The above approach, however, appears to contain problems that may result in mis-identification of homologous codons. First, it is implicitly assumed that the unit of insertions and deletions (indels) in protein-coding nucleotide sequences is the codon (Fletcher and Yang, 2010), which is not always the case in reality because frame-shifts can occur at least partially (Mills et al., 2006). When a frame-shift occurs at some codons in a sequence, they are no longer homologous to codons in other sequences. Yet, non-homologous codons may be identified as homologous in the above approach, because it is difficult to infer the occurrence of frame-shifts only from the comparison of amino acid sequences. Second, even when the unit of indels was the codon, aligning amino acid sequences itself is not always easy, and non-homologous amino acids may be aligned especially at variable sites, which may lead to mis-identification of homologous codons (Liu et al., 2009; Fletcher and Yang, 2010). It has been reported that excluding the sites with gaps from the alignment of amino acid sequences was insufficient to reduce alignment errors in the real data analysis, suggesting that mis-alignment of amino acid sequences may be common (Wong et al., 2008; Fletcher and Yang, 2010).

In general, majority of amino acid sites in proteins are under functional constraint with $d_N/d_S < 1$ (Suzuki and Gojobori, 2001; Suzuki, 2006). However, the d_N/d_S ratio is known to be inflated at mis-aligned codon sites (Wong et al., 2008; Mallick et al., 2009; Schneider et al., 2009). It has also been reported that as more non-homologous codons are aligned, more amino acid sites are falsely identified as positively selected (Vamathevan et al., 2008; Wong et al., 2008; Mallick et al., 2009; Schneider et al., 2009; Fletcher and Yang, 2010). The purpose of the present study was to examine the effect of reverse-translating aligned amino acid sequences on the estimation of d_N/d_S ratio, through a large-scale analysis of protein-coding nucleotide sequences from vertebrate species.

MATERIALS AND METHODS

Sequence data The entire sets of protein-coding nucleotide sequences for 10 vertebrate species (human [*Homo sapiens*; GRCh37], chimpanzee [*Pan troglodytes*; CHIMP2.1], orangutan [*Pongo pygmaeus abelii*; PPYG2], macaque [*Macaca mulatta*; MMUL_1.0], mouse [*Mus musculus*; NCBIM37], cow [*Bos taurus*; Btau_4.0], opossum [*Monodelphis domestica*; monDom5], chicken [*Gallus gallus*; WASHUC2], frog [*Xenopus tropicalis*; JGI4.1], and zebrafish [*Danio rerio*, Zv8]) (Nei et al., 2010) were retrieved from Ensembl Genes 57 through BioMart (Durinck et al., 2005). Distantly related species, such as chicken, frog, and zebrafish, were included in the analysis of d_N/d_S ratio in the present study for the following reasons. First, the d_S and d_N values did not appear to be saturated but increased linearly along with time for these species (Nei et al., 2010). Second, even for distantly related species, the d_S and d_N values are considered to be estimated reliably when the number of codon sites analyzed is relatively large (Nei and Kumar, 2000). Third, natural selection has been detected based on the d_N/d_S ratio even when distantly related species, such as chicken and frog, were included (Uddin et al., 2008; Goodman et al., 2009; Goodman and Sterner, 2010).

The possible orthology data for human sequences to sequences of other vertebrate species were also available in BioMart. It should be noted, however, that the possible orthology data in BioMart were generated based on the topology of the phylogenetic tree constructed from the multiple alignment of nucleotide sequences that was obtained by reverse-translating aligned amino acid sequences (Vilella et al., 2009). If homologous nucleotides were mis-identified in this process, the number of nucleotide substitutions may be overestimated for some pairs of sequences, which may result in construction of incorrect topology. It was therefore possible that orthologues were identified as paralogues, and *vice versa*. Although the identification of paralogues as orthologues may be problematic in the present study, the probability for the occurrence of mis-identification appears to be small, because it is unlikely that a particular topology (species tree) is generated by random effects. In addition, the probability may be further reduced by focusing only on one-to-one possible orthologues between species.

Data processing Using the possible orthology data, a list of one-to-one possible orthologues was generated between human and other vertebrate species. Nine lists of one-to-one possible orthologues obtained were combined using human sequences as the reference, to generate 4,313 sets of possible orthologues that were shared by 10 vertebrate species. The sets of possible orthologues whose member sequence contained a premature termination codon or an ambiguous nucleotide were discarded, and 3,878 sets of possible orthologues were retained for the next step.

For each of 3,878 sets of possible orthologues, multiple alignments of amino acid and nucleotide sequences for 10 vertebrate species were made by using the computer program CLUSTAL W (version 1.8) (Thompson et al., 1994) with the default parameter settings. The alignment of amino acid sequences was reverse-translated into codon sequences, and the alignment of codon sequences obtained was compared with the alignment of nucleotide sequences. The codon sites that were aligned consistently in these alignments for all of 10 vertebrate species (class-1 codon sites) were extracted to construct another alignment of codon sequences. It should be noted that the class-1 sites represent the codon sites that were aligned consistently using the amino acid and nucleotide

sequences. The sets of possible orthologues for which the number of class-1 codon sites was < 100 were discarded to reduce the possibility that they do not encode a real protein, and 3,325 sets of possible orthologues were retained for the next step.

For each of 3,325 sets of possible orthologues, two alignments of codon sequences generated above, by reverse-translating aligned amino acid sequences and by extracting class-1 sites, were used for estimating the d_S and d_N values between human and other vertebrate species by the method of Nei and Gojobori (1986) taking into account the transition/transversion rate ratio (Kondo et al., 1993; Zhang et al., 1998; Suzuki et al., 2009), which has been estimated to be 4 in mammals (Rosenberg et al., 2003; Jiang and Zhao, 2006; Zhang et al., 2007). The codon sites shared by all of 10 vertebrate species without gaps were used for the estimation. The sets of possible orthologues for which the d_S or d_N value between human and any of other vertebrate species was incalculable were discarded to reduce the possibility that they contained paralogous sequences. Finally, the remaining 3,222 sets of possible orthologues were considered as orthologues and used for the analysis of d_N/d_S ratio.

Analysis of d_N/d_S ratio Two alignments of codon sequences generated above, by reverse-translating aligned amino acid sequences and by extracting class-1 sites, were concatenated, after eliminating the codon sites with gaps, for 3,222 sets of orthologues to make the alignments of codon sequences with 1,318,081 codon sites and 1,128,326 codon sites, respectively. Using these alignments, the d_S and d_N values as well as the d_N/d_S ratio were estimated between human and other vertebrate species, as described above.

RESULTS AND DISCUSSION

d_N/d_S ratio by reverse-translating aligned amino acid sequences The d_S and d_N values as well as the d_N/d_S ratio estimated between human and other vertebrate species using the alignment of codon sequences constructed by reverse-translating aligned amino acid sequences are summarized in Table 1. The d_N/d_S ratio between human and non-human primates ranged from 0.260 to 0.272 (The Chimpanzee Sequencing and Analysis Consortium, 2005; Bakewell et al., 2007; Rhesus Macaque Genome Sequencing and Analysis Consortium, 2007). In contrast, smaller d_N/d_S ratios, ranging from 0.131 to 0.217, were observed between human and non-primate mammals and non-mammalian vertebrates (Mouse Genome Sequencing Consortium, 2002; International Chicken Genome Sequencing Consortium, 2004; Rat Genome Sequencing Project Consortium, 2004).

It should be noted that the effect of natural selection is positively correlated with the effective population size of organisms (Kimura, 1983). Therefore, the difference in the d_N/d_S ratio observed above appears to reflect the fact that effective population sizes of primates are smaller than those of other vertebrate species analyzed in the present study. In fact, the effective population size has been estimated to be ~10,000 for human (Takahata, 1993), ~25,000 for chimpanzee (Won and Hey, 2005), ~15,000 for orangutan (Becquet and Przeworski, 2007), and ~25,000 for macaque (Bonhomme et al., 2009), which are smaller than the estimates of ~400,000 for mouse (Geraldes et al., 2008) and ~90,000 for cow (MacEachern

Table 1. The d_S and d_N values and the d_N/d_S ratio between human and other vertebrate species

Species	Reverse-translated[a]			Class-1[b]			Class-2-1[c]			Class-2-2[d]			Estimated[e]		
	d_S	d_N	d_N/d_S	d_S	d_N	d_N/d_S	d_S	d_N	d_N/d_S	d_S	d_N	d_N/d_S	d_S	d_N	d_N/d_S
Chimpanzee	0.0125[f]	0.00337	0.270	0.0114	0.00195	0.171	0.0122	0.00451	0.369	0.0989	0.101	1.02	0.0115	0.00232	0.201
Orangutan	0.0354	0.00921	0.260	0.0321	0.00496	0.154	0.0339	0.0114	0.337	0.265	0.277	1.05	0.0324	0.00589	0.182
Macaque	0.0678	0.0185	0.272	0.0616	0.0101	0.164	0.0635	0.0223	0.351	0.478	0.525	1.10	0.0619	0.0118	0.191
Mouse	0.436	0.0571	0.131	0.431	0.0430	0.0997	0.442	0.124	0.280	0.668	0.375	0.561	0.433	0.0541	0.125
Cow	0.311	0.0497	0.160	0.303	0.0359	0.118	0.320	0.0962	0.301	0.654	0.499	0.762	0.306	0.0443	0.145
Opossum	0.666	0.105	0.158	0.655	0.0783	0.120	0.675	0.209	0.310	1.04	0.797	0.764	0.658	0.0957	0.146
Chicken	0.909	0.158	0.173	0.899	0.121	0.135	0.907	0.286	0.315	1.18	0.992	0.844	0.900	0.143	0.159
Frog	1.27	0.232	0.183	1.27	0.177	0.140	1.20	0.382	0.319	1.34	1.34	1.00	1.26	0.203	0.162
Zebrafish	1.39	0.303	0.217	1.40	0.238	0.171	1.30	0.369	0.284	1.42	1.48	1.04	1.38	0.256	0.185

[a]Alignment of codon sequences was constructed by reverse-translating aligned amino acid sequences.
[b]Alignment of codon sequences was constructed by extracting the codon sites that were aligned consistently using the amino acid and nucleotide sequences for all of 10 vertebrate species.
[c]Codon sites that were aligned inconsistently using the amino acid and nucleotide sequences for any of 10 vertebrate species but consistently for the pairwise comparison of human and other vertebrate species.
[d]Codon sites that were aligned inconsistently using the amino acid and nucleotide sequences for the pairwise comparison of human and other vertebrate species.
[e]Estimated values of d_S and d_N as well as the d_N/d_S ratio by correcting the d_N/d_S ratio for the class-2-2 sites.
[f]Standard errors were mostly more than two orders of magnitude smaller than the estimates.

et al., 2009). These observations suggest that functional constraint has operated less effectively in primates compared to other vertebrate species.

d_N/d_S ratio by extracting the codon sites aligned consistently using the amino acid and nucleotide sequences In the above analysis, the alignment of codon sequences was constructed by reverse-translating aligned amino acid sequences. In this method, however, non-homologous codons may be aligned when frame-shifts occurred or amino acid sequences were mis-aligned, which may lead to overestimation of the d_N/d_S ratio, as discussed above. It may be difficult to measure the degree of overestimation accurately in the real data analysis, because the correct alignment of codon sequences is usually unknown. However, the codon sites that are

Table 2. The numbers of synonymous and nonsynonymous sites and differences, and the proportions of different sites between human and other vertebrate species

