Lineage diversification and recombination in type-4 human astroviruses

Vito Martella, Maria C. Medici, Valentina Terio, Cristiana Catella, Giancarlo Bozzo, Fabio Tummolo, Adriana Calderaro, Floriana Bonura, Marzia Di Franco, Krisztían Bánya, Giovanni M. Giammanco, Simona De Grazia

**Abstract**

Human astroviruses (HAstVs) are important enteric pathogens and can be classified genetically and antigenically into eight types. During surveillance of HAstVs in Italy, type-4 HAstVs were detected only sporadically and found to cluster into two distinct genetic groups. Upon sequence analysis of the 3' end of the polymerase gene (ORF1b) and of the full-length ORF2, the 2008 type-4 HAstV strains were characterised as a novel ORF2 genetic lineage, designated as 4c. The 2008 type-4 HAstVs also shared the ORF1b gene with similar HAstV-4c strains detected globally, thus displaying a conserved ORF1b/ORF2 asset. By interrogation of the databases, this novel lineage 4c accounted for 60.8% of the type-4 strains identified worldwide and the vast majority of recent type-4 HAstVs. The 2002 type-4 HAstVs displayed a type-4b ORF2, whereas in the ORF1b they resembled type-1 HAstVs. This inconsistency suggests a possible recombinant origin, with the RNA switch taking place upstream the ORF1b/ORF2 junction region. Also, recombination likely played a role in the diversification of the ORF2 of the three type-4 lineages. Multi-target analysis is required for appropriate characterisation and identification of recombinant HAstVs.

**1. Introduction**

Astroviruses (AstVs), family *Astroviridae*, are enteric viruses associated with enteric and extra-intestinal diseases in several mammalian and avian species (Mendez and Arias, 2007). AstVs have a single-stranded positive sense RNA genome containing three open reading frames (ORFs). ORF1a and ORF1b, at the 5' end of the genome, encode the non-structural viral proteins, including the RNA-dependent RNA-polymerase, while ORF2, at the 3' end, encodes the capsid protein precursor. Human AstVs (HAstVs) are a major cause of gastroenteritis in young children, elderly people and immunocompromised adults (Mendez and Arias, 2007). By sequence comparison of HAstVs, the capsid protein precursor can be divided into a highly conserved N-terminal domain (aa 1–424), a hypervariable domain (HVR) (aa 425–688) and a highly acidic C-terminal domain (Wang et al., 2001). The mature infectious virion contains three predominant protein species derived from the N-terminal domain and from the HVR after intra- and extra-cellular processing (Bass and Qiu, 2000; Sanchez-Fauquier et al., 1994). Mapping of neutralising monoclonal antibodies (Sanchez-Fauquier et al., 1994) and structural analysis have revealed that AstV capsid spike is formed by dimerisation of a polypeptide spanning the HVR and a putative binding receptor site with affinity for polysaccharide molecules has been predicted in the spike structure (Dong et al., 2011).

Early in the study of HAstVs, marked antigenic differences were noted by cross-neutralisation among some HAstV isolates (Kurtz and Lee, 1984). Subsequent studies in immune electron microscopy, immunofluorescence, ELISA and plaque neutralisation assays (Hudson et al., 1989; Jakab et al., 2004; Koopmans et al., 1998) revealed that the extent of HAstV antigenic diversity was greater, allowing for distinction of eight serotypes, HAstV-1 to -8. Sequence analysis of short fragments at either the 5' or 3' end of ORF2 (D5' and D3' regions) and RT-PCR genotyping protocols with type-specific primers have been used for genetic characterisation of HAstV-1 to -8 (Jakab et al., 2004; Mustafa et al., 2000; Noel et al., 1995). In addition, upon molecular analysis of the ORF2, discrete sequence variation has been observed within some HAstV types, allowing for distinction of genetic lineages, with two distinct genetic lineages (4a and 4b) being described in type-4 HAstVs (Colomba et al., 2006; De Grazia et al., 2011; Gabbay et al., 2007; Guix et al., 2002; Victoria et al., 2007).
Surveillance studies for HAstV in Italy identified the circulation of type-4 HAstV strains in 2002 (De Grazia et al., 2011) and again in 2008 (De Grazia et al., 2012, 2011; Medici et al., 2012). Based on a small sequence generated by the diagnostic primers Mon269–Mon270 (Noel et al., 1995) in the D5 region, the 2008 Italian type-4 HAstVs were clearly distinguishable by phylogenetic analysis from the 2002 type-4 HAstVs (De Grazia et al., 2011). In order to investigate further the extent of the genetic heterogeneity observed in type-4 HAstV strains, a 3.2 kb portion at the 3’ end of the genome was sequenced for two representative strains (ITA/2008/BA393/08-65 and ITA/2002/PA73) and compared with type-4 HAstVs retrieved from the databases. The 2008 type-4 HAstVs differed markedly from the older type-4 HAstVs and were classified within a novel, yet unrecognised, type-4 lineage, along with similar strains detected globally. By converse, the 2002 type-4 HAstVs were found to have a recombinant ORF1b derived from type-1 HAstVs.

