Archives of Virology (2022) 167:1235–1245
https://doi.org/10.1007/s00705-022-05396-9

ORIGINAL ARTICLE

Genomic stratification and differential natural selection signatures among human norovirus genogroup II isolates

Sehrish Kakakhel[1] · Hizbullah Khan[1] · Kiran Nigar[1] · Asifullah Khan[1]

Received: 5 October 2021 / Accepted: 12 January 2022 / Published online: 23 March 2022

Abstract

Noroviruses (NoVs), which are members of the family *Caliciviridae*, are the most common cause of gastroenteritis in humans. Ten NoV genogroups have been reported so far. Of these, genogroup II (GII) is the most prevalent, and it causes serious infections worldwide. The complete genome sequences of NoV GII isolates from different geographical regions were retrieved from the public database. The model-based clustering approach, implemented in the STRUCTURE resource, was employed for assessment of genetic composition. The MEGA X and IQ Tree tools were used for phylogenetic analysis. Genome-wide natural selection analysis was performed using maximum-likelihood-based methods. The demographic features of NoV GII genome sequences were assessed using the BEAST package. All of the NoV GII sequences initially clustered into two main subpopulations at significant K = 2, where the genotype GII.4 samples clearly split from the rest of the genotypes. This indicates a marked genetic distinction between norovirus GII.4 and non-GII.4 samples. Phylogenetic analysis showed the presence of five distinct subclades for genotype GII.2 and seven subclades for GII.4 samples. Several isolates with admixed ancestry were identified that constituted distinct subclusters in the phylogenetic tree. No continental-specific genetic distinctions were observed among the NoV GII samples. Significant genomic signatures of both positive and negative natural selection were identified across the NoV GII genes. A differential pattern of positive selection signals was inferred between the GII.4 and non-GII.4 genotypes. The demographic analysis revealed an increase in the effective population size of NoV GII during 2009-2010, followed by a rapid fall in 2015.

Keywords norovirus genogroup II · genomic diversity · spatiotemporal · selection pressure

Introduction

Noroviruses (NoVs) are the most important pathogens causing viral gastroenteritis in humans, causing about 50% of all acute gastroenteritis cases [1, 2]. The World Health Organization (WHO) has estimated that NoV caused 684 million cases and 212,489 deaths worldwide in 2010 [3]. NoV infections are most prevalent in children and the elderly, causing severe symptoms and prolonged shedding [4]. NoV infection spreads among humans through multiple routes, including waterborne, foodborne, and person-to-person transmissions [5].

NoVs are small positive-strand RNA viruses belonging to the genus *Norovirus*, family *Caliciviridae*. The approximately 7.5-kb genome of NoVs contains three open reading frames (ORFs) [6]. ORF1 encodes six nonstructural proteins, including NS1/2 (p48), NS3 (NTPase), NS4 (p22), NS5 (vpg), NS6 (3C-like protease), and RNA-dependent RNA polymerase (RdRp) [7]. ORF2 encodes a major structural protein (VP1), which forms the virus capsid, whereas ORF3 encodes a minor structural protein (VP2). The VP1 is comprised of a conserved shell (S) domain and two protruding (P) domains, P1 and P2. The P1 domain contributes to capsid stability, while the P2 domain facilitates binding of this protein to histoblood group antigens (HBGAs) [8]. NoVs are classified into 10 genogroups based on VP1 protein diversity [9]. These genogroups are further classified into 49 genotypes based on the VP1 coding region and 60 genotypes based on the RdRp coding region. The

Handling Editor: Akbar Dastjerdi

✉ Asifullah Khan
asif@awkum.edu.pk

[1] Department of Biochemistry, Abdul Wali Khan University Mardan, Mardan, Pakistan

genogroups I, II, IV, GVIII, and GIX have been reported to be associated with human diseases [10], with GII being most commonly responsible for outbreaks worldwide [11]. Within genogroup II (GII), genotype GII.4 is predominant, and the emergence of new genetic variants has resulted in a pandemic [12]. The major global variants characterized so far include the Sydney_2012, Den Haag_2006, and New Orleans_2009 variants [10, 4]. The predominance of genotype GII.4 has persisted for over two decades due to its fast rate of mutation and evolution [13]. Non-GII.4 genotypes have also caused massive epidemics and transiently surpassed genotype GII.4. These include the recently emerged GII.17 and GII.2 lineages. A novel GII.17 variant, termed the Kawasaki genotype, appeared as the primary cause of outbreaks in some Asian countries and replaced the Sydney_2012 variant [14]. However, among children, genotype GII.3 commonly causes irregular NoV infections [15]. The NoV genetic repertoire substantially expands within and between genotypes through recombination events [16].

There are no antiviral medicines or vaccines available so far to combat NoV infection [6, 17]. The complete genome sequences of NoV GII isolates from different geographical regions are available in the public genome sequence repositories. The high prevalence of NoV GII along with the recent emergence of novel strains provoked us to examine the complete genome sequences of this genogroup to understand their genetic composition and distinguishing features and the extent of possible genetic admixtures. In addition, natural selection and recombination analyses were performed to understand the possible role of these events in shaping the genetic structure of NoV GII. The findings of the current study, based on genomic features of NoV GII isolates worldwide, may have implications for devising effective vaccines against NoV infection.

Methodology

Retrieval of genome sequences

Complete genome sequences of human NoV GII isolates were obtained from the Virus Pathogen Resource database hosted by NCBI [18]. Several NoV genome sequences are present in public databases with no genotype information. The genotype information for such sequences was obtained using the NoV automated online genotyping tool (version 2.0) [19]. Genome sequences that were submitted without location, host, or sampling time information were excluded. Finally, a dataset comprising 822 complete genome sequences of NoV GII was generated (Supplementary Table S1).

Multiple sequence alignment and identification of parsimony-informative sites

A multiple sequence alignment (MSA) was performed using Clustal Omega 3 [20]. MEGA X was used to extract parsimony-informative (PI) sites from the alignment. A total of 4069 PI sites were acquired from aligned data.

Linkage analysis

The LIAN v3.5 tool was employed to examine the null hypothesis and the linkage equilibrium within NoV GII genomic data [21]. This program calculates the standardized index of association (I^SA) to quantify the haplotype-wide linkage derived from the dataset. In addition, |D'| and r^2 were computed via DnaSpv6.0 [22] to measure the linkage disequilibrium (LD). |D'| represents the absolute value of the difference between the observed and expected haplotype frequency in the absence of LD. The variance of the allele frequency between the observed and expected haplotype is represented by r^2 [23].