Species		Reverse-translated[a]			Class-1[b]			Class-2-1[c]			Class-2-2[d]		
		Site	Difference	Proportion	Site	Difference	Proportion	Site	Difference	Proportion	Site	Difference	Proportion
Chimpanzee	Synonymous	1,119,714	13,870	0.0124[f]	956,930	10,831	0.0113	149,568	1,814	0.0121	13,215	1,225	0.0927
		(1.00)[e]	(1.00)		(0.855)	(0.781)		(0.134)	(0.131)		(0.0118)	(0.0883)	
	Nonsynonymous	2,697,860	9,075	0.00336	2,310,381	4,503	0.00195	355,994	1,601	0.00450	31,485	2,972	0.0944
		(1.00)	(1.00)		(0.856)	(0.496)		(0.132)	(0.176)		(0.0117)	(0.327)	
Orangutan	Synonymous	1,119,651	38,722	0.0346	956,874	30,100	0.0315	145,807	4,839	0.0332	16,971	3,784	0.223
		(1.00)	(1.00)		(0.855)	(0.777)		(0.130)	(0.125)		(0.0152)	(0.0977)	
	Nonsynonymous	2,697,941	24,705	0.00916	2,310,419	11,430	0.00495	347,169	3,942	0.0114	40,353	9,332	0.231
		(1.00)	(1.00)		(0.856)	(0.463)		(0.129)	(0.160)		(0.0150)	(0.378)	
Macaque	Synonymous	1,119,917	72,613	0.0648	957,001	56,605	0.0591	142,108	8,649	0.0609	20,808	7,358	0.354
		(1.00)	(1.00)		(0.855)	(0.780)		(0.127)	(0.119)		(0.0186)	(0.101)	
	Nonsynonymous	2,697,811	49,235	0.0183	2,310,356	23,146	0.0100	338,014	7,421	0.022	49,442	18,669	0.378
		(1.00)	(1.00)		(0.856)	(0.47)		(0.125)	(0.151)		(0.0183)	(0.379)	
Mouse	Synonymous	1,121,273	370,520	0.330	958,166	314,142	0.328	145,223	48,473	0.334	17,884	7,906	0.442
		(1.00)	(1.00)		(0.855)	(0.848)		(0.130)	(0.131)		(0.0160)	(0.0213)	
	Nonsynonymous	2,697,505	148,365	0.055	2,310,105	96,527	0.0418	345,144	39,373	0.114	42,257	12,465	0.295
		(1.00)	(1.00)		(0.856)	(0.651)		(0.128)	(0.265)		(0.0157)	(0.0840)	
Cow	Synonymous	1,121,675	285,406	0.254	958,461	239,172	0.250	141,899	36,929	0.26	21,316	9,305	0.437
		(1.00)	(1.00)		(0.854)	(0.838)		(0.127)	(0.129)		(0.0190)	(0.0326)	
	Nonsynonymous	2,697,163	129,798	0.0481	2,309,825	81,054	0.0351	337,044	30,428	0.0903	50,294	18,316	0.364
		(1.00)	(1.00)		(0.856)	(0.624)		(0.125)	(0.234)		(0.0186)	(0.141)	
Opossum	Synonymous	1,116,628	492,847	0.441	954,113	416,835	0.437	131,183	58,364	0.445	31,333	17,647	0.563
		(1.00)	(1.00)		(0.854)	(0.846)		(0.117)	(0.118)		(0.0281)	(0.0358)	
	Nonsynonymous	2,699,964	265,535	0.0983	2,312,037	171,929	0.0744	313,841	57,252	0.182	74,086	36,355	0.491
		(1.00)	(1.00)		(0.856)	(0.647)		(0.116)	(0.216)		(0.0274)	(0.137)	
Chicken	Synonymous	1,117,171	588,412	0.527	954,779	499,984	0.524	118,277	62,244	0.526	44,115	26,184	0.594
		(1.00)	(1.00)		(0.855)	(0.85)		(0.106)	(0.106)		(0.0395)	(0.0445)	
	Nonsynonymous	2,701,419	384,051	0.142	2,313,190	259,366	0.112	284,564	67,644	0.238	103,666	57,040	0.550
		(1.00)	(1.00)		(0.856)	(0.675)		(0.105)	(0.176)		(0.0384)	(0.149)	
Frog	Synonymous	1,113,482	680,724	0.611	951,870	582,089	0.612	87,910	52,603	0.598	73,702	46,033	0.625
		(1.00)	(1.00)		(0.855)	(0.855)		(0.0790)	(0.0773)		(0.0662)	(0.0676)	
	Nonsynonymous	2,704,662	539,354	0.199	2,315,566	365,360	0.158	212,404	63,605	0.299	176,693	110,388	0.625
		(1.00)	(1.00)		(0.856)	(0.677)		(0.0785	(0.118)		(0.0653)	(0.205)	
Zebrafish	Synonymous	1,119,773	708,664	0.633	957,172	606,291	0.633	58,697	36,209	0.617	103,904	66,164	0.637
		(1.00)	(1.00)		(0.855)	(0.856)		(0.0524)	(0.0511)		(0.0928)	(0.0934)	
	Nonsynonymous	2,700,835	672,851	0.249	2,312,432	471,966	0.204	140,763	40,990	0.291	247,639	159,894	0.646
		(1.00)	(1.00)		(0.856)	(0.701)		(0.0521)	(0.0609)		(0.0917)	(0.238)	

[a] Alignment of codon sequences was constructed by reverse-translating aligned amino acid sequences.
[b] Alignment of codon sequences was constructed by extracting the codon sites that were aligned consistently using the amino acid and nucleotide sequences for all of 10 vertebrate species.
[c] Codon sites that were aligned inconsistently using the amino acid and nucleotide sequences for any of 10 vertebrate species but consistently for the pairwise comparison of human and other vertebrate species.
[d] Codon sites that were aligned inconsistently using the amino acid and nucleotide sequences for the pairwise comparison of human and other vertebrate species.
[e] Proportions of sites and differences in the alignment of codon sequences constructed by reverse-translating aligned amino acid sequences.
[f] Standard errors were mostly more than two orders of magnitude smaller than the estimates.

aligned consistently using the amino acid and nucleotide sequences for all of 10 vertebrate species (class-1 codon sites) may be more likely to be composed of homologous codons than those that are aligned inconsistently for any of 10 vertebrate species (class-2 codon sites).

Therefore, another alignment of codon sequences was constructed by extracting the class-1 sites, and the d_S and d_N values as well as the d_N/d_S ratio were estimated between human and other vertebrate species (Table 1). Compared to the case for reverse-translating aligned amino acid sequences, the d_N/d_S ratio for the class-1 sites dropped to 0.154–0.171 between human and other primates and 0.0997–0.171 between human and non-primate mammals and non-mammalian vertebrates. These results indicate that the d_N/d_S ratio was large for the class-2 sites. However, it should be noted that the codon sites under weak functional constraint or positive selection, where the d_N/d_S ratio is intrinsically high, are more difficult to be aligned compared to those under strong functional constraint, where the d_N/d_S ratio is low (Liu et al., 2009; Fletcher and Yang, 2010). Therefore, the large d_N/d_S ratio for the class-2 sites may be due to mis-alignment of homologous codons or intrinsically high d_N/d_S ratio.

Overestimation of d_N/d_S ratio by reverse-translation of aligned amino acid sequences To distinguish the above possibilities, the class-2 codon sites were further classified into those that were aligned consistently (class-2-1 codon sites) and inconsistently (class-2-2 codon sites) using the amino acid and nucleotide sequences for the pairwise comparison of human and other vertebrate species. The d_S and d_N values as well as the d_N/d_S ratio were estimated for these classes of sites separately (Table 1). For the class-2-1 sites, the d_S value was similar to that obtained for the class-1 sites, which were considered to be composed of homologous codons, suggesting that the codons in the class-2-1 sites were largely homologous. However, the d_N/d_S ratio for the class-2-1 sites was greater than that for the class-1 sites, indicating that the former and latter sites were relatively variable and conservative at the amino acid sequence level, respectively. In contrast, for the class-2-2 sites, the d_S value was much greater than that for the class-1 and class-2-1 sites, suggesting that non-homologous codons were included in this class of sites. The d_N/d_S ratio for the class-2-2 sites was also unduly high.

It should be noted, however, that the proportion of class-2-2 sites in the entire alignment of codon sequences constructed by reverse-translating aligned amino acid sequences was only 1–9% (Table 2). The proportion appeared to be positively correlated with the sequence divergence between vertebrate species, reflecting the fact that aligning sequences is more difficult when sequences are more variable (Liu et al., 2009; Fletcher and Yang, 2010). To examine the effect of class-2-2 sites on the estimation of d_N/d_S ratio, the actual ratio for the entire alignment was estimated by correcting the ratio for the class-2-2 sites, under the assumption that the ratio for this class of sites was similar to that for the class-2-1 sites. This assumption is based on the fact that class-2-1 sites in some pairs of vertebrate species may be classified as class-2-2 sites in other pairs, and *vice versa*. It was observed that the d_N/d_S ratio obtained without correction was 5–43% greater than that obtained with correction (Table 1). The uncorrected ratio was still 0.3–39% greater than the corrected ratio even when the d_N/d_S ratio for the class-2-2 sites was assumed to be twice as great as that for the class-2-1 sites (data not shown). The degree of overestimation for the d_N/d_S ratio appeared to be positively correlated with the ratios of the numbers of synonymous and nonsynonymous differences for the class-2-2 sites to those for other classes of sites (Table 2).

These results suggest that even if the proportion of mis-aligned codon sites is small, they cause significant overestimation of the d_N/d_S ratio for the entire alignment of codon sequences constructed by reverse-translating aligned amino acid sequences (Wong et al., 2008; Mallick et al., 2009; Schneider et al., 2009). These codon sites may also be falsely identified as positively selected (Vamathevan et al., 2008; Wong et al., 2008; Mallick et al., 2009; Schneider et al., 2009; Fletcher and Yang, 2010). Therefore, caution should be exerted in the study of natural selection using the d_N/d_S ratio by reverse-translating aligned amino acid sequences. It may be necessary to add information from nucleotide sequences to that from amino acid sequences for constructing reliable alignments of codon sequences (Fletcher and Yang, 2010). In addition, since the alignment of codon sequences is not an observation but an inference, it may also be useful to take into account alignment errors for obtaining reliable estimates of the d_N/d_S ratio (Wong et al., 2008).

The author thanks Masafumi Nozawa for technical comments on the retrieval and processing of protein-coding nucleotide sequences for preparing orthologues from 10 vertebrate species analyzed in the present study. The author is also indebted to Yuki Kobayashi and two anonymous reviewers for valuable comments. The present study was supported by KAKENHI 20570008.

REFERENCES

Bakewell, M. A., Shi, P., and Zhang, J. (2007) More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc. Natl. Acad. Sci. USA **104**, 7489–7494.

Becquet, C., and Przeworski, M. (2007) A new approach to estimate parameters of speciation models with application to apes. Genome Res. **17**, 1505–1519.

Bonhomme, M., Cuartero, S., Blancher, A., and Crouau-Roy, B. (2009) Assessing natural introgression in 2 biomedical model species, the rhesus macaque (*Macaca mulatta*) and

the long-tailed macaque (*Macaca fascicularis*). J. Hered. **100**, 158–169.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics **21**, 3439–3440.

Fletcher, W., and Yang, Z. (2010) The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol. Biol. Evol. **27**, 2257–2267.

Geraldes, A., Basset, P., Gibson, B., Smith, K. L., and Harr, B. (2008) Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. Mol. Ecol. **17**, 5349–5363.

Goodman, M., and Sterner, K. N. (2010) Phylogenomic evidence of adaptive evolution in the ancestry of humans. Proc. Natl. Acad. Sci. USA **107**, 8918–8923.

Goodman, M., Sterner, K. N., Islam, M., Uddin, M., Sherwood, C. C., Hof, P. R., Hou, Z.-C., Lipovich, L., Jia, H., Grossman, L. I., and Wildman, D. E. (2009) Phylogenomic analyses reveal convergent patterns of adaptive evolution in elephant and human ancestries. Proc. Natl. Acad. Sci. USA **106**, 20824–20829.

Hughes, A. L., and Nei, M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335**, 167–170.

International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature **432**, 659–716.

Jiang, C., and Zhao, Z. (2006) Mutational spectrum in the recent human genome inferred by single nucleotide polymorphisms. Genomics **88**, 527–534.

Kimura, M. (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature **267**, 275–276.

Kimura, M. (1983) The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, New York, Melbourne.

Kondo, R., Horai, S., Satta, Y., and Takahata, N. (1993) Evolution of hominoid mitochondrial DNA with special reference to the silent substitution rate over the genome. J. Mol. Evol. **36**, 517–531.

Liu, K., Raghavan, S., Nelesen, S., Linder, C. R., and Warnow, T. (2009) Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. Science **324**, 1561–1564.

MacEachern, S., Hayes, B., McEwan, J., and Goddard, M. (2009) An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (*Bos taurus*) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in Domestic cattle. BMC Genomics **10**, 181.

Mallick, S., Gnerre, S., Muller, P., and Reich, D. (2009) The difficulty of avoiding false positives in genome scans for natural selection. Genome Res. **19**, 922–933.

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., and Devine, S. E. (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. **16**, 1182–1190.

Miyata, T., and Yasunaga, T. (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J. Mol. Evol. **16**, 23–36.

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature **420**, 520–562.

Nei, M., and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**, 418–426.

Nei, M., and Kumar, S. (2000) Molecular Evolution and Phylogenetics. Oxford University Press, Oxford, New York.

Nei, M., Suzuki, Y., and Nozawa, M. (2010) The neutral theory of molecular evolution in the genomic era. Annu. Rev. Genomics Hum. Genet. **11**, 265–289.

Nielsen, R., and Yang, Z. (2003) Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. Mol. Biol. Evol. **20**, 1231–1239.

Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R., and Dodgson, J. (1980) The evolution of genes: the chicken preproinsulin gene. Cell **20**, 555–566.

Rat Genome Sequencing Project Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature **428**, 493–521.

Rhesus Macaque Genome Sequencing and Analysis Consortium (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science **316**, 222–234.

Rosenberg, M. S., Subramanian, S., and Kumar, S. (2003) Patterns of transitional mutation biases within and among mammalian genomes. Mol. Biol. Evol. **20**, 988–993.

Schneider, A., Souvorov, A., Sabath, N., Landan, G., Gonnet, G. H., and Graur, D. (2009) Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. Genome Biol. Evol. **1**, 114–118.

Suyama, M., Torrents, D., and Bork, P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. **34**, W609–W612.

Suzuki, Y. (2006) Natural selection on the influenza virus genome. Mol. Biol. Evol. **23**, 1902–1911.

Suzuki, Y., and Gojobori, T. (2001) Positively selected amino acid sites in the entire coding region of hepatitis C virus subtype 1b. Gene **276**, 83–87.

Suzuki, Y., and Gojobori, T. (2003) Analysis of coding sequences. In: The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny (eds.: M. Salemi, and A.-M. Vandamme), pp. 283–311. Cambridge University Press, Cambridge.

Suzuki, Y., Gojobori, T., and Kumar, S. (2009) Methods for incorporating the hypermutability of CpG dinucleotides in detecting natural selection operating at the amino acid sequence level. Mol. Biol. Evol. **26**, 2275–2284.

Takahata, N. (1993) Allelic genealogy and human evolution. Mol. Biol. Evol. **10**, 2–22.

The Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437**, 69–87.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple-sequence alignment through sequence weighting, position-specific gap penalties, and weight-matrix choice. Nucleic Acids Res. **22**, 4673–4680.

Uddin, M., Goodman, M., Erez, O., Romero, R., Liu, G., Islam, M., Opazo, J. C., Sherwood, C. C., Grossman, L. I., and Wildman, D. E. (2008) Distinct genomic signatures of adaptation in pre- and postnatal environments during human evolution. Proc. Natl. Acad. Sci. USA **105**, 3215–3220.

Vamathevan, J. V., Hasan, S., Emes, R. D., Amrine-Madsen, H., Rajagopalan, D., Topp, S. D., Kumar, V., Word, M.,

Simmons, M. D., Foord, S. M., et al. (2008) The role of positive selection in determining the molecular cause of species differences in disease. BMC Evol. Biol. **8**, 273.

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. **19**, 327–335.

Won, Y.-J., and Hey, J. (2005) Divergence population genetics of chimpanzees. Mol. Biol. Evol. **22**, 297–307.

Wong, K. M., Suchard, M. A., and Huelsenbeck, J. P. (2008) Alignment uncertainty and genomic analysis. Science **319**, 473–476.

Zhang, J., Rosenberg, H. F., and Nei, M. (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. USA **95**, 3708–3713.

Zhang, W., Bouffard, G. G., Wallace, S., Bond, J. P., and NISC Comparative Sequencing Program (2007) Estimation of DNA sequence context-dependent mutation rates using primate genomic sequences. J. Mol. Evol. **65**, 207–214.

Genes Genet. Syst. (2011) **86**, p. 287–294

Positive selection for gains of N-linked glycosylation sites in hemagglutinin during evolution of H3N2 human influenza A virus

Yoshiyuki Suzuki*

Graduate School of Natural Sciences, Nagoya City University, 1 Yamanohata, Mizuho-cho, Mizuho-ku, Nagoya-shi, Aichi-ken 467-8501, Japan

(Received 9 October 2011, accepted 15 November 2011)

The number of N-linked glycosylation sites in the globular head of hemagglutinin (HA) has increased during evolution of H3N2 human influenza A virus. Here natural selection operating on the gains of N-linked glycosylation sites was examined by using the single-site analysis and the single-substitution analysis. In the single-site analysis, positive selection was not inferred at the amino acid sites where the substitutions generating N-linked glycosylation sites were observed, but was detected at antigenic sites. In contrast, in the single-substitution analysis, positive selection was detected for the amino acid substitutions generating N-linked glycosylation sites. The single-site analysis and the single-substitution analysis appeared to be suitable for detecting recurrent and episodic natural selection, respectively. The gains of N-linked glycosylation sites were likely to be positively selected for the function of shielding antigenic sites from immune responses. At the antigenic sites, positive selection appeared to have operated not only on the radical substitution but also on the conservative substitution in terms of the charge of amino acids, suggesting that the antigenic drift is not a by-product of the evolution of receptor binding avidity in HA of human H3N2 virus.

Key words: influenza A virus, N-linked glycosylation site, positive selection, single-site analysis, single-substitution analysis

INTRODUCTION

Influenza A virus is classified as the genus *Influenzavirus A* in the family *Orthomyxoviridae* (Carstens, 2010). The genome of this virus is an eight-segmented and negative-stranded RNA, encoding envelope glycoproteins, matrix proteins, nonstructural proteins, nucleoproteins, and polymerase subunits. The envelope glycoproteins include hemagglutinin (HA) and neuraminidase (NA), with HA existing 4–5 times more abundantly than NA in virions. According to the antigenicity of HA and NA, influenza A virus is classified into subtypes H1–H16 and N1–N9, respectively (World Health Organization, 1980).

Influenza A virus is an etiological agent of influenza (Shope, 1931). Genomic sequence data of this virus are available for the strains circulating since 1918. In the human population, H1N1 virus circulated in 1918–1956, followed by H2N2 virus in 1957–1967. Subsequently, H3H2 virus has been circulating in 1968-present. In addition, H1N1 virus closely related to that observed in early 1950s reappeared in 1977 and circulated thereafter, and another H1N1 virus (A(H1N1)pdm09) emerged in 2009 (World Health Organization, 2011). A(H1N1)pdm09 virus has been circulating as the predominant H1N1 virus in 2009-present.

HA is a homotrimeric type I transmembrane glycoprotein. The HA gene encodes HA0, consisting of a signal peptide (amino acid sites [–16]–[–1] in H3N2 virus), HA1 (1–328), and HA2 (330–550) (Skehel and Wiley, 2000). Signal peptide directs the co-translational transport of HA into the endoplasmic reticulum (ER). HA1 is the sialic acid receptor-binding protein and the major target of humoral immunity. HA2 is an anchor protein to the envelope and mediates fusion of the envelope and the endosomal membrane.

The ectodomain of HA is composed of the globular head (amino acid sites 58–272 of HA1) and the fibrous stem (1–57 and 273–328 of HA1 and 330–514 of HA2) (Wilson et al., 1981). Antigenic sites are distributed in the globular head, constituting 5 epitopes (A–E) in H3N2 virus (Suzuki, 2004b). In the ectodomain, asparagine [N] of the sequon, which is defined as the sequence of N, any amino acid except for proline [P], and serine [S] or threo-

Edited by Fumio Tajima

* Corresponding author. E-mail: yossuzuk@nsc.nagoya-cu.ac.jp

nine [T], mostly serves as an N-linked glycosylation site, where high-mannose or complex oligosaccharide is attached with various forms (Hebert et al., 1997; Schulze, 1997; Daniels et al., 2003; Abe et al., 2004; Blackburne et al., 2008; Das et al., 2010). O-linked glycosylation has not been reported for influenza A virus.

The N-linked glycan is involved in the folding of ectodomain in the ER lumen by binding to lectin chaperones (Hebert et al., 1997; Daniels et al., 2003; Cui et al., 2009). N-linked glycans in the fibrous stem are involved in the fusion activity of HA. In the globular head, N-linked glycosylation sites usually overlap with antigenic sites. N-linked glycans may be involved in shielding of antigenic sites from binding by antibodies (Skehel et al., 1984; Tsuchiya et al., 2002; Das et al., 2010; Wei et al., 2010; Wanzeck et al., 2011) and major histocompatibility complex (Jackson et al., 1994), interference with proteolytic activity of HA, and recognition by collectins for neutralization (Vigerust et al., 2007). In addition, structural complexity of N-linked glycans is positively and negatively correlated with HA-receptor binding specificity and affinity, respectively (Tsuchiya et al., 2002; Wang et al., 2009; de Vries et al., 2010).

N-linked glycosylation sites in the fibrous stem are usually conserved among influenza A viruses (Sun et al., 2011). However, in the globular head of H1N1 virus, the number of N-linked glycosylation sites, which was 1 in 1918, increased up to 6 and is 4 at present (Igarashi et al., 2008; Das et al., 2010; Sun et al., 2011). Although the number of N-linked glycosylation sites remained 1 in 1957–1967 for H2N2 virus (Tsuchiya et al., 2001, 2002; Abe et al., 2004), the number increased from 2 to 6 or 7 in 1968-present for H3N2 virus (Abe et al., 2004).