2. Materials and methods

2.1. Samples origin

During surveillance activity on viral gastroenteritis conducted in Italy, two different lineages of type-4 HAstVs were identified by sequence analyses of 348-nucleotide (nt) portion at the 5’ end of ORF2. The prevalence of HAstV circulation ranged from 4.2% to 7.4%. In particular, the analyses of stool samples of children, aged less than 5 years, hospitalised with acute gastroenteritis revealed that three of the five HAstVs detected in Palermo in 2002 and three of the 32 HAstVs detected in Bari and Parma in 2008 were type-4 HAstV strains. The 2002 type-4 HAstVs markedly differed genetically from the 2008 viruses, falling into two distinct lineages. Two samples, ITA/2002/PA73 and ITA/2008/BA393/08-65, were selected as representatives of these lineages for further analysis.

2.2. RNA extraction and amplification

Viral RNA was extracted from 140 μl of stool suspension using the QIAmp viral RNA kit (Qiagen, GmbH, Hilden, Germany). A 3’ RACE-PCR protocol (Wang et al., 2005) was used to generate a 3.2 kb amplicon encompassing the 3’ end of ORF2, the full-length ORF2, the 3’ untraslated region (UTR) through the poly-A tail. Briefly, cDNA was synthesised by SuperScript III First-Strand cDNA synthesis kit (Invitrogen Ltd, Paisley, UK) with primer VN3T20 (5’-GAGTGACCGCGGCCGCT) upstream the ORF2 start codon (Walter et al., 2001). A 5–8ATGNC-3’ region, the 2008 Italian strains along with a set of 19 type-4 HAstVs retrieved from the databases allowed identifying three distinct genetic lineages, 4a–4c (Fig. 1A). The clades were monophyletic, with high strict and liberal probability values and high Bayesian posterior probability values, as defined using Species Delimitation software. Five strains clustered in lineage 4a (prototype AY720891/DEU/2004/Dresden), four strains in lineage 4b (prototype strain DQ070852/BRa/1995/Goiania/GO/12) and 12 strains in lineage 4c (prototype strain DQ344027/CHN/2005/Guangzhou). In the ORF2, the nt identity between strains of different lineages ranged between 89.1% and 93.5% while the nt identity within each lineage was not lower than 93.8%. The strain ITA/2002/PA73 displayed the highest nt identity (97.3–98.1%) to type-4b HAstV strains while the strain ITA/2008/BA393/08-65 was most similar to type-4c HAstV strains (95.8–98.2%). Nucleotide (difference) cut-off values of 6.5% were found to determine individual lineages in the full-length ORF2. These values are similar to values calculated in the D5’ region in other studies (Gabbay et al., 2007; Guix et al., 2002).

The deduced amino acid (aa) sequence of the capsid precursor was determined and an alignment generated to assess the rate of aa variation across the various ORF2 regions (Walter et al., 2001). The aa variation within type-4 HAstVs reached 6.8% in the N-terminal domain (aa 1–424), 8.1% in the HVR hypervariable region (aa 425–688) and 23.7% in the highly acidic C-terminal domain. In the HVR, intra-lineage aa variation reached 3.4%, 2.7% and 7.2% for 4a, 4b and 4c strains, respectively, while inter-lineage aa variation ranged between 5.7% and 8.1%. The cell-adapted strain GBR/Oxford-S4 (accession AB000296) presented a nucleotide deletion at the end of the HVR, which altered the frame of the capsid precursor throughout the C-terminal domain. For this strain, aa variation was higher, reaching values of 7.8%, 14.0% and 85.5% in the N-domain, HVR and acidic domain of the capsid precursor, respectively.

