Genomics stratification and differential natural selection signatures among the human *Norovirus* genogroup-II isolates

Sehrish Kakakhel
Abdul Wali Khan University Mardan, Pakistan
Hizbullah Khan
Abdul Wali Khan University Mardan
Kiran Nigar
Abdul Wali Khan University Mardan
Asifullah Khan (asif@awkum.edu.pk)
Abdul Wali Khan University Mardan https://orcid.org/0000-0002-1444-4249

Research Article

Keywords: Norovirus genogroup II, Genomic diversity, spatiotemporal, selection pressure

Posted Date: October 20th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-992229/v1

Version of Record: A version of this preprint was published at Archives of Virology on March 23rd, 2022.
See the published version at https://doi.org/10.1007/s00705-022-05396-9.

Abstract

The Norovirus (NoV) from the family *Caliciviridae* is the most common cause of gastroenteritis diseases in human. There are ten NoV genogroups are reported so far. Among these, the genogroup II (GII) is commonly prevalent and causes serious infection worldwide. The complete genome sequences of NoV GII isolates from different continental origin 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 were tools used for phylogenetic analyses. Genome-wide natural selection analyses were pursued via the maximum likelihood based methods. The demography features of NoV GII genome sequences were assessed using the BEAST package. All the NoV GII sequences initially clustered into two main subpopulations at significant K=2. The genotype GII.4 samples clearly split from the rest of all the genotypes. This indicate marked genetic distinction between norovirus GII.4 and non-GII.4 samples. The phylogenetic analyses depicted five distinct sub-clades for genotype GII.2 and seven sub-clades for GII.4 samples, speculate about the emergence of new lineages from these genotypes. Several isolates with admixed ancestry were identified, that constituted distinct sub-clusters. No continental-specific genetic distinction was observed among the NoV GII isolates. Significant genomic signatures of both positive and negative natural selection were identified across the NoV GII genes. Differential pattern of positive selection signal inferred between the GII.4 and non-GII.4 genotypes. The demographic analyses unveiled a rise in effective population size of NoV GII during 2009-2010, followed by a rapid fall in 2015.

1. Introduction

Noroviruse (NoV) is the utmost pathogenic agent of viral gastroenteritis disease in humans. About 50% of all acute gastroenteritis is caused by NoV [1, 2]. World health organization (WHO) estimated that NoV causes 684 million cases worldwide and 212,489 deaths in the 2010 [3]. The infection is more prevalent in children and old age people with severe symptoms and prolonged shedding [4]. The NoV infection spreads among humans through multiple routes, including, waterborne, foodborne, and person-to-person transmissions [5].

NoV is a small positive RNA virus that belongs to the family *Calcivirade* and genus *Norovirus*. The species have approximately 7.5 kb genome mainly comprises of three open reading frames (ORF) [6]. The 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]. The ORF2 encodes a major structural protein (VPI) that constitutes the virus capsid. Whereas, the ORF3 encodes a minor structural protein (VP2). The VP1 is comprised of the conserved shell (S) and two protruding (P), i.e. P1 and P2 domains. The P1 domain improves partial stability, while the P2 domain facilitates the binding to histoblood group antigens (HBGAs) [8]. The NoV is classified into 10 genogroups based on VP1 protein diversity (9). These genogroups are further classified into 49 genotypes based on VP1 coding loci and 60 genotypes based on RdRp loci information. The genogroup I, II, IV, GVIII, and GIX are reported so far with respect to human diseases [10]. The GII is majorly responsible for worldwide outbreaks [11]. Among

genogroup II (GII), the genotype GII.4 mediated infection is predominant, where new genetic variants emerged and outbreak a pandemic [12]. The major global variants characterized so far include are Sydney_2012, Den Haag_2006, and New Orleans_2009 [10, 4]. The genotype GII.4 mediated infection predominantly persisted for over two decades due to fast mutation and evolutionary rates [13]. The non-GII.4 genotypes have also caused massive epidemics and transiently surpassed the genotype GII.4. This includes the recently emerging GII.17 and GII.2 lineages. A novel genotype GII.17 variant, termed as Kawasaki, