SCIENTIFIC REPORTS

OPEN

Molecular evolution of the capsid gene in human norovirus genogroup II

Received: 18 January 2016
Accepted: 20 June 2016
Published: 07 July 2016

Miho Kobayashi[1], Yuki Matsushima[2], Takumi Motoya[3], Naomi Sakon[4], Naoki Shigemoto[5], Reiko Okamoto-Nakagawa[6], Koichi Nishimura[7], Yasutaka Yamashita[8], Makoto Kuroda[9], Nobuhiro Saruki[1], Akihide Ryo[10], Takeshi Saraya[11], Yukio Morita[12], Komei Shirabe[6], Mariko Ishikawa[2], Tomoko Takahashi[13], Hiroto Shinomiya[8], Nobuhiko Okabe[2], Koo Nagasawa[14], Yoshiyuki Suzuki[15], Kazuhiko Katayama[16] & Hirokazu Kimura[10,14]

Capsid protein of norovirus genogroup II (GII) plays crucial roles in host infection. Although studies on capsid gene evolution have been conducted for a few genotypes of norovirus, the molecular evolution of norovirus GII is not well understood. Here we report the molecular evolution of all GII genotypes, using various bioinformatics techniques. The time-scaled phylogenetic tree showed that the present GII strains diverged from GIV around 1630CE at a high evolutionary rate (around 10^{-3} substitutions/site/year), resulting in three lineages. The GII capsid gene had large pairwise distances (maximum > 0.39). The effective population sizes of the present GII strains were large ($>10^2$) for about 400 years. Positive (20) and negative (over 450) selection sites were estimated. Moreover, some linear and conformational B-cell epitopes were found in the deduced GII capsid protein. These results suggested that norovirus GII strains rapidly evolved with high divergence and adaptation to humans.

Norovirus (NoV) is a pathogenic agent of acute gastroenteritis in humans[1]. It has led to pandemics of acute gastroenteritis around the world[1]. In Japan, half of acute gastroenteritis cases in the winter season may be caused by NoV infection[2,3]. Furthermore, large outbreaks of food poisoning involving NoV have been reported in many countries[4,5]. Thus, NoV is a major causative agent of acute viral gastroenteritis worldwide, and NoV infection is a major disease burden in many countries[1,6].

NoV belongs to the genus *Norovirus* and the family *Caliciviridae* and, at present, is classified into seven genogroups (GI–GVII), based on phylogenetic analysis of the capsid gene[7]. Among them, NoV belonging to genogroups I, II, and IV may infect humans[7]. Furthermore, the NoV GI and GII strains can be classified into 9 and 22 genotypes, respectively[8].

Previous epidemiological studies suggested that specific genogroup/genotype viruses (e.g., GII.2, GII.3, GII.4, and GII.6) caused more recent large outbreaks of gastroenteritis than other GII and GI genotypes[9–11]. In particular,

[1]Gunma Prefectural Institute of Public Health and Environmental Science, Maebashi-shi, Gunma 371-0052, Japan. [2]Kawasaki City Institute for Public Health, Kawasaki-shi, Kanagawa 210-0821, Japan. [3]Ibaraki Prefectural Institute of Public Health, Mito-shi, Ibaraki 310-0852, Japan. [4]Osaka Prefectural Institute of Public Health, Osaka-shi, Osaka 537-0025, Japan. [5]Hiroshima Prefectural Technology Research Institute, Public Health and Environment Center, Hiroshima-shi, Hiroshima 734-0007, Japan. [6]Yamaguchi Prefectural Institute of Public Health and Environment, Yamaguchi-shi, Yamaguchi 753-0821, Japan. [7]Kumamoto Prefectural Institute of Public-Health and Environmental Science, Uto-shi, Kumamoto 869-0425, Japan. [8]Ehime Prefectural Institute of Public Health and Environmental Science, Matsuyama-shi, Ehime 790-0003, Japan. [9]Pathogen Genomics Center, National Institute of Infectious Diseases, Musashimurayama-shi, Tokyo 208-0011, Japan. [10]Department of Microbiology, Yokohama City University Graduate School of Medicine, Yokohama-shi, Kanagawa 236-0027, Japan. [11]Department of 1st Internal Medicine, Kyorin University School of Medicine, Mitaka-shi, Tokyo 181-0004, Japan. [12]Department of Food and Nutrition, Tokyo Kasei University, Itabashi-ku, Tokyo 173-0003, Japan. [13]Iwate Prefectural Meat Inspection Center, Shiwa-cho, Iwate 020-3311, Japan. [14]Infectious Disease Surveillance Center, National Institute of Infectious Diseases, Musashimurayama-shi, Tokyo 208-0011, Japan. [15]Division of Biological Science, Nagoya City University, Nagoya-shi, Aichi 467-0000, Japan. [16]Department of Virology II, National Institute of Infectious Diseases, Musashimurayama-shi, Tokyo 208-0011, Japan. Correspondence and requests for materials should be addressed to K.K. (email: katayama@nih.go.jp) or H.K. (email: kimhiro@nih.go.jp)

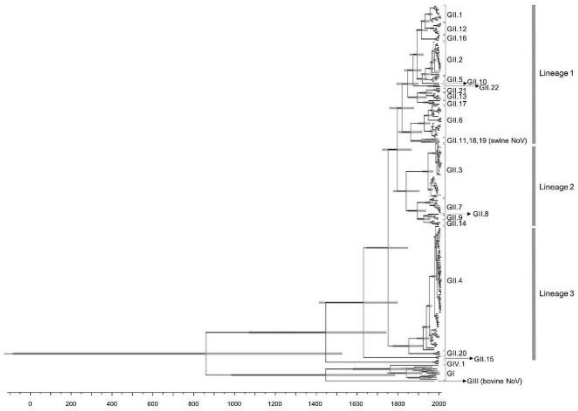

Figure 1. Phylogenetic tree of the capsid gene on NoV constructed by the Bayesian MCMC method. 203 strains of human GII, three strains of swine GII, nine strains of GI, one strain of GIII, and three strains of GIV were included in this tree. Grey bars show 95% HPDs. The scale bar represents actual time (year). The time of the most recent common ancestor of this tree was around 854 CE. GII strains were divided from GIV around 1630 CE. NoV GII was formed three lineages.

endemics of gastroenteritis caused by GII.4 have been recognized for at least 20 years[12–14]. Furthermore, another genotype, GII.P17-GII.17 virus, emerged in 2013 and spread rapidly as GII.4[15].

To gain a better understanding of antigenic variations in the molecular evolution of NoV, it is essential to analyze the capsid gene. The capsid protein, encoded by the second of three open reading frames[1], is crucial for viral adsorption and entry and the production of neutralizing antibodies[16–19]. Thus, predicting the common epitopes in the capsid protein (major antigen) may aid the development of an effective vaccine against NoV.