Population structure analysis

The genetic structure of NoV GII was analyzed using a Bayesian model-based clustering program, i.e., STRUCTURE v2.3.4 [24]. The STRUCTURE program identifies the genetically distinct subpopulation in a given dataset based on differences in allelic frequency and probabilistically assigns individuals to subpopulations. STRUCTURE operates via an admixture model with the correlated allele frequency. The admixture model accounts for the individual holding mixed ancestry and allocates such admixed strains to their specific subpopulations probabilistically [25]. The analysis was performed with a burn-in length of 100,000, followed by 100,000 MCMC iterations with default parameters (i.e., Dirichlet parameter α and allele frequency parameter). Five independent runs were performed for each value of K (1 to 15). The K_{opt} (optimum number of subclusters) was determined by the Evanno ΔK approach using the STRUCTURE HARVESTER resource [26, 27]. A plot of K vs ΔK was used to determine K_{opt}. The value of K_{opt} was confirmed using various combinations of burn-ins burn-in lengths, including 50,000-50,000, 70,000-70,000, and 100,000-100,000.

F-statistics and PCA analysis

The NoV GII genetic composition estimates were additionally corroborated by F-statistics known as fixation index (F_{ST}) calculation and principal component analysis (PCA). The F_{ST} was calculated by analysis of molecular variance

(AMOVA) implemented in ARLEQUINv3.11 with 1,000 permutations [28]. AMOVA calculates the partitioning variance at different levels of population subdivision and yields F_{ST}. PCA was performed using PLINKv1.9 [29], and the output results were visualized with the built-in function "prcomp".

Phylogenetic analysis

A neighbor-joining (NJ)-based tree was constructed using MEGA X with a minimum of 1,000 bootstrap replicates. A maximum-likelihood (ML) tree was constructed using IQ tree [30], employing the GTR + I + G substitution model and ultrafast bootstrap replicates [31]. The tree topology was visualized and annotated using FigTree.v1.4.4 [32].

Recombination analysis

The aligned complete genome dataset was used for the identification of potential recombination events using the seven different methods implemented in the RDP4 package, including RDP [33], GENECONV [34], BOOTSCAN [35], MaxChi [36], CHIMAERA [37], SiSCAN [38], and 3SEQ [39]. A recombination event was considered likely if it was identified by at least three of these methods, with a *p*-value of 0.00001.

Demography estimation of NoV GII

Fluctuations in the effective population size of NoV GII with respect to time were inferred for the available isolates using the Bayesian skyline model [40] in BEAST2 [41]. The selection of the best-fit nucleotide substitution model was achieved using jModelTest [42]. GTR+I+G was chosen as the best model of nucleotide substitution. The best clock model was determined using path sampling (PS) and stepping stone sampling (SS), implemented in the BEAST v1.10.4 program by calculating marginal likelihood values. A relaxed uncorrelated clock model was selected as the best-fit model. The MCMC steps were run for a chain length of 300 million generations to ensure convergence. The convergence of the MCMC log output files and effective sample size (ESS) > 200 was analyzed using the Tracerv1.7 program [43].

Natural selection analysis

A dataset of 538 NoV GII sequences was prepared for natural selection analysis. Potentially recombinant samples were excluded from the analyses to avoid inferential biases. A total of eight datasets were generated, corresponding to eight protein coding sequences, i.e., p48, NTPase, p22, vpg, protease, RdRp, ORF2, and ORF3. The accuracy of the selection pressure calculation mainly depends upon the quality of the MSA. Therefore, the quality of the MUSCLE-generated MSA was checked using the GUIDANCE server, which identifies the unreliable alignment regions within an MSA using a confidence threshold score of ~1 [44]. All eight datasets were analyzed separately using different ML-based methods with a default value of 0.1. The methods included single-likelihood ancestor counting (SLAC) [45], internal fixed effects likelihood (IFEL) [46], and fixed effects likelihood (FEL) [47], accessible through the Datamonkey webserver in the HYPHY package [48, 49]. These three methods identify sites that are under the influence of pervasive positive selection across all the lineages in a phylogenetic tree. The run for the identification of the best model was carried out using an automated model selection tool on the Datamonkey server. Episodic positive selection signatures were detected using the MEME (mixed-effects model of evolution) method available on the Datamonkey server. Episodic positive selection affects a few lineages even when the majority of the lineages undergo purifying selection [50].

Results

Linkage analysis

In order to assess the genetic composition using the STRUCTURE program, it is first necessary to evaluate the pattern of linkage of loci. LD is the nonrandom association of alleles at different polymorphic sites. In the case of free recombination, the value of $I^S A$, calculated using LIAN 3.5, is assumed to be zero. The $I^S A$ value obtained for NoV GII sequences was 0.0000 ($P < 10^{-4}$, 10,000 replicates), indicating a signal of linkage equilibrium and weak LD. To confirm the low LD, plots of |D| and r^2 were computed using DnaSP v5. D is the function of LD measurement. The average value of |D| and r^2 was found to be 0.8206 and 0.0522, respectively, indicating that the loci were weakly linked and that the STRUCTURE program was therefore appropriate for analysis of the NoV dataset.

Genetic composition analysis

Clustering analysis using STRUCTURE

The admixture model implemented in STRUCTURE was built for K = 1 to 13, with five independent simulation runs to confirm the consistency of parameter estimates and the reproducibility of the clusters (see Methodology). A K_{opt} of 2 was obtained from the plot of K vs. ΔK (Fig. 1A). This

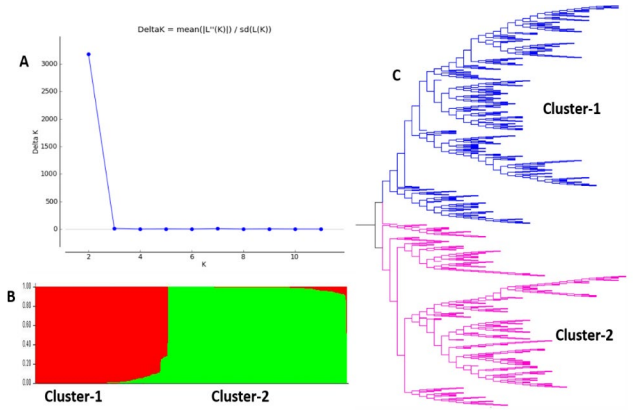


Fig. 1 [A] Determination of K_{opt} for NoV GII. The graph shows a plot of K versus delta K, which defines the optimum number of clusters K_{opt} in the NoV GII population. K represents the number of clusters, while delta K is the rate of change in likelihood posterior probability for the given subcluster K. The plot was executed using large values for simulation burn-ins (100,000) and burn-in length (100,000). The major peak at $K = 2$ shows that the NoV genetic structure is grouped into two main subpopulations. (B) Estimate of the population genetic structure of NoV at a K_{opt} of 2 using an admixture model in STRUCTURE software. C-1 comprises genomic entries from genotypes GII.1, GII.2, GII.3, GII.5, GII.6, GII.7, GII.8, GII.12, GII.13, GII.17, and GII.26, and C-2 comprises genotype GII.4 strains [C] The initial clustering pattern of the phylogenetic tree results is consistent with the STRUCTURE result.

