SHORT COMMUNICATION

Time-series oligonucleotide count to assign antiviral siRNAs with long utility fit in the big data era

K Wada[1], Y Wada[1], Y Iwasaki[1,2] and T Ikemura[1]

Oligonucleotides are key elements of nucleic acid therapeutics such as small interfering RNAs (siRNAs). Influenza and Ebolaviruses are zoonotic RNA viruses mutating very rapidly, and their sequence changes must be characterized intensively to design therapeutic oligonucleotides with long utility. Focusing on a total of 182 experimentally validated siRNAs for influenza A, B and Ebolaviruses compiled by the siRNA database, we conducted time-series analyses of occurrences of siRNA targets in these viral genomes. Reflecting their high mutation rates, occurrences of target oligonucleotides evidently fluctuate in viral populations and often disappear. Time-series analysis of the one-base changed sequences derived from each original target identified the oligonucleotide that shows a compensatory increase and will potentially become the ‘awaiting-type oligonucleotide’; the combined use of this oligonucleotide with the original can provide therapeutics with long utility. This strategy is also useful for assigning diagnostic reverse transcription-PCR primers with long utility.

Gene Therapy (2017) **24,** 668–673; doi:10.1038/gt.2017.76; published online 14 September 2017

INTRODUCTION

Viruses have always posed significant threats to public health, as highlighted by the recent Ebolavirus (EBOV) outbreak in West Africa[1–5] and the emerging and re-emerging nature of influenza viruses.[6] To face the worldwide threats caused by zoonotic RNA viruses unpredictably transmitted from nonhumans, we must understand the details of molecular evolutionary changes in highly mutable genomes, including changes in oligonucleotide occurrences. Notably, even for large numbers of the genome sequences currently available, a time-series word count for oligonucleotides can be conducted without difficulty and specialized assumptions. In addition, the obtained results for viral evolutions are intuitively and easily understandable.

Viruses depend on many host factors for their growth (for example, nucleotide pools, proteins and RNAs) and must escape from host antiviral mechanisms (for example, interferon-induced systems).[7–9] Because human cells may not provide ideal growth conditions for the viruses transmitted from nonhumans, a certain level of directional change in viral sequences should occur during human-to-human transmission after the invasion from nonhumans. In fact, the G+C% of influenza A viruses isolated from humans is lower than that of strains isolated from avian and the CG dinucleotide in an A/U context was preferentially eliminated from classical H1N1 influenza viruses during human-to-human transmission.[10–13] By focusing on short oligonucleotide (2- to 5-mer) occurrences, we previously found[14–16] time-dependent unidirectional changes for a wide range of oligonucleotides in genomes of influenza viruses,[17–19] Ebolaviruses and MERS (Middle East respiratory syndrome) coronaviruses.[20] Notably, some oligonucleotides showed the time-dependent unidirectional changes occurring commonly across three influenza A subtypes that invaded independently from nonhumans at long intervals (for example, several decades).[16] We thus proposed that such changes should reoccur at a high probability after future invasions from their natural reservoir hosts.

Longer oligonucleotides than the above-mentioned short oligonucleotides are key elements of nucleic acid therapeutics (for example, antisense RNA, small interfering RNA (siRNA) and microRNA)[21–24] and diagnostic PCR primers.[25] Because RNA viruses mutate very rapidly, their sequence changes should be intensively characterized for designing therapeutic and diagnostic oligonucleotides with long utility. Our previous small-scale analyses of 20-mer oligonucleotides in influenza A genomes[16] showed that one oligonucleotide that showed a unidirectional decrease in common for three A subtypes after their independent invasions corresponded to the target sequence for an experimentally validated siRNA,[26] pointing out importance of time-series analysis for identifying siRNAs with long utility. Here, by focusing on ~12 000 genomes of influenza A, B and Ebolaviruses and on 182 experimentally validated siRNAs compiled by VIRsiRNAdb (a curated database of experimentally validated viral siRNA/short hairpin RNA),[26] we analyzed the time-series changes in occurrences of siRNA targets in the respective viral populations to develop strategies to predict siRNAs with long and wide utility.

