Meta Gene 23 (2020) 100629

Contents lists available at ScienceDirect

Meta Gene

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

Methods for reducing the number of sequences in molecular evolutionary analyses

Yoshiyuki Suzuki[a,*], Maho Nishimura[a], Tatsuya Inoue[a], Yuki Kobayashi[b]

[a] Graduate School of Natural Sciences, Nagoya City University, Nagoya-shi, Aichi-ken 467-8501, Japan
[b] Nihon University Veterinary Research Center, Fujisawa-shi, Kanagawa-ken 252-0880, Japan

ARTICLE INFO

Keywords:
Distance matrix
Multiple alignment
Nucleotide diversity
Phylogenetic tree
Total branch length

ABSTRACT

Due to the progress in the sequencing technology, the number of nucleotide sequences for pathogens deposited in the public databases has been increasing rapidly. Consequently, in the molecular evolutionary analyses of pathogens, it may occasionally be difficult to include all the available sequences and necessary to reduce the number of sequences to accomplish computation within a realistic time frame. Here several methods for reducing the number of sequences were evaluated using the amount of evolutionary information contained in the retained sequences, which was measured as the total branch length of the phylogenetic tree (L). In the REA (random elimination in alignment) method, each of sequences was eliminated with equal probability. In the phylogenetic tree-based methods, the sequences associated with short exterior branches were eliminated; the sequences to be eliminated were required to constitute neighbors with another sequence in the CNT (closest neighbor in tree) method, whereas no such restriction was imposed in the SET (shortest exterior branch in tree) method. In the distance matrix-based methods, the sequences with small average distances to other sequences were eliminated; the sequences to be eliminated were required to be closely related to another sequence in the CPM (closest pair in matrix) method, whereas no such restriction was imposed in the SDM (smallest average distance in matrix) method. From the analyses of 2113 sequences for viral protein 1 of norovirus and 13,063 sequences for hemagglutinin of influenza A virus, it was observed that the CPM method was the most useful to obtain large L, in which the exterior branch length (L_E) tended to be elongated. In contrast, the interior branch length (L_I) tended to be elongated such that the L_I/L_E was heightened in the SDM method, which may be suitable for the phylogenetic analysis. The nucleotide diversity (π), the synonymous diversity (π_S), the nonsynonymous diversity (π_N), and the π_N/π_S were almost constant in the REA method, whereas they increased in other methods, suggesting that the REA method may be appropriate for the analyses of population diversity and natural selection.

1. Introduction

The nucleotide sequence data deposited in the public databases are important resource in the study of life science (Stevens, 2018). For instance, in the molecular evolutionary analyses for elucidating the evolutionary mechanisms and histories of pathogens such as human norovirus (HuNoV) and human influenza A virus (HuIAV), it is customary to include all nucleotide sequences for particular genes available in the public databases (e.g., Rambaut et al., 2008; Russell et al., 2008; Parra et al., 2017; Suzuki et al., 2019).

Due to the progress in the sequencing technology, the number of nucleotide sequences deposited in the public databases has been increasing rapidly (Karsch-Mizrachi et al., 2018). Theoretically, this situation is preferable for molecular evolutionary analyses because a large number of nucleotide sequences is expected to contain a large amount of evolutionary information, which is beneficial to conducting statistical tests (Nei et al., 2010; Suzuki, 2010). Practically, however, molecular evolutionary analyses of a large number of nucleotide sequences require a large amount of computation and are occasionally not accomplished within a realistic time frame (Ozaki et al., 2018). To avoid this problem, it may be necessary to reduce the number of nucleotide sequences to be included in molecular evolutionary analyses (Yonezawa et al., 2013; Kobayashi et al., 2018).

A variety of methods may be conceivable for reducing the number of nucleotide sequences (Vinod, 1969; Zaslavsky et al., 2008; Yonezawa et al., 2013). Thus, it may be important to develop appropriate criteria for evaluating the methods. In the previous study, it was proposed that the number of nucleotide sequences should be reduced so that the

* Corresponding author at: Graduate School of Natural science, Nagoya City University, 1 Yamanohata, Mizuho-ku, Nagoya-shi, Aichi-ken 467-8501, Japan.
E-mail address: yossuzuk@nsc.nagoya-cu.ac.jp (Y. Suzuki).

https://doi.org/10.1016/j.mgene.2019.100629
Received 19 July 2019; Received in revised form 26 September 2019; Accepted 28 October 2019
Available online 31 October 2019