revealed a basic stratification of all of the NoV isolates samples into two subpopulations. In addition, an AMOVA test suggested marked genetic distinction, i.e., $F_{ST} = 0.53293$ ($P = 0000$), between the two subpopulation genetic components. Cluster 1 (C-1) acquired at K_{opt} of 2, comprises all of the NoV genotype samples except GII.4. The GII.4 samples comprised a separate subpopulation (C-2) (Fig. 1B). Several admixed strains were observed in both the C-1 and C-2 clusters obtained at $K_{opt} = 2$. This observed genetic stratification of NoV samples was not congruent with the isolates' geographical origin.

Further analysis was performed to further investigate the genetic stratification in each of the major genetic components. The C-1 cluster was stratified with a significant peak of $K_{opt} = 3$, followed by minor peaks of $K_{opt} = 4$ and $K_{opt} = 5$ (Fig. 2A). The K_{opt} value of 3 reveals the diversification of C-1 into three further subpopulations/lineages, designated as C-1.1, C-1.2, and C-1.3 (Fig. 2B). C-1.1 consists of genotype GII.2 strains. The UK strains from GII.2 genotype were observed to be admixed, with significant membership scores ranging from 0.500 to 0.434 for the clusters C-1.2 and C-1.3. Cluster C-1.2 consists of genotype GII.17 strains, while the C-1.3 cluster consists of GII.3, GII.5, GII.6, GII.7, GII.8, GII.12, and GII.13 strains. At $K_{opt} = 4$, C-1.3 further stratified into two subclusters (i.e., C-1.3a and C-1.3b) (Fig. 2B). The GII.3 samples constitute C-1.3a, while the samples from genotypes GII.5, GII.6, GII.7, GII.12, GII.13, and GII.26 constitute C-1.3b. Likewise, at a K_{opt} of 5, the subclustering of the GII.2 genotype, i.e. formerly comprising the C-1.1 cluster at K_{opt} 3, further stratified into two lineages, i.e. C-1.1a and C-1.1b (Fig. 2B). The overall clustering pattern of samples obtained at K_{opt} of 3, 4, and 5 did not reveal any geography-based distinction among the NoV GII isolates, and genetic stratification was mainly based on genotype identity.

Genetic stratification of GII.4 samples

The GII.4 samples, initially split at $K = 2$, stratified further during subsequent Bayesian clustering analysis (Fig. 2). The samples of genotype GII.4 were stratified into two major lineages and five minor lineages (Fig. 2C). The C-2.1 cluster corresponds to the GII.4-Sydney_2012, GII.4-New Orleans_2009, and GII.4-Apeldoorn_2007 strains, whereas, the C-2.2 cluster corresponds to the Den Haag_2006b strain. At $K = 5$, C-2.1 stratified into three lineages, i.e., C-2.1a, C-2.1b, and C-2.1c, while C-2.2 split into two sublineages, i.e. C-2.2a and C-2.2b (Fig. 2D). The C-2.1a subpopulation includes the Sydney_2012 strains, and C-2.1b includes the Sydney_2012 and New Orleans_2009 GII.4 strains, whereas the C-2.1c cluster includes the GII.4 Sydney_2012 samples. The GII.4-Den Haag_2006b strains initially clustered in C-2.2, but this stratified into two additional lineages, designated as C-2.2a and C-2.2b (Supplementary Table S2).

Principal component analysis

PCA was used to confirm the genetic composition and stratification pattern of NoV GII isolates. The PCA estimated 26.4% of the total genetic variance, with 9.09% of the first PC and 17.31% of the second PC. The principal components (PCs) split the GII.4 samples from the rest of non-GII.4 genotypes (Fig. 3). The genotype GII.4, GII.2, GII.3, and GII.12 samples clustered separately, while the GII.26, GII.17, GII.6, and GII.7 samples clustered closely. The stratification pattern observed in the PCA plot is consistent with the STRUCTURE results.

Phylogenetic analysis

ML- and NJ-based phylogenetic analysis produced similar tree topologies. The phylogenetic tree results were examined according to the clustering pattern obtained using STRUCTURE. The NoV GII samples grouped into ten independent clades in the NJ tree (Fig. 4A), which corresponds to the ten clusters (C-1.1a, C-1.1b, C-1.2, C-1.3a, C-1.3b, C-2.2a, C-2.2b, C-2.1a, C-2.1b, C-2.1c) obtained by STRUCTURE analysis. Some admixed strains were observed in the phylogenetic tree clade represented by the C-1.b*, C-1.2b*,

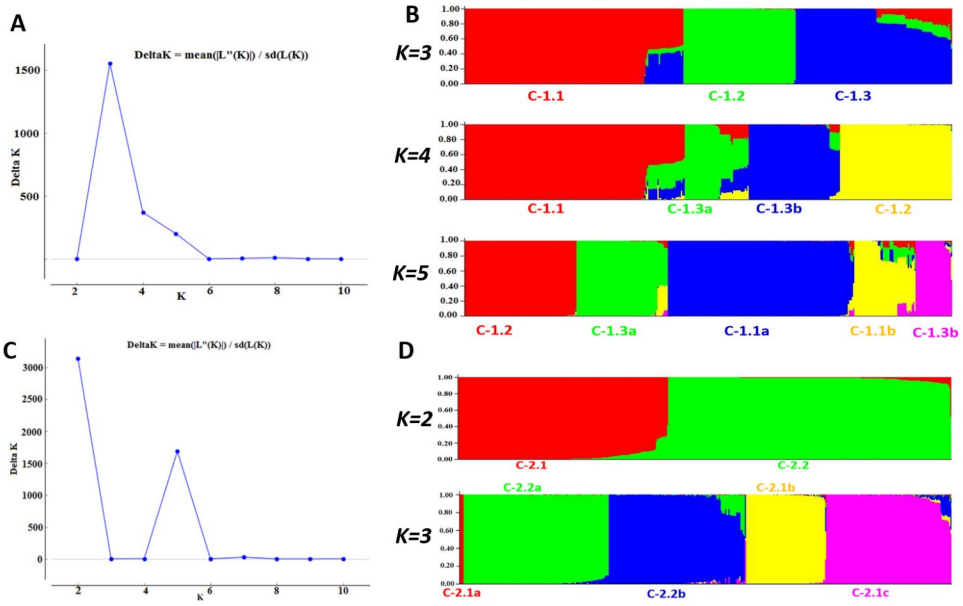
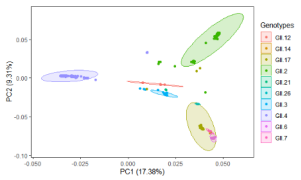