Recently, various bioinformatics technologies have enabled estimations of the phylogenies and genetic properties of diverse viruses, including NoV[20,21]. For example, the Bayesian Markov Chain Monte Carlo (MCMC) method was used to estimate the evolutionary time-scale of the capsid gene in NoV GI[22]. Siebenga *et al.* and Eden *et al.* reported the molecular evolution of GII.4[20,21]. Furthermore, *in silico* methods may be able to predict the linear and conformational epitopes in the antigens of NoV[23]. Studies on the molecular evolution of NoV GII have been performed in part for some genotypes[20,21]. However, NoV GI and GII are genetically quite different, although they are classified in the same family and genus[1,8]. Moreover, a detailed understanding of the molecular evolution of the capsid gene is an open issue. Therefore, in the present study, we conducted a comprehensive study into the molecular evolution of the capsid gene for all GII genotype strains, using bioinformatics algorithms similar to a previous work[22].

Results

Phylogenetic analysis of NoV capsid gene using Bayesian Markov chain Monte Carlo methods. We constructed a phylogenetic tree, based on the capsid gene by the Bayesian MCMC method (Fig. 1). To gain an understanding of the time scale of the phylogeny of the full-length capsid gene, we used 206 strains of all genotypes of NoV GII (22 genotypes) and 13 strains of other genogroups/genotypes (total 219 strains).

First, the MCMC phylogenetic tree showed that the 22 genotypes of NoV GII strains could be classified into three lineages: lineage 1 (GII.1, 2, 5, 6, 10–13, 16–19, 21 and 22), lineage 2 (GII.3, 7, 8, 9 and 14), and lineage 3 (GII.4, 15 and 20; Fig. 1). Each lineage contained one or two major genotypes (lineage 1, GII.2 and GII.6; lineage 2, GII.3; and lineage 3, GII.4).

Next, the MCMC tree showed that the most recent common ancestor of the tree was around 854 CE (95% highest posterior densities [HPDs] 53 BCE–1537 CE; Fig. 1). The ancestor of the GII strain diverged around 1630 CE (95% HPDs 1409–1796 CE). Three major lineages and the common ancestor of GIV date back to around 1445 CE (95% HPDs 1065–1739 CE). The years of divergence of each lineage, genotype, and genogroup are presented in Supplementary Table S1. Lineage 3 diverged in 1630 CE, lineage 1 in 1819 CE, and lineage 2 in 1839 CE (Fig. 1 and Supplementary Table S1). The mean evolutionary rate of the present human GII strains was estimated to be 3.76×10^{-3} substitutions/site/year (95% HPDs 3.21×10^{-3}–4.30×10^{-3} substitutions/site/year). The results suggested that the present GII strains formed three major lineages at a high evolutionary rate (around 10^{-3} substitutions/site/year) and the common ancestor dates back over 500 years.

Pairwise distances (*p*-distances) among genogroups and lineages. We analyzed the distribution of *p*-distances among the present strains (Supplementary Fig. S1a–d). Human NoV GII had a large *p*-distance (mean ± standard deviation [SD]; 0.286 ± 0.094), based on the nucleotide sequences of the capsid gene (Supplementary Fig. S1a). The maximum pairwise distance was 0.398. The *p*-distance values of lineages 1, 2, and

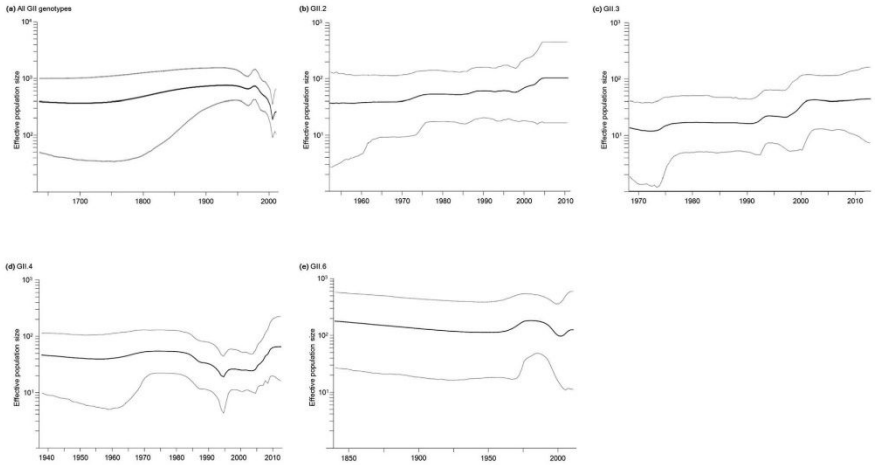

Figure 2. Bayesian skyline plots of all NoV GII (**a**) GII.2 (**b**) GII.3 (**c**) GII.4 (**d**) and GII.6 (**e**). The x-axis represents actual time (years) and starts at mean tree model root height. The y-axis represents the effective population size. Mean effective population size is shown as a black line. HPDs of 95% are shown as grey lines.

3 were 0.283 ± 0.081 (mean $\pm$ SD), 0.205 ± 0.117, and 0.119 ± 0.089, respectively (Supplementary Fig. S1b–d). The results suggested that the capsid gene of NoV GII has a high degree of genetic divergence.

Phylodynamics of human NoV GII strains. We estimated the effective population sizes of the capsid gene of human NoV GII strains in Bayesian skyline plots (BSPs; Fig. 2a). In the present human NoV GII strains, the mean effective population size remained constant until the 1960s. Thereafter, it decreased temporally and increased again around 2000 CE. We also performed BSP analysis of the major prevalent genotypes, such as GII.2, 3, 4, and 6[9–11]. Although the mean effective population sizes of GII.2 and GII.3 grew slowly after the 1970s, those of GII.4 and GII.6 remained unstable throughout the plotted times (1937–2013 for GII.4, 1839–2012 for GII.6) (Fig. 2b–e). Notably, the effective population sizes of GII.4 declined from the 1980s to the middle of the 1990s, but these values increased during the past 15 years (Fig. 2d). The GII.6 values reached a small peak around 1990 and decreased slightly thereafter (Fig. 2e). The GII.2 and GII.3 values increased slightly after 2000 (Fig. 2b,c), and the GII.6 values increased in the 1970/80s and decreased thereafter (Fig. 2e). Overall, the effective population sizes of all NoV GII strains were estimated to be 10^2 for about 400 years. The results suggested that NoV GII strains have become highly adapted to humans over a long period.