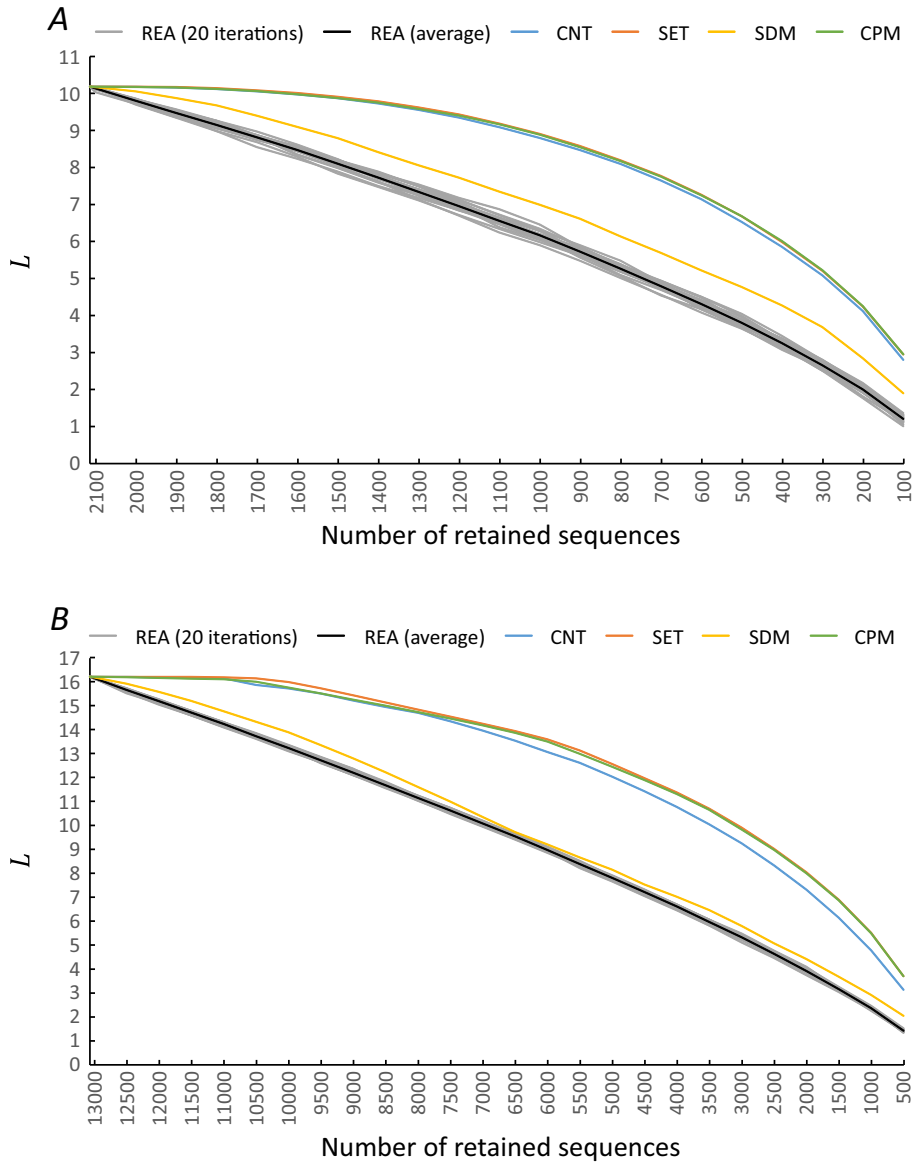

Fig. 1. The trajectories of the *L* for the retained sequences along with the reduction of the number of nucleotide sequences in the random elimination (REA) method, the phylogenetic tree-based (CNT and SET) methods, and the distance matrix-based (SDM and CPM) methods. *A*. GII.4 HuNoV VP1. *B*. H3N2 HuIAV HA.

average number of nucleotide differences between each of the eliminated sequences and its most closely related counterpart in the retained sequences would be small and the nucleotide diversity (π) of the retained sequences would be large (Yonezawa et al., 2013). These criteria were based on the notion that the nucleotide sequences adding a small amount of evolutionary information should be eliminated so that the retained sequences would contain a large amount of evolutionary information. However, it was not clear whether the amount of evolutionary information contained in the retained sequences could be appropriately measured in these criteria.

In molecular evolutionary analyses, the amount of evolutionary information is usually measured as the number of nucleotide substitutions (Nei and Kumar, 2000). Thus, the amount of evolutionary information contained in nucleotide sequences may be measured as the total number of nucleotide substitutions that have accumulated during evolution, which can be represented as the total branch length of the phylogenetic tree (*L*). The purpose of the present study was to evaluate the methods for reducing the number of nucleotide sequences using the *L* for the retained sequences.

2. Materials and methods

2.1. Methods

2.1.1. Random elimination method

One of the simplest methods for reducing the number of nucleotide sequences may be to eliminate each of *n* sequences in the original dataset with equal probability until the number of retained sequences reaches *m*. This protocol, called the REA ("random elimination in alignment") method in the present study, was used as the reference for evaluating performances of other methods as described below.

2.1.2. Phylogenetic tree-based methods

In the method proposed by Yonezawa et al. (2013), a phylogenetic tree is first constructed for the original dataset of *n* sequences, and the following procedure is repeated until the number of retained sequences reaches *m*. That is, in the phylogenetic tree, the neighbor connected with the smallest distance is identified, and the member of the neighbor associated with shorter exterior branch is eliminated. It was observed that by using this method the average number of nucleotide differences