Fig. 2 Additional genetic structure analysis of C-1 and C-2 using an admixture model in STRUCTURE. (A) Plot of *K* versus delta *K* showing the optimum number of subpopulations in C-1. The plot shows a significant peak at K = 3, followed by two minor peaks at K = 4 and K = 5. [B] Sublevel genetic structure of C-1 (non-GII.4 genotypes) obtained using the STRUCTURE program, applying an admixture model. The analysis suggests the presence of three clusters at K = 3, represented as a color bar plot. At K = 3, C-1.1 contains genotype GII.2, C-1.2 contains genotypes GII.17, and C-1.3 consists of genotype GII.3, GII.5, GII.6, GII.7, GII.8, GII.12, and GII.26. At K = 4, four clusters were observed: C-1.1 (GII.2), C-1.2 (GII.17), C-1.3a (GII.3), and C-1.3b (GII.5, GII.6, GII.7, GII.8, GII.12, and GII.26). At K = 5, possible clustering of genotype GII.2 into two lineages (C-1.1a and C-1.1b) can be seen. (C) Genetic structure of genotype GII.4 (cluster 2). At K = 2, there is clustering of C-2.1 (GII.4 Sydney_2012, New Orleans_2009, and Apeldoorn) and C-2.2 (Den_Haag_2006b), and at K = 5, C-2.1a (Sydney_2012), C-2.1b (Sydney_2012 and New Orleans), C-2.1c (Sydney_2012), and C-2.2a (Den_Haag_2006), C-2.2b (Den_Haag_2006) can be seen. (D) Plot of *K* versus delta *K* showing a major peak at K = 2, dividing C-2 (GII.4) into two subpopulations. The second peak was found at K = 5, which shows additional diversification of GII.4 into five lineages.

Fig. 3 The two-dimensional PCA analysis of NoV GII samples



C-1.3b*, and C-2.1a* clusters. The ML phylogenetic tree revealed that cluster C-1.1 further stratified into five minor clades (Fig. 4B). Contrary to the phylogenetic tree stratification, STRUCTURE failed to split GII.2 into additional lineages and identified only the two main subpopulations among the genotype GII.2 samples (Fig. 2A and B). The ML-based tree inferred a total of 15 clades with high bootstrap (>90%) support (Fig. 4B). Each main clade in the ML tree was observed to stratify into additional small variants/clades. In order to prevent possible bias during phylogenetic inferences, the analysis was performed after filtering out the recombinant sequences. However, a similar clustering pattern was observed in a phylogenetic tree constructed without recombinant sequences, except that the GII.4 Apeldoorn_2007 variant formed a separate lineage (Supplementary Fig. S1).

Recombination patterns and distinction

Recombination analysis was performed to confirm the admixed samples observed in STRUCTURE and phylogenetic tree analysis. A total of 40 recombinant strains from 822 sequences of NoV GII were identified, with a threshold of $p < 0.00001$ (Supplementary Table S3). The STRUCTURE and RDP4 results were found congruent in the case of many admixed and recombinant strains, with a few exceptions. Different recombination breakpoints were observed in the GII.4 and non-GII.4 genotypes. In the non-GII.4 genotypes, the majority of recombination breakpoints were detected at the junction of ORF2 and ORF3, while in the GII.4 genotypes, recombination breakpoints were mostly detected in the ORF1 region. A few strains were found to have multiple recombination breakpoints. For instance, the sample MH218571.1 was observed to have undergone three recombination events. RDP4 also identified this as a recombinant with probable minor and major parents. Both inter-genotypic and intra-genotypic recombination events were observed in NoV GII genotypes. For example, a Chinese strain (MG745991.1) of genotype GII.2 had undergone intra-genotype recombination with both the major

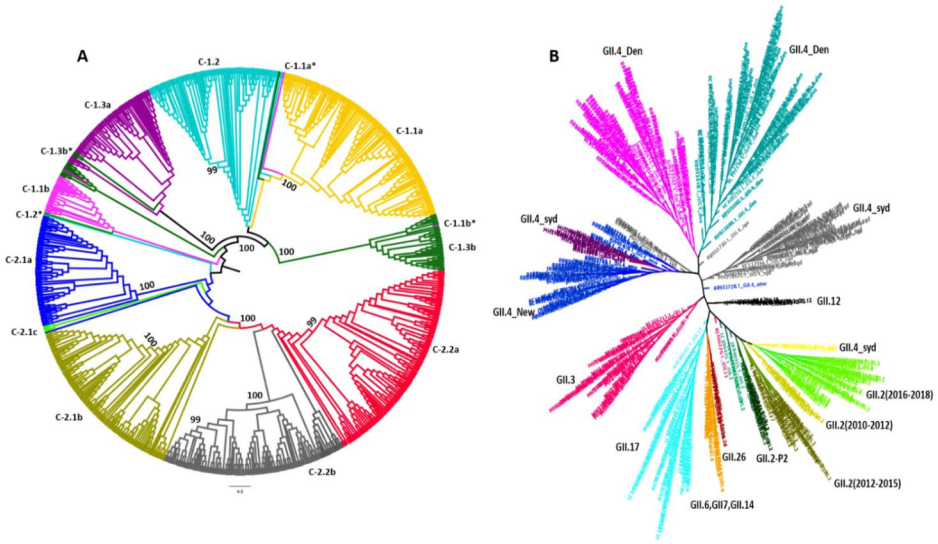


Fig. 4 Phylogenetic analysis of complete genome sequences of NoV GII isolates. (A) NJ-based tree of 822 strains constructed using MEGA X. Ten clades were observed in the phylogenetic tree, which is congruent with the clustering pattern observed using STRUCTURE, i.e., C-1.1a (GII.2), C-1.1b (GII.2), C-1.2(GII.17), C-1.3a (GII.3), C-1.3b (GII.5, 6, 7, 12, 13 and 26), C-2.1a (GII.4- Sydney_2012/P31), C-2.1b (GII.4-Sydney_2012/P4, and New Orleans_2009), C-2.1c (Sydney_2012/p16) C-2.2a (Den Haag_2006b), and C-2.2b (Den Haag_2006b). The branches of recombinant/admixed strains are indicated by an asterisk (*). (B) Maximum-likelihood tree of NoV GII.2. All major clades of NoV are colored and labeled.

