Comparison of Insertion, Deletion, and Point Mutations in the Genomes of Human Adenovirus HAdvC-2 and SARS-CoV-2

Tetsuya Akaishi [1,2,3]

[1]Division of General Internal Medicine, Tohoku University Hospital, Sendai, Miyagi, Japan
[2]Department of Education and Support for Regional Medicine, Tohoku University Hospital, Sendai, Miyagi, Japan
[3]COVID-19 Screening Test Center, Tohoku University, Sendai, Miyagi, Japan

Virus genome mutation profiles with insertion, deletion, and point mutations have recently been revealed to differ remarkably between viruses. In RNA viruses like human coronaviruses or influenza viruses, genome samples collected over two to three decades usually show point mutations in 10-20% of the bases, while the rate of insertion and/or deletion mutations (indels) largely depends on the virus. This study evaluates the mutation profiles of DNA viruses by comparing a recently sampled genome of human adenovirus species C type 2 (isolate SG06/HAdvC2/2016) with a genome of the same species sampled in the 1970s. It was found insertions of 23 bases at seven sites and deletions of 22 bases at nine sites. The longest indels were 6-base insertions in *E2B* and *L4*. All indels in the coding regions were in-frame mutations with base lengths in multiples of three. In the non-coding regions, the lengths of the indels ranged from 1-4 consecutive bases. Long indels with more than 10 consecutive bases, which comprise nearly half of indels in the SARS-CoV-2 genome, were absent. In other sites, the point mutation rate was approximately 0.3%, which was significantly lower than in RNA viruses. In summary, the estimated point mutation rate in human adenovirus species C type 2 (HAdvC-2) was over 10 times lower than in RNA viruses. Unlike the relatively long indels in the SARS-CoV-2 genome, the indels in HAdvC-2 were short, with 6 or fewer consecutive bases.

Keywords: deletion; human adenovirus; insertion; point mutation; severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)
Tohoku J. Exp. Med., 2022 September, **258** (1), 23-27.
doi: 10.1620/tjem.2022.J049

Introduction

The genome of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been recently reported to include several medium-to-long insertion and/or deletion mutations (indels) with 10 or more consecutive bases (Andersen et al. 2020; Akaishi 2022). The long indels in the SARS-CoV-2 genome are unevenly distributed across the genome and concentrated in the non-structural protein 3 and spike (S) genes. Among these indels, one approximately 40 consecutive bases long occurring at the S1/S2 junction has been considered a key player in the acquisition of enhanced transmissibility of the virus to humans by creating a discriminative polybasic cleavage motif (Wrobel et al. 2020; Sasaki et al. 2021; Naveca et al. 2022). Although indels may have contributed significantly to the evolutionary process of SARS-CoV-2, the characteristics of the indels, such as their length and distribution in the genome of other DNA and RNA viruses, remain generally unknown. In two RNA viruses of the same species sampled two to three decades apart, it is thought that 10-20% of bases will display point mutations and 0-5% of bases will have indels (Akaishi 2022). However, the exact rate and size of indels in DNA viruses, including the human adenovirus (HAdv), are currently unknown. Therefore, this study compared two HAdv virus genomes with the aim of clarifying the viral mutation profile and estimating the potential role of indels in the evolutionary process of DNA viruses.

Methods

Genome sequences

The sequences of the two human adenovirus species C

Received May 20, 2022; revised and accepted June 5, 2022; J-STAGE Advance online publication June 16, 2022

Correspondence: Tetsuya Akaishi, Department of Education and Support for Regional Medicine, Tohoku University Hospital, 1-1 Seiryo-machi, Aoba-ku, Sendai, Miyagi 980-8574, Japan.
e-mail: t-akaishi@med.tohoku.ac.jp

©2022 Tohoku University Medical Press. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC-BY-NC-ND 4.0). Anyone may download, reuse, copy, reprint, or distribute the article without modifications or adaptations for non-profit purposes if they cite the original authors and source properly.
https://creativecommons.org/licenses/by-nc-nd/4.0/

type 2 (HAdvC-2) genomes were obtained from the GenBank database of the US National Institutes of Health (https://www.ncbi.nlm.nih.gov/genbank/). The selected isolate of the more recent HAdvC-2, SG06/HAdvC2/2016 (GenBank: MN513342.1) (Coleman et al. 2020), was sampled in Singapore in the years 2012-2015. The earlier reference sequence was sampled and combined in the 1970s (GenBank: J01917.1) (Zain et al. 1979a, b; Hérissé et al. 1981; Gingeras et al. 1982).