Estimation of positive selection sites and negative selection sites in human NoV GII. The selection pressures on each site in the capsid gene were analyzed for the present GII strains. Positively selected sites were estimated by four methods: single likelihood ancestor counting (SLAC), fixed effects likelihood (FEL), internal fixed effects likelihood (IFEL), and mixed effects model of evolution (MEME)[24,25]; 20 sites under positive selection were detected (Table 1). Common sites under positive selection estimated by the four methods occurred after amino acid changes at two sites: Ser6Asn and Asn6Ser/Lys/Ile and Arg435Thr/His, Thr435Pro/Val, Pro435His/Ser, His435Ala/Arg/Gln, Ala435Arg/Ser/His/Val, and Gln435Pro. The mean dN/dS ratio (0.106) obtained by the SLAC method was relatively low (95% confidential intervals; 0.103–0.109). We also detected 489, 498, and 460 sites under negative selection by the SLAC, FEL, and IFEL methods, respectively.

Furthermore, we mapped the 20 positively selected sites in Table 1 in purple and orange on the dimer of the capsid protein (Fig. 3 and Supplementary Fig. S2). Most of the sites were located within the surface of the capsid protein. The results suggested that selective pressure from host causes amino acid substitution of the virus.

Epitopes predicted on the deduced capsid protein in human NoV GII. Previous reports studied B-cell epitope predictions with two distinct definitions: linear and conformational epitopes[26–32]. In this study, we predicted both linear and conformational epitopes of the capsid protein (VP1) in the standard strains of each genotype. Linear epitopes were predicted by combination analysis with seven tools: LEPS[26], Epitopia[27], BCPRED[28], FBCPRED[28], Bepipred[29], Antigenic[30], and LBtope[31], according to a previous report[33]. GII.6 and GII.12 could not be analyzed. The protein sequences of GII.6 (accession No. AJ277620) and GII.12 (accession No. AJ277618) have unknown amino acids (X) because of including mixed nucleotide sequences.

The linear epitopes predicted are shown in Table 2. Notably, a common sequence of 11 amino acids (DPTXXXPAPXG or similar sequence to this) was found in almost all GII genotypes, apart from GII.6 and GII.12. The common epitope motif was located in the protruding 2 (P2) domain, which corresponds to the positions at

Amino acid change	SLAC	FEL	IFEL	MEME
Ser6Asn Asn6Ser, Lys, Ile	○	○	○	○
Asn9Thr, Ser Thr9Ala Ala9Thr Ser9Asn Asn9Thr,Lys,Ser Ala9Val, Thr		○		○
Thr16Ala, Ser Ala16Ser, Thr				○
Val23Ile, Ala Ile23Val, Ser, Ala Ala23Gly			○	
Asn25Ser, Thr, His, Gln, Mix Ser25Asn				○
Glu64Mix, His				○
Cys268Ser, Ala, Val Val268Cys, Ala Ser268Thr				○
Asp297His, Asn, Ser, Gly, Val, Glu His297Pro, Gln, Asp Pro297Ser Gln297His Asn297Ile, Ser Gly297Ser, Pro, Arg, Ala Ser297Asn, Ala			○	
Gly298Asp, Arg, Ala, Ile, Gln, Asn, Lys Asp298Gly, Asn, Glu Arg298Ser				○
Ala303Val, Ile, Thr Thr303Val				○
Asp359Ala Thr359Ser Ala359Ser, Val Ser359Asn, Gly Pro359Thr Ser359Asn			○	○
Ala360Thr, Ser		○		
Gly370Ala, Ser, Mix Ala370Ser, Gly				○
Ser379Thr, Asp, Ala, Gly, Asn, Pro Asp379Asn Gly379Ser, Asp Ala379Ser Asn379Asp Thr379Ser, Ala				○
Asn397Ser, Asp, Glu, Gly, Thr, Gln Ser397Arg, Asp Gly397Ser, Asp, Asp397Glu, Asn His397Arg Thr397Pro Gln397Asp				○
Gly416Asp, Ala, Ser Asp416Gly, Ser, Asn, Glu Glu416Asp His416Asn, Gln, Arg Asn416Arg, Thr, Asp Thr416Pro, Ala Ser416Thr				○
Asp416Asn, Gln, Gly, Ser, Glu, Ala Asn416Asp, Ser, Gly Ser416Ala Gly416Ser				○
Ala419Thr Thr419Asn, Ala Asn419Asp, Ala Asp419Gly Thr419Ala Asp419Pro	○	○	○	○
Continued				

Amino acid change	SLAC	FEL	IFEL	MEME
Arg435Thr, His Thr435Pro, Val Pro435His, Ser His435Ala, Arg, Gln Ala435Arg, Ser, His, Val Gln435Pro				○
Trp485Phe				○

Table 1. Positive selection sites on capsid gene in human NoV GII. mean $dN/dS = 0.106$ (95% CI = 0.103–0.109). Cut off p-value < 0.05.

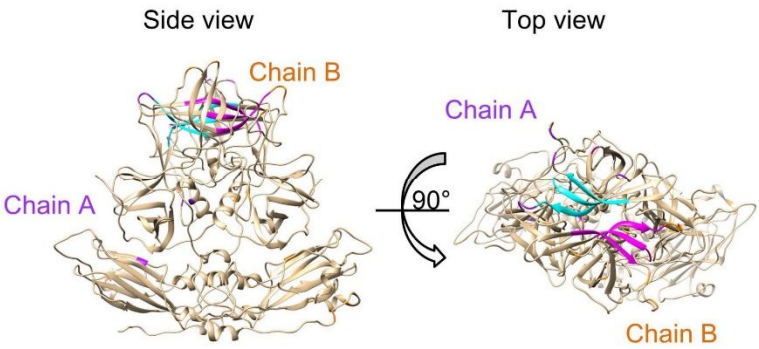

Figure 3. Location of positive selection sites on predicted structure of capsid protein in GII.4/Bristol/1993/UK. To construct the model, we used five suitable templates of NoV capsid sequences (PDB ID: 1IHM, 3ONU, 4RLZ, 3PUM, and 4X07). Twenty positively selected sites on chains A and B are colored purple and orange, respectively. The HBGA binding sites[45] are colored blue and pink. These sites were located within the surface of the protein.

amino acids (aa) 312–322 in the capsid protein of GII.4/Bristol/1993/UK strain. Figure 4 and Supplementary Fig. S3 show the common linear epitopes on the predicted capsid protein structure (dimer) in green and blue.

Next, we predicted the conformational epitopes using CBtope[32]. For each genotype, 4–36 sites were estimated to be conformational epitopes (Supplementary Table S2). The epitopes were mainly located in the P1 and P2 domains on the capsid protein (Fig. 5 and Supplementary Fig. S4).