(i.e. MG746023.1) and minor (MG745990.1) parents of the GII.2 genotype, while a Japanese strain (LC209439.1) of genotype GII.2 had undergone inter-genotypic recombination and originated from a GII.2 major parent (LC209463.1) and a minor parent (KJ196283.1) of the GII.4 genotype.

Phylodynamics of NoV GII

The complete GII NoV genome sequences obtained from the public database spanned a period of almost two decades. BSP plot analysis generally showed a consistent pattern of the effective population size of GII NoV. However, a slight increase in the population size was observed from 2009 to 2010, followed by a sudden decrease in the effective population size in 2015 (Fig. 5).

Episodic positive selection signatures across NoV genomes

The signature of episodic positive selection was found in all of the coding genes of NoVs. The MEME method identified a total of 72 codons that had possibly evolved under significant episodic diversifying selection (Supplementary Table S4). Most of these codons are found in the VP1 and RdRp coding genes. The VP2 and NTPase have 25 and 11 codons, respectively, that are under selection pressure. The protease

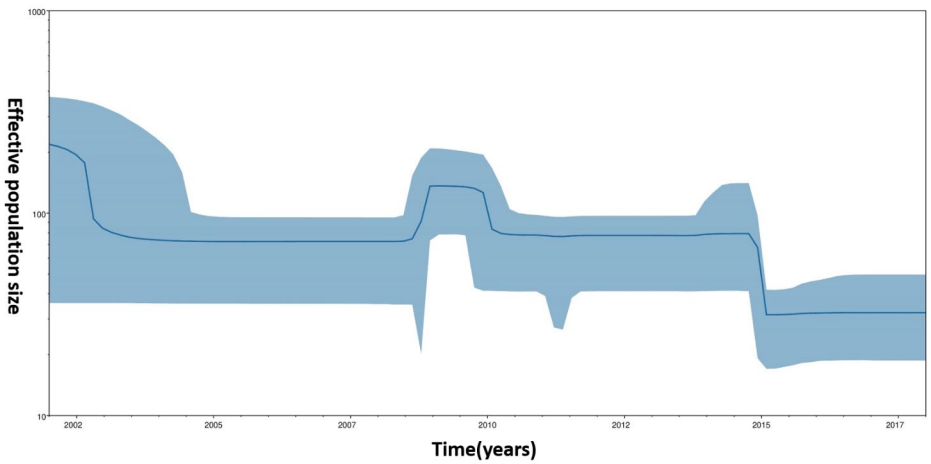


Fig. 5 Bayesian skyline plot of Human NoV GII.2. The *y*-axis represents the effective population size, and the *x*-axis represents time in years. The solid black line indicates the mean posterior value, and the blue shaded area represents the 95% HPD interval.

gene has six sites with evidence of episodic positive selection despite the fact that the coding region is comparatively short.

Footprints of pervasive positive selection

The analysis conducted using FEL, IFEL, and SLAC identified limited signals of pervasive positive selection in the non-structural proteins p48, NTPase, p22, vpg, and RdRp (Table 1). However, in the case of the VP1 and VP2 structural protein coding genes, many codons appeared to be under the influence of pervasive positive selection (Table 1). Notably, a large number of codons are evolving under the influence of strong negative selection (Table 2). The evidence of purifying selection indicates a highly adapted phenotype, probably caused by constraints imposed by protein structure and function.

Differential evolutionary pressure across NoVs genotypes

We then investigated whether the NoV GII population clusters and genotypes contain differential or homogeneous natural selection signatures, which highlight the differences in the antigenicity and dispersal pattern of the pathogen. The analysis revealed differential positive selection signatures in the structural and non-structural proteins of NoV GII. In the VP1 protein, 19 distinct sites with features of episodic positive selection were detected, specifically in the GII.4 strains (Supplementary Table S5). Likewise, in the case of VP2 protein, 14 distinct codons had undergone episodic positive selection only in the GII.4 genotypes (Supplementary Table S5).

Among the nonstructural proteins NS1/2, NS3, NS4, NS5, NS6, and NS7, evidence of differential episodic positive selection was observed among different genotypes. A total of 13 codons in the NS1/2(p48) gene were found to be under positive natural selection. Among these, six were positively selected in the GII.4 genotype specifically, while the other seven were selected in the non-GII.4 genotypes samples (Supplementary Table S5). Codon 44 of NS1/2 codes for serine (S) in the UK isolates of genotype GII.3 and phenylalanine (F) in the Asian GII.3 strains. Similarly, in the NS3 (NTPase) gene, two codons were under positive selection among the GII.4 isolates, while among all non-GII.4 genotypes, 11 distinct sites were under positive selection pressure (Supplementary Table S5). Histidine 224 of the NTPase was substituted by lysine (K) in the GII.6, GII.7, and GII.14 genotypes and by glutamine (Q) in the GII.17 genotype, (Supplementary Table S5). A selection signature was observed across seven codons in the NS4 (P22) gene, differentially selected among the GII.4, GII.2, and GII.3 samples (Supplementary Table S5). In the case of the NS5 (Vpg) protein, codon 127, coding for asparagine (N), is under the influence of episodic positive selection and is mutated to histidine (H) in GII.17 genotype samples. Likewise, in the case of the NS6 (protease) protein, only one codon is under episodic positive selection in GII.4 strains, but seven different codons are under episodic positive

Table 1 Pervasive positive selection sites across the NoV genome identified by the FEL, IFEL, and SLAC methods with a default *p*-value of 0.1. * indicates a selection signal at the respective codon site detected by one of the methods.