RESULTS

siRNAs for influenza A viruses

Because of the social importance of influenza viruses, wide varieties of siRNAs have been designed, and from VIRsiRNAdb, 96 siRNAs for influenza A viruses, which were experimentally validated in various cell lines, could be downloaded on 23 September 2016. When focusing only on target sequences, 47 oligonucleotides were obtained; 42 were 19 nucleotide (nt), and the other 5 were 21 nt in length. To search siRNAs with long utility, we analyzed the time-series change in an abundance ratio of

[1]Department of Bioscience, Nagahama Institute of Bio-Science and Technology, Nagahama, Japan and [2]National Institute of Genetics, Mishima, Japan. Correspondence: Professor T Ikemura, Department of Bioscience, Nagahama Institute of Bio-Science and Technology, Tamura 1266, Nagahama, Shiga 526-0829, Japan.
E-mail: t_ikemura@nagahama-i-bio.ac.jp

Received 17 February 2017; revised 18 June 2017; accepted 31 July 2017; published online 14 September 2017

strains having the target oligonucleotides in the viral population, as follows. From the NCBI (National Center for Biotechnology Information) Influenza Virus Resource,[27] we first selected ~10 000 influenza A strains isolated from humans who had a full set of eight segment sequences, and grouped the strains according to the serological subtype and the isolated year (or month). Specifically, as conducted in the previous time-series study,[16] we first selected a year with at least 10 strains of one subtype and then a subtype with >5 years fulfilling the threshold (⩾10 strains). We thus selected H1N1, H3N2 and H1N1/09[refs 17–19] (starting from 2009 and abbreviated pH1N1 in the database). In the case of pH1N1, a sufficient number of strains fulfilling the above threshold (⩾10 strains) were available even per month; thus, the data for the conformable months were analyzed that allowed us to study the changes occurred within one outbreak.

The above selection resulted in focusing on major (but not minor) populations of viruses spreading among humans. Time-series analyses of all 47 siRNA targets, which are listed in Supplementary Figures S1–S3, showed that approximately half of the targets were absent at least in one subtype. Here, to search siRNAs with wide utility, we focus on the residual half, each of which is present in all three subtypes, although the occurrence significantly differs among subtypes. Because of high mutation rate of RNA viruses, most targets clearly fluctuate in their occurrence and often disappear after or during a certain period (Figure 1). Three subtypes are distinguished by different colors: H1N1 (blue), H3N2 (brown) and pH1N1 (green). Disappearance of targets in the viral population denotes a loss or reduction in usefulness of the respective siRNA. The most desirable target for long and wide utility should maintain an abundance near 1.0 in viral populations of all three subtypes during the most epidemic period; that is, almost all strains of the three subtypes always have the target in their genomes. There was no such target 19-mers, but there were two 19-mers maintaining the abundance near 1.0 for two subtypes, although the abundance for the residual subtype was very low (Figures 1a and b). The data specified by green X-marks for pH1N1 in Figure 1b are explained later.

Occurrences of other 19-mers significantly fluctuate in different ways. One type fluctuates in an irregular way (Figures 1c and d and Supplementary Figures S1–S3), but several others show time-dependent, unidirectional changes. The most interesting examples are the two 19-mers that show compensatory change: CAGCAGAGUGCUGUGGAUG and ACAGCAGAAUGCUGUGGAU (Figures 1e and f). These two are located in almost the same area in M gene (one base difference in their start position) but have one base change (G −> A) on the inside that is responsible for the time-dependent compensatory change. The former is maintained by most strains in the early stage after invasion for all three subtypes, which have occurred independently at long intervals, but unidirectionally reduces its occurrence during human–human transmission (Figure 1e). In contrast, the latter (Figure 1f) is almost absent in the early stage after invasion but begins to increase unidirectionally in H1N1 and H3N2 after a certain period. A strain that undergoes the G −> A mutation progressively increases the abundance of its descendants. Because this unidirectional increase is observed for two independently invaded subtypes, the change may have a certain biological significance.

The time-series analysis shown in Figure 1 covers ~10 000 viral genomes simultaneously and visualizes the compensatory changes of two 19-mers in an easily understandable way (Figures 1e and f); the pooled use of the respective two siRNAs should provide therapeutic oligonucleotides with long and wide utility. Notably, the present word-count analysis requires no specific assumption or model, and the obtained result is simple and straightforward.

One-base changed 19-mers

The above two 19-mers showing compensatory changes were found within the original siRNA targets. There is no other pair with this characteristic within the original targets, but several show a time-dependent unidirectional decrease (Figures 2a–c). If we can find a 19-mer that shows a compensatory increase, this will provide a candidate for the ‘awaiting-type siRNA’, and its pooled use with the original will provide a therapeutic oligonucleotide with long and wide utility. We thus generated one-base changed 19-mers from each original. Most of the 19-mers thus generated are not used in viral genomes, but the following show the unidirectional increase, compensatory to the decrease of their originals; the 19-mers shown in Figures 2d–f are the compensatory pair to those in Figures 2a–c, respectively. Another one-base changed 19-mer can compensate for low occurrence of the