Discussion

We completed a comprehensive study on the molecular evolution of the capsid gene in all genotypes of NoV (GII). As a result, we estimated that the common ancestor of the present GII strains diverged from a GIV strain with a high evolutionary rate (around 10^{-3} substitutions/site/year) around 1630 CE and formed three major lineages. The capsid gene in the present GII strains shows a high level of divergence (maximum p-distance >0.39). Furthermore, some significant findings were made. 1) The effective population sizes of the present GII strains were relatively large (over 10^2) during 400 years. 2) Some positive (20 sites) and many negative (over 450 sites) selection sites were estimated. 3) Some linear and conformational B-cell epitopes were found in the predicted capsid protein of GII.

The results suggest that NoV GII strains rapidly evolved with high levels of genetic divergence and adaptation to humans. However, since we obtained the GII capsid gene sequences from GenBank alone, the present data may be subject to selection bias. In addition, the present alignment data of the nucleotide sequences may have a sequence length bias, because these strains belonging to various genogroups show the different nucleotide lengths of the capsid genes. This may reflect on the accuracy of the data. Thus, the bias may limit the present study.

We conducted phylogenetic analyses by the Bayesian MCMC method. The results showed that GII strains formed three major lineages and 22 genotypes with high genetic divergence (Fig. 1). Moreover, the MCMC tree estimated that the common ancestor GII diverged from another genogroup, GIV, about 380 years ago (1630 CE; Fig. 1 and Supplementary Table S1). Thereafter, the present GII strains formed 22 genotypes (Fig. 1). Previous studies reported the molecular evolution of some genotypes/genogroups of NoV[20,22,34]. For example, Kobayashi *et al.* showed that the evolutionary rate of the GI was estimated as 1.26×10^{-3} substitutions/site/years, and GI strains divided into two lineages about 750 years ago[22]. Siebenga *et al.*[20] estimated the most recently common ancestor year of GII.4 as 1982. Rackoff *et al.*[34] reported that the evolutionary rate of GI.3 NoV was 1.25×10^{-3} substitutions/site/year. Furthermore, other ssRNA virus, such as HIV or H3N2 influenza virus, evolved with similar evolutionary rates of about 10^{-3} substations/sites/year[35,36]. In this study, we found that the evolutionary rate

Genotype	Strain (Accession No.)	Position	Predicted epitopes
GII.1	Hawaii virus/1971/US (U07611)	305–326	VTNTNGTPF**DPTEDVPAPLG**TP
		357–366	PKFTPKLGSV
GII.2	Melksham/1994/UK (X81879)	4–15	AS**N**DA**A**PSTDG**A**
		313–326	FD**PSEDIPAPLG**VP
		359–373	VPTY**T**AQYTPKLGQI
		531–541	PMGTGNGRRRV
GII.3	Toronto24/1991/CA (U02030)	59–68	APGG**E**FTVSP
		294–307	T**SR**ASDQ**A**DTPTPR
		325–338	**YDPAEDIPAPLG**TP
		387–400	FDPNQPTKFTPVGV
GII.4	Bristol/1993/UK (X76716)	64–74	FTVSPRNAPGE
		125–135	PPNFPTEGLSP
		251–263	TGPSSAFVVQPQN
		309–326	SNY**DPTEEIPAPLG**TPDF
		436–445	TMPGCSGYPN
GII.5	Hillingdon/1990/UK (AJ277607)	64–73	FTVSPKNSPG
		213–222	TYLVPPTVES
		313–327	F**DLTDDVPAPLG**VPD
		337–351	SQRNRGESNPANR**A**H
		374–385	WNTNDVENQPTK
		439–448	PLKGGFGNPA
GII.7	Leeds/1990/UK (AJ277608)	306–328	ITNTDGTPI**DPTEDTPGPIG**SPD
		338–349	SQRNKNEQNPAT
		358–368	TGGDQYAPKLA
		390–401	VGVAGD**PS**HPFR
GII.8	Amsterdam/98-18/1998/NET (AF195848)	59–72	APAG**E**FTVSPRNAP
		308–327	NLDGSPV**DPTDEVPAPLG**TP
		369–383	FKSPSTDFSDNEPIK
GII.9	VA97207/1997/USA (AY038599)	4–15	AS**N**DA**A**PSSDG**A**
		59–72	APAG**E**FTVSPRNAP
		308–326	LDGSPI**DPTDDTPGPLG**CP
		336–380	ASQRGPGDATR**A**HEARIDTG**S**DTFAPKIGQVRFYSTS**S**DFETNQP
GII.10	Erfurt/546/2000/DE (AF427118)	4–15	AS**N**DA**A**PSSDG**A**
		204–223	TRPTPDFDFTYLVPPTVESK
		295–304	QDEHRG**T**HWN
		310–329	LNGTPF**DPTEDVPAPLG**TPD
		340–353	QRNTNT**VP**GEGDLP
		384–396	**Q**DVSSGQPTKFTP
GII.13	Fayetteville/1998/US (AY113106)	217–230	PPSVESKTKPFTLP
		250–263	YTAPNETNVVQCQN
		308–325	PNGASY**DPTDEVPAPLG**T
GII.14	M7/1999/US (AY130761)	307–325	LDGSPI**DPTDDMPAPLG**TP
		363-385	IGQVRFKSSS**D**DFDLHDPTKFTP
		455–466	EHFYQEAAPSQS
GII.15	J23/1999/US (AY130762)	20–30	VPE**SQQ**EVLPL
		316–336	EPDGEEF**SPTGPNPAPVG**TPD
		349–359	NTGGAGQNSNR
		427–440	AGKLAPPVAPNYPG
GII.16	Tiffin/1999/USA (AY502010)	310–325	GTPF**DPTDDVPAPLG**M
		338–349	QRDTGTNPANR**A**
		359–378	AKYTPKLGSVQIGTWD**TE**DL
		380–389	ERQPVKFTPV
		434–447	FRSYIPLKGGHGDP
GII.17	CS-E1/2002/USA (AY502009)	7–17	DA**A**PSNDG**A**TG
		314–328	F**DPTEDVPAPLG**TPD
		341–351	NVGSNPNTTR**A**
		365–379	PKLGSVNFGSTSTDF

Continued

Genotype	Strain (Accession No.)	Position	Predicted epitopes
		420–432	PPIAPNFPGEQLL
GII.20	Luckenwalde591/2002/DE (EU373815)	59-69	APGG**E**FTVSPR
		125–135	PPNFPPENLSP
		308–323	NGSAY**DPTEDIPAVLG**
		337–346	QRSPNNSTR**A**
		350–361	TLNTGSPRYTPK
GII.21	IF1998/2003/IR (AY675554)	2–12	AS**K**DA**A**PSNDG
		211–222	TYLVPPSVESKT
		248–261	YTSPNADVVVQPQN
		310–323	TY**DPTEDVPAPFG**T
		335–348	TQNPRASGDEAAN**S**
		374–384	GHHSQHQQSKF
		457–468	HFYQESAPSQSD
GII.22	YURI2002/JP (AB083780)	159–172	PDVRNQFFHYNQVN
		217–226	PPTVESRTKP
		315–328	**DPTEDVPAPLG**TPD
		341–369	NDYN**D**GSQGPANR**A**HDAVVPTT**S**AKFTPK
		441–450	LKGGHGNPAI

Table 2. Predicted linear B-cell epitopes of standard strains for each genotype Linear epitopes of GII.6 and GII.12 could not be predicted. The positions of the amino acids correspond to each strain. Common epitopes sequences are shown in the bold letters. Positive selection sites are shown in underlined text.