Codon	FEL	IFEL	SLAC
NS1/2			
81	*	*	*
89	*	*	
9	*		*
150		*	
305	*	*	*
NS4/P22			
66	*	*	
NS5/vpg			
8	*	*	
NS7/RdRP			
497			*
VP1			
6		*	
22		*	
61		*	
297		*	
302	*		
VP2			
81	*	*	*
109	*	*	*
145	*	*	*
151	*	*	*
165	*	*	*
171	*		*
176	*		*
182	*		*
229	*		*
242	*	*	*
269	*		*

Table 2 Total number of negatively selected codons in each gene, detected by the FEL, IFEL, and SLAC methods

Gene	FEL	IFEL	SLAC
NS1	190	183	165
NTPase	347	335	340
P22	139	129	133
Polymerase	452	437	444
Protease	164	157	162
VPG	125	117	121
VP1	441	398	424
VP2	260	218	240

selection in strains non-GII4 genotypes. Similarly, the NS7 (RdRp) protein coding gene was also observed to be target of episodic diversifying selection, and 17 codons are specifically selected in the genotype GII.4. Different codons in the NS7 coding gene were found to be under selection pressure in the GII.2, GII.3, and GII.17 genotypes (Supplementary Table S5). Overall, major differential natural selection features were observed between the GII.4 and non-GII.4 genotypes, and marked differential selection signatures were observed among the non-GII.4 samples, including the GII.2, GII.3, and G.II.17 genotypes.

Discussion

A fast evolutionary rate, selection pressure, and recombination act as prodigious evolutionary forces to intensify the genetic diversity of noroviruses [50]. Owing to their small genome size, high mutation rate, short generation time, and large population size, RNA viruses are suitable models to study evolution in the context of population genetics. Previous studies have focused on specific NoV genotypes, part of the genome, or a specific geographic region [1, 51]. In the current study, we performed a genome-wide comprehensive analysis of NoV GII isolates from different continents to gain a better understanding of their genetic structure, recombination events, and natural selection pattern.

The genetic structure analyses in the current study did not reveal any geographically based distinctions among the NoV GII isolates. Due to the high degree of mobility and frequent travel in the modern world, NoV GII isolates might have been disseminated worldwide, and hence, no regional distinctions were observed among the GII isolates. However, Tohma et al. recently reported some non-typeable genotypes of NoV GII circulating in South America that exhibited marked genetic divergence from other NoV genotypes [66]. Genetic structure analysis revealed that the genotype GII.4 strains differ from those of the other NoV GII genotypes (Fig. 1). The stratification of the NoV GII samples into two main subpopulations was also supported by a branching pattern in a phylogenetic tree and PCA analysis (Figs. 3 and 4). However, Kobayshi et al. reported three clusters in the NoV population based on analysis of ORF2. Moreover, additional analysis of GII.4 sequences suggested extra clustering at K = 2 and 5 (Fig. 2C). At K = 5, the GII.4 Sydney_2012 variant stratified into three lineages. This stratification pattern of the Sydney_2012 variant was also reported earlier based on ORF2 gene sequences [53].

We identified admixture strains using the admixture model/linkage model implemented in the STRUCTURE program. The admixture model fails to take into account the physical relationship between loci, and the proportion of admixed strains may sometimes be under- or overestimated. Therefore, to optimize the membership scores given to the admixed strains, a linkage-correlated model was applied that accounts for potential linkages. Admixed isolates were observed in the C-1.1b, C-1.2b, C-1.3b, and C-2.1a clusters. The majority of the admixed and recombinant strains belong to the non-GII.4 genotypes. A few of these admixed strains, such as GII.3 [P33], GII.13 [P33], and GII.12 [P33], have been reported to be prevalent globally [54]. Recombination among NoV strains occurs at high frequency and acts as a major driving force in viral evolution. Recombination allows the virus to increase its genetic fitness and spread in the host population by escaping the host immune response [55]. The admixture in NoV is possibly responsible for the genetic diversification of the C-1.2b and C-1.3b clusters. Likewise, the |D'|, r^2, and $I^S A$ statistics indicated weak linkage of norovirus GII isolates in the current study, indirectly suggesting a role of recombination in shaping the evolution of norovirus GII isolates.

The BSP plot generated based on markers throughout the genome suggests a stable effective population size for the NoV GII isolates originating from human hosts (Fig. 5). The BSP plot implies a rapid increase in the effective population size during 2009-10. This might have been accompanied by the large outbreaks and epidemicity of the GII.4 New Orleans_2009 variant [57]. Likewise, a novel GII.12 strain also emerged during this period and caused several outbreaks [58]. The effective population size fell sharply in 2015, and this is likely to correspond to a gain of host immunity against the dominant NoV variants.

Substantial signals of episodic diversifying selection were observed in all of the proteins, including both the structural and non-structural proteins. However, few pervasive positive selection signals were identified in the VP1 and VPG genes. Xingguang et al. reported a lack of episodic positive selection in genotype GII.2 isolates and suggested genetic drift as a possible mechanism for NoV GII.2 evolution [59]. However, significant positive selection signatures were identified for the GII.2 strains in the current study (Supplementary Tables S4 and S5). This suggests that selection pressure is a possible driving force in GII.2 evolution. Other studies have also shown a small number of positive selection sites in the VP1 protein of NoV GII isolates [52, 60]. The VP1 protein plays a fundamental role in the interaction of NoVs with their host cells and is considered to be a key site for immune recognition and receptor binding. Therefore, this protein could be a potential target for vaccine development [61]. We identified several sites that are under positive selection pressure in the P1, P2, and shell domains of the VP1 protein. Mutations at positions 282 to 395 of VP1 (Supplementary Table S5), which are part of the P2 domain, have been reported to play an important role in

its interaction with human blood group antigens (HBGAs) [62]. The S domain is highly conserved across different genotypes, and the antigenic sites within this domain are mostly cross-reactive [63]. In addition to positive selection, a large number of sites were also found to be under the influence of negative selection, indicating that purifying selection has occurred. In general, positive selection sites may be influenced by immune pressure, leading to escape mutations, whereas negative selection sites may prevent deterioration of antigenic function and structures [64]. The sites under positive selection could provide markers for vaccine design. The identification of negatively selected sites in NoV GII genes might help to identify highly conserved regions that will be useful in new diagnostic protocols [65]. A marked difference in the positive selection signature pattern was observed between the GII.4 samples and the other GII genotypes, and this might have shaped the genetic composition of the GII.4 genotype.