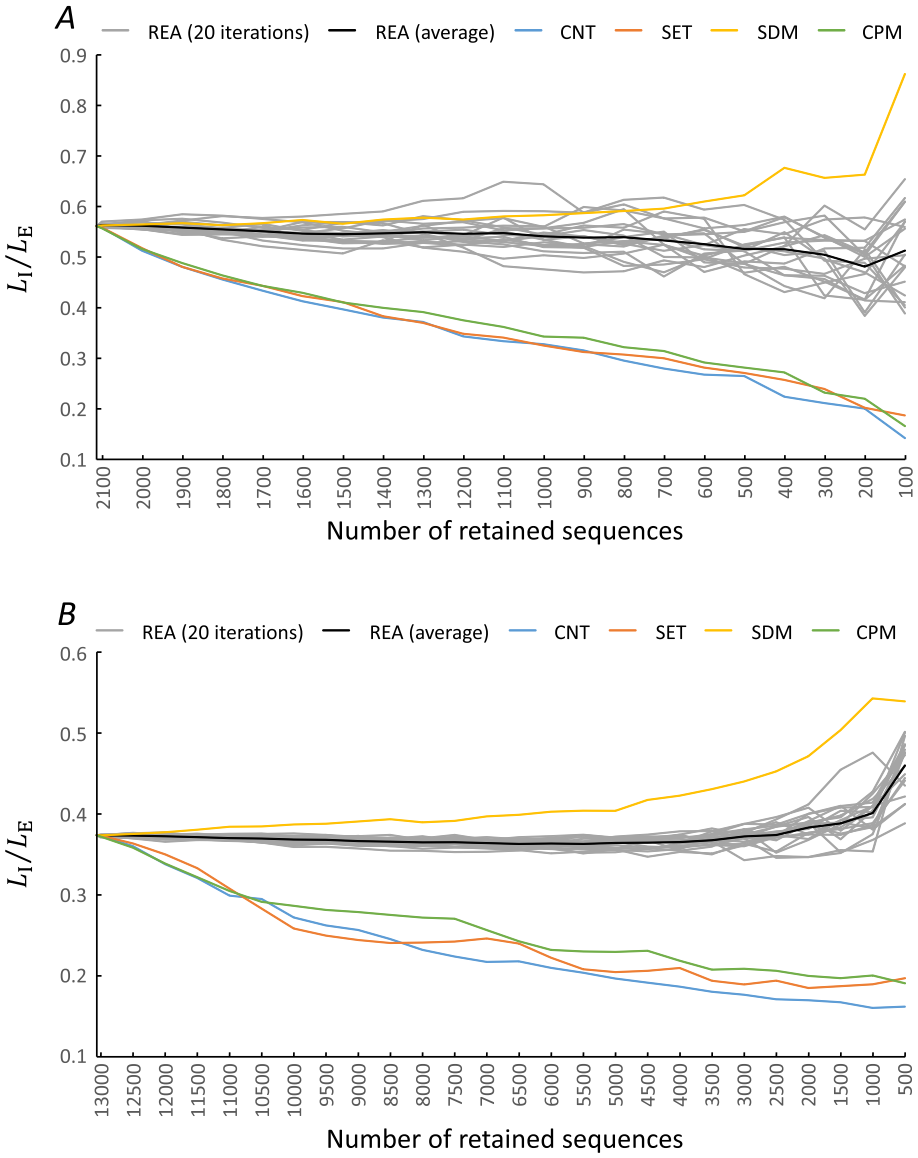

Fig. 2. The trajectories of the L_I/L_E for the retained sequences along with the reduction of the number of nucleotide sequences in the random elimination (REA) method, the phylogenetic tree-based (CNT and SET) methods, and the distance matrix-based (SDM and CPM) methods. *A*. GII.4 HuNoV VP1. *B*. H3N2 HuIAV HA.

between each of the eliminated sequences and its most closely related counterpart in the retained sequences tended to be small and the π of the retained sequences tended to be large (Yonezawa et al., 2013). This protocol is called the CNT ("closest neighbor in tree") method in the present study.

In the CNT method, the nucleotide sequences associated with short exterior branches are eliminated only when they constitute neighbors with another sequence in the phylogenetic tree. However, the nucleotide sequences that do not constitute neighbors may also be associated with short exterior branches. Therefore, in the protocol called the SET ("shortest exterior branch in tree") method in the present study, the nucleotide sequence associated with the shortest exterior branch in the phylogenetic tree is eliminated repeatedly, regardless of whether it constitutes a neighbor or not.

2.1.3. Distance matrix-based methods

In the phylogenetic tree-based methods as described above, it is necessary to construct a phylogenetic tree for the original dataset. However, this process itself may not be accomplished within a realistic time frame when the number of nucleotide sequences in the original dataset is large (Yonezawa et al., 2013). To avoid this problem, the distance matrix, which can be generated as an intermediate product in the course of tree construction, may be directly used to reduce the number of nucleotide sequences. Thus, in the protocol called the SDM ("smallest average distance in matrix") method in the present study, only the distance matrix is produced for the original dataset of n sequences. Then, the nucleotide sequence associated with the smallest average distance to other sequences is eliminated repeatedly until the number of retained sequences reaches m.