2.3. Sequence and phylogenetic analyses

The amplicons were purified and cloned using TOPO XL Cloning Kit (Invitrogen Ltd, Paisley, UK). Additional primers were designed to determine the complete 3.2-kb sequence by an overlapping strategy. Sequence editing and multiple codon-based (translation) alignments were performed with Geneious software v6.2 (Drummond et al., 2011). Phylogenetic analysis was conducted by using Genious software (Drummond et al., 2011). MrBayes version 3.2.1 (Huelsenbeck and Ronquist, 2001) was used to infer Bayesian analysis. Species Delimitation software (Masters et al., 2011) was used to assess clade monophyly and robustness of the phylogenetic analysis. SimPlot software (version 3.2) (Lole et al., 1999) was used to identify cross-over sites due to recombination. The accession number of the strains ITA/2002/PA73 and ITA/2008/BA393/08-65 are KC915035 and KC915034, respectively. A total of 19 full-length ORF2 sequences of type-4 HAstV were available in the databases. The sequences with their accession numbers are listed in Fig. 1A. In addition, a total of 54 partial (~350 bp) ORF2 sequences of type-4 HAstV strains spanning the D5’ region (nt 4573–4920 of prototype strain AY720891/DEU/2004/Dresden) and 6 partial (~200 bp) ORF2 sequences of type-4 HAstV strains spanning the D3’ region (nt 6427–6642 of prototype strain AY720891/DEU/2004/Dresden) were selected from the databases.

3. Results

A ~750-nt-long fragment of ORF1b was sequenced for strains ITA/2002/PA73 and ITA/2008/BA393/08-65. The full-length ORF2 of both viruses was of 2316 nt in length and the 3’ UTR was 81 nt long. There was a 8-nt overlap between the 3’ end of ORF1b and the start of ORF2. The highly conserved nt stretch 5’-ATT-TGGAGNGNGACTGAAATNGC-39, believed to be part of a promoter region for synthesis of subgenomic RNA, was retained upstream the ORF2 start codon (Walter et al., 2001).

3.1. Sequence analysis of the ORF2 of type-4 HAstVs

Analysis of the full-length ORF2 of the two Italian strains along with a set of 19 type-4 HAstVs retrieved from the databases allowed identifying three distinct genetic lineages, 4a–4c (Fig. 1A). The clades were monophyletic, with high strict and liberal probability values and high Bayesian posterior probability values, as defined using Species Delimitation software. Five strains clustered in lineage 4a (prototype AY720891/DEU/2004/Dresden), four strains in lineage 4b (prototype strain DQ070852/BRa/1995/Goiania/GO/12) and 12 strains in lineage 4c (prototype strain DQ344027/CHN/2005/Guangzhou). In the ORF2, the nt identity between strains of different lineages ranged between 89.1% and 93.5% while the nt identity within each lineage was not lower than 93.8%. The strain ITA/2002/PA73 displayed the highest nt identity (97.3–98.1%) to type-4b HAstV strains while the strain ITA/2008/BA393/08-65 was most similar to type-4c HAstV strains (95.8–98.2%). Nucleotide (difference) cut-off values of 6.5% were found to determine individual lineages in the full-length ORF2. These values are similar to values calculated in the D5’ region in other studies (Gabbay et al., 2007; Guix et al., 2002).

The deduced amino acid (aa) sequence of the capsid precursor was determined and an alignment generated to assess the rate of aa variation across the various ORF2 regions (Walter et al., 2001). The aa variation within type-4 HAstVs reached 6.8% in the N-terminal domain (aa 1–424), 8.1% in the HVR hypervariable region (aa 425–688) and 23.7% in the highly acidic C-terminal domain. In the HVR, intra-lineage aa variation reached 3.4%, 2.7% and 7.2% for 4a, 4b and 4c strains, respectively, while inter-lineage aa variation ranged between 5.7% and 8.1%. The cell-adapted strain GBR/Oxford-S4 (accession AB000296) presented a nucleotide deletion at the end of the HVR, which altered the frame of the capsid precursor throughout the C-terminal domain. For this strain, aa variation was higher, reaching values of 7.8%, 14.0% and 85.5% in the N-domain, HVR and acidic domain of the capsid precursor, respectively.

3.2. Recombinant origin of strain ITA/2008/BA393/08-65

In the ORF1b, strain ITA/2008/BA393/08-65/type4c displayed the highest nt identity (96.5%) to the type-4c HAstV strain CHN/2005/Guangzhou (accession DQ344027). On the opposite, strain ITA/2002/PA73/type4b displayed the highest nt identity (97.6–97.8%) to Indian and Chinese type 1 HAstVs (accessions
AB308374, JF327666 and FJ375759), while the nt identity to the prototype strain BRA/1995/Goiania/GO/12/type4b (accession DQ070852) was 92.0%. These findings are consistent with a recombination event, occurring in the ORF1b/ORF2 junction region for the type-4b strain ITA/2002/PA73 (Fig. 2A). By Simplot analysis (Fig. 2B) the crossover site was mapped to the ORF1b/ORF2 region, upstream the highly conserved promoter region for synthesis of subgenomic RNA (Walter et al., 2001).