It was unclear whether the increase in the number of N-linked glycosylation sites observed in human H1N1 and H3N2 viruses was due to neutral evolution (Zhang et al., 2004) or positive selection. In general, the number of N-linked glycosylation sites is negatively correlated with the GC-content, because N, which is included in the sequon, is encoded by GC-poor (AAY) codons and P, which is not included in the sequon, is encoded by GC-rich (CCN) codons (Cui et al., 2009). Although the GC-content of influenza A virus has decreased in humans because of mutation bias (Rabadan et al., 2006) and natural selection against CpG dinucleotides (Greenbaum et al., 2008), the observed increase in the number of N-linked glycosylation sites appeared to exceed the expectation from the decrease in the GC-content (Cui et al., 2009).

In the phylogenetic tree for HA of H3N2 virus, it was observed that the gain/loss ratio of N-linked glycosylation sites was greater in the branches more proximal to the root (i.e., trunk branches > non-trunk interior branches > non-trunk exterior branches) (Cherry et al., 2009). From this observation, positive selection was inferred for gains of N-linked glycosylation sites. However, since the occurrences of gains and losses are interdependent (e.g., N-linked glycosylation sites should be gained in proximal branches to be lost in distal branches), the null hypothesis of the equal gain/loss ratio for proximal and distal branches without positive selection may not hold.

Positive selection was also inferred for the amino acid substitutions generating N-linked glycosylation sites by DEPS, which is the method for detecting directional evolution of protein sequences (Kosakovsky Pond et al., 2008). In this method, asymmetrical substitution rates between pairs of amino acids are identified at each site, assuming that the pattern of substitution is the same at all sites without positive selection, which appears to be unrealistic. For detecting asymmetrical substitution rates, it is required that a number of substitutions occurred between pairs of amino acids at single sites. DEPS has been reported to produce many false-positives (Nozawa et al., 2009).

The purpose of the present study was to examine natural selection operating on the gains of N-linked glycosylation sites in the globular head of HA during evolution of human H3N2 virus by using the single-site analysis and the single-substitution analysis. The human H3N2 virus was analyzed because of the clinical importance as currently circulating in the human population, and of the availability of sequence data containing a sufficiently large amount of genetic variation for the statistical analysis.

MATERIALS AND METHODS

Sequence data A total of 3,206 nucleotide sequences for the entire protein-coding region of HA for human H3N2 virus, excluding laboratory and vaccine strains, were retrieved from the Influenza Virus Resource at the National Center for Biotechnology Information (Bao et al., 2008) as of May 19, 2011. After eliminating sequences for the same strains as others, sequences identical to others, sequences derived from incidental human infections of swine strains, and sequences with minor gaps, ambiguous nucleotides, and premature termination codons, 2,043 sequences were used in the following analysis. A sequence from duck H3N8 virus was added as the outgroup to identify the position of the root for the phylogenetic tree of human sequences. Each sequence consisted of 1,688 nucleotide sites.

Phylogenetic analysis Multiple alignment of 2,044 human and duck sequences was made by using the computer program MAFFT (version 6.853b) (Katoh et al., 2002), which did not contain any gaps. Phylogenetic trees were constructed by the neighbor-joining method (Saitou and Nei, 1987) with the p distance (Nei and Kumar, 2000) and the maximum composite likelihood (MCL) distance (Tamura et al., 2004), which are known to produce reliable phylogenetic trees when a large num-

ber of closely related sequences is analyzed, using MEGA (version 5.05) (Tamura et al., 2011). The nucleotide sequence at each interior node of the phylogenetic tree was inferred by the maximum parsimony method (Fitch, 1971; Hartigan, 1973) using PAML (version 4.4b) (Yang, 2007).

Single-site analysis of natural selection: d_N-d_S test Natural selection operating at the amino acid sequence level can be detected by comparing the rates of synonymous (r_S) and nonsynonymous (r_N) substitutions under the assumption that synonymous mutations are neutral or nearly neutral; the relationships $r_S < r_N$, $r_S > r_N$, and $r_S = r_N$ indicate positive, negative, and no selection, respectively (Kimura, 1977; Hughes and Nei, 1988). Nonsynonymous substitutions may be divided into conservative and radical substitutions according to whether they retain or alter a property of amino acids, respectively. If charge is considered, conservative and radical substitutions may be defined as nonsynonymous substitutions within and between charge categories (Hughes et al., 1990). Arginine [R], histidine [H], and lysine [K] are positively charged; aspartic acid [D] and glutamic acid [E] are negatively charged; and N, P, S, T, alanine [A], cysteine [C], glutamine [Q], glycine [G], isoleucine [I], leucine [L], methionine [M], phenylalanine [F], tryptophan [W], tyrosine [Y], and valine [V] are neutral (Arinaminpathy and Grenfell, 2010). Natural selection operating on conservative and radical substitutions can be detected separately by comparing the rates of these substitutions with r_S, in a similar manner to the comparison of r_S and r_N (Hughes et al., 1990; Suzuki, 2007).

For examining natural selection at single amino acid sites of HA, the numbers of synonymous and nonsynonymous differences and sites were computed at single codon sites for each branch of the phylogenetic tree by comparing the nucleotide sequences at the ancestral and descendant nodes (Suzuki and Gojobori, 1999). Here the transition/transversion rate ratio of nucleotide mutation (κ) was required for computing the numbers of synonymous and nonsynonymous sites. Using the ratio of the transitional/transversional nucleotide diversity at 96 four-fold degenerate sites of 2,043 human sequences, κ was estimated to be 4.057. Therefore, κ was assumed to be 4 in the computation. For each codon site, the numbers of synonymous and nonsynonymous differences were summed and the numbers of synonymous and nonsynonymous sites were averaged with the weight proportional to the branch length over all branches of the phylogenetic tree, to obtain the total numbers of synonymous (c_S) and nonsynonymous (c_N) differences and the average numbers of synonymous (s_S) and nonsynonymous (s_N) sites (Suzuki and Gojobori, 1999; Suzuki, 2004a). Although multiple substitutions were not corrected in this method, the degree of underestimation for c_S and c_N appeared to be negligible for the data set analyzed in the present study, because the branch lengths were generally very small (Saitou, 1989). The total numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions over the phylogenetic tree were computed as c_S/s_S and c_N/s_N, respectively. The r_N/r_S value was estimated as d_N/d_S, and the null hypothesis of no selection ($d_S = d_N$) was tested by computing the probability (p) of obtaining the observed or more biased values for c_S and c_N, which were assumed to follow a binomial distribution with the probabilities of occurrence of synonymous and nonsynonymous substitutions given by $s_S/(s_S + s_N)$ and $s_N/(s_S + s_N)$, respectively.

When the test is conducted for multiple codon sites, it is necessary to correct for multiple testing. In the Bonferroni correction, the family-wise significance level ($\alpha = 0.05$ in the present study) is divided by the number (n) of tests to obtain the corrected significance level (α_c) for individual tests. In this approach, α_c may become unrealistically small when n is large. It should be noted, however, that a number of nucleotide differences ($c_S + c_N$) is required for detecting a significant difference between d_S and d_N at a codon site. For example, if $s_S = 1$ and $s_N = 2$ with $\alpha_c = 0.05$, at least 9 ($c_S = 0$ and $c_N = 9$) and 3 ($c_S = 3$ and $c_N = 0$) nucleotide differences are required for detecting positive and negative selection, respectively (Suzuki, 2008a; Nozawa et al., 2009). In other words, the codon sites with $(c_S + c_N) < 9$ and $(c_S + c_N) < 3$ are not testable for positive and negative selection, respectively, and may be eliminated from the test to reduce n. Since the numbers of nucleotide differences required for detecting positive and negative selection may differ, the tests of positive and negative selection may be conducted separately using different α_c. In the test of positive selection, the probability (p_0) of observing 0 synonymous and $(c_S + c_N)$ nonsynonymous differences or more biased values is computed at each codon site as indicated above. The codon sites are ranked (r) according to p_0 in ascending order, and those with $p_0 < \alpha/r$ are considered to be detectable as positively selected with correction. The test is conducted only for these (n_c) sites; positive selection is inferred when $p < \alpha_c$ ($= \alpha/n_c$) and $d_S < d_N$. Negative selection can be inferred in a similar manner.

Single-site analysis of natural selection: interior-exterior test In the phylogenetic tree of individuals sampled from a population, advantageous and deleterious mutations tend to be accumulated on the branches more proximal and distal to the root, respectively (McDonald and Kreitman, 1991). Therefore, d_N/d_S for proximal branches may be greater and smaller than that for distal branches at the codon sites under positive and negative selection, respectively. Since interior branches are usually more proximal than exterior branches in the phylogenetic tree, positive and negative selection may be inferred when d_N/d_S for the former branches is greater and smaller than that for the latter branches, respectively

(Pybus et al., 2007).

For examining natural selection at single amino acid sites of HA using this approach, c_S, c_N, s_S, and s_N were computed separately for interior (c_{Sint}, c_{Nint}, s_{Sint}, and s_{Nint}) and exterior (c_{Sext}, c_{Next}, s_{Sext}, and s_{Next}) branches. The expected values of c_{Sint}, c_{Nint}, c_{Sext}, and c_{Next} ($E[c_{Sint}]$, $E[c_{Nint}]$, $E[c_{Sext}]$, and $E[c_{Next}]$) were obtained under the null hypothesis of equal d_N/d_S for interior and exterior branches, fixing the total numbers of nucleotide differences for interior and exterior branches. The goodness-of-fit of the null hypothesis was examined by using p for the χ^2 value computed from 4 classes with 1 degree of freedom. The correction for multiple testing can be conducted in a similar manner to the d_N-d_S test. In the test of positive selection, p_0 for the χ^2 value is computed at each codon site under the assumption that the numbers of synonymous and nonsynonymous differences are 0 and c_{Sint} + c_{Nint} for interior branches and c_{Sext} + c_{Next} and 0 for exterior branches, respectively. The codon sites are ranked according to p_0 in ascending order, and those with $p_0 < \alpha/r$ are considered to be detectable as positively selected with correction. The test is conducted only for these sites; positive selection is inferred when $p < \alpha_c$ (= α/n_c) with $c_{Nint} > E[c_{Nint}]$. Negative selection can be inferred in a similar manner.

It should be noted that the χ^2 value may be unreliable when the expected value is < 5 for any class (Sokal and Rohlf, 1995). In this case, c_{Sint} + c_{Next} and c_{Sext} + c_{Nint} may be compared with $E[c_{Sint}]$ + $E[c_{Next}]$ and $E[c_{Sext}]$ + $E[c_{Nint}]$, respectively, to obtain p for the χ^2 value computed from 2 classes with 1 degree of freedom. Positive and negative selection may be inferred when $p < \alpha_c$ (= α/n_c) with (c_{Sext} + c_{Nint}) > ($E[c_{Sext}]$ + $E[c_{Nint}]$) and (c_{Sint} + c_{Next}) > ($E[c_{Sint}]$ + $E[c_{Next}]$), respectively, in a similar manner as above.

Single-substitution analysis of natural selection The fitness effect of amino acid substitutions may be advantageous, deleterious, or neutral. Under the assumption that natural selection operating at each amino acid site did not change to any large extent during evolution, which may be the case for human H3N2 virus with relatively short evolutionary history after transmission into the human population, the fitness effect of reverse substitutions may be deleterious, advantageous, and neutral if that of original substitutions was advantageous, deleterious, and neutral, respectively (Bazykin and Kondrashov, 2011). Therefore, natural selection operating on single amino acid substitutions may be inferred by detecting natural selection operating on reverse substitutions; positive, negative, and no selection are inferred for original substitutions when negative, positive, and no selection are detected for reverse substitutions, respectively (two-tailed test). If the fitness effect of original substitutions is assumed to be only advantageous or neutral (McDonald and Kreitman, 1991), positive and no selection are inferred for original substitutions when negative and no selection are detected for reverse substitutions, respectively (one-tailed test). The two-tailed test is adopted in the present study.