Theoretically, the SDM method is equivalent to reducing the number of nucleotide sequences so that the nucleotide sequences associated with the shortest exterior branches of the star phylogeny are eliminated and the π of the retained sequences is maximized (Saitou and Nei, 1987). However, the phylogenetic tree for the actual sequences is highly unlikely to be the star phylogeny, and it is not clear whether the nucleotide sequences eliminated in the SDM method are really associated with short exterior branches in the phylogenetic tree. Nevertheless, the nucleotide sequences that are closely related to another sequence are likely to be associated with short exterior branches in the phylogenetic tree. Thus, in the protocol called the CPM ("closest pair in

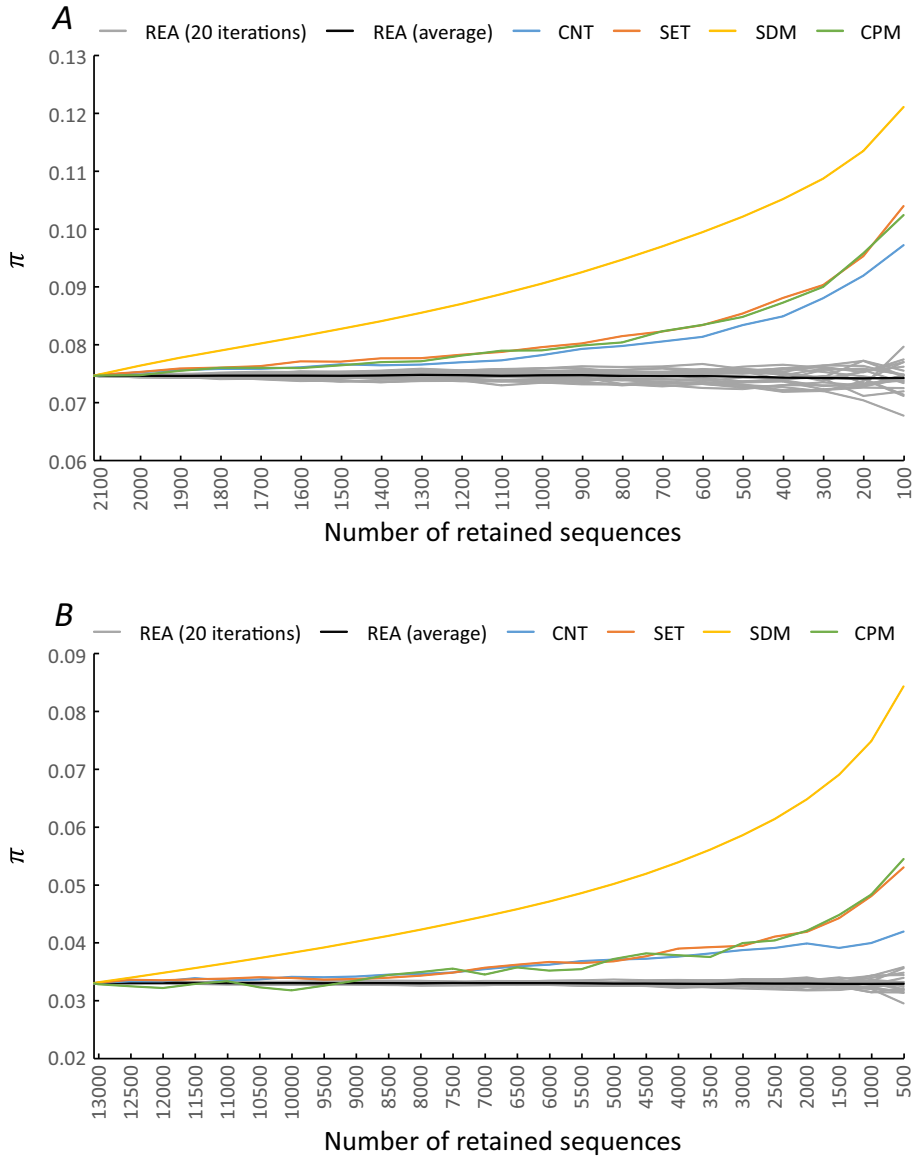

Fig. 3. The trajectories of the π for the retained sequences along with the reduction of the number of nucleotide sequences in the random elimination (REA) method, the phylogenetic tree-based (CNT and SET) methods, and the distance matrix-based (SDM and CPM) methods. *A*. GII.4 HuNoV VP1. *B*. H3N2 HuIAV HA.

matrix") method in the present study, the pair of nucleotide sequences with the smallest distance in the distance matrix is identified and the member of the pair associated with shorter average distance to other sequences is eliminated repeatedly.

2.2. Sequence data

All the nucleotide sequences encoding the entire region (1617 sites) of viral protein 1 (VP1) for genotype GII.4 HuNoV, provided with the information on the isolation year, were retrieved from the International Nucleotide Sequence Database (INSD) through the DNA Data Bank of Japan (https://www.ddbj.nig.ac.jp/) on January 19, 2019 (Stevens, 2018). In addition, all the nucleotide sequences encoding the entire region (1698 sites) of hemagglutinin (HA) for subtype H3N2 HuIAV provided with the information on the isolation year, were retrieved from the Influenza Research Database (https://www.fludb.org/) on January 23, 2019 (Zhang et al., 2017). Reportedly, intra-genic recombination does not occur within these genes (Katayama et al., 2002; Boni et al., 2008).