3.3. D5₀ and D3₀ regions are predictors of lineage specificity

A total of 54 type-4 HAstV partial ORF2 sequences of various length (250–350 nt) spanning the D5₀ diagnostic region were retrieved from the NCBI databases, and analysed with the 19 full-length ORF2 sequences selected from the databases. The 73 sequences were retrieved form strains detected over a nearly 3-decade period. Nine (12.3%) sequences were characterised as lineage 4a, 19 (26%) as lineage 4b and 45 (61.6%) as lineage 4c. Also, a total of six partial ORF2 sequences is spanning the D3₀ diagnostic region were retrieved and analysed with the 19 full-length ORF2 sequences. Six sequences (24%) were characterised as 4a, 6 (24%) as 4b and 45 (61.6%) as lineage 4c. Also, a total of six partial ORF2 sequences is spanning the D3₀ diagnostic region were retrieved and analysed with the 19 full-length ORF2 sequences. Six sequences (24%) were characterised as 4a, 6 (24%) as 4b and 45 (61.6%) as lineage 4c. When considering the whole data set, 10 (12.6%), 21 (26.6%) and 48 (60.8%) of the sequences were characterised as type 4a, 4b and 4c, respectively (Fig. 1B). A type-8 sequence in Genbank (accession AY007591) was erroneously notated as type-4. No peculiar temporal or geographical pattern could be inferred from the data set. However, the vast majority of the recent type-4 HAstVs could be classified into lineage 4c.

In order to assess the reliability of the D5₀ region for accurate characterisation within each of the three lineages, a selection of 63 partial (~350 bp) ORF2 sequences of type-4 HAstV strains spanning the D5₀ region were analysed. In this analysis, intra-lineage identity for lineage 4a, 4b and 4c was ≥ 95.7%, 96.0% and 96.8% nt, respectively. Inter-lineage identity was 91.1–94.8% nt between lineages 4b and 4c, and 92.8–95.5% nt between lineages 4a and 4b, while it was 93.1–97.4% nt between lineages 4a and 4c. In the D5₀-based tree, the 4c lineage appeared to descend from type-4a HAstVs, with type-4b HAstVs forming a clear distinct cluster (Fig. 3B). In the D3₀-based tree, also the three lineages were highly supported statistically, although an inversion in the clustering pattern was observed, with the type-4c strains segregating a part from type-4a and -4b strains (Fig. 3C). By Simplot analysis, a possible crossover point among the three lineages was identified at the 3₀ end of the region coding for the N-terminal domain, thus encompassing all the HVR region downstream. The possible recombinant origin of the ORF2 of three lineages is schematically depicted in Fig. 3A.

4. Discussion

Epidemiological investigations for HAstVs worldwide have shown that type-1 HAstV is predominant, while the prevalence of other HAstVs can vary markedly (De Grazia et al., 2011; Medici et al., 2012). Type-4 HAstV accounts for 3–26% of the HAstV strains detected in various settings worldwide, representing the second most common type in several studies (De Grazia et al., 2012;
Gabbay et al., 2007; Guix et al., 2002; Mustafa et al., 2000; Victoria et al., 2007. Interestingly, type-4 HAstV, along with type-8 HAstV, seems to be associated with infection of older children and with longer duration of diarrhoea (>7 days) (Guix et al., 2002; Silva et al., 2006).

Accumulation and positive selection of point mutations is regarded as a powerful mechanism that steadily generates virus diversification. The HVR of AstV is believed to form the capsid spike (Dong et al., 2011) and to control binding to cell receptors as neutralising epitopes have been mapped inside this capsid portion. This region is subjected to impressive diversification across type-1 to -8 HAstVs, with the aa identity ranging from 44% to 69% (Sanchez-Fauquier et al., 1994; Wang et al., 2001). Intra-type variation among the various lineages of type-4 HAstVs appeared also appreciable. By analysis of a set of 21 full-length ORF2 sequences, at least three main genetic lineages, 4a–4c, were recognised. Strains within each lineage differed by at least 6.5% nt from strains of other lineages. In the HVR variation ranged between 5.7% and 8.1% aa among type-4 HAstVs of various lineages. Classification of HAstVs into lineages has been adopted and developed for epidemiological purpose by researchers using short capsid sequences generated with consensus primers Mon260–Mon270 targeting the D50 region, in the highly conserved N-terminus of the capsid protein (Noel et al., 1995). Although other diagnostic primers are available in the ORF1a, ORF1b, ORF2 and 3' UTR (Belliot et al., 1997; Guix et al., 2005; Jakab et al., 2004; Noel et al., 1995) the Mon260–Mon270 primer pair has been adopted and used in several epidemiological studies worldwide as it is able to amplify all HAstV types and the sequence information of the amplified region is suitable to predict serotype specificity.