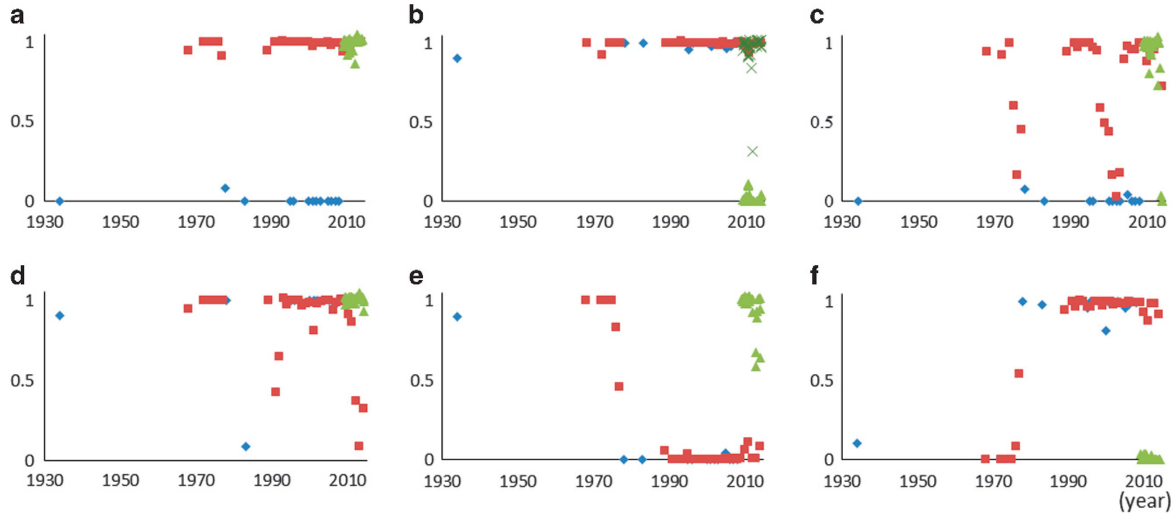

Figure 1. Time-series changes in abundance ratios of strains having the siRNA target in the viral population of three influenza A subtypes. The ratio is plotted according to the elapsed year (or month for pH1N1) from 1930 by color distinguishing subtypes: H1N1 (blue), H3N2 (brown) and pH1N1 (green). (**a**) UGGCAAAUGUUGUGAGAAA in PB1 gene. (**b**) GCAAUUGAGGAGUGCCUGA in PA gene. Occurrence of one-base changed 19-mer, GCAAUCGAGGAGUGCCUGA, for pH1N1 is specified by green X-marks. (**c**) GGACCAAACUUAUACAAUA in PB1 gene. (**d**) CGGGACUCUAGCAUACUUA in PB2 gene. (**e**, **f**) CAGCAGAGUGCUGUGGAUG and ACAGCAGAAUGCUGUGGAU in M gene, respectively. The base change is underlined.