Point mutation rate

The point mutation rate in the SG06/HAdvC2/2016 genome was calculated by dividing the number of bases with point mutations by the total number of bases, after excluding all confirmed sites with indels. To evaluate the difference in the point mutation rates by their position in the genome, the rolling average of point mutations (± 50 bases) at each base position of the genome was calculated and displayed as a line graph of the point mutation rate across the whole genome of the virus. The point mutation rate in HAdvC-2 was then compared with SARS-CoV-2 and influenza A viruses (Akaishi 2022).

Statistical analyses

Distribution of the size of indels in HAdvC-2 or SARS-CoV-2 was described using the median and interquartile range (IQR; 25-75 percentiles). Distributions of the size of indels between the two viruses were compared by the Mann-Whitney U test. The point mutation rate and the ratio of indels to point mutations were evaluated through either a chi-square test or Fisher's exact test according to the number of bases with each type of mutation. Statistical significance was set at $p < 0.05$. Statistical analyses were performed using R Statistical Software (version 4.0.5; R Core Team, Vienna, Austria).

Results

Indels in HAdvC-2

Compared with the HAdvC-2 genome from the 1970s, the genome of SG06/HAdvC2/2016 had a total of 16 sites with indels: 7 sites with insertions for a total of 23 bases and 9 sites with deletions for a total of 22 bases. The base sequences of each indel site are listed in Table 1. The longest indels were 6-base insertions in *E2B* and *L4*. All indels in the coding regions were bases in multiples of three that could avoid frameshift mutations. The length of indels in the non-coding regions ranged from 1-4 consecutive bases, and there were no indels longer than 6 consecutive bases across the entire HAdvC-2 genome. Some of the indels were associated with repeated sequences, such as 5′-cttcttcttctt-3′ (base 9,375-9,386; deletion of "ctt"), 5′-gatgatgatgat-3′ (base 16,661-16,672; deletion of "gat"), or 5′-ccaccacca-3′ (base 28,437-28,445; deletion of "cca"). The median (IQR) of the size of indels (i.e., insertion, deletion, or insertion-and-deletion) in SARS-CoV-2 was 7.5 (3-23) bases, whereas that in HAdvC-2 was 2 (2-3) bases. The distributions of the size of indels in HAdvC-2 and SARS-CoV-2 are shown in Fig. 1a. The size of indels was

Table 1. List of indels in the genome of human adenovirus species C type 2 (HAdvC-2).

Base position in SG06/HAdvC2/2016	Genes	Sequence in SG06/HAdvC2/2016	Type of indels	Corresponding sequences in reference HAdvC-2
(Base# 1,113/1,114)	Non-coding	–	Deletion	5′- a -3′
Base# 1,209	Non-coding	5′- t -3′	Insertion	–
Base# 1,607-1,608	Non-coding	5′- aa -3′	Insertion	–
(Base# 1,628/1,629)	Non-coding	–	Deletion	5′- tg -3′
(Base# 9,386/9,387)	*E2B*	–	In-frame deletion	5′- ctt -3′
Base# 9,391-9,396	*E2B*	5′- cgstgg -3′	In-frame insertion	–
(Base# 16,672/16,673)	*L2*	–	In-frame deletion	5′- gat -3′
Base# 26,298-26,303	*L4*	5′- tgggac -3′	In-frame insertion	–
(Base# 26,333/26,334)	*L4*	–	In-frame deletion	5′- gaggag -3′
(Base# 28,445/28,446)	*E3A*	–	In-frame deletion	5′- cca -3′
(Base# 35,103/35,104)	Non-coding	–	Deletion	5′- t -3′
Base# 35,132-35,133	Non-coding	5′- aa -3′	Insertion	–
Base# 35,601-35,602	Non-coding	5′- at -3′	Insertion	–
(Base# 35,632/35,633)	Non-coding	–	Deletion	5′- ta -3′
(Base# 35,785/35,786)	Non-coding	–	Deletion	5′- a -3′
Base# 35,850-35,853	Non-coding	5′- gcac -3′	Insertion	–