For examining natural selection operating on the amino acid substitutions generating N-linked glycosylation sites in HA, natural selection operating on reverse substitutions was inferred by comparing the rate of this substitution (r_{Nrev}) with r_S, focusing on the branches where the ancestral amino acid was the substituted form in the phylogenetic tree. For example, if an amino acid substitution D → N generated an N-linked glycosylation site, r_S and r_{Nrev} (causing N → D) were compared at the codon site using the branches where the ancestral amino acid was N. The numbers of reverse nonsynonymous differences (c_{Nrev}) and sites (s_{Nrev}) were computed in a similar manner to the computation of c_N and s_N. The null hypothesis of no selection for reverse substitution ($d_S = d_{Nrev}$) was tested with correction in a similar manner to the comparison of d_S and d_N. Although the single-substitution analysis is conceptually different from the single-site analysis, they are methodologically similar to each other. It has been shown that the latter analysis is generally conservative and reliable in the computer simulation and real data analysis (Suzuki and Gojobori, 1999; Suzuki, 2004a, 2007). It should be noted that in the single-substitution analysis, positive selection for original substitutions can be detected even when the number of (reverse) nonsynonymous substitutions is 0.

RESULTS

Identification of amino acid substitutions generating N-linked glycosylation sites When the phylogenetic tree was constructed for HA of 2,043 human H3N2 viruses and a duck H3N8 virus with the p distance and the ancestral sequences were inferred at interior nodes, N-linked glycosylation sites were observed in some sequences of the human lineage at amino acid sites 6, 7, 8, 22, 38, 45, 276, 278, 285, 483, and 498 in the fibrous stem and at sites 63, 81, 122, 126, 133, 144, 165, 171, and 246 in the globular head. The amino acid substitutions generating N-linked glycosylation sites were observed at amino acid sites 63 (D → N), 124 (G → S), 126 (D → N and T → N), 133 (D → N), 144 (D → N, I → N, and T → N), 173 (K → T), and 248 (N → S and N → T) in the globular head (Table 1).

Single-site analysis of natural selection The single-site analysis of natural selection using the d_N-d_S test and the interior-exterior test was conducted for the amino acid sites where the substitutions generating N-linked glycosylation sites were observed, as indicated above. No site was inferred as positively or negatively selected by either test (Table 1). Natural selection was not detected even

Table 1. The p values obtained from the single-site analysis and the single-substitution analysis for the amino acid sites and substitutions generating N-linked glycosylation sites using the phylogenetic tree constructed with the p distance

Position	Substitution	Single-site analysis						Single-substitution analysis	
		d_N-d_S test		Interior-exterior test					
				4 classes		2 classes			
		Individual	Combined	Individual	Combined	Individual	Combined	Individual	Combined
63	D → N	*0.0498*[a]		N.A.[b]		0.527		*0.0625*	
124	G → S	0.829		N.A.[b]		*0.957*		*0.625*	
126	D → N	*0.0220*		N.A.[b]		0.733		*0.151*	
	T → N								
133	D → N	*0.0578*		N.A.[b]		*0.278*		*0.267*	
144	D → N	*0.0108*	*0.209*	N.A.[b]	0.244	0.0460	0.291	0.218	*0.0477*
	I → N								
	T → N								
173	K → T	0.516		N.A.[b]		0.505		N.A.[d]	
248	N → S	*0.160*		N.A.[b]		N.A.[b]		*1*	
	N → T								
p_c	Positive	0.0125	0.05	N.A.[c]	0.05	0.00833	0.05	0.0167	0.05
p_c	Negative	*0.00714*	*0.05*	N.A.[c]	*0.05*	*0.00833*	*0.05*	*0.0167*	*0.05*

[a] Values are indicated in plain text and in italic when the configuration of test statistics was in favor of positive and negative selection, respectively.
[b] Not applicable because the expected number of differences was < 5 for some classes.
[c] Not applicable because all individual sites were not testable for natural selection.
[d] Not applicable because no substitution was observed.

when the c_S, c_N, s_S, and s_N values for these sites were combined to increase the sensitivity of the tests. To examine the property of the amino acid sites that are identified as positively selected by these tests, the tests were performed using all sites of HA. Although the interior-exterior test failed to detect natural selection at any site, the d_N-d_S test identified negative selection at many sites and positive selection at sites 53 and 138, which were antigenic sites included in epitopes C and A, respectively. When nonsynonymous substitutions were divided into conservative and radical substitutions according to whether they retain or alter the charge of amino acids, positive selection was inferred to have operated on conservative and radical substitutions at sites 138 and 53 by the d_N-d_S test, respectively.

Single-substitution analysis of natural selection The single-substitution analysis of natural selection was conducted for the amino acid substitutions generating N-linked glycosylation sites, as identified above. Specifically, natural selection was examined for the reverse substitutions, N → D at amino acid site 63, S → G at site 124, N → D and N → T at site 126, N → D at site 133, N → D, N → I, and N → T at site 144, T → K at site 173, and S → N and T → N at site 248, focusing on the branches where the ancestral amino acid was the substituted form in the phylogenetic tree. No selection was inferred for each of the reverse substitutions. However, negative selection was detected when the c_S, c_{Nrev}, s_S, and s_{Nrev} values for these substitutions were combined (Table 1), suggesting that the original amino acid substitutions generating N-linked glycosylation sites were positively selected.

Similar results were obtained when all of the above analyses were repeated using the phylogenetic tree constructed with the MCL distance (data not shown). Although it was possible that the mutation bias affected the results of single-substitution analysis, negative selection for the reverse substitutions was also identified when the pattern of nucleotide substitution obtained from the analysis of PB2 (Rabadan et al., 2006) was used (data not shown).

DISCUSSION

Positive selection for gains of N-linked glycosylation sites In the single-substitution analysis of HA for human H3N2 virus, positive selection was detected for the amino acid substitutions generating N-linked glycosylation sites. However, all of the sites where these substitutions occurred were antigenic sites and some of the substitutions increased the positive charge of HA, which can also be the target of positive selection (Suzuki, 2006; Hensley et al., 2009). Nevertheless, the effect of antigenic variation at these sites may be shielded by the N-linked glycans attached to the newly generated and closely located N-linked glycosylation sites (Skehel et al., 1984; Jackson et al., 1994; Tsuchiya et al., 2002; Das et al., 2010; Wei et al., 2010; Wanzeck et al., 2011), and the effect of increment of positive charge by some of these

substitutions may be cancelled by the increment of negative charge in the sialic acid or sulfuric acid, which may be added to the N-linked glycans (Spiro and Spiro, 2000). Therefore, positive selection appears to have operated for gains of N-linked glycosylation sites.

The N-linked glycans of HA are involved in several functions, such as folding of ectodomain, fusion activity of HA, shielding of antigenic sites, proteolytic activity of HA, recognition by collectins, and receptor binding, as discussed above. However, since most of the gains of N-linked glycosylation sites apparently occurred around the receptor-binding pocket in the three-dimensional structure of HA (Abe et al., 2004; Kobayashi and Suzuki, in preparation), it is likely that positive selection has operated on shielding of antigenic sites or receptor binding. It should be noted that the number of N-linked glycosylation sites has increased only in the influenza A viruses circulating in human. In particular, the number stayed constant in influenza A viruses circulating in swine, which apparently expresses similar distributions of the receptors with sialic acid α2,3-galactose and α2,6-galactose linkages in organs (Nelli et al., 2010; Sriwilaijaroen et al., 2011), but weaker immune responses against influenza A viruses (Nerome et al., 1995) compared with human. These observations suggest that the target of positive selection was not the receptor binding but the shielding of antigenic sites by the gains of N-linked glycosylation sites in human H3N2 virus.

It has been proposed that the gain and loss of N-linked glycosylation sites may not have been involved in antigenic changes of HA in human H3N2 virus, because the gain and loss were not found to be coincided with the transition of antigenic clusters (Smith et al., 2004), as well as the increase in the rate of change in the substitution pattern at amino acid sites (Blackburne et al., 2008). However, the existence of antigenic clusters in human H3N2 virus itself has been questioned (Shih et al., 2007; Suzuki, 2008b; Bhatt et al., 2011). In addition, the antigenic change appeared to occur continuously even within antigenic clusters (Suzuki, 2008b). Therefore, the gains of N-linked glycosylation sites may be involved in antigenic changes of HA in human H3N2 virus.

Properties of single-site analysis and single-substitution analysis In the study of natural selection for HA of human H3N2 virus, the single-site analysis failed to detect positive selection at the amino acid sites where the substitutions generating N-linked glycosylation sites were observed, but identified positive selection at antigenic sites. In contrast, the single-substitution analysis succeeded in detecting positive selection for the amino acid substitutions generating N-linked glycosylation sites. At the antigenic sites, positive selection is considered to operate recurrently, so that the virus can escape from immune responses continuously (Suzuki, 2008b). Many amino acid substitutions may be accumulated at the antigenic sites, and the excess of nonsynonymous substitutions over synonymous substitutions is detected as the signature of positive selection by the single-site analysis. In contrast, positive selection for gains of N-linked glycosylation sites is considered to be directional. Once a substitution generating an N-linked glycosylation site occurs, further substitutions may be suppressed at the site, so that the advantageous effect of shielding antigenic sites is maintained unless natural selection changes. The suppression of further amino acid substitutions, especially that of the reverse substitution, is detected as the signature of positive selection for the original substitution by the single-substitution analysis. These observations indicate that the single-site analysis is suitable for detecting recurrent natural selection (Suzuki, 2010), whereas the single-substitution analysis is suitable for detecting episodic natural selection.

In the single-site analysis, the d_N/d_S test appeared to be more efficient than the interior-exterior test in detecting natural selection. This is partly because in the latter test the c_S and c_N values were divided into c_{Sint}, c_{Sext}, c_{Nint}, and c_{Next} at single codon sites, which may be too small for obtaining statistical significance. It should also be noted that in the interior-exterior test the difference in the d_N/d_S value between interior and exterior branches can occur not only when positive or negative selection operated but also when natural selection was weakened or strengthened during evolution.

Antigenic drift is not a by-product of the evolution of receptor binding avidity In the single-site analysis of HA for human H3N2 virus, the number of amino acid sites identified as positively selected was only 2, which appeared to be relatively small compared with the envelope glycoproteins of other viruses, such as hepatitis C virus (HCV) and human immunodeficiency virus type 1 (HIV-1) (Suzuki and Gojobori, 2001; Yang et al., 2003). This is probably because the antigenic sites, which are usually the targets of positive selection, were shielded from immune responses by N-linked glycans after the gains of N-linked glycosylation sites in HA of human H3N2 virus, although N-linked glycans are also known to be attached to the envelope glycoproteins of HCV and HIV-1 (Zhang et al., 2004). In fact, a reduction in d_N/d_S has been observed at the antigenic sites of HA that are likely to be shielded by N-linked glycans after the gains of N-linked glycosylation sites during evolution of human H3N2 virus (Kobayashi and Suzuki, in preparation).