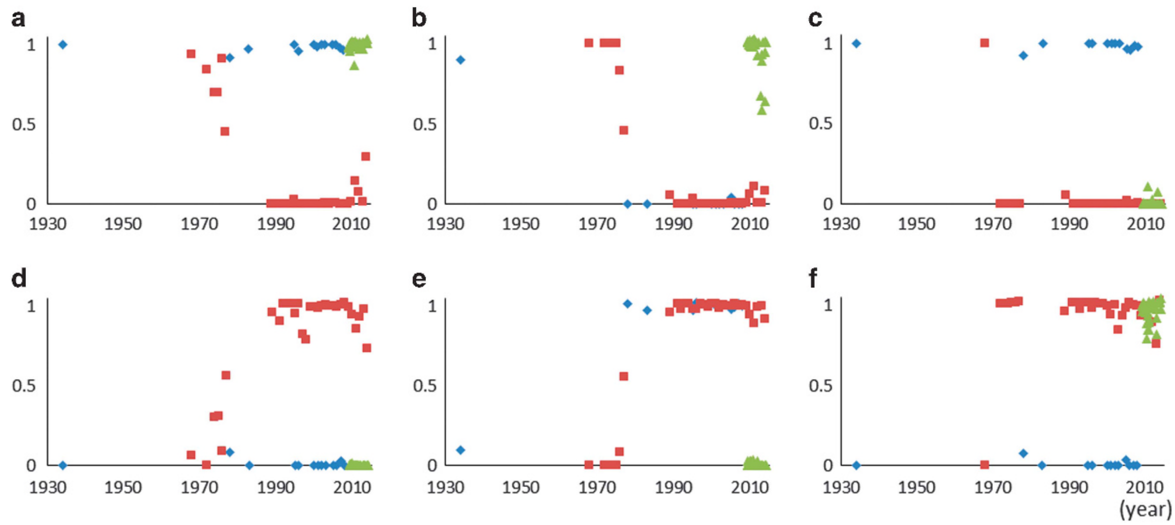

Figure 2. Time-series changes for strains having siRNA target 19-mers (**a**–**c**) and their one-base changed 19-mers (**d**–**f**). The abundance of strains having the respective 19-mer in the population of three influenza A subtypes is presented as described in Figure 1. (**a**, **d**) GAGACAAUGCAGAGGAGUA (original 19-mer) and GAGACAAUGCAGAAGAGUA (19-mer with the G –> A change) in NP gene, respectively. (**b**, **e**) CAGCAGAGUGCUGUGGAUG (original) and CAGCAGAAUGCUGUGGAUG (G –> A change) in M gene. (**c**, **f**) CGGCUACAUUGAGGGCAAG (original) and CGGCUGCAUUGAGGGCAAG (A –>G change) in PA gene.

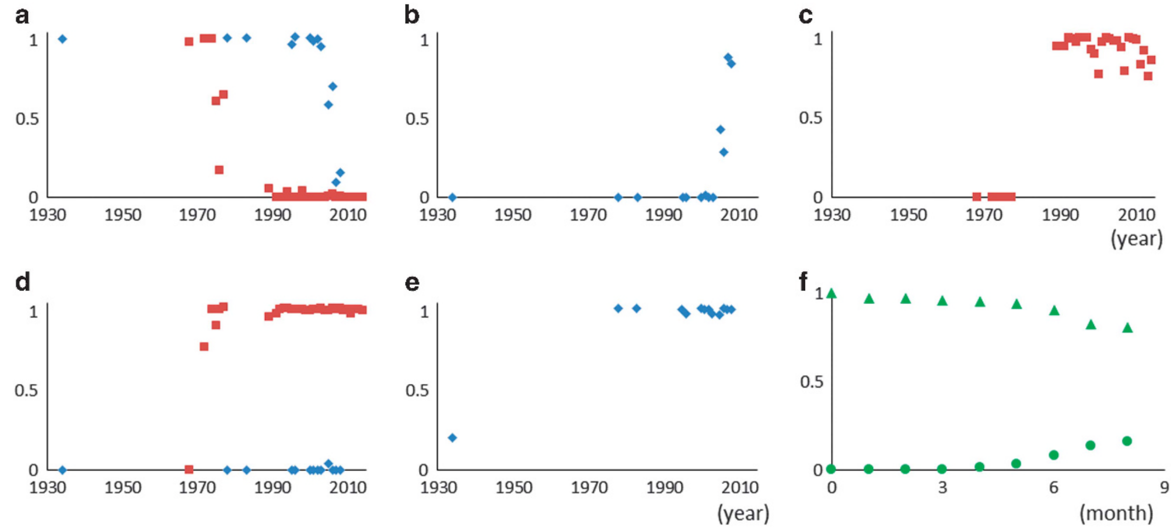

Figure 3. Time-series changes for siRNA targets and their one-base changed oligonucleotides. The abundance of strains having the respective 21- or 19-mer in the viral population of three influenza A subtypes is presented as described in Figure 1a AAUUUGCAGGCCUAUCAGAAA is the original 21-mer in M gene. (**b**) AAUUUACAGGCCUAUCAGA. The G –> A change at the sixth position from the original (listed in (**a**)) compensates for the decrease of the original in H1N1. (**c**) AAUUUGCAGACCUAUCAGA. The G –> A change at the tenth position from the original (listed in (**a**)) compensates for the decrease of the original in H3N2. (**d**, **e**) The original 21-mer GGUCGAAACGUAUGUUCUCUC and its one-base changed 21-mer GGUCGAAACGUACGUUCUCUC, respectively. Their additive occurrence maintains an approximate 1.0 level for H1N1 and H3N2 in the entire period, except 1930. (**f**) Abundance of pH1N1 strains having CGGCAAAGGCUAUGGAACA (original 19-mer) and CAGCAAAGGCUAUGGAACA (19-mer with the G –> A change) is plotted for the first 9 months from March 2009 with triangle and circle symbols, respectively.

original 19-mer for pH1N1 (Figure 1b); this one-base changed 19-mer is specified by green X-marks.

When additive occurrence for the two 19-mers maintains a level near 1.0 in major epidemic periods for all three types, the pooled use of the respective two siRNAs will provide therapeutic oligonucleotides with broad utility. Experimental validation of siRNA candidates thus predicted should rigorously define siRNAs with broad utility.