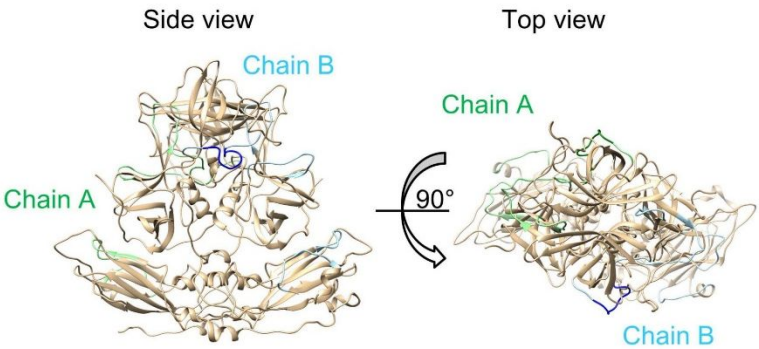

Figure 4. Predicted linear B-cell epitopes mapping on the capsid protein of GII. 4. The predicted structure of capsid protein is the same as in Fig. 3. Linear B-cell epitopes on chain A and B are shown in green and blue, respectively. Common locations among all genotypes are represented by deeper tones. These sites consist of 11 amino acids (DPTXXXPAPXG or similar sequence to this).

of the GII capsid gene was as rapid as that of the GI capsid gene[22]. To our knowledge, these are first descriptions of the evolution of the all genotypes of GII capsid gene.

Our previous study suggested that human NoV GI also had high genetic divergence (maximum *p*-distance values >0.39). The present MCMC tree suggested that all genogroups of NoV have high genetic divergence. These findings may, therefore, indicate the biological divergence of capsid function and host specific infectivity.

Next, the effective population size may reflect virus genome populations in the host during the periods analysed[37]. The effective population size of the present NoV GII strains was relatively large (over 10^2) for 350 years (Fig. 2a). Our previous study indicated that NoV GI had a large effective population size (about 10^3) for 500 years[22]. Therefore, like the NoV GI strains, GII strains have become highly adapted to humans because of the effects of natural selection rather than genetic drift. We analyzed the BSP of the major prevalent genotypes, including GII.2, GII.3, GII.4, and GII.6 (Fig. 2b–e). Previous molecular epidemiological reports suggested that these genotypes appeared within the last 20 years[9–11]. Among them, GII.4 is the most dominant[9–11]. Specifically, this genotype has been detected in patients with acute gastroenteritis in various countries since the 1990s[12–14]. Some variants of GII.4 emerged and spread around these countries[1,12–14,20,21]. The BSP data from the present study show that the effective population size of GII.4 increased since 2000 (Fig. 2d). The periods of increased effective

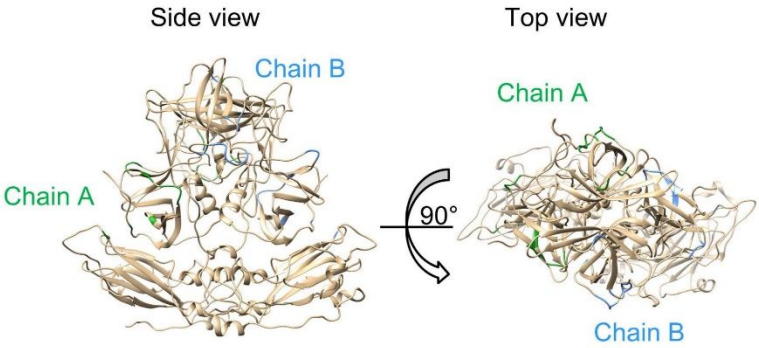

Figure 5. Predicted conformational B-cell epitopes mapping on the capsid protein of GII. 4. The predicted structure of capsid protein is the same as in Fig. 3. These sites on chain A and B are shown in green and blue, respectively. Most of conformational epitopes were located in the P1 and P2 domains.

population size were preceded by periods of prevalence; such fluctuations in BSP data may help predict the prevalence of NoV. However, we did not exactly examine these relationships among the genogroups, because the data are scarce at present[9–11]. Hence, further and larger studies of each genotype and predictions of their prevalence may be needed.

Host defense mechanisms may affect viral antigens and lead to virus escape mutations[38]. Such substitutions are thought to represent positive selection[38]. In the present GII strains, positive selection was estimated at 20 sites of amino acid substitutions, though the SLAC method estimated two sites (Table 1). The sites under positive selection were mainly located in the P2 domain. In our previous study of NoV GI capsid gene evolution, 19 sites under positive selection were estimated by the MEME method, and no sites were estimated, by the SLAC method, even in the P2 domain[22]. The SLAC method is appropriate for detecting non-neutral evolution[24] and may be a stricter algorithmic model for estimating positive selection sites. On the other hand, the MEME method considers lineage-to-lineage variations by a nonsynonymous (dN) and synonymous (dS) substitutions ratio (dN/dS)[25]. This method is suitable for estimating episodic selective pressure[25]. Thus, the difference of the algorithm reflected the numbers of positive selection sites in the present GII strains. Together, host defence mechanisms and immunity are more effective against the GII capsid protein. The antigenicity of the GII strains may be stronger than that of the GI capsid protein, because the capsid protein in the P2 domain may largely reflect the antigenicity of NoV[1,17].

In the present study, over 450 sites under negative selection were confirmed in the NoV GII capsid protein. Mahar *et al.*[39] reported many sites under negative selection in the GII capsid protein. Moreover, our previous data showed a large number (over 400 sites) in NoV GI capsid protein, although the locations of the sites under negative selection were different[22]. Negative selection may rephrase stabilising selection[38]. This type of selection may act to eliminate variant genomes, leading to adaptation to an environment, because most of these mutations are deleterious[38]. Thus, negative selection in the present GII strains may prevent deteriorations of capsid protein functions, including infectivity. Furthermore, it may be important to clarify the roles of the negative selections in NoV capsid proteins, although numerous codon substitutions as negative selection sites are inferred in the NoV GII capsid protein. However, regarding each substitution, it may be difficult to computationally and experimentally examine the stability and folding of NoV capsid protein.