Multiple alignment of nucleotide sequences was made for each of the datasets of GII.4 HuNoV VP1 and H3N2 HuIAV HA using the computer program MAFFT (version 7.305b) (Katoh et al., 2002). From each multiple alignment, the sequences including ambiguous nucleotides or singleton gaps as well as the sequences identical to another sequence with the same isolation year were excluded, and the codon sites containing gaps in any sequence were omitted. Consequently, the multiple alignments for GII.4 HuNoV VP1 consisting of 2113 sequences with 1605 sites (Supplementary Table S1) and H3N2 HuIAV HA consisting of 13,063 sequences with 1689 sites (Supplementary Table S2) were used for the following analyses.

2.3. Data analysis

Computer programs were written for executing the REA, CNT, SET, SDM, and CPM methods. In the REA method, each of the nucleotide sequences in the multiple alignment was eliminated with equal probability using pseudo-random numbers. In the phylogenetic tree-based methods, the distance matrix of the proportion of different sites (p distance) (Nei and Kumar, 2000) was first generated from the multiple alignment. Then, the phylogenetic tree was constructed from the

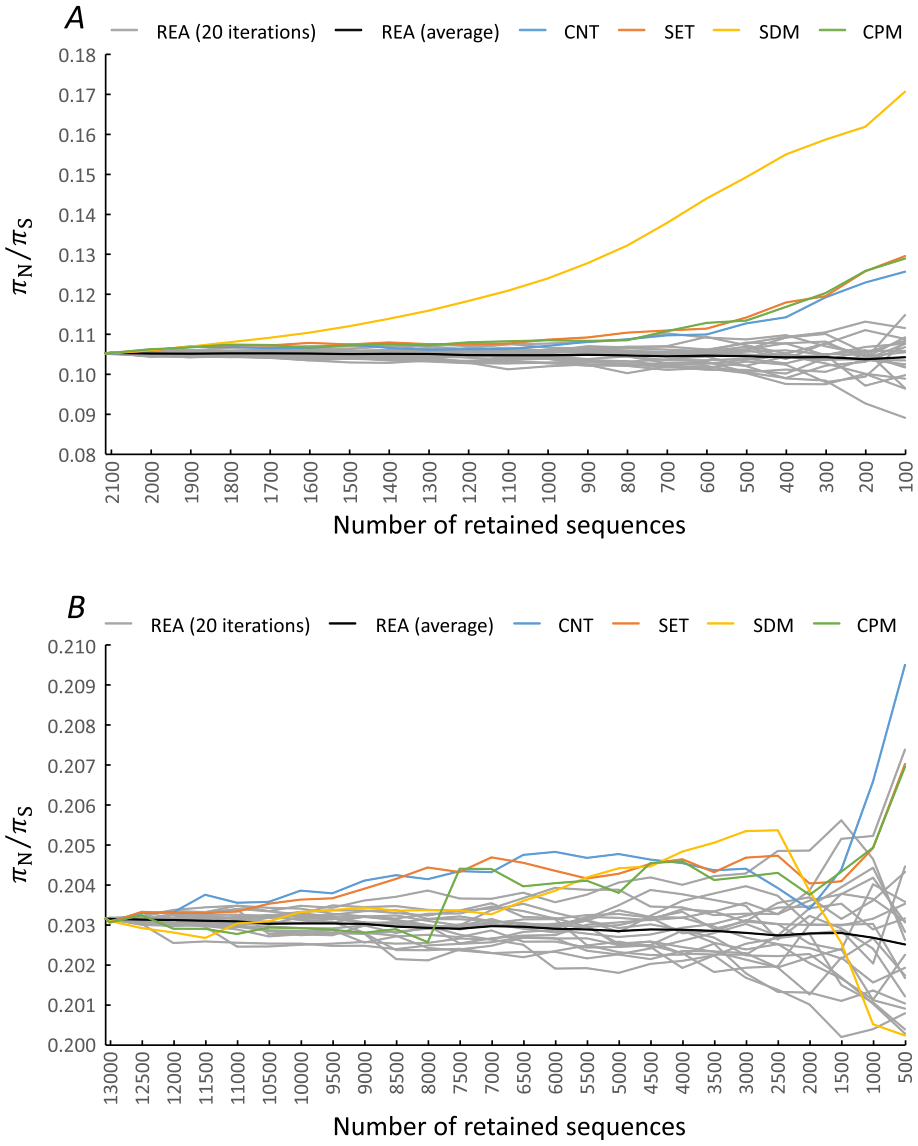

Fig. 4. The trajectories of the π_N/π_S for the retained sequences along with the reduction of the number of nucleotide sequences in the random elimination (REA) method, the phylogenetic tree-based (CNT and SET) methods, and the distance matrix-based (SDM and CPM) methods. *A*. GII.4 HuNoV VP1. *B*. H3N2 HuIAV HA.

distance matrix by the neighbor-joining (NJ) method (Saitou and Nei, 1987) using MEGA (version 6.06) (Tamura et al., 2013), and was adopted to reduce the number of nucleotide sequences. In the distance matrix-based methods, the distance matrix generated above was directly used to reduce the number of nucleotide sequences.