21-mer siRNAs

We next analyzed the aforementioned five siRNA-target 21-mers that are located in M gene but are separate from each other within the gene. One was not used at a significant level (< 0.1) in all three subtypes and omitted from the analysis. Among the residual four, only one shows a significant occurrence for two subtypes, but even in the two subtypes, it progressively reduces the occurrence to zero (Figure 3a). To search the compensatory pair, we generated one-base changed 21-mers but could not find a pair with an increasing trend. As an alternative approach, we first generated three 19-mers derived from each original 21-mer and successively one-base changed 19-mers. Among the one-base changed 19-mers, we found two 19-mers, each of which can compensate for the decrease in one subtype (Figures 3b and c).

The residual three 21-mers showed a high occurrence only in one subtype and changed time-dependently; Figure 3d shows that one 21-mer maintains its high occurrence in H3N2 except for the early stage after invasion, but is almost absent in H1N1. One-base changed 21-mer (U–> C change) derived from this original can compensate for low occurrence in H1N1 (Figure 3e).

Importance of changes occurring after new invasions

Viruses are inevitably dependent on many host factors for growth and must avoid host antiviral mechanisms. Concerning the time-dependent directional changes in oligonucleotide occurrences commonly found for three A subtypes after their independent invasions, we previously proposed that the changes should be related at least in part with viral adaptation to the new growth in human cells and will reoccur at a significant probability in future invasions.[16] Invasions of new influenza A subtypes and other emerging RNA viruses from nonhumans are highly probable and may cause worldwide threats. Understanding the changes that occur during the early stage after a new invasion is undoubtedly important for preventing its worldwide pandemic, and the important issue is to know how quickly a major population of newly invaded viruses will change their oligonucleotide sequences and how long it takes to detect the directional decreasing trend of the interested oligonucleotide and assign its compensatory pair with the increasing trend.

Notably, sequence data of pH1N1 can provide valuable information about changes that occur within the first outbreak cycle after a new invasion, because sequences of >200 strains per month are available for 9 months from March 2009 that correspond to the first outbreak after the new invasion. Among the 47 siRNA targets, 16 show high occurrences and maintain a high level (>0.9) during this period, except for one 19-mer. Figure 3f shows the unidirectional decrease of this exceptional 19-mer (green triangle: Pearson's correlation coefficient = −0.93) and the compensatory increase of its one-base changed pair (green circle: Pearson's correlation coefficient = 0.89). The null hypothesis (that is, no correlation) was rejected for these two 19-mers at the significant level of 0.01.

Even including a lag of a few months possibly required for occurrence of the respective mutation, 4 or 5 months appears appropriate for detecting the decreasing trend and predicting awaiting-type oligonucleotide; when omitting the lag period, a few months may be appropriate for this prediction. Undoubtedly, the required period for the prediction depends on individual oligonucleotides and viruses, as well as the period required for getting the respective mutation in the highly mutable genomes. As the number of available genome sequences increases, the required period for detecting a directional change should decrease, and the probabilistic reliability of the prediction should increase. Therefore, in the era of big data accumulation, the time-series analysis of oligonucleotide occurrences is increasingly important for preparing for the emerging virus threats.

Influenza B virus

From VIRsiRNAdb, 81 siRNAs for influenza B virus were available on 23 September 2016. All 42 target sequences were 19 nt in length, and their occurrences in the human influenza B population were analyzed by using ~2000 strains obtained from the NCBI Influenza Virus Resource.[27] When years with >10 strains were selected, as above done, only data from 1995 were available; if the threshold was lowered to 5, data for 1940 could be included. Because B type can currently infect only humans and the probability of a new invasion from nonhumans is very low, information about changes after 1995 may be sufficient for assigning siRNAs with long utility. However, understanding changes from an earlier stage may provide much comprehensive information about this viral evolution. Therefore, we mainly focused on data after 1995, but the additional data (⩾5 strains) before 1995 are also shown but separated by a broken line in each panel (Figures 4a–c).

Because the occurrences of eighteen 19-mers were very low or confined to a limited period, they were omitted from the analysis.