A total of 16 indel sites were confirmed in the genome of SG06/HAdvC2/2016. The reference viral genome was obtained from HAdvC-2 sampled in the 1970s (GenBank: J01917.1). All indels in the coding regions were in-frame mutations formed of multiples of three bases. All indels were short with 6 or fewer consecutive bases. There were no medium-to-large indels.

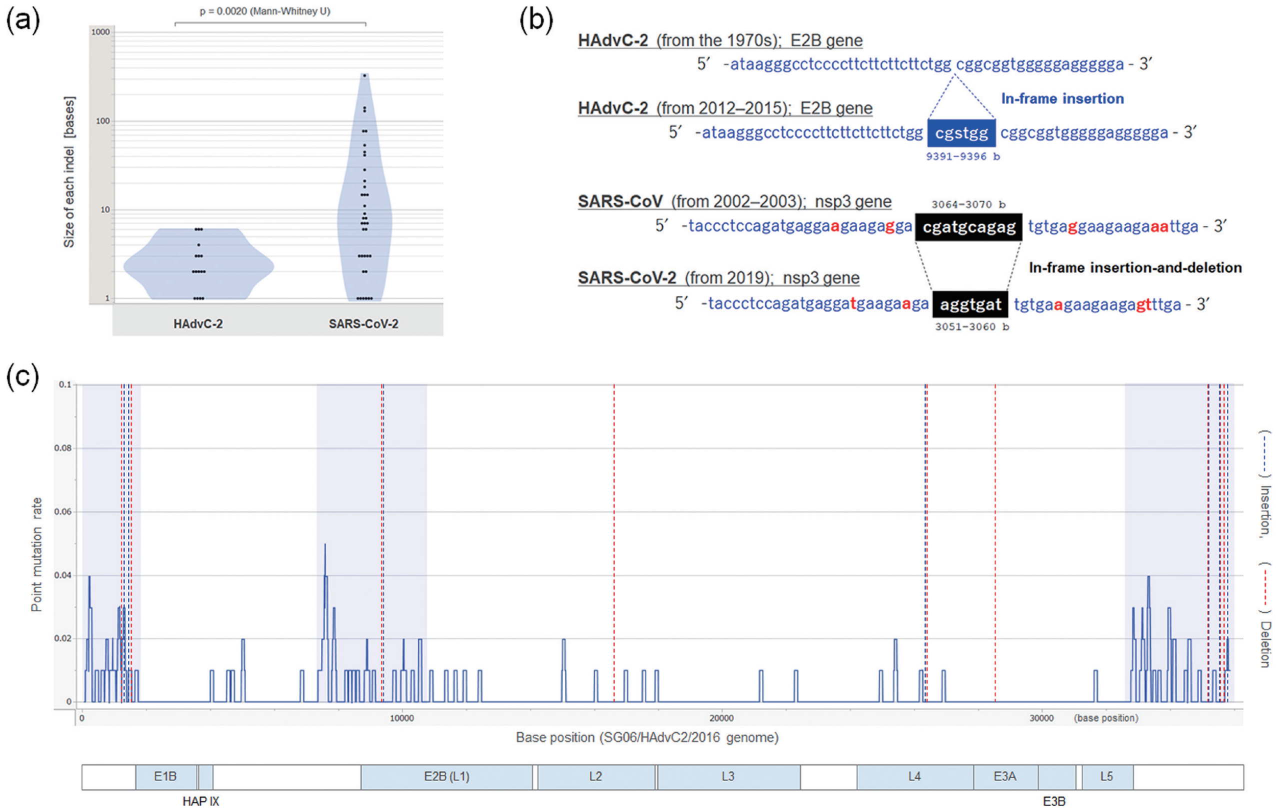

Fig. 1. Indels and point mutation rate in the genomes of HAdvC-2 and SARS-CoV-2.