In this study, we used four methods (i.e., FEL, IFEL, SLAC, and MEME) to make a candidate list of positively and negatively selected amino acid sites. Based on these analyses, we showed that the biological significance of these sites was validated with the structural data. However, these methods may have advantages and disadvantages[40]. Thus, further and larger studies, including the fitting of the bioinformatics technology, may be needed to understand the roles of the negative selection in the capsid protein.

In addition, we predicted both linear and conformational B-cell epitopes in the capsid protein in GII for all genotype strains. Some epitopes were confirmed for each genotype strain (Table 2 and Supplementary Table S2) by both methods. First, the common location of linear epitopes, apart from GII.6 and GII.12, were confirmed, and the common motif was DPTXXXPAPXG in GII.1, 4, 8, 10, 13, 14, 16, 17, 21, and 22 (Table 2), located at the side of the P2 domain as shown in a deeper tone (Fig. 4 and Supplementary Fig. S3). Moreover, some conformational epitopes were confirmed in each genotype (Supplementary Table S2). Most of the predicted epitopes, however, did not overlap with the blockade epitopes A, D, and E amino acid residues and locations of the capsid protein that predicted with GII.4 NoV[41] (Fig. 5 and Supplementary Fig. S4). In particular, the common motif DPTXXXPAPXG may not relate to blocking of the HBGA binding. However, it may have an important function that is related to an internalising receptor binding because it is highly conserved among the NoV genotypes.

Previous studies suggest that different NoV genotype strains infect humans[42]. Furthermore, humoral immunity against NoV may not persist for long[42]. Thus, the protective (neutralising) antibodies against the common epitopes in NoV GII strains may not be produced in the host. Alternatively, if antibodies against the common

epitopes are produced, they cannot prevent NoV infection of host cells. Further studies on common epitopes in NoV are needed.

Next, histo-blood group antigens (HBGAs) in the host cells may be associated with the binding of NoV GII capsid protein to the P2 domain[43], and this association may be important for viral attachment to host cells[44]. For example, Cao *et al.*[45] showed that aa336, aa345, and aa374 in the P2 domain of GII.4/VA387/1998/US strain could bind HBGA, and these were associated with NoV GII infections in the host. Furthermore, host defence mechanisms (i.e., humoral immunity) produce protective antibodies against NoV. If amino acid substitutions occur around HBGA binding sites, the antibodies that block HBGA binding cannot protect the host efficiently against NoV infection[42]. Amino acid substitutions under positive selection were observed at residues 370 and 397, adjacent to the HBGA binding sites (Table 1). In addition, B-cell epitopes may be associated with sites under positive selection[46]. Thus, these substitutions might protect against host immunity.

In conclusion, the common ancestor of GII diverged from GIV around 1630 CE at a high evolutionary rate. The GII capsid gene had very high divergence. In addition, the effective population sizes of GII strains had relatively large values during a prolonged period. NoV GII may have been affected by natural selection and strong selective pressure from the host and may have adapted to humans through these evolutionary processes affecting the capsid gene. These results will be a basis of prediction of escape mutants or novel genotype. While our data should be helpful for developing vaccines or for preventing epidemics, further study is needed.

Methods

Strains used in this study. We obtained a comprehensive range of the full-length nucleotide sequences (1620 nt for GII.4/Bristol/1993/UK, Genbank accession No. X76716) of human NoV GII capsid gene, excluding ORF1/2 recombinant strains from GenBank in August 2014. A total of 1582 strains were obtained, and the year in which they were detected was clearly described. These sequences were aligned by Clustal W2[47]. Strains with more than 97.5% identity were excluded from the dataset. Ultimately, 203 strains were used in this study. The average nucleotide divergence in the dataset was 0.54.

Phylogenetic tree constructed by Bayesian MCMC method. We used Bayesian MCMC method in BEAST package v1.8.2 to estimate the time-scaled phylogenies[48]. To estimate the ancestor of various genogroups of NoV, we added 13 outgroups of NoV, including NoV GI (human type), GII (porcine type), GIII (bovine type), and GIV (human type). Detailed data of the strains are shown in Supplementary Table S3.

First, the substitution model was selected using KAKUSAN 4[49] with GTR-Γ model. Next, three clock models (strict clock, uncorrelated lognormal relaxed clock, and uncorrelated exponential relaxed clock) and four demographic models (constant size, exponential growth, expansion growth, and logistic growth) were calculated by generating 100,000,000 steps with sampling every 20,000 steps. These models were compared by Akaike's Information Criterion through MCMC (AICM) using Tracer[50,51]. The lowest AICM value was used. Finally, 219 strains were analysed using exponential clock and exponential growth models with coalescent tree prior. The MCMC chain length was 500,000,000 steps with sampling every 20,000 steps. Convergence was evaluated by the effective sample size by Tracer[51], and values more than 200 were acceptable. The maximum clade credibility tree was obtained after 10% burn-in using TreeAnnotator v1.8.2[48]. The MCMC phylogenetic tree was constructed by FigTree v 1.4.0[48]. The reliability of branches is supported by 95% HPDs.

The evolutionary rate of human NoV GII was also estimated. In this calculation, 203 strains were tested under the best-fit model (GTR-Γ + lognormal relaxed clock + constant size). The MCMC chain length was set at 100,000,000 steps with sampling every 20,000 steps.

Calculation of pairwise distance (*p*-distance). We analyzed *p*-distances to assess the genetic distances between human GII strains. The *p*-distance values of intergenogroup and interlineages were calculated using MEGA 6.0[52].

Bayesian skyline plot analysis. BSP analysis was performed to estimate the phylodynamics in human GII strains. Human GII (203 strains) were analysed with the BSP coalescent prior using BEAST v1.8.2[48]. The substitution and clock models were selected using AICM, as mentioned earlier. Datasets were analysed using a GTR-Γ exponential clock model. MCMC chains were run for 1,000,000,000 steps with sampling every 20,000 steps. BSP was constructed using Tracer[51]. We also estimated the effective population sizes of the major genotypes such as GII.2, 3, 4, and 6. Calculations of these genotypes were performed as described earlier. The detailed conditions of analysis are shown in Supplementary Table S4.