Conclusion

The complete-genome-based population genetic analysis presented here revealed significant differences between of the GII.4 genotype and the other NoV GII genotypes, which might be due to specific positive selection signatures. The genetic stratification of GII.4 samples suggests the emergence of additional GII.4 lineages. The analysis did not reveal geographical variations in the genetic composition of the NoV GII strains. The data also suggest that recombination and selection pressure are major factors driving the genetic diversification of NoV GII strains and the emergence of new lineages. These findings might be useful for planning effective strategies to combat NoV GII infections.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00705-022-05396-9.

Acknowledgments The authors acknowledge the National Center of Physics, Islamabad, for providing access to high-performance computing (HPC) for data analysis.

Authors' contributions S.K. and A.K. conceived the research plan. S.K. H.K., and K.N. performed the data analyses. S.K. wrote the initial draft of the manuscript. A.K. supervised the study, critically reviewed the analyses and finalized the draft preparation.

Funding The study was conducted without any specific funding or financial grant support.

Declarations

Availability of supplementary data Supplementary data relevant to this study are provided.

Conflict of interest The authors declare that they have no competing interests.

References

1. Qiao N, Ren H, Liu L (2017) Genomic diversity and phylogeography of norovirus in China. BMC Med Genom 10(3):51
2. Kapikian AZ, Wyatt RG, Dolin R, Thornhill TS, Kalica AR, Chanock RM (1972) Visualization by immune electron microscopy of a 27-nm particle associated with acute infectious nonbacterial gastroenteritis. *Journal of virology*, *10*(5), pp. 1075-1081.3
3. Havelaar AH, Kirk MD, Torgerson PR, Gibb HJ, Hald T, Lake RJ, Praet N, Bellinger DC, De Silva NR, Gargouri N, Speybroeck N (2015) World Health Organization global estimates and regional comparisons of the burden of foodborne disease in 2010. PLoS Med 12(12):e1001923
4. Mans J (2019) Norovirus Infections and Disease in Lower-Middle-and Low-Income Countries, 1997–2018. *Viruses*, *11*(4), p. 341
5. Nasheri N, Petronella N, Ronholm J, Bidawid S, Corneau N (2017) Characterization of the genomic diversity of norovirus in linked patients using a metagenomic deep sequencing approach. *Frontiers in microbiology*, *8*, p. 73
6. Jung J, Grant T, Thomas DR, Diehnelt CW, Grigorieff N, Joshua-Tor L (2019) High-resolution cryo-EM structures of outbreak strain human norovirus shells reveal size variations. *Proceedings of the National Academy of Sciences*, *116*(26), pp. 12828-12832
7. Cotten M, Petrova V, Phan MV, Rabaa MA, Watson SJ, Ong SH, Kellam P, Baker S (2014) Deep sequencing of norovirus genomes defines evolutionary patterns in an urban tropical setting. J Virol 88(19):11056–11069
8. Ford-Siltz LA, Mullis L, Sanad YM, Tohma K, Lepore CJ, Azevedo M, Parra GI (2019) Genomics analyses of GIV and GVI noroviruses reveal the distinct clustering of human and animal viruses. *Viruses*, *11*(3), p. 204
9. Chhabra P, de Graaf M, Parra GI, Chan MCW, Green K, Martella V, Wang Q, White PA, Katayama K, Vennema H, Koopmans MP (2019) Updated classification of norovirus genogroups and genotypes. *The Journal of general virology*, *100*(10), p. 1393
10. Chen C, Yan JB, Wang HL, Li P, Li KF, Wu B, Zhang H (2018) Molecular epidemiology and spatiotemporal dynamics of norovirus associated with sporadic acute gastroenteritis during 2013–2017, Zhoushan Islands, China. PLoS ONE 13:e0200911
11. Gaythorpe KAM, Trotter CL, Lopman B, Steele M, Conlan AJK (2018) Norovirus transmission dynamics: a modelling review. Epidemiol Infect 146(2):147–158
12. Bok K, Abente EJ, Realpe-Quintero M, Mitra T, Sosnovtsev SV, Kapikian AZ, Green KY (2009) Evolutionary dynamics of GII. 4 noroviruses over a 34-year period. J Virol 83(22):11890–11901
13. Hasing ME, Lee BE, Qiu Y, Xia M, Pabbaraju K, Wong A, Tipples G, Jiang X, Pang XL (2019) Changes in norovirus genotype diversity in gastroenteritis outbreaks in Alberta, Canada: 2012–2018. *BMC infectious diseases*, *19*(1), pp.1-9
14. das Costa N, Teixeira LCP, Portela DM, de Lima ACR, da Silva Bandeira ICG, Júnior R, Siqueira ECS, Resque JAM, da Silva HR, Gabbay YB (2019) Molecular and evolutionary characterization of norovirus GII. 17 in the northern region of Brazil. BMC Infect Dis 19(1):1–11
15. Boon D, Mahar JE, Abente EJ, Kirkwood CD, Purcell RH, Kapikian AZ, Green KY, Bok K (2011) Comparative evolution of GII. 3 and GII. 4 norovirus over a 31-year period. J Virol 85(17):8656–8666
16. Eden JS, Hewitt J, Lim KL, Boni MF, Merif J, Greening G, Ratcliff RM, Holmes EC, Tanaka MM, Rawlinson WD, White PA