Although influenza A virus was believed to escape from immune responses by changing the antigenicity gradually through mutations at antigenic sites (antigenic drift) and abruptly through reassortment of genomic segments encoding HA and NA (antigenic shift), it was proposed that the antigenic drift is a by-product of repeated natural

selection for increased and decreased receptor binding avidity of virus in immune and naïve individuals, which is caused by the amino acid substitutions increasing and decreasing the positive charge of HA, respectively (Hensley et al., 2009). The receptor binding avidity of virus was considered to be positively correlated with the positive charge of HA, because the sialic acid receptor and the host cell membrane are negatively charged. According to this hypothesis, it is expected that positive selection has operated only on the radical substitution in terms of the charge of amino acids at antigenic sites. However, in the single-site analysis of HA for human H3N2 virus, positive selection was identified not only on the radical substitution but also on the conservative substitution at antigenic sites, suggesting that the antigenic drift is not a by-product of the evolution of receptor binding avidity of HA, but the evolutionary mechanism of influenza A virus where amino acid substitutions inhibit recognition of antigenic sites by immune responses.

The author thanks Yuki Kobayashi and two anonymous reviewers for valuable comments.

REFERENCES

Abe, Y., Takashita, E., Sugawara, K., Matsuzaki, Y., Muraki, Y., and Hongo, S. (2004) Effect of the addition of oligosaccharides on the biological activities and antigenicity of influenza A/H3N2 virus hemagglutinin. J. Virol. **78**, 9605–9611.

Arinaminpathy, N., and Grenfell, B. (2010) Dynamics of glycoprotein charge in the evolutionary history of human influenza. PLoS One **5**, e15674.

Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J., and Lipman, D. (2008) The Influenza Virus Resource at the National Center for Biotechnology Information. J. Virol. **82**, 596–601.

Bazykin, G. A., and Kondrashov, A. S. (2011) Detecting past positive selection through ongoing negative selection. Genome Biol. Evol. **3**, 1006–1013.

Bhatt, S., Holmes, E. C., and Pybus, O. G. (2011) The genomic rate of molecular adaptation of the human influenza A virus. Mol. Biol. Evol. **28**, 2443–2451.

Blackburne, B. P., Hay, A. J., and Goldstein, R. A. (2008) Changing selective pressure during antigenic changes in human influenza H3. PLoS Pathog. **4**, e1000058.

Carstens, E. B. (2010) Ratification vote on taxonomic proposals to the International Committee on Taxonomy of Viruses (2009). Arch. Virol. **155**, 133–146.

Cherry, J. L., Lipman, D. J., Nikolskaya, A., and Wolf, Y. I. (2009) Evolutionary dynamics of N-glycosylation sites of influenza virus hemagglutinin. PLoS Curr. **1**, RRN1001.

Cui, J., Smith, T., Robbins, P. W., and Samuelson, J. (2009) Darwinian selection for sites of Asn-linked glycosylation in phylogenetically disparate eukaryotes and viruses. Proc. Natl. Acad. Sci. USA **106**, 13421–13426.

Daniels, R., Kurowski, B., Johnson, A. E., and Hebert, D. N. (2003) N-linked glycans direct the cotranslational folding pathway of *influenza* hemagglutinin. Mol. Cell **11**, 79–90.

Das, S. R., Puigbo, P., Hensley, S. E., Hurt, D. E., Bennink, J. R., and Yewdell, J. W. (2010) Glycosylation focuses sequence variation in the influenza A virus H1 hemagglutinin globular domain. PLoS Pathog. **6**, e1001211.

de Vries, R. P., de Vries, E., Bosch, B. J., de Groot, R. J., Rottier, P. J. M., and de Haan, C. A. M. (2010) The influenza A virus hemagglutinin glycosylation state affects receptor-binding specificity. Virology **403**, 17–25.

Fitch, W. M. (1971) Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. **20**, 406–416.

Greenbaum, B. D., Levine, A. J., Bhanot, G., and Rabadan, R. (2008) Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathog. **4**, e1000079.

Hartigan, J. A. (1973) Minimum mutation fits to a given tree. Biometrics **29**, 53–65.

Hebert, D. N., Zhang, J.-X., Chen, W., Foellmer, B., and Helenius, A. (1997) The number and location of glycans on influenza hemagglutinin determine folding and association with calnexin and calreticulin. J. Cell Biol. **139**, 613–623.

Hensley, S. E., Das, S. R., Bailey, A. L., Schmidt, L. M., Hickman, H. D., Jayaraman, A., Viswanathan, K., Raman, R., Sasisekharan, R., Bennink, J. R., et al. (2009) Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. Science **326**, 734–736.

Hughes, A. L., and Nei, M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335**, 167–170.

Hughes, A. L., Ota, T., and Nei, M. (1990) Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Mol. Biol. Evol. **7**, 515–524.

Igarashi, M., Ito, K., Kida, H., and Takada, A. (2008) Genetically destined potentials for *N*-linked glycosylation of influenza virus hemagglutinin. Virology **376**, 323–329.

Jackson, D. C., Drummer, H. E., Urge, L., Otvos, L. Jr., and Brown, L. E. (1994) Glycosylation of a synthetic peptide representing a T-cell determinant of influenza virus hemagglutinin results in loss of recognition by $CD4^+$ T-cell clones. Virology **199**, 422–430.

Katoh, K., Misawa, K, Kuma, K.-i., and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. **30**, 3059–3066.

Kimura, M. (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature **267**, 275–276.

Kosakovsky Pond, S. L., Poon, A. F. Y., Leigh Brown, A. J., and Frost, S. D. W. (2008) A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. Mol. Biol. Evol. **25**, 1809–1824.

McDonald, J. H., and Kreitman, M. (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature **351**, 652–654.

Nei, M., and Kumar, S. (2000) Molecular Evolution and Phylogenetics, pp. 165–186. Oxford University Press, Oxford, New York.

Nelli, R. K., Kuchipudi, S. V., White, G. A., Perez, B. B., Dunham, S. P., and Chang, K.-C. (2010) Comparative distribution of human and avian type sialic acid influenza receptors in the pig. BMC Vet. Res. **6**, 4.

Nerome, K., Kanegae, Y., Shortridge, K. F., Sugita, S., and Ishida, M. (1995) Genetic analysis of porcine H3N2 viruses originating in southern China. J. Gen. Virol. **76**, 613–624.

Nozawa, M., Suzuki, Y., and Nei, M. (2009) Reliabilities of iden-

tifying positive selection by the branch-site and the site-prediction methods. Proc. Natl. Acad. Sci. USA **106**, 6700–6705.

Pybus, O. G., Rambaut, A., Belshaw, R., Freckleton, R. P., Drummond, A. J., and 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.

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

Saitou, N. (1989) A theoretical study of the underestimation of branch lengths by the maximum parsimony principle. Syst. Zool. **38**, 1–6.

Saitou, N., and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**, 406–425.

Schulze, I. T. (1997) Effects of glycosylation on the properties and functions of influenza virus hemagglutinin. J. Infect. Dis. **176**, S24–S28.

Shih, A. C.-C., Hsiao, T.-C., Ho, M.-S., and Li, W.-H. (2007) Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. Proc. Natl. Acad. Sci. USA **104**, 6283–6288.

Shope, R. E. (1931) Swine influenza. III. Filtration experiments and etiology. J. Exp. Med. **54**, 373–380.

Skehel, J. J., and Wiley, D. C. (2000) Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. Annu. Rev. Biochem. **69**, 531–569.

Skehel, J. J., Stevens, D. J., Daniels, R. S., Douglas, A. R., Knossow, M., Wilson, I. A., and Wiley, D. C. (1984) A carbohydrate side chain on hemagglutinins of Hong Kong influenza viruses inhibits recognition by a monoclonal antibody. Proc. Natl. Acad. Sci. USA **81**, 1779–1783.

Smith, D. J., Lapedes, A. S., de Jong, J. C., Bestebroer, T. M., Rimmelzwaan, G. F., Osterhaus, A. D. M. E., and Fouchier, R. A. M. (2004) Mapping the antigenic and genetic evolution of influenza virus. Science **305**, 371–376.

Sokal, R. R., and Rohlf, F. J. (1995) Biometry. 3rd edition, pp. 685–793. W. H. Freeman and Company, New York.

Spiro, M. J., and Spiro, R. G. (2000) Sulfation of the *N*-linked oligosaccharides of influenza virus hemagglutinin: temporal relationships and localization of sulfotransferases. Glycobiology **10**, 1235–1242.

Sriwilaijaroen, N., Kondo, S., Yagi, H., Takemae, N., Saito, T., Hiramatsu, H., Kato, K., and Suzuki, Y. (2011) *N*-glycans from porcine trachea and lung: predominant NeuAcα2-6Gal could be a selective pressure for influenza variants in favor of human-type receptor. PLoS One **6**, e16302.

Sun, S., Wang, Q., Zhao, F., Chen, W., and Li, Z. (2011) Glycosylation site alteration in the evolution of influenza A (H1N1) viruses. PLoS One **6**, e22844.

Suzuki, Y. (2004a) New methods for detecting positive selection at single amino acid sites. J. Mol. Evol. **59**, 11–19.

Suzuki, Y. (2004b) Three-dimensional window analysis for detecting positive selection at structural regions of proteins. Mol. Biol. Evol. **21**, 2352–2359.

Suzuki, Y. (2006) Natural selection on the influenza virus genome. Mol. Biol. Evol. **23**, 1902–1911.

Suzuki, Y. (2007) Inferring natural selection operating on conservative and radical substitution at single amino acid sites. Genes Genet. Syst. **82**, 341–360.

Suzuki, Y. (2008a) False-positive results obtained from the branch-site test of positive selection. Genes Genet. Syst. **83**, 331–338.

Suzuki, Y. (2008b) 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., and Gojobori, T. (1999) A method for detecting positive selection at single amino acid sites. Mol. Biol. Evol. **16**, 1315–1328.

Suzuki, Y., and Gojobori, T. (2001) Positively selected amino acid sites in the entire coding region of hepatitis C virus subtype 1b. Gene **276**, 83–87.

Tamura, K., Nei, M., and Kumar, S. (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc. Natl. Acad. Sci. USA **101**, 11030–11035.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. **28**, 2731–2739.

Tsuchiya, E., Sugawara, K., Hongo, S., Matsuzaki, Y., Muraki, Y., Li, Z.-N., and Nakamura, K. (2001) Antigenic structure of the haemagglutinin of human influenza A/H2N2 virus. J. Gen. Virol. **82**, 2475–2484.

Tsuchiya, E., Sugawara, K., Hongo, S., Matsuzaki, Y., Muraki, Y., Li, Z.-N., and Nakamura, K. (2002) Effect of addition of new oligosaccharide chains to the globular head of influenza A/H2N2 virus haemagglutinin on the intracellular transport and biological activities of the molecule. J. Gen. Virol. **83**, 1137–1146.

Vigerust, D. J., Ulett, K. B., Boyd, K. L., Madsen, J., Hawgood, S., and McCullers, J. A. (2007) N-linked glycosylation attenuates H3N2 influenza viruses. J. Virol. **81**, 8593–8600.

Wang, C.-C., Chen, J.-R., Tseng, Y.-C., Hsu, C.-H., Hung, Y.-F., Chen, S.-W., Chen, C.-M., Khoo, K.-H., Cheng, T.-J., Cheng, Y.-S. E., et al. (2009) Glycans on influenza hemagglutinin affect receptor binding and immune response. Proc. Natl. Acad. Sci. USA **106**, 18137–18142.

Wanzeck, K., Boyd, K. L., and McCullers, J. A. (2011) Glycan shielding of the influenza virus hemagglutinin contributes to immunopathology in mice. Am. J. Respir. Crit. Care Med. **183**, 767–773.