Selective pressure analysis. To find candidates of positive/negative selected sites in capsid protein on human NoV GII, nonsynonymous (dN) and synonymous (dS) substitutions rates at every codon were calculated using Datamonkey[24]. To multilaterally analyze the selective pressure of NoV capsid gene, we used the following four methods: SLAC, FEL, IFEL, and MEME. SLAC, the fastest method, is appropriate for large (>50) datasets[40]. FEL and IFEL are suitable for intermediate alignments[40]. FEL method directly estimates site-by-site substitutions[40]. Although IFEL method is similar to FEL, it only calculates along the internal branches of the tree[40]. SLAC, FEL and IFEL may appear to underestimate the number of positive selectionsites[25]. MEME method is suitable for estimating episodic positive selections at each site[25]. Sites under positive selection (dN > dS) were determined by a p-value of <0.05. We also estimated negative selection sites (dN < dS) using SLAC, FEL, and IFEL methods. The dN/dS ratio was estimated under the MG94 model in the Datamonkey. The cut off p-value was at 0.05.

B-cell epitope prediction of human NoV GII. We predicted both linear and conformational epitopes in the capsid protein, using the deduced amino acid sequences of the standard strains of each genotype. Linear B-cell

epitopes were predicted using the following seven tools: LEPS[26], Epitopia[27], BCPRED[28], FBCPRED[28], BepiPred[29], Antigenic[30], and LBtope[31]. These tools were used in default conditions and amino acids estimated by four or more tools with >10 consecutive sites were considered linear B-cell epitopes[33]. In addition, conformational epitopes were predicted using CBtope[32]. The threshold of the support vector machine score was set at 0.0.

Mapping of positive selection sites and predicted epitopes. A structural model of the standard strains in each genotype was predicted using MODELLER v9.15[53]. Homology modelling was based on the crystal structure of five strains (PDB ID: 1IHM, 3ONU, 4RLZ, 3PUM and 4X07). The capsid structure of GI (PDB ID: 1IHM) was used to construct the whole structure of the VP1 dimer, including the P1 and shell domains. The structures of five templates and the standard strains were aligned by MAFFTash[54,55]. To surely provide the structures, the sequence identities of templates and targets were 45.3–100%[56]. The constructed models were minimized by GROMOS96[57], implemented in Swiss PDB Viewer v4.1[58] and evaluated by Ramachandran plots through the RAMPAGE server[59]. Final models were modified and coloured by Chimera v1.10.2[60]. Positive selection sites and linear and conformational epitopes of each genotype were mapped on the structures.

References

1. Green, K. Y. In *Fields Virology* 6th edn, Vol. 1 (eds Knipe, D. M. *et al.*) Ch. 20, 582–608 (Lippincott Williams & Wilkins, 2013).
2. Hamano, M. *et al.* Epidemiology of acute gastroenteritis outbreaks caused by Noroviruses in Okayama, Japan. *J. Med. Virol.* **77,** 282–289 (2005).
3. Chan-It, W. *et al.* Emergence of a new norovirus GII.6 variant in Japan, 2008–2009. *J. Med. Virol.* **84,** 1089–1096 (2012).
4. Bernard, H. *et al.* Outbreak Investigation Team. Large multistate outbreak of norovirus gastroenteritis associated with frozen strawberries, Germany, 2012. *Euro. Surveill.* **19,** 20719 (2014).
5. Zomer, T. P. *et al.* A foodborne norovirus outbreak at a manufacturing company. *Epidemiol. Infect.* **138,** 501–506 (2010).
6. Belliot, G., Lopman, B. A, Ambert-Balay, K. & Pothier, P. The burden of norovirus gastroenteritis: an important foodborne and healthcare-related infection. *Clin. Microbiol. Infect.* **20,** 724–730 (2014).
7. Vinjé, J. Advances in laboratory methods for detection and typing of norovirus. *J. Clin. Microbiol.* **53,** 373–381 (2015).
8. Kroneman, A. *et al.* Proposal for a unified norovirus nomenclature and genotyping. *Arch. Virol.* **158,** 2059–2068 (2013).
9. Centers for Disease Control and prevention. *CaliciNet Data.* Available at: http://www.cdc.gov/norovirus/reporting/calicinet/data.html (Accessed: September 17, 2015) (2015).
10. Infectious Disease Surveillance Center, National Institute of Infectious Diseases. Epidemiology of Norovirus in Japan, 2010/11–2013/14 seasons. *IASR.* **35,** 161–163 (2014).
11. Infectious Disease Surveillance Center, National Institute of Infectious Diseases. Norovirus epidemic in Japan during 2006/07–2009/10 seasons. *IASR.* **31,** 312–314 (2010).
12. Bull, R. A. *et al.* Emergence of a new norovirus genotype II.4 variant associated with global outbreaks of gastroenteritis. *J. Clin. Microbiol.* **44,** 327–333 (2006).
13. Noel, J. S. *et al.* Identification of a distinct common strain of "Norwalk-like viruses" having a global distribution. *J. Infect. Dis.* **179,** 1334–1344 (1999).
14. Vinjé, J., Altena, S. A. & Koopmans, M. P. The incidence and genetic variability of small round-structured viruses in outbreaks of gastroenteritis in The Netherlands. *J. Infect. Dis.* **176,** 1374–1378 (1997).
15. Han, J. *et al.* Emergence and predominance of norovirus GII.17 in Huzhou, China, 2014–2015. *Virol. J.* **12,** 139 (2015).
16. Harrison, S. C. In *Fields Virology* 6th edn, Vol. 1 (eds Knipe, D. M. *et al.*) Ch. 3, 52–86 (Lippincott Williams & Wilkins, 2013).
17. Chakravarty, S., Hutson, A. M., Estes, M. K. & Prasad, B. V. Evolutionary trace residues in noroviruses: importance in receptor binding, antigenicity, virion assembly, and strain diversity. *J . Virol.* **79,** 554–568 (2005).
18. Prasad, B. V. *et al.* X-ray crystallographic structure of the Norwalk virus capsid. *Science.* **286,** 287–290 (1999).
19. Tan, M., Hegde, R. S. & Jiang, X. The P domain of norovirus capsid protein forms dimer and binds to histo-blood group antigen receptors. *J. Virol.* **78,** 6233–6242 (2004).
20. Siebenga, J. J. *et al.* Phylodynamic reconstruction reveals norovirus GII.4 epidemic expansions and their molecular determinants. *PLoS Pathog.* **6,** e1000884 (2010).
21. Eden, J. S. *et al.* Recombination within the pandemic norovirus GII.4 lineage. *J. Virol.* **87,** 6270–6282 (2013).
22. Kobayashi, M. *et al.* Molecular Evolution of the Capsid Gene in Norovirus Genogroup I. *Sci. Rep.* **5,** 13806 (2015).
23. Chen, L. *et al.* Bioinformatics analysis of the epitope regions for norovirus capsid protein. *BMC Bioinformatics.* **14,** S5 (2013).
24. Pond, S. L. & Frost, S. D. Datamonkey: Rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics.* **21,** 2531–2533 (2005).
25. Murrell, B. *et al.* Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* **8,** e1002764 (2012).
26. Wang, H. W., Lin, Y. C., Pai, T. W. & Chang, H. T. Prediction of B-cell linear epitopes with a combination of support vector machine classification and amino acid propensity identification. *J. Biomed. Biotechnol.* **2011,** 432830 (2011).
27. Rubinstein, N. D., Mayrose, I., Martz, E. & Pupko, T. Epitopia: a web-server for predicting B-cell epitopes. *BMC Bioinformatics.* **10,** 287 (2009).
28. EL-Manzalawy, Y., Dobbs, D. & Honavar, V. Predicting linear B-cell epitopes using string kernels. *J. Mol. Recognit.* **21,** 243–255 (2008).
29. Larsen, J. E., Lund, O. & Nielsen, M. Improved method for predicting linear B-cell epitopes. *Immunome Res.* **2,** 2 (2006).
30. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends Genet.* **16,** 276–277 (2000).
31. Singh, H., Ansari, H. R. & Raghava, G. P. Improved method for linear B-cell epitope prediction using antigen's primary sequence. *PLoS One* **8,** e62216 (2013).
32. Ansari, H. R. & Raghava, G. P. Identification of conformational B-cell Epitopes in an antigen from its primary sequence. *Immunome. Res.* **6,** 6 (2010).
33. Kim, Y. J. *et al.* Rapid replacement of human respiratory syncytial virus A with the ON1 genotype having 72 nucleotide duplication in *G* gene. *Infect. Genet. Evol.* **26,** 103–112 (2014).
34. Rackoff, L. A., Bok, K., Green, K. Y. & Kapikian, A. Z. Epidemiology and evolution of rotaviruses and noroviruses from an archival WHO Global Study in Children (1976–79) with implications for vaccine design. *PLoS One.* **8,** e59394 (2013).
35. Westgeest, K. B. *et al.* Genomewide analysis of reassortment and evolution of human influenza A(H3N2) viruses circulating between 1968 and 2011. *J. Virol.* **88,** 2844–2857 (2014).
36. Roy, C. N., Khandaker, I. & Oshitani, H. Evolutionary Dynamics of Tat in HIV-1 Subtypes B and C. *PLoS One.* **10,** e0129896 (2015).
37. Holmes, E. C. In *Fields Virology* 6th edn, Vol. 1 (eds Knipe, D. M. *et al.*) Ch. 11, 286–313 (Lippincott Williams & Wilkins, 2013).
38. Domingo, E. In *Fields Virology* 5th edn, Vol. 1 (eds Knipe, D. M. *et al.*) Ch. 12, 389–421 (Lippincott Williams & Wilkins, 2007).
39. Mahar, J. E., Bok, K., Green, K. Y. & Kirkwood, C. D. The importance of intergenic recombination in norovirus GII.3 evolution. *J. Virol.* **87,** 3687–3698 (2013).