In all methods, the number of nucleotide sequences for GII.4 HuNoV VP1, which was originally 2113, was reduced one-by-one, and the amount of evolutionary information contained in the retained sequences was monitored when the number of retained sequences reached multiples of 100, that is, 2100, 2000, …, and 100. Similarly, the number of nucleotide sequences for H3N2 HuIAV HA, which was originally 13,063, was reduced one-by-one, and the amount of evolutionary information contained in the retained sequences was monitored when the number of retained sequences reached multiples of 500, that is, 13,000, 12,500, …, and 500. The amount of evolutionary information contained in the retained sequences was measured as the L, which was obtained by summing the lengths of all branches in the phylogenetic tree constructed by the NJ method (Saitou and Nei, 1987) with the p distance (Nei and Kumar, 2000) using MEGA (version 6.06) (Tamura et al., 2013). The L was further decomposed into the interior branch length (L_I) and the exterior branch length (L_E) to monitor the ratio of the former to the latter (L_I/L_E) (Fu and Li, 1993).

In addition to the L, the π for the retained sequences was also monitored using MEGA (version 6.06) (Tamura et al., 2013). The π was further decomposed into the synonymous diversity (π_S) and the non-synonymous diversity (π_N) to monitor the ratio of the latter to the former (π_N/π_S) (Nei et al., 2010; Suzuki, 2010). The entire process was iterated 20 times in the REA method.

3. Results

3.1. Total branch length of the phylogenetic tree (L)

The amount of evolutionary information contained in the original dataset measured as the L was 10.198 for GII.4 HuNoV VP1 and 16.203 for H3N2 HuIAV HA (Fig. 1). As expected, the L decreased as the number of nucleotide sequences was reduced in any of the random elimination (REA) method, the phylogenetic tree-based (CNT and SET) methods, and the distance matrix-based (SDM and CPM) methods (Wakeley, 2009). The L varied according to the method. However, the relative order of the L among the methods was similar between GII.4 HuNoV VP1 and H3N2 HuIAV HA.

When nucleotide sequences were randomly eliminated in the REA method, the trajectory of L along with the reduction of the number of nucleotide sequences fluctuated to some extent among 20 iterations. Overall, however, the L in the REA method appeared to be smaller than that in other (CNT, SET, SDM, and CPM) methods ($P < 0.05$) (Fig. 1). These results suggested that on average the nucleotide sequences associated with short exterior branches of the phylogenetic tree were eliminated in the CNT, SET, SDM, and CPM methods, as designed above.

In the phylogenetic tree-based methods, the nucleotide sequences associated with short exterior branches were eliminated from the phylogenetic tree constructed for the original dataset. The nucleotide sequences to be eliminated were required to constitute neighbors with another sequence in the CNT method, whereas no such restriction was imposed in the SET method. In Fig. 1, the L in the SET method appeared to be larger than that in the CNT method, suggesting that the nucleotide sequences associated with short exterior branches did not necessarily constitute neighbors in the phylogenetic tree.

In the distance matrix-based methods, the nucleotide sequences with small average distances to other sequences were eliminated from the distance matrix. The nucleotide sequences to be eliminated were required to be closely related to another sequence in the CPM method, whereas no such restriction was imposed in the SDM method. In Fig. 1, the L in the SDM method was smaller than that in the CPM method as well as the CNT and SET methods, suggesting that the nucleotide sequences associated with relatively long exterior branches were also eliminated in the SDM method, as long as elimination of these sequences maximized the π. In contrast, the L in the CPM method was similar to that in the SET method, which exhibited better performance than the CNT method as described above, suggesting that the nucleotide sequences associated with short exterior branches without constituting neighbors in the phylogenetic tree could also be eliminated in the CPM method.

When the L was decomposed into the L_I and L_E, it was observed that the L_I in the SDM method was greater than that in the CNT, SET, and CPM methods, which was similar to that in the REA method (Supplementary Fig. S1). In contrast, the L_E in the CNT, SET, and CPM methods was greater than that in the SDM method, which was similar to that in the REA method (Supplementary Fig. S2). As a result, the L_I/L_E in the SDM method was greater than that in the REA method ($P < 0.05$), which in turn was greater than that in the CNT, SET, and CPM methods ($P < 0.05$) (Fig. 2). These results suggested that the nucleotide sequences associated with the exterior branches connected to other exterior branches constituting neighbors tended to be eliminated in the CNT, SET, and CPM methods, whereas those connected to interior branches tended to be eliminated in the SDM method.

3.2. Nucleotide diversity (π)

The property of the retained sequences in each of the methods for reducing the number of nucleotide sequences was further examined by monitoring the π. In the REA method, the π of the retained sequences fluctuated to some extent among 20 iterations. Overall, however, the π was almost constant (Fig. 3). In contrast, the π appeared to increase in other (CNT, SET, SDM, and CPM) methods ($P < 0.05$), because the nucleotide sequences associated with short exterior branches of the phylogenetic tree were eliminated in these methods, as described above.