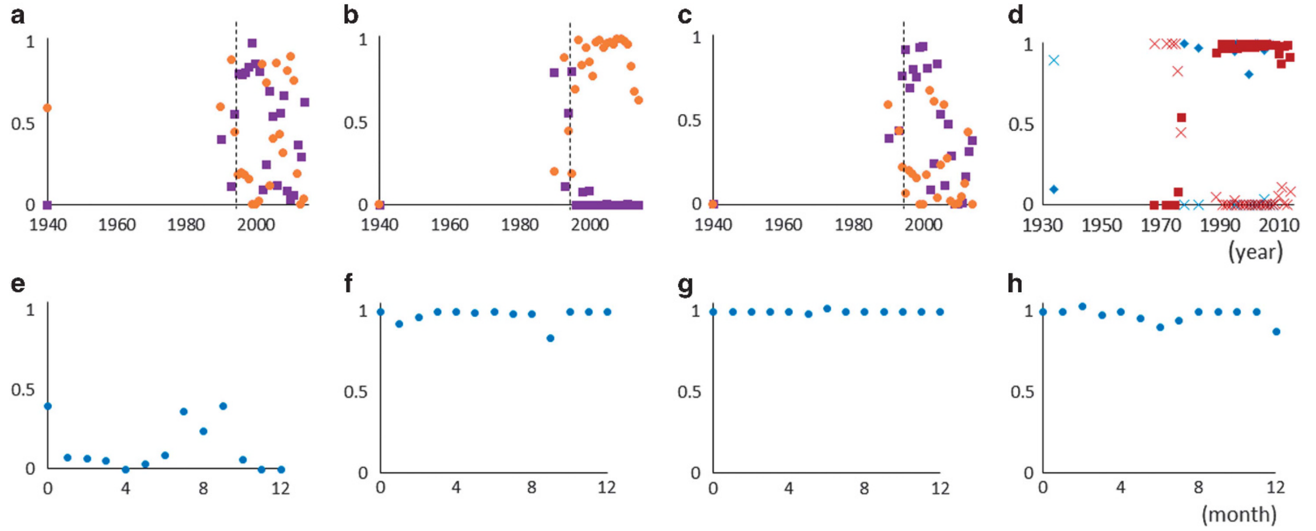

Figure 4. Time-series changes for strains having siRNA targets in the viral population of influenza B virus (**a**–**c**) and EBOV (**d**–**g**). (**a**) Abundance of influenza B strains having GCAACGGCACUAAACACAA (original 19-mer) and GCAACGGCACUGAACACAA (19-mer with the A –> G change) is plotted according to the elapsed year from 1940 with violet squares and orange circles, respectively. The data before and after 1995 are separated by a broken line. (**b**) GAGAGACAAUUAGACUGGU (original) and GGGAGACAAUUAGACUGGU (with the A –> G change). (**c**) GAUCUGUUUAGCAUACCAU (original) and GAUCUGUUUAGUAUACCAU (with the C –> U change). (**d**) Time-series changes for influenza A H1N1 (blue) and H3N2 (brown) for CAGCAGAGUGCUGUGGAUG (rhombus for H1N1 and square for H3N2) and ACAGCAGAAUGCUGUGGAU (x mark) previously listed in Figure 1e, f are relisted. (**e**) Abundance of EBOV strains having CGGACACACAAAAAGAAAG is plotted according to the elapsed month from March 2014. We focused on the month in which >10 strains were available. (**f**–**h**) The data for GGCAUCAGUGUGCUCAGUUGA, GGGCUCAUAUUGUUAUUGAUA and GUCUGGGCUCAUAUUGUUAUU are presented, respectively, as described in (**e**). The last two 21-mers are derived from adjacent positions in NP gene and harbored within one 25-mer (GUCUGGGCUCAUAUUGUUAUUGAUA).

Among the residual 24 (Supplementary Figures S4 and S5), CCAGAUGAUGGUCAAAGCU and GGAUGAAGAAGAUGGCCAU maintain a high level (> 0.9) from 1940 to 1995, respectively; and the other 6 maintain a relatively high level (> 0.5) from 1940 to 1995 (Supplementary Figures S4 and S5). Occurrences of the residual 16 evidently fluctuated and often neared zero, and therefore we generated one-base changed 19-mers. In contrast to the A type, the 19-mers showing the unidirectional change were rather rare, and those irregularly fluctuating were prominent. Among the latter cases, several pairs maintained their additive occurrence above 0.8 during a major period after 1995. Figures 4a–c show the time-series change for each compensatory pair. For comparison with the pattern of influenza A, the time-series change for two 19-mers in Figures 1a and b is relisted in Figure 4d.

The siRNA target harboring multiple neutral sites should be easily lost in the viral population, and the compensatory pair may not be uniquely assigned. In contrast, in the cases listed in Figures 4a–c, the changed base responsible for the compensatory change can be confined, indicating that there are a few neutral sites and the compensatory pair is suitable for pooled use despite their non-unidirectional changes. Time-series analysis can visualize these evolutionary details in an easily understandable way by analyzing large numbers of genomes simultaneously.

The above findings are consistent with the finding of Greenbaum *et al.*[28] that B type strains have already adapted well to growth in human cells, and thus the unidirectional changes responsible for adaptation to a new cellular environment have become less prominent. Other evolutionary processes, such as a bottleneck effect occurring at the onset of each seasonal outbreak, should become prominent.