Fig. 2. (A) Recombinant nature of the type-4b HAstV strain ITA/2002/PA73. Schematic representation of recombination between type-4b and type-1 HAstVs. (B) Simplot (Lole et al., 1999) analysis of the lineage 4b strain ITA/2002/PA73. The virus was plotted against the type-4b strain BRA/1995/Goi/12 and the type-1 strain CHN/2008/SH1. Sequences were analysed with Simplot using a window size of 200 and step size of 20 with gap strip off and Hamming correction on. The recombination break point is evidenced with a vertical dashed line. The genome organisation of HAstV (the 3' end of ORF1b and the 5' end of ORF2) is also shown.

Classification of HAstVs into lineages has been adopted and developed for epidemiological purpose by researchers using short capsid sequences generated with consensus primers Mon260–Mon270 targeting the D5' region, in the highly conserved N-terminus of the capsid protein (Noel et al., 1995). Although other diagnostic primers are available in the ORF1a, ORF1b, ORF2 and 3' UTR (Belliot et al., 1997; Guix et al., 2005; Jakab et al., 2004; Noel et al., 1995) the Mon260–Mon270 primer pair has been adopted and used in several epidemiological studies worldwide as it is able to amplify all HAstV types and the sequence information of the amplified region is suitable to predict serotype specificity. Taking advantage of the high conservation in the 3' end of ORF2 and in the 3' UTR, other consensus primers (prBEG/JWT4–Mon2) (Jakab et al., 2004) have been designed at the 3' end of ORF2 (D3' region). Interrogation of the databases was accomplished and an additional 60 partial ORF2 sequences spanning either the D5' (54 sequences) or D3' (6 sequences) regions were retrieved. All the retrieved sequences could be correctly characterised via phylogenetic analysis and bar-coding within the three lineages defined on the basis of...
full-length ORF2 classification. Neither temporal nor geographical patterns could be defined by data mining in the databases, although the number of lineage 4c strains was greater (60.8% of all the type-4 sequences) and the viruses originated from several geographical areas (Europe, South and North America, Africa, Asia, Australia), suggesting that this lineage is widespread. It is remarkable that, although several sequences of this novel HAstV lineage 4c were already available in the databases (with the oldest one dating back to late 1980s), this novel lineage had not been described yet, as only two type-4 lineages, namely 4a and 4b, were described thus far in the literature.

By analysing in parallel the ORF1b and ORF2, inconsistencies were observed in the clustering patterns for strain ITA/2002/PA73. The virus was grouped with type-4b HAstV strains in the ORF2 and with type-1 HAstVs in the ORF1b. By Simplot analysis, the cross over site was mapped upstream the conserved ORF1b/ORF2 junction region. The exchange of genome fragments via recombination is common in single-stranded RNA viruses and appears to occur at higher frequency in highly conserved genomic regions among genetically related strains (Bull et al., 2007). Molecular recombination of genomic RNA can affect phylogenetic groupings, increase the virulence of the virus, confuse molecular
epidemiological studies and have major implications in vaccine design (Bull et al., 2007). As seen in noroviruses, the highly conserved junction region connecting the non-structural and structural genes of HAstVs seems to be a recombination hot spot (De Grazia et al., 2012; Wolfaardt et al., 2011). On the opposite, strain ITA/2008/D5 in this type-4 lineage. In addition, a detailed analysis of the ORF2 D5’ and D3’ regions of type-4 HAstVs revealed a possible cross-over point among the three type-4 lineages, with type 4a and 4c strains being more related in the 1-kb 5’ end of ORF2 and type 4a and 4b strains being more related in the 1.3-kb 3’ end. Altogether, these findings clearly suggest puzzling pathways of evolution for the various lineages of type-4 HAstVs, with recombination playing a major role in generating novel strains.

5. Conclusions

The lack of structured molecular epidemiological studies for HAstVs still limits the knowledge on the dynamics of HAstV circulation and the pathways followed by these viruses in their evolution. Investigating systematically the genetic variability of HAstVs is important to understand if periodic genotype shifts and genetic/antigenic drifts occur and if these changes may fit a model of evolution analogous to those observed for other antigenically/genetically heterogeneous viruses.

Acknowledgements

This investigation was supported by the grant “MicroMap (PON01_02589)” and by the grant “Sorveglianza epidemiologica e molecolare delle gastroenteriti acute ad eziologia virale”, Ricerca Finalizzata Ministero della Salute 2006.

References
