmes, 2003). Unfortunately, the available approaches to analyse disparity in codon usage were created for DNA organisms which have a G content that equals their C content, and could therefore not account for the high C content in AiV. In addition, due to the high C content specifically at the third codon position, AiV exhibited notable disparity in dinucleotide frequencies in codon positions 2–3 and 3–1, which was also ignored by common analysis methods. A set of sequence scrambling methods was therefore developed to discern distinct evolutionary pressures on the RNA sequence in AiV and other picornaviruses to explore the origin of CUB in AiV and other picornaviruses.

The results indicated that the CUB in all picornaviruses could be almost entirely explained by nucleotide or dinucleotide bias, indicating absence of significant translation-based selective pressure, as also suggested for other
RNA viruses (D’Andrea et al., 2011; Jenkins & Holmes, 2003; Liu et al., 2011; Woo et al., 2007), but not in higher organisms that feature a notable translational bias (Chamary et al., 2006; Hershberg & Petrov, 2008). Interestingly, our results demonstrated that the level of mutational pressure among picornaviruses could be different. In AiV, a high C content alone explained its dinucleotide bias, high AI and a strong CUB, suggesting mutation pressure different from those affecting most other picornaviruses, which displayed evidence of an additional dinucleotide pressure on synonymous codon preference.

The fourth distinctive feature was the low synonymous variation observed in AiV. Within the family Picornaviridae, different degrees of synonymous variation exist, ranging from 45% in HAV to 80% in enteroviruses. Our analysis indicated that approximately 30% synonymous diversity in AiV was the lowest described so far among picornaviruses. While the latter value might grow upon analysis of further AiV complete genomic sequences, the increase would not likely be huge. For example, comparable variation graphs could be obtained for AiV, HAV and HEV-A with complete datasets and with only two sequences (data not shown). In addition, despite a small number of unique sequences, AiV genomes used in this study provide a global coverage for over 20 years, a significant time frame for an RNA virus. Notably, in addition to an extreme nucleotide bias, another explanation for low synonymous variability in AiV could be the presence of a strong secondary RNA structure predicted previously (Davis et al., 2008; Simmonds et al., 2004).

As for the non-synonymous variation, AiV showed a comparable degree of variability in structural and non-structural genome regions, which was in contrast to other common picornaviruses, except for HAV. In most picornaviruses, e.g. enteroviruses, the difference between multiple (sero)types in the capsid-encoding region, especially in the parts encoding antigenic determinants, provides very high apparent non-synonymous variation. Non-structural proteins of these picornaviruses are much more conserved. So far, only HAV, represented by a sole serotype (Stapleton & Lemon, 1987), stood out from other picornaviruses by having a low non-synonymous variation over the genome. The similar variability profile in AiV in addition to the common occurrence of recombination events in the structural protein-encoding genome region, usually seen in picornaviruses only between strains of the same (sero)type (Drexler et al., 2010; Huang et al., 2008), might be compatible with an absence of serotypes in AiV, which has been assumed previously (Reuter et al., 2011), but never proven experimentally.

The absence of serotypes, however, does not exclude the existence of multiple AiV taxonomic units, possibly species. One of the features of a picornavirus species is its incompatibility of viral proteins, suggested by the absence of natural recombination between different species (Lukashev, 2010). We did not observe recombination between genotypes A and B of AiV despite co-circulation of the two genotypes at least within the clinical cohort the samples analysed here originated from (Drexler et al., 2011). This might indicate reproductive isolation between AiV genotypes as a sixth and final distinctive AiV feature, in analogy to HAV genotypes (Belalov et al., 2011). The utility of distance-based criterion in VP1 genome region for distinguishing AiV subtypes requires further confirmation using a much bigger sampling; however, we hypothesize that in absence of recombination the subtypes would inevitably diverge into distinct species. At the current stage, only an experimental study of compatibility between AiV types A and B can support formal taxonomic conclusions.