Wei, C.-J., Boyington, J. C., Dai, K., Houser, K. V., Pearce, M. B., Kong, W.-P., Yang, Z.-y., Tumpey, T. M., and Nabel, G. J. (2010) Cross-neutralization of 1918 and 2009 influenza viruses: role of glycans in viral evolution and vaccine design. Sci. Transl. Med. **2**, 24ra21.

Wilson, I. A., Skehel, J. J., and Wiley, D. C. (1981) Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 Å resolution. Nature **289**, 366–373.

World Health Organization (1980) A revision of the system of nomenclature for influenza viruses: a WHO memorandum. Bull. W. H. O. **58**, 585–591.

World Health Organization (2011) Review of the 2010–2011 winter influenza season, northern hemisphere. Wkly. Epidemiol. Rec. **86**, 222–227.

Yang, W., Bielawski, J. P., and Yang, Z. (2003) Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. J. Mol. Evol. **57**, 212–221.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24**, 1586–1591.

Zhang, M., Gaschen, B., Blay, W., Foley, B., Haigwood, N., Kuiken, C., and Korber, B. (2004) Tracking global patterns of *N*-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin. Glycobiology **14**, 1229–1246.

＜特集関連情報＞

パンソルビン・トラップ法による食品からのウイルス検出法

1997（平成9）年に食品衛生法の中に初めてウイルスが登場し，ウイルス性食中毒という概念が確立したのは，歴史の一つの節目と言える。しかし一方で，カキ以外の一般的な食品からウイルスを検出する方法として，一応の標準としてポリエチレングリコール（PEG）沈澱法が存在していたものの，手探りの状態が続いてきた。そこで2007（平成19）年から厚生労働科学研究費補助金（食品の安心・安全確保推進研究事業）による研究の一環として，食品中のウイルスを検出するための実践的手法の開発に関する研究がスタートした。その結果，固形，液状，練り物，油物などの多種・多様な食品からノロウイルス（NoV）に代表される食中毒起因ウイルスを検出することができるパンソルビン・トラップ法（パントラ法）を開発し，ルーチンの食品検査として実施可能な段階に達してきたため，本稿にてその概要を紹介する。なお，パンソルビンとは黄色ブドウ球菌をホルマリン固定・熱処理したもので，メルク社から製造・販売されているが，相当品を自作することも可能である。

糞便検体と違って，食品検体の場合は含まれるウイルス量が極めて少ない（広く拡散した状態）ため，何らかの濃縮手段が必要となる。しかし，食品検体を適当な緩衝液に懸濁して乳剤とした場合，その量は少なく見積もっても50ml程度になる。一方，PCRで用いる検体（RNA抽出液）は50μl程度であり，1,000倍に相当する減量濃縮が必要である。食品検体の質的な問題に目を向けると，表面が平滑な固形食品では，緩衝液で洗滌することで，比較的濁質の少ない形でウイルスを回収できる可能性がある。しかし，ほとんどの

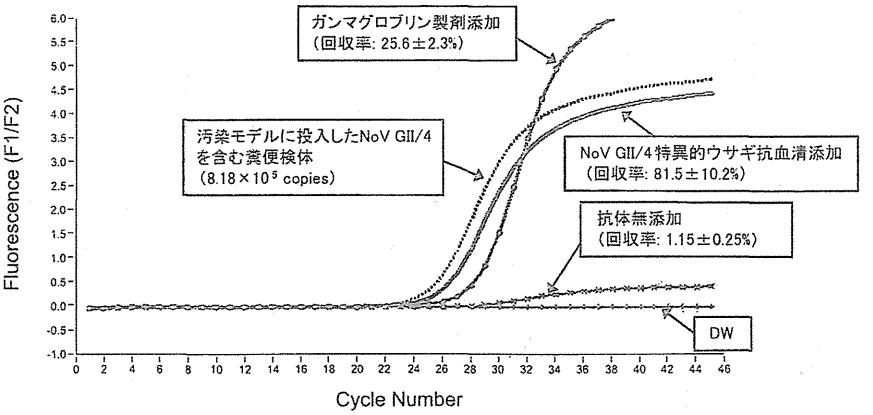

図. 汚染モデル食品からのウイルス回収試験の一例(NoV GII/4で汚染した焼きそばにおける増幅曲線)

ケースでは遠心後も上清は濁ったままであり，フィルターを用いたろ過では目詰まりを起こしてうまくいかない。この状態でPEG沈澱法を用いると，ウイルスとは無関係の大量の沈澱が生じて手に負えないことが多い。さらに，PEG沈澱法では原則として一夜放置の工程が必要であり，その後10,000rpm程度の遠心によって沈澱を回収しなければならない。しかし，50m*l*の容量を10,000rpmで遠心するには高速冷却遠心機のような大型機器を必要とし，遠心チューブも専用品を用いるようになっている。こうした専用遠心チューブはディスポーザブル使用を前提としていないため高価であり，洗って再利用するのはPCRを行う関係上不安が残る。また，一般に“抽出キット”として市販されている試薬は0.2m*l*程度の検体量を想定して作られているため，50m*l*の食品乳剤をそのまま適用するのは困難である。

パントラ法の基本原理は，食品乳剤中にウイルスに対する抗体を添加することにより，抗原抗体複合体を形成させ，それを黄色ブドウ球菌表面のプロテインAに吸着させることで，菌体とともにウイルス粒子を沈澱・回収することである。最大の特長は，食品乳剤が濁ったままでよいという点にある。最初に食品を50m*l*の緩衝液に懸濁した後で3,000rpm，30分の遠心を行うが，このとき用いる遠心機は一般の検査室にある普通のものであり，チューブもプラスチックのディスポーザブル製品である。この遠心条件で沈澱する固形物だけを除去しておけば，上清は濁っていてもかまわない。その後，抗体とパンソルビンを添加して再び3,000rpm，20分の遠心を行うが，この条件で沈澱する食品由来の固形成分はすでに最初の遠心の際に除去されているため，結果としてウイルス粒子を吸着した黄色ブドウ球菌だけが沈澱してくる。この段階で上清は濁ったままであることが多いが，ウイルス粒子は菌体と一緒に沈澱物の方に移行しているため，上清は捨ててよい。沈澱した菌体を少量の緩衝液で再懸濁してから市販のキットを用いてRNA抽出を行えば，50m*l*の食品乳剤から50μlのPCRスケールのRNA溶液まで，効率良く減量濃縮できることになる。添加する抗体として，開発時にはGII/4型のNoVウイルス様粒子（VLPs）を免疫して作製したウサギ抗血清が用いられた。その後，実践使用のための安定的な抗体供給源として，市販ガンマグロブリン製剤を用いる汎用プロトコールが考案された。図に例を示したとおり，回収率はウサギで作製した特異的抗血清の1/3程度となるが（PCRでは2サイクル以内の差），NoVの他の型や，サポウイルス（SaV），A型肝炎ウイルス（HAV）などへも幅広く対応できるという利点がある。これまでのところ，ガンマグロブリン製剤は，NoVでは13遺伝子型（GI/3, GI/4, GI/8, GI/9, GI/14, GII/2, GII/3, GII/4, GII/5, GII/6, GII/12, GII/13, GII/18），SaVではヒトに感染する4種類すべての型，他にHAVとアデノウイルス41型において有効であることを確認している。また，流通食品が疑われる大規模・広域事例では，最初に感染して回復した人の血清を用いることも可能である。当事者の協力が得られることが大前提となるが，不幸にして社会問題に発展した場合には原因究明のための選択肢の一つとなるであろう。

以上のとおり，パントラ法は食品検体から調製された乳剤を濃縮・精製してRNA抽出液を得る段階までを担保するものであり，それ以降の逆転写反応やPCRについては既報に従うことになる。多くのケースでは逆転写反応前にDNase I処理を行っているが，α-Amylaseも同時に添加することで検出効率が向上する。また，ランダムプライマーよりも，特異的プライマーを用いて逆転写反応を行った方が，検出効率は高い。この場合，cDNAの種類が増えて煩雑となるため，ホットスタート対応のone-step PCRキットを用いるなどの工夫が有効である。ポテトサラダと焼きそばをNoV GII/4で汚染したモデル食品を作製（様々な汚染レベルのもの）し，ガンマグロブリン製剤を用いたパントラ法で抽出したRNAからnested PCR（1st. PCR: COG2F/G2-SKR, 2nd.PCR: COG2F/COG2Rによるリアルタイム PCR）による検出を試みたところ，いずれも食品1g当たり35コピーの汚染レベルの

ものまで検出できた。2nd.PCRの段階でリアルタイムPCRを用いると結果は定性扱いとなるが，ゲル電気泳動をハイブリダイゼーションで確認したのと同義であることから，タイムプレッシャーの中で高感度を求められる局面においては効果的と考えられる。プロトコールの詳細については，「日本食品微生物学会雑誌，Vol.29，No.2，2012」に掲載が予定されているため，以後の引用文献として利用されたい。また，最適反応条件の検討など，開発過程におけるデータについては，「秋田県健康環境センター年報，No.4～6」と「福井県衛生環境研究センター年報，No.7」の記述が参考となる。

秋田県健康環境センター　斎藤博之
福井県衛生環境研究センター　東方美保
（現福井県健康福祉部医薬食品・衛生課）
国立感染症研究所　岡　智一郎　片山和彦
堺市衛生研究所　田中智之
国立医薬品食品衛生研究所　野田　衛

＜特集関連情報＞

非晶性リン酸カルシウム微粒子を用いた食品からのウイルス検出法

ノロウイルスは，冬季を中心に発生するウイルス性食中毒の主要な病原体である。しかし，食品中のウイルス汚染量は一般に微量であること，食品成分がウイルス濃縮や遺伝子増幅反応等を阻害することなどから，食品からのウイルスの検出は極めて困難であり，その検出報告例も少ない。我々は，短時間で簡便に実施でき，かつ特殊な試薬や装置を必要としない食品からのウイルス検出方法の構築を目的として，非晶性リン酸カルシウム（Amorphous calcium phosphate; ACP）微粒子を用いたウイルス濃縮方法（ACP微粒子濃縮法）を検討しているので，これまで得られた結果の概要を報告する。

ACP微粒子はハイドロキシアパタイト（HAP）の前駆体であり，$Ca_3(PO_4)_2 \cdot n(OH)_2$ (n=1-2) を主成分とする，多孔質の白色微粒子である。HAPと比較し，比表面積が大きいため，タンパク質，脂肪酸，ウイルス等の吸着能が大きいと考えられる。

ACP微粒子濃縮法は，食品からのウイルス粒子の誘出，ACP微粒子へのウイルス粒子の吸着，およびACP微粒子の収集・溶解の3つのステップからなる。具体的には，食品10gをストマッカーバッグに入れ，食品洗浄液としてPBS（－）あるいはTris-glycine液（pH9.5）40m*l*を加えて10分間振とう後，メッシュを用いてのろ過および3,000rpm，30分間の遠心により食品残渣を除去する。遠心上清をフラスコに移し，ACP微粒子0.3gを添加して1時間撹拌した後，3,000rpm，10分間の遠心によりACP微粒子を集め，3.3Mクエン酸3m*l*で溶解する。以下定法に従い，この溶解液140μlからQIAamp Viral RNA Mini Kit（Qiagen）を用いて，ウイルスRNAを抽出後，逆転写反応，リアルタイムPCRを実施する。

本法によるウイルス回収率を千切りキャベツ，ちぎりレタス，スライスハムを対象食品として，ネコカリシウイルス（FCV）の添加回収実験により評価した。ゲノムコピー数の測定には，森ら[1)]が報告したリアルタイムPCR法を用いた。