The genome-wide mutation profiles in the genomes of HAdvC-2 and SARS-CoV-2 are shown. (a) Violin plots reporting the size of each indel in the genomes of HAdvC-2 and SARS-CoV-2 are shown. The size of indels was significantly larger in SARS-CoV-2 than in HAdvC-2. (b) Examples of indels in the genomes of HAdvC-2 and SARS-CoV-2 are shown. The former one in HAdvC-2 genome is a 6-base in-frame insertion, and the latter one in the genome of SARS-CoV-2 is an in-frame insertion-and-deletion mutation with the base size of the involved sequence decreased from 10 to 7 nucleotides. (c) Genome-wide point mutation rate and distribution of indels in the HAdvC-2 genome are shown. The solid blue line graph indicates the base-position-oriented point mutation rate, which was calculated as the rolling average of the point mutations in nearby (± 50) bases at each position. Transparent blue areas show the gene regions with relatively high incidences of point mutations. The broken blue and red lines show the distributions of insertion and deletion mutations, respectively.

b, base; E, early; HAdvC-2, human adenovirus species C type 2; HAP, hexon-associated protein; indels, insertion and/or deletion mutations; L, late; nsp3, non-structural protein 3; SARS-CoV, severe acute respiratory syndrome coronavirus; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

significantly larger in SARS-CoV-2 than in HAdvC-2 (p = 0.0020, Mann-Whitney U test). The actual sites of indels in the genomes of HAdvC-2 and SARS-CoVs are shown in Fig. 1b. The former one in HAdvC-2 genome is a 6-base in-frame insertion, and the latter one in SARS-CoV-2 genome is an in-frame insertion-and-deletion mutation with an exchange of consecutive bases with the base size of involved sequence decreased from 10 to 7 nucleotides.

Point mutations in HAdvC-2

After excluding the 23 bases at the indel sites, the point mutation rate in the overall genome of SG06/HAdvC2/2016 was 0.3% (91 of 35,901 bases). As shown in Table 2, the point mutation rate did not differ significantly between the genes or between the coding and non-coding regions, and the rate was significantly lower than in the SARS-CoV-2 and H1N1 influenza A genomes ($p < 0.0001$ for both, per chi-square test or Fisher's exact test). The ratio of substituted bases with indels to bases with point mutations in HAdvC-2 was not significantly different from that in SARS-CoV-2 (p = 0.2467, chi-square test: Cramer's V = 0.014) but was significantly higher than in the H1N1 influenza A virus ($p < 0.0001$, Fisher's exact test; Cramer's V = 0.439).

The base-position-oriented rolling average of the point mutation rate (± 50 bases) across the SG06/HAdvC2/2016 genome is shown in Fig. 1c. Although the overall point mutation rate was significantly lower than in the RNA viruses, non-coding regions of HAdvC-2 had a higher incidence of point mutations and indels than the coding regions. Across the whole genome, the rolling average of point mutation rates did not surpass 5.0%, suggesting a much lower mutation frequency in DNA viruses than RNA viruses.

Table 2. Mutation profiles in human adenovirus species C type 2 (HAdvC-2) from 2012-2015 compared with the 1970s.