40. Pond, S. L. & Frost, S. D. Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. *Mol. Biol. Evol.* **22,** 1208–1222 (2005).
41. Lindesmith, L. C. *et al.* Broad blockade antibody responses in human volunteers after immunization with a multivalent norovirus VLP candidate vaccine: immunological analyses from a phase I clinical trial. *PLoS Med.* **12,** e1001807. (2015).
42. Pringle, K. *et al.* Noroviruses: epidemiology, immunity and prospects for prevention. *Future Microbiol.* **10,** 53–67 (2015).
43. Choi, J. M., Hutson, A. M., Estes, M. K. & Prasad, B. V. Atomic resolution structural characterization of recognition of histo-blood group antigens by Norwalk virus. *Proc. Natl. Acad. Sci. USA* **105,** 9175–9180 (2008).
44. Murakami, K. *et al.* Norovirus binding to intestinal epithelial cells is independent of histo-blood group antigens. *PLoS One.* **8,** e66534 (2013).
45. Cao, S. *et al.* Structural basis for the recognition of blood group trisaccharides by norovirus. *J. Virol.* **81,** 5949–5957 (2007).
46. Chen, P. *et al.* Computational evolutionary analysis of the overlapped surface (S) and polymerase (P) region in hepatitis B virus indicates the spacer domain in P is crucial for survival. *PLoS One.* **8,** e60098 (2013).
47. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics.* **23,** 2947–2948 (2007).
48. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7,** 214 (2007).
49. Tanabe, A. S. Kakusan4 and Aminosan: two programs for comparing nonpartitioned, proportional and separate models for combined molecular phylogenetic analyses of multilocus sequence data. *Mol. Ecol. Resour.* **11,** 914–921 (2011).
50. Suchard, M. A., Weiss, R. E. & Sinsheimer, J. S. Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol.* **18,** 1001–1013 (2001).
51. Rambaut, A. & Drummond, A. J. *Tracer.* (2013) Available at: http://tree.bio.ed.ac.uk/software/tracer. (Accessed: 11th December 2014).
52. Tamura, K. *et al.* MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **30,** 2725–2729 (2013).
53. Webb, B. & Sali, A. Protein structure modeling with MODELLER. *Methods. Mol. Biol.* **1137,** 1–15 (2014).
54. Standley, D. M., Toh, H. & Nakamura, H. ASH structure alignment package: sensitivity and selectivity in domain classification. *BMC Bioinformatics*. **8,** 116 (2007).
55. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucl. Acids Res.* **30,** 3059–3066 (2002).
56. Dolan, M. A., Noah, J. W. & Hurt, D. Comparison of common homology modeling algorithms: application of user-defined alignments. *Methods Mol. Biol.* **857,** 399–414 (2012).
57. van Gunsteren W. F. *et al.* In *Biomolecular Simulation: The GROMOS96 Manual and User Guide*, 1–1042 (Vdf Hochschulverlag AG an der ETH, 1996).
58. Guex, N. & Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis.* **18,** 2714–2723 (1997).
59. Lovell, S. C. *et al.* Structure validation by Calpha geometry: phi, psi and Cbeta deviation. *Proteins*. **50,** 437–450 (2003).
60. Pettersen, E. F. *et al.* UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).

Acknowledgements

This work was partly supported by a commissioned project for Research on Emerging and Re-emerging Infectious Diseases from the Japanese Ministry of Health, Labour and Welfare and Japan Agency for Medical Research and Development.

Author Contributions

H.K. and K.K. designed the study. M.K., Y.M. and M.I. analysed the data. T.M., N.S., N.S., R.O., K.N., Y.Y., M.K., N.S., A.R., T.S., Y.M., K.S., T.T., H.S., N.O., K.N. and Y.S. contributed analysis tools. H.K., M.K., Y.M. and K.K. wrote the paper.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Kobayashi, M. *et al.* Molecular evolution of the capsid gene in human norovirus genogroup II. *Sci. Rep.* **6**, 29400; doi: 10.1038/srep29400 (2016).