When the π of the retained sequences was decomposed into the π_S (Supplementary Fig. S3) and π_N (Supplementary Fig. S4), they were observed to behave in a similar manner to the π; i.e., they were almost constant in the REA method, whereas they increased in other methods ($P < 0.05$). Accordingly, the π_N/π_S was almost constant in the REA method (Fig. 4). Notably, however, the π_N/π_S increased in other methods particularly for GII.4 HuNoV VP1 ($P < 0.05$). Generally, in the phylogenetic tree, advantageous and deleterious mutations are accumulated more proximal and distal to the root, respectively, and interior and exterior branches are located more proximal and distal to the root, respectively (McDonald and Kreitman, 1991; Pybus et al., 2007; Suzuki, 2011). In the above analysis, it was observed that the L_I/L_E was lowered in the CNT, SET, and CPM methods and heightened in the SDM method. Therefore, the increase in the π_N/π_S may represent enrichment of deleterious mutations on distal branches in the CNT, SET, and CPM methods and advantageous mutations on proximal branches in the SDM method.

Although the trajectory of π_N/π_S was not clearly discriminated between 20 iterations of the REA method and other methods for H3N2 HuIAV HA, the π_N/π_S in the latter methods tended to be greater than that in the former method, similarly to the case for GII.4 HuNoV VP1 (Fig. 4). However, the π_N/π_S in the SDM method fluctuated to a relatively large extent. Reportedly, positive selection has operated on multiple lineages for GII.4 HuNoV VP1 (Tohma et al., 2019) and on a single trunk lineage for H3N2 HuIAV HA (Fitch et al., 1991; Wolf et al., 2006; Suzuki, 2008). These observations suggested that the difference in the trajectory of π_N/π_S in the SDM method between GII.4 HuNoV VP1 and H3N2 IAV HA may reflect difference in the pattern of accumulation of advantageous mutations on the phylogenetic tree.

4. Discussion

In the present study, performances of the random elimination (REA) method, the phylogenetic tree-based (CNT and SET) methods, and the distance matrix-based (SDM and CPM) methods for reducing the number of nucleotide sequences were evaluated using the amount of evolutionary information contained in the retained sequences measured as the L. The results obtained from the analyses of 2113 sequences for GII.4 HuNoV VP1 and 13,063 sequences for H3N2 HuIAV HA were largely consistent with each other. The L for the retained sequences in the CNT, SET, SDM, and CPM methods was greater than that in the REA method. Although the SET and CPM methods exhibited better performances than the CNT and SDM methods, the CPM method was considered to be the most useful to obtain large L for the retained sequences, because the amount of computation required in the CPM method was smaller than that in the SET method.

From the comparison of L_I and L_E, it was suggested that the nucleotide sequences associated with the exterior branches connected to other exterior branches constituting neighbors tended to be eliminated in the CNT, SET, and CPM methods to elongate the L_E, whereas those connected to interior branches tended to be eliminated in the SDM method to elongate L_I. It should be noted that the L_I but the L_E contributes to resolving distinct clusters in the phylogenetic analysis (Nei and Kumar, 2000). Therefore, in the phylogenetic analysis, the SDM method may be suitable for reducing the number of nucleotide sequences, in which the amount of computation required was also relatively small.

Along with the reduction of the number of nucleotide sequences, the π, π_S, and π_N for the retained sequences increased in the CNT, SET, SDM, and CPM methods, because the nucleotide sequences associated with short exterior branches of the phylogenetic tree were eliminated in these methods. Notably, however, the π_N/π_S also increased in the CNT, SET, SDM, and CPM methods, although the increase in the π_N/π_S may represent enrichment of deleterious mutations in the CNT, SET, and CPM methods and advantageous mutations in the SDM method. In contrast, the π, π_S, π_N, and π_N/π_S appeared to be almost constant in the REA method. The π, π_S, and π_N can be used as indicators of the degree of population diversity (Nei and Li, 1979; McDonald and Kreitman, 1991). In addition, the π_N/π_S can be used as an indicator of the direction and magnitude of natural selection (Nei et al., 2010; Suzuki, 2010). These observations suggested that the REA method, in which the amount of computation required was the smallest, may be appropriate for reducing the number of nucleotide sequences in the analyses of population diversity and natural selection.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.mgene.2019.100629.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgements

The authors thank two anonymous reviewers for valuable comments. This work was supported by JSPS KAKENHI Grant Number JP19K12221 and AMED Grant Number JP19fk0108033h0003 to Y.S.