	Point mutation [bases]	Point mutation rate*	Indels [bases]	Indel rate	No mutation [bases]	Total [bases]
SG06/HAdvC2/2016 (vs HAdvC-2 in the 1970s)						
E1B	0	0.0%	0	0.0%	1,793	1,793
HAP IX	0	0.0%	0	0.0%	423	423
E2B	15	0.3%	6	0.1%	5,469	5,490
L2	5	0.1%	0	0.0%	3,760	3,765
L3	2	0.0%	0	0.0%	4,391	4,393
L4	5	0.1%	6	0.2%	3,780	3,791
E3A	0	0.0%	0	0.0%	1,872	1,872
E3B	0	0.0%	0	0.0%	1,050	1,050
L5	1	0.1%	0	0.0%	1,748	1,749
Non-coding regions	63	0.5%	11	0.1%	11,524	11,598
Total	91	0.3%	23	0.1%	35,810	35,924
SARS-CoV-2 (vs. SARS-CoV from 2002-2003)						
Coding regions	5,515	19.5%	1,024	3.5%	22,721	29,260
Non-coding regions	33	5.5%	44	6.8%	566	643
Total	5,548	19.2%	1,068	3.6%	23,287	29,903
H1N1 influenza A virus during 2009 pandemic (vs. swine H1N1 influenza A virus in the 1980s)						
Total	1,915	14.6%	0	0.0%	11,195	13,110

Mutation profiles in RNA viruses, SARS-CoV-2 and H1N1 influenza A virus, are shown for comparison. The point mutation rates in RNA viruses were over 10 times higher than those in the double-stranded DNA virus HAdvC-2. In contrast to the SARS-CoV-2 genome, all the indels in the HAdvC-2 genome were short, with 6 or fewer consecutive bases.

*The point mutation rate was calculated after excluding sites with indels.

E, early; H, hemagglutinin; HAdvC-2, human adenovirus species C type 2; HAP, hexon-associated protein; indel, insertion and/or deletion; L, late; N, neuraminidase; SARS-CoV, severe acute respiratory syndrome coronavirus; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

Discussion

The results of the present study clearly demonstrate a much lower point mutation rate in the genome of HAdvC-2, a double-stranded DNA virus, than in the genomes of RNA viruses, such as SARS-CoV-2 or H1N1 influenza A. Furthermore, in this study there were no medium-to-long indels with 10 or more consecutive bases across the entire HAdvC-2 genome. This implies that medium-to-long indels, which are frequently observed in the SARS-CoV-2 genome, are not a universal or ubiquitous phenomenon in other viruses and could be a rare phenomenon specific to a few viruses, including betacoronaviruses. Another notable finding of the present study is that some indels were associated with repeated sequences. Because such short tandem repeats are sparse in virus genomes, they could be hot spots for the occurrence of indels in both prokaryotes and eukaryotes, as has been previously suggested (Darvasi and Kerem 1995; McDonald et al. 2011). In particular, when the tandem repeat unit is three bases long, the resultant indels do not cause frameshift mutations, and the indels have a higher probability of avoiding removal through natural selection. Additionally, the results of the present study demonstrate that indels 3-6 consecutive bases long are, for some viruses including HAdvC-2, not rare in the natural environment. Such relatively long indels may be more prevalent than previously thought, and may play a role in the evolutionary process of these viruses (Brown 2002). Future studies elucidating the mechanisms of relatively long indels in virus genomes are certainly warranted.

The current study had several limitations. First, this study selected only HAdvC-2 to evaluate double-stranded DNA virus mutation profiles and did not evaluate other DNA viruses. Therefore, it remains uncertain whether the findings of the present study can be generalized to other DNA viruses. Another limitation was that the recently reported medium-to-long indels in the SARS-CoV-2 genome have not yet been verified, and the existence of long indels has been confirmed in only the SARS-CoV-2 and bat coronavirus RaTG13 genomes. As a result, it remains unknown whether long indels are universally observed across all betacoronaviruses.

In conclusion, the point mutation rate in HAdvC-2, a double-stranded DNA virus, was over 10 times lower than in RNA viruses, such as SARS-CoV-2 or influenza A. All indels in the HAdvC-2 genome were short, with the involvement of 6 or fewer consecutive bases, and all indels in the coding regions had in-frame insertions or in-frame deletions. Some indels occurred within the repeated sequence of a 3-base unit, and such short tandem repeat sequences may be hot spots for the occurrence of indels in HAdvC-2.