(2014) The emergence and evolution of the novel epidemic norovirus GII. 4 variant Sydney 2012. Virology 450:106–113
17. Petronella N, Ronholm J, Suresh M, Harlow J, Mykytczuk O, Corneau N, Bidawid S, Nasheri N (2018) Genetic characterization of norovirus GII. 4 variants circulating in Canada using a metagenomic technique. BMC Infect Dis 18(1):1–11
18. Pickett BE, Greer DS, Zhang Y, Stewart L, Zhou L, Sun G, Gu Z, Kumar S, Zaremba S, Larsen CN, Jen W (2012) Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community. Viruses 4(11):3209–3226
19. Kroneman A, Vennema H, Deforche KVD, Avoort HVD, Peñaranda S, Oberste MS, Vinjé J, Koopmans M (2011) An automated genotyping tool for enteroviruses and noroviruses. J Clin Virol 51(2):121–125
20. Sievers F, Higgins DG (2018) Clustal Omega for making accurate alignments of many protein sequences. Protein Sci 27(1):135–145
21. Haubold B, Hudson RR (2000) LIAN 3.0: detecting linkage disequilibrium in multilocus data. Bioinformatics 16:847–849
22. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, RamosOnsins SE, Sánchez-Gracia A (2017) DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol biology Evol 34:3299–3302
23. Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29(2):311–322
24. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155(2):945–959
25. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587
26. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–2620
27. Earl DA (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour 4(2):359–361
28. Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evolutionary Bioinf 1:117693430500100003
29. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575
30. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32(1):268–274
31. Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS (2018) UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol 35(2):518–522
32. Rambaut A, Drummond AJ (2018) FigTree v1. 4.4. Institute of Evolutionary Biology. University of Edinburgh, Edinburgh
33. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B (2015) RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol 1:1–5
34. Padidam M, Sawyer S, Fauquet CM (1999) Possible emergence of new geminiviruses by frequent recombination. Virology 265:218–225
35. Martin DP, Posada D, Crandall KA, Williamson C (2005) A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res Hum Retroviruses 21(1):98–102
36. Smith JM (1992) Analyzing the mosaic structure of genes. J Mol Evol 34:126–129
37. Posada D, Crandall KA (2001) Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proceedings of the National Academy of Sciences*, *98*(24), pp. 13757-13762
38. Gibbs MJ, Armstrong JS, Gibbs AJ (2000) Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics 16(7):573–582
39. Boni MF, Posada D, Feldman MW (2007) An exact nonparametric method for inferring mosaic structure in sequence triplets. Genetics 176(2):1035–1047
40. Drummond AJ, Rambaut A, Shapiro BETH, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22(5):1185–1192
41. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol 10(4):e1003537
42. Posada D (2008) jModelTest: phylogenetic model averaging. Mol Biol Evol 25(7):1253–1256
43. Rambaut A, Drummond AJ, .Tracer (2013) Available at http://tree.bio.ed.ac.uk/software/tracer
44. Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T (2010) GUIDANCE: a web server for assessing alignment confidence scores. Nucleic Acids Res 38:23–28
45. Kosakovsky Pond SL, Frost SD (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol biology Evol 22:1208–1222
46. Pond SLK, Frost SD, Grossman Z, Gravenor MB, Richman DD, Brown AJL (2006) Adaptation to different human populations by HIV-1 revealed by codon-based analyses. PLoS Comput biology 2:1–9
47. Pond SLK, Frost SD (2005a) Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21:2531–2533
48. Pond SLK, Muse SV (2005b) HyPhy: hypothesis testing using phylogenies. Bioinformatics, 21:676–9
49. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL (2012) Detecting individual sites subject to episodic diversifying selection. PLoS Genet 8:1–12
50. Xue L, Wu Q, Dong R, Cai W, Wu H, Chen M, Chen G, Wang J, Zhang J (2017) Comparative phylogenetic analyses of recombinant noroviruses based on different protein-encoding regions show the recombination-associated evolution pattern. Sci Rep 7(1):1–10
51. Fioretti JM, Bello G, Rocha MS, Victoria M, Leite JPG, Miagostovich MP (2014) Temporal dynamics of norovirus GII. 4 variants in Brazil between 2004 and 2012. PLoS ONE 9(3):e92988
52. Kobayashi M, Matsushima Y, Motoya T, Sakon N, Shigemoto N, Okamoto-Nakagawa R, Nishimura K, Yamashita Y, Kuroda M, Saruki N, Ryo A (2016) Molecular evolution of the capsid gene in human norovirus genogroup II. Sci Rep 6(1):1–11
53. Hernandez JM, Silva LD, Sousa Júnior EC, Cardoso JF, Reymão TKA, Portela ACR, de Lima CPS, Teixeira DM, Lucena MSS, Nunes MRT, Gabbay YB (2020) Evolutionary and molecular analysis of complete genome sequences of norovirus from Brazil: emerging recombinant strain GII. P16/GII. 4. *Frontiers in microbiology*, *11*, p. 1870
54. White PA (2014) Evolution of norovirus. Clin Microbiol Infect 20(8):741–745
55. Wu X, Han J, Chen L, Xu D, Shen Y, Zha Y, Zhu X, Ji L (2015) Prevalence and genetic diversity of noroviruses in adults with acute gastroenteritis in Huzhou, China, 2013–2014. Arch Virol 160(7):1705–1713
56. Yen C, Wikswo ME, Lopman BA, Vinje J, Parashar UD, Hall AJ (2011) Impact of an emergent norovirus variant in 2009 on norovirus outbreak activity in the United States. Clin Infect Dis 53(6):568–571

57. Vega E, Vinjé J, Novel GII (2011) 12 norovirus strain, United States, 2009–2010. *Emerging infectious diseases*, *17*(8), p.1516
58. Li X, Liu H, Magalis BR, Pond SLK, Volz EM (2021) Molecular evolution of human norovirus GII. 2 clusters. *Frontiers in microbiology*, *12*
59. Parra GI, Squires RB, Karangwa CK, Johnson JA, Lepore CJ, Sosnovtsev SV, Green KY (2017) Static and evolving norovirus genotypes: implications for epidemiology and immunity. *PLoS pathogens*, *13*(1), p.e1006136
60. Campillay-Véliz CP, Carvajal JJ, Avellaneda AM, Escobar D, Covián C, Kalergis AM, Lay MK (2020) Human norovirus proteins: implications in the replicative cycle, pathogenesis, and the host immune response. *Frontiers in Immunology*, *11*, p.961
61. Hardy MJ, Kuczera G, Coombes PJ (2005) Integrated urban water cycle management: the UrbanCycle model. Water Sci Technol 52(9):1–9
62. Parra GI, Azure J, Fischer R, Bok K, Sandoval-Jaime C, Sosnovtsev SV, Sander P, Green KY (2013) Identification of a broadly cross-reactive epitope in the inner shell of the norovirus capsid. PLoS ONE 8(6):e67592.
63. Domingo E (2007) Virus Evolution In: Knipe, D.M. and Howley, P.M., Eds., Fields Virology, 5th Edition, Lippincott Williams & Wilkins, Philadelphia, 389–421.
64. Presti AL, Rezza G, Stefanelli P (2020) Selective pressure on SARS-CoV-2 protein coding genes and glycosylation site prediction. Heliyon 6(9):e05001
65. Tohma K, Lepore CJ, Martinez M, Degiuseppe JI, Khamrin P, Saito M et al (2021) Genome-wide analyses of human noroviruses provide insights on evolutionary dynamics and evidence of coexisting viral populations evolving under recombination constraints. PLoS Pathog 17(7):e1009744