食品10gに4.5×10^4～7.5×10^1コピーのFCVを添加し，1時間乾燥後，ACP微粒子濃縮法によりウイルスの回収を試みたところ，キャベツでは平均32%，レタスで平均50%の回収率であった。ハムでは全く回収できなかったが，ハム洗浄液にアスコルビン酸1.0g（最終濃度約2.5%）を添加することにより，ウイルス回収が可能となり，45%の回収率が得られた。

検出感度の検討では食品10gに4.5×10^3コピーおよび7.5×10^2コピーのFCVを添加し，1時間乾燥後，同様に濃縮操作を行った結果，各食品ともFCV 4.5×10^3コピーの添加までリアルタイムPCRで検出が可能であった。

本法を各種食品に応用した結果を表に示した。野菜類，食肉・魚肉類，穀物類からは効率良くウイルスを回収することができたが，冷凍ラズベリーや油脂を多

表. 食品群ごとのウイルス回収状況

食品群	平均FCV回収率(%)			
	0 - 4	5 - 9	10 - 19	20 -
野菜・果物		冷凍ラズベリー	ポテトサラダ 春雨サラダ つぼ漬	キャベツ　レタス 中華風きゅうりの和え物
穀物		ロールパン マカロニサラダ フライドポテト	白飯 うずら豆煮物	ゆでうどん ソース焼きそば ミートソーススパゲティ*
肉・魚		焼鮭	皮なしウインナー つくね	ハム*　マグロの刺身* 肉団子の甘酢あん
その他	小豆あん　シュークリーム ひじき煮物　卵の花 白和え　ハムカツ 卵フィリング　ツナフィリング			

*これらの食品では若干の変更を加えたACP濃縮法を使用した

く含む食品（ミートソーススパゲティ，フライドポテトなど）では回収率は低かった。油脂を多く含む食品として，回収率が１％であったミートソーススパゲティについて検討した結果，洗浄液に Tris-glycine 液（pH9.5）を用いることおよび油脂分除去のために遠心前にイソアミルアルコール10m*l* を添加し，遠心後に食品油脂をイソアミルアルコール層とともに除去することにより，回収率が32％に向上した。

食中毒の原因となる食品は多種多様で，その性状もさまざまである。これまでに報告されている濃縮法も食品の種類を問わずに十分な回収率を得ることは困難な場合が多い。本法においても，食品によっては一部操作法の改良を行う必要があると予測された。一方，本法は ACP 微粒子へのウイルスの非特異的な吸着を利用しているため，粒子溶解時の酸性条件に耐性のウイルスであれば，ウイルスの種類に依存することなく検査が可能であると考えられる。また，試薬等が非常に安価（ACP 微粒子は１検体あたり15円程度）であり，簡便な操作により２時間以内に検査ができるという利点もある。今後，本法を種々の食品に応用し，適用可能な食品群を増やしていくことにより，食中毒発生時の食品検査や輸入食品等のウイルスモニタリングへの利用を検討していきたい。

参考文献

1) 感染症学雑誌, 80: 496-500, 2006

埼玉県衛生研究所
篠原美千代　富岡恭子　峯岸俊貴　内田和江
鈴木典子　島田慎一　河橋幸恵　岸本　剛
国立医薬品食品衛生研究所　野田　衛

＜特集関連情報＞

二枚貝関連の食中毒疑い事例における各種胃腸炎ウイルスの関与 ―北海道

食中毒疑い事例の原因究明において，ウイルス検査の対象は主にノロウイルス（NoV）であり，その他の胃腸炎ウイルスの食中毒への関与については十分には把握されていない。そこで，二枚貝の喫食がみられた食中毒疑い事例を対象に，胃腸炎ウイルス感染の実態調査を行った。

二枚貝関連事例における胃腸炎ウイルスの検出状況

検査対象は，2000年７月～2010年12月の間に北海道で発生した42事例の患者糞便307検体である。NoVと，それ以外の胃腸炎ウイルスとしてサポウイルス（SaV），アストロウイルス（AstV），アイチウイルス（AiV），A群・C群ロタウイルス（A, C-RV），アデノウイルス（AdV），パレコウイルス（PeV），エンテロウイルス（EntV）の９種類についてPCR法による検索を行ったところ，41事例からウイルスが検出された。このうちウイルス１種類のみが検出された事例は61％に過ぎず（NoV：24事例，SaV：１事例），39％にあたる16事例からは複数（２～５種類）のウイルスが

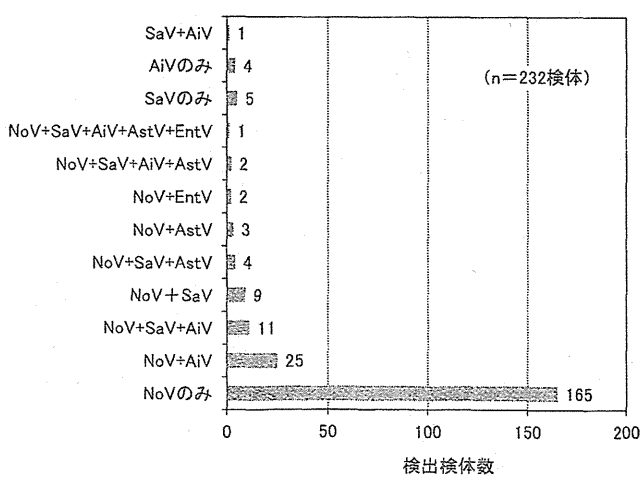

図. 検出された胃腸炎ウイルスの組み合わせ

表1. 食中毒事例における患者とカキからの検出ウイルス

検体		検出ウイルス※1 (コピー数※2)									
		NoV GI	NoV GII	SaV	AiV	AstV	A-RV	C-RV	AdV	PeV	EntV
患者糞便	1	GI/8 (7.5E+04)	- (-)	- (-)	typeA	-	-	-	-	-	-
	2	GI/4 (1.0E+08)	GII/12 (3.5E+08)	GI/3 (-)	typeA	-	-	-	-	-	-
	3	GI/4 (2.0E+09)	GII/2 (6.2E+06)	- (-)	-	-	-	-	-	-	-
	4	GI/4 (1.1E+08)	GII/14 (2.9E+09)	- (-)	-	-	-	-	-	-	-
	5	- (-)	GII/14 (2.7E+06)	- (-)	-	-	-	-	-	-	-
	6	- (-)	GII/14 (8.9E+08)	- (-)	typeA	-	-	-	-	-	-
	7	GI/4 (6.6E+07)	GII/2 (1.4E+09)	- (-)	-	-	-	-	-	-	-
	8	- (-)	GII/2 (2.2E+09)	- (-)	-	-	-	-	-	-	-
カキ	oys1	GI/4, 11 (-)	GII/6 (-)	- (-)	typeA	type8	G1	-	-	-	NT
	oys2	- (-)	- (-)	- (-)	typeA	-	-	-	-	-	NT
	oys3	- (-)	- (-)	- (-)	-	-	-	-	-	-	NT
	oys4	GI/4 (-)	GII/4, 6 (-)	- (-)	typeA	type8	G1	-	-	-	NT
	oys5	GI/4 (-)	GII/2 (1.9E+03)	GI/1 (-)	typeA	+	G1	-	-	-	NT
	oys6	GI/7 (-)	GII/2,3,4,13 (3.1E+03)	- (-)	typeA	type8	G1	-	-	-	NT
	oys7	GI/4 (-)	GII/3, 4 (-)	- (-)	typeA	type8	-	-	-	-	NT
	oys8	GI/1, 4 (-)	GII/4,6,13 (-)	- (-)	typeA	+	-	-	-	-	NT

※1 検出されたウイルスの遺伝子型を記入、型別に使用する領域の塩基配列が確認できなかった検体は「+」とした

※2 糞便は1gあたり、カキは1個あたりの換算値を示し、実測値が定量下限値 (NoV:10、SaV:25) 未満の検体は (-) とした

NT:not tested

検出された。この検出ウイルスの組み合わせはすべて「NoVと他のウイルス」であり，いずれの事例もNoVの検出率が最も高かった。NoV以外のウイルスの検出事例数は，AiV：14，SaV：13，AstV：3，EntV：2事例であった。検体ごとにみた検出ウイルスの組み合わせを前ページ図に示した。ウイルス陽性232検体のうち複数ウイルスの検出例は「NoVと他のウイルスの組み合わせ」が57検体（25%），「SaVとAiV」が1検体であった。最も多いものでは1検体から5種類のウイルスが検出された。

原因食品と患者からの検出ウイルスの比較

混合感染例が高頻度に認められたことから，原因二枚貝のウイルス汚染状況の把握が必要と考えた。そこで，今回の対象事例のうち原因食品の原材料が確保できた1事例について，食品と患者からの検出ウイルスを比較した（表1）。この事例は，加熱用冷凍カキが加熱不十分な状態で提供されたことが原因と推定されており，同一ロットのカキからNoV GI，GII，AiV，AstV，A-RVが高率に，さらにSaVも1検体から検出された。患者糞便ではNoV GI，GIIの検出率が高く，AiVとSaVも認めた。しかし，カキからの検出率の高かったAstVとA-RVは患者からは検出されず，NoV，SaV，AiVに比べて感染リスクが低い可能性が示唆された。リアルタイムPCR法によりカキのNoVおよびSaVコピー数の測定[1, 2]を行ったが，ほとんどが定量下限値未満であり，その把握は困難であった。

混合感染時におけるNoVとSaVの増殖動態

今回の調査では一人から最大5種類の胃腸炎ウイルスが検出されたが，混合感染時にすべてのウイルスが発症に関与しているとは限らない。各々のウイルスの

表2. SaV陽性検体のSaVコピー数と他ウイルスの検出状況

No.	SaVコピー数 (copies/g 糞便)	SaV以外のウイルスの検出状況				
		NoV (GII)	NoV (GI)	AiV	AstV	EntV
1	1.9E+10	-	-	+	-	-
2	1.6E+10	-	-	-	-	-
3	1.2E+10	-	+	-	+	-
4	4.6E+09	-	-	-	-	-
5	3.6E+09	-	-	-	-	-
6	2.2E+09	-	-	-	-	-
7	9.5E+08	-	+	-	-	-
8	3.8E+07	+	-	-	-	-
9	3.3E+07	+	+	+	-	-
10	1.3E+07	+	-	+	-	-
11	7.6E+06	+	-	-	-	-
12	3.1E+06	+	+	-	-	-
13	9.1E+05	+	-	+	-	-
14	7.4E+05	+	+	-	-	-
15	5.4E+05	+	-	+	-	-
16	3.5E+05	+	+	+	+	-
17	3.4E+05	+	-	-	-	-
18	2.3E+05	+	+	+	-	-
19	2.1E+05	+	+	-	-	-
20	1.3E+05	-	-	-	-	-
21	1.3E+05	+	-	+	-	-
22	1.3E+05	+	+	+	-	-
23	8.7E+04	+	+	-	+	-
24	< 8.6E+04 *	+	-	-	-	-
25	< 8.6E+04	+	+	+	-	-
26	< 8.6E+04	+	+	+	-	-
27	< 8.6E+04	+	+	+	+	+
28	< 8.6E+04	+	-	+	+	-
29	< 8.6E+04	+	+	-	+	-
30	< 8.6E+04	+	+	-	-	-
31	< 8.6E+04	+	+	-	+	-
32	< 8.6E+04	+	-	+	-	-
33	< 8.6E+04	+	+	+	-	-

*：定量下限値未満