Author Contributions

T.A. conceived, analyzed data, and drafted the manuscript.

Conflict of Interest

The author declares no conflict of interest.

References

Akaishi, T. (2022) Insertion-and-deletion mutations between the genomes of SARS-CoV, SARS-CoV-2, and bat coronavirus RaTG13. *Microbiol. Spectr.*, **10**, e0071622.

Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C. & Garry, R.F. (2020) The proximal origin of SARS-CoV-2. *Nat. Med.*, **26**, 450-452.

Brown, T.A. (2002) Chapter 14, Mutation, Repair and Recombination. In *Genomes,* 2nd ed., Wiley-Liss, Oxford, UK.

Coleman, K.K., Wong, C.C., Jayakumar, J., Nguyen, T.T., Wong, A.W.L., Yadana, S., Thoon, K.C., Chan, K.P., Low, J.G., Kalimuddin, S., Dehghan, S., Kang, J., Shamsaddini, A., Seto, D., Su, Y.C.F., et al. (2020) Adenoviral infections in Singapore: should new antiviral therapies and vaccines be adopted? *J. Infect. Dis.*, **221**, 566-577.

Darvasi, A. & Kerem, B. (1995) Deletion and insertion mutations in short tandem repeats in the coding regions of human genes. *Eur. J. Hum. Genet.*, **3**, 14-20.

Gingeras, T.R., Sciaky, D., Gelinas, R.E., Bing-Dong, J., Yen, C.E., Kelly, M.M., Bullock, P.A., Parsons, B.L., O'Neill, K.E. & Roberts, R.J. (1982) Nucleotide sequences from the adenovirus-2 genome. *J. Biol. Chem.*, **257**, 13475-13491.

Hérissé, J., Rigolet, M., de Dinechin, S.D. & Galibert, F. (1981) Nucleotide sequence of adenovirus 2 DNA fragment encoding for the carboxylic region of the fiber protein and the entire E4 region. *Nucleic Acids Res.*, **9**, 4023-4042.

McDonald, M.J., Wang, W.C., Huang, H.D. & Leu, J.Y. (2011) Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. *PLoS Biol.*, **9**, e1000622.

Naveca, F.G., Nascimento, V., Souza, V., Corado, A.L., Nascimento, F., Silva, G., Mejía, M.C., Brandão, M.J., Costa, Á., Duarte, D., Pessoa, K., Jesus, M., Gonçalves, L., Fernandes, C., Mattos, T., et al. (2022) Spread of Gamma (P.1) sublineages carrying spike mutations close to the furin cleavage site and deletions in the N-terminal domain drives ongoing transmission of SARS-CoV-2 in Amazonas, Brazil. *Microbiol. Spectr.*, **10**, e0236621.

Sasaki, M., Toba, S., Itakura, Y., Chambaro, H.M., Kishimoto, M., Tabata, K., Intaruck, K., Uemura, K., Sanaki, T., Sato, A., Hall, W.W., Orba, Y. & Sawa, H. (2021) SARS-CoV-2 bearing a mutation at the S1/S2 cleavage site exhibits attenuated virulence and confers protective immunity. *mBio*, **12**, e0141521.

Wrobel, A.G., Benton, D.J., Xu, P., Roustan, C., Martin, S.R., Rosenthal, P.B., Skehel, J.J. & Gamblin, S.J. (2020) SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. *Nat. Struct. Mol. Biol.*, **27**, 763-767.

Zain, S., Gingeras, T.R., Bullock, P., Wong, G. & Gelinas, R.E. (1979a) Determination and analysis of adenovirus-2 DNA sequences which may include signals for late messenger RNA processing. *J. Mol. Biol.*, **135**, 413-433.

Zain, S., Sambrook, J., Roberts, R.J., Keller, W., Fried, M. & Dunn, A.R. (1979b) Nucleotide sequence analysis of the leader segments in a cloned copy of adenovirus 2 fiber mRNA. *Cell*, **16**, 851-861.