Ebolaviruses

Sequences for more than 1000 genomes of human EBOV strains, isolated from March 2014, were available from the NCBI Virus Variation Database.[29] From VIRsiRNAdb, four siRNA targets for EBOV were obtained on 21 September 2016; one was 19-mer and three were 21-mers. The 19-mer (CGGACACACAAAAAGAAAG) shows a low occurrence (Figure 4e). In contrast, three 21-mers (GGCAUCAGUGUGCUCAGUUGA, GGGCUCAUAUUGUUAUUGAUA and GUCUGGGCUCAUAUUGUUAUU) maintain high occurrences (Figures 4f–h), showing that most viral strains keep the target sequences during the current epidemic.

DISCUSSION

The experimental measurement of siRNA efficacy should be the next important process to decide proper siRNAs, for the candidates obtained by one-base change from the experimentally validated siRNAs: for example, GAGACAAUGCAGAAGAGUA derived from GAGACAAUGCAGAGGAGUA; CAGCAGAAUGCUGUGGAUG derived from CAGCAGAGUGCUGUGGAUG; and CGGCUGCAUUGAGGGCAAG derived from CGGCUACAUUGAGGGCAAG (for details and additional examples, see the legends to Figures 2 and 3). Before conducting the experimental verification, it may be useful to predict their efficacy by using several informatic methods, such as VIRsiRNApred.[30]

Considering the pooled use, the pair composed of totally independent sequences, such as those listed in Figures 1a and b, is expected, and for selection of this pair, an overview of time-dependent changes of a wide range of oligonucleotides for a long period is required by analyzing large numbers of viral genomes simultaneously. Taking this into account, time-series patterns of 60 experimentally validated siRNA targets are listed in Supplementary Figures S1–S5.

Huge numbers of genome sequences from disease-causing microorganisms have rapidly accumulated because of revolutionary developments in sequencing technologies and social importance. In this era of big data accumulation, the participation of experts in big data analysis is increasingly important for interdisciplinary efforts against worldwide threats posed by infectious microorganisms. In this study, we focused on zoonotic RNA viruses, whose reservoir hosts are vertebrates, but it is undoubtedly important to analyze other RNA viruses spread by bugs (for example, ticks and mosquitoes) that have caused serious emerging infectious diseases, such as Dengue fever and Zika virus disease.[31,32] Notably, the present strategy is also useful for designing diagnostic reverse transcription-PCR primers with long utility.

SUBJECTS AND METHODS

From VIRsiRNAdb,[26] target sequences for 96 and 81 experimentally validated siRNAs for influenza A and B viruses, respectively, were downloaded on 23 September 2016. From the NCBI Influenza Virus Resource,[27] a total of ~ 200 000 segment sequences derived from ~ 25 000 influenza strains were obtained on 1 September 2015, and grouped according to host, serological type and isolated year (or month). We calculated occurrences of siRNA target sequences in each influenza virus genome. To prevent possible misassignment from a large number of pH1N1 strains, relatively small numbers of human classical H1N1 strains isolated from 2009 were omitted from the present analysis. From VIRsiRNAdb, target sequences for 4 siRNAs for EBOV were downloaded on 30 September 2016. EBOV genome sequences were downloaded from the NCBI Virus Variation Database[29] on 3 January 2016. Computer codes for the word count are available from KW (k_wada@nagahama-i-bio.ac.jp).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by a Grant-in-Aid for Scientific Research (C: no. 26330334) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

REFERENCES

1 WHO Ebola Response Team. Ebola virus disease in West Africa - The first 9 months of the epidemic and forward projections. *N Engl J Med* 2014; **371**: 1481–1495.

2 Tong YG, Shi WF, Liu D, Qian J, Liang L, Bo XC *et al.* Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. *Nature* 2015; **524**: 93–96.

3 Gatherer D. The 2014 Ebola virus disease outbreak in West Africa. *J Gen Virol* 2014; **95**: 1619–1624.

4 Park DJ, Dudas G, Wohl S, Goba A, Whitmer SL, Andersen KG *et al.* Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. *Cell* 2015; **161**: 1516–1526.

5 Carroll MW, Matthews DA, Hiscox JA, Elmore MJ, Pollakis G, Rambaut A *et al.* Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa. *Nature* 2015; **524**: 97–101.

6 Nichol ST, Arikawa J, Kawaoka Y. Emerging viral diseases. *Proc Natl Acad Sci USA* 2000; **97**: 12411–12412.

7 Garcı´a-Sastre A. Inhibition of interferon-mediated antiviral responses by influenza A viruses and other negative-strand RNA viruses. *Virology* 2001; **279**: 375–384.

8 Voinnet O. Induction and suppression of RNA silencing: insights from viral infections. *Nat Rev Genet* 2005; **6**: 206–220.

9 Randall RE, Goodbourn S. Interferons and viruses: an interplay between induction, signalling, antiviral responses and virus countermeasures. *J Gen Virol* 2008; **89**: 1–47.