**METHODS**

AiV-positive faecal specimens were sampled in Bremen, Northern Germany in 2004 from patients with gastroenteritis (Table 1). To determine the complete genome sequences, conserved elements of the 5’-NTR, VP1 and 3D genomic regions were amplified by one-step RT-PCR (Drexler et al., 2011; Yamashita et al., 2000) and sequenced on an ABI 3100 automated sequencer (Applied Biosystems). cDNA was generated using the SuperScriptIII kit (Invitrogen) and long range PCR-based amplification of 4 kb fragments was done using the Expand High Fidelity Plus kit (Roche). Where this was unsuccessful, further PCR assays were designed to amplify islets spanning approximately 200 bp along the whole genome. Smaller fragments of up to 1 kb were then successfully amplified and sequenced from five of ten initially positive samples. The 3’ end of the genome was completed using the GeneRacer kit (Invitrogen). The 5’ terminal 64 nt of the genome were not sequenced because they were not relevant for our analyses.

The obtained sequences (GenBank accession numbers GQ927704–GQ927706, GQ927711 and GQ927712) were aligned with the five complete genome sequences available at GenBank (Table 1). Sequence handling was done using BioEdit v7.0.5.2 (Hall, 1999). The polyprotein-encoding sequences were aligned with CLUSTAL W (Thompson et al., 1994). Two reference alignments were created for comparison, using complete coding sequences of all HAV field isolates (complete list available in Belalov et al., 2011) and all available HEV-A viruses, excluding highly abundant enterovirus 71 sequences to avoid a sampling bias towards a single serotype. In both reference alignments, sequences sharing more than 98% nt identity were omitted, leaving 33 and 21 unique genomes for the HEV-A and HAV-based alignments, respectively (alignments available upon request). Alignments of all available AiV sequences in 3CD (n=103) and VP1 (n=33) genome regions (genome positions nt 6294–6752 and 3005–3838 according to GenBank #NC_001918, respectively) were used for distance matrices that were calculated using MEGA 5.0 (Tamura et al., 2011). In addition, eight datasets representing diverse and well-characterized species among the family Picornaviridae were analysed for dinucleotide bias. To this end, all available complete coding sequences were extracted, and genomes that differed by less than 5% of nucleotide sequence were discarded. Single picornavirus sequences representing six Picornaviridae genera used in the analyses of CUB were poliovirus type 3 (PV3, GenBank accession number X04468), enterovirus 71 (EV71, AF119795), human parechovirus type 1 (HPeV1, S45208), human rhinovirus 89 (HRV89, M16248), saffold virus type 1 (SafV1, EF165067), HAV (AF119795), foot-and-mouth disease virus (FMDV, NC_002554) and SapV (NC_003987). Analysis of RNA sequence polymorphism was done using DNAsp v5 (Librado & Rozas, 2009). Phylogenetic compatibility matrices were created with Simmonics 1.6 (Simmonds & Smith, 2004).


1999). The ENC was calculated with the chips module in the EMBOSS package (http://emboss.sourceforge.net/). Python scripts were developed for dinucletide content analysis (dint-stat), as well as sequence randomization and scrambling (cub-stat). A general rationale for analysing the impact of (di)nucleotide bias on ENC is randomizing or shuffling of synonymous codons in the original genome sequence using a constraint under study. If generated sequences have ENC close to the original sequence, the constraint (e.g. the nucleotide content) fully explains the CUB, i.e. at this nucleotide content ENC could not possibly be higher than it is. On the other hand, any differences between the ENC in the original and scrambled sequences shows the extent of CUB that could not be explained by the studied factor. Sequence randomization and scrambling algorithms are summarized in Table 2, and the scripts are available at [http://www.poliomielit.ru/images/doc/enc-stat.rar]. Bootscanning was performed with SimPlot v.3.5.1 (Lole et al., 1999).

ACKNOWLEDGEMENTS

This study was funded by the EU FP7 projects EMPERIE (contract number 223498) and EVA (contract number 228292). We are grateful to H. Hilbig-Hanl, H. Kocken, U. Krause, G. Mueseler, and E. Voß from the Institute of Hygiene and the Environment for excellent technical assistance.

REFERENCES