References

Boni, M.F., Zhou, Y., Taubenberger, J.K., Holmes, E.C., 2008. Homologous recombination is very rare or absent in human influenza A virus. J. Virol. 82, 4807–4811.
Fitch, W.M., Leiter, J.M., Li, X.Q., Palese, P., 1991. Positive Darwinian evolution in human influenza A viruses. Proc. Natl. Acad. Sci. U. S. A. 88, 4270–4274.
Fu, Y.-X., Li, W.-H., 1993. Statistical tests of neutrality of mutations. Genetics 133, 693–709.
Karsch-Mizrachi, I., Takagi, T., Cochrane, G., The International Nucleotide Sequence Database Collaboration, 2018. The international nucleotide sequence database collaboration. Nucleic Acids Res. 46, D48–D51.
Katayama, K., Shirato-Horikoshi, H., Kojima, S., Kageyama, T., Oka, T., Hoshino, F., Fukushi, S., Shinohara, M., Uchida, K., Suzuki, Y., Gojobori, T., Takeda, N., 2002. Phylogenetic analysis of the complete genome of 18 Norwalk-like viruses. Virology 299, 225–239.
Katoh, K., Misawa, K., Kuma, K.-i., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066.
Kobayashi, Y., Pybus, O.G., Itou, T., Suzuki, Y., 2018. Conserved secondary structures predicted within the 5′ packaging signal region of influenza A virus PB2 segment. Meta Gene 15, 75–79.
McDonald, J.H., Kreitman, M., 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351, 652–654.
Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, Oxford, New York.
Nei, M., Li, W.-H., 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. U. S. A. 76, 5269–5273.
Nei, M., Suzuki, Y., Nozawa, M., 2010. The neutral theory of molecular evolution in the genomic era. Annu. Rev. Genomics Hum. Genet. 11, 265–289.
Ozaki, K., Matsushima, Y., Nagasawa, K., Motoya, T., Ryo, A., Kuroda, M., Katayama, K., Kimura, H., 2018. Molecular evolutionary analyses of the RNA-dependent RNA polymerase region in norovirus genogroup II. Front. Microbiol. 9, 3070.
Parra, G.I., Squires, R.B., Karangwa, C.K., Johnson, J.A., Lepore, C.J., Sosnovtsev, S.V., Green, K.Y., 2017. Static and evolving norovirus genotypes: implications for epidemiology and immunity. PLoS Pathog. 13, e1006136.
Pybus, O.G., Rambaut, A., Belshaw, R., Freckleton, R.P., Drummond, A.J., Holmes, E.C., 2007. Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. Mol. Biol. Evol. 24, 845–852.
Rambaut, A., Pybus, O.G., Nelson, M.I., Viboud, C., Taubenberger, J.K., Holmes, E.C., 2008. The genomic and epidemiological dynamics of human influenza A virus. Nature 453, 615–619.
Russell, C.A., Jones, T.C., Barr, I.G., Cox, N.J., Garten, R.J., Gregory, V., Gust, I.D., Hampson, A.W., Hay, A.J., Hurt, A.C., de Jong, J.C., Kelso, A., Klimov, A.I., Kageyama, T., Komadina, N., Lapedes, A.S., Lin, Y.P., Mosterin, A., Obuchi, M., Odagiri, T., Osterhaus, A.D., Rimmelzwaan, G.F., Shaw, M.W., Skepner, E., Stohr, K., Tashiro, M., Fouchier, R.A., Smith, D.J., 2008. The global circulation of seasonal influenza A (H3N2) viruses. Science 320, 340–346.
Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–435.
Stevens, H., 2018. Globalizing genomics: the origins of the International Nucleotide Sequence Database Collaboration. J. Hist. Biol. 51, 657–691.
Suzuki, Y., 2008. Positive selection operates continuously on hemagglutinin during evolution of H3N2 human influenza A virus. Gene 427, 111–116.
Suzuki, Y., 2010. Statistical methods for detecting natural selection from genomic data. Genes Genet. Syst. 85, 359–376.
Suzuki, Y., 2011. Positive selection for gains of N-linked glycosylation sites in hemagglutinin during evolution of H3N2 human influenza A virus. Genes Genet. Syst. 86, 287–294.
Suzuki, Y., Doan, Y.H., Kimura, H., Shinomiya, H., Shirabe, K., Katayama, K., 2019. Predicting directions of changes in genotype proportions between norovirus seasons in Japan. Front. Microbiol. 10, 116.
Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30, 2725–2729.
Tohma, K., Lepore, C.J., Gao, Y., Ford-Siltz, L.A., Parra, G.I., 2019. Population genomics of GII.4 noroviruses reveal complex diversification and new antigenic sites involved in the emergence of pandemic strains. MBio 10, e02202–e02219 bioRxiv 668772.
Vinod, H.D., 1969. Integer programming and the theory of grouping. J. Am. Stat. Assoc. 64, 506–519.
Wakeley, J., 2009. Coalescent Theory. Roberts & Company Publishers, Colorado.
Wolf, Y.I., Viboud, C., Holmes, E.C., Koonin, E.V., Lipman, D.J., 2006. Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. Biol. Direct 1, 34.
Yonezawa, K., Igarashi, M., Ueno, K., Takada, A., Ito, K., 2013. Resampling nucleotide sequences with closest-neighbor trimming and its comparison to other methods. PLoS One 8, e57684.
Zaslavsky, L., Bao, Y., Tatusova, T.A., 2008. Visualization of large influenza virus sequence datasets using adaptively aggregated trees with sampling-based subscale representation. BMC Bioinforma. 9, 237.
Zhang, Y., Aevermann, B.D., Anderson, T.K., Burke, D.F., Dauphin, G., Gu, Z., He, S., Kumar, S., Larsen, C.N., Lee, A.J., Li, X., Macken, C., Mahaffey, C., Pickett, B.E., Reardon, B., Smith, T., Stewart, L., Suloway, C., Sun, G., Tong, L., Vincent, A.L., Walters, B., Zaremba, S., Zhao, H., Zhou, L., Zmasek, C., Klem, E.B., Scheuermann, R.H., 2017. Influenza Research Database: an integrated bioinformatics resource for influenza virus research. Nucleic Acids Res. 45, D466–D474.