10 Rabadan R, Levine AJ, Robins H. Comparison of avian and human influenza A viruses reveals a mutational bias on the viral genomes. *J Virol* 2006; **80**: 11887–11891.

11 Jimenez-Baranda S, Greenbaum B, Manches O, Handler J, Rabadán R, Levine A *et al.* Oligonucleotide motifs that disappear during the evolution of influenza in humans increase IFN-α secretion by plasmacytoid endritic cells. *J Virol* 2011; **85**: 3893–3904.

12 Karlin S, Doerfler W, Cardon LR. Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? *J Virol* 1994; **68**: 2889–2897.

13 Greenbaum BD, Levine AJ, Bhanot G, Rabadan R. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. *PLoS Pathog* 2008; **64**: e1000079.

14 Iwasaki Y, Abe T, Wada K, Itoh M, Ikemura T. Prediction of directional changes of influenza A virus genome sequences with emphasis on pandemic H1N1/09 as a model case. *DNA Res* 2011; **18**: 125–136.

15 Iwasaki Y, Abe T, Wada Y, Wada K, Ikemura T. Novel bioinformatics strategies for prediction of directional sequence changes in influenza virus genomes and for surveillance of potentially hazardous strains. *BMC Infect Dis* 2013; **13**: 386.

16 Wada Y, Wada K, Iwasaki Y, Kanaya S, Ikemura T. Directional and reoccurring sequence change in zoonotic RNA virus genomes visualized by time-series word count. *Scientific Reports* 2016; **6**: 36197.

17 Novel Swine-Origin Influenza A (H1N1) Virus Investigation Team. Emergence of a novel swine-origin influenza A (H1N1) virus in humans. *N Engl J Med* 2009; **360**: 2605–2615.

18 Neumann G, Noda T, Kawaoka Y. Emergence and pandemic potential of swine-origin H1N1 influenza virus. *Nature* 2009; **459**: 931–939.

19 Smith GJ, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature 2009* **459**: 1122–1125.

20 Azhar EI, El-Kafrawy SA, Farraj SA, Hassan AM, Al-Saeed MS, Hashem AM *et al.* Evidence for camel-to-human transmission of MERS coronavirus. *N Engl J Med* 2014; **370**: 2499–2505.

21 Crooke ST. Progress toward oligonucleotide therapeutics: pharmacodynamic properties. *FASEB J* 1993; **7**: 533–539.

22 Opalinska JB, Gewirtz AM. Nucleic-acid therapeutics: basic principles and recent applications. *Nat Rev Drug Discov* 2002; **1**: 503–514.

23 Meister G, Tuschl T. Mechanisms of gene silencing by double-stranded RNA. *Nature* 2004; **431**: 343–349.

24 Bennett CF, Swayze EE. RNA targeting therapeutics: molecular mechanisms of antisense oligonucleotides as a therapeutic platform. *Annu Rev Pharmacol Toxicol* 2010; **50**: 259–293.

25 Kageyama T. Broadly reactive and highly sensitive assay for Norwalk-like viruses based on real-time quantitative reverse transcription-PCR. *J Clin Microbiol* 2003; **41**: 1548–1557.

26 Thakur N, Qureshi A, Kumar M. VIRsiRNAdb: a curated database of experimentally validated viral siRNA/shRNA. *Nucleic Acids Res* 2012; **40** (Database issue): D230–D236.

27 Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T *et al.* The Influenza Virus Resource at the National Center for Biotechnology Information. *J Virol* 2008; **82**: 596–601.

28 Greenbaum BD, Cocco S, Levine AJ, Monasson R. Quantitative theory of entropic forces acting on constrained nucleotide sequences applied to viruses. *Proc Natl Acad Sci USA* 2014; **111**: 5054–5059.

29 Brister JR, Bao Y, Zhdanov SA, Ostapchuck Y, Chetvernin V, Kiryutin B *et al.* Virus Variation Resource - recent updates and future directions. *Nucleic Acids Res* 2014; **42** (Database issue): D660–D665.

30 Qureshi A, Thakur N, Kumar M. VIRsiRNApred: a web server for predicting inhibition efficacy of siRNAs targeting human viruses. *J Transl Med* 2013; **11**: 305.

31 Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL *et al.* The global distribution and burden of dengue. *Nature* 2013; **496**: 504–507.

32 Paixão ES, Barreto F, da Glória Teixeira M, da Conceição N, Costa M, Rodrigues LC. History, epidemiology, and clinical manifestations of Zika: a systematic review. *Am J Public Health* 2016; **106**: 606–612.

Supplementary Information accompanies this paper on Gene Therapy website (http://www.nature.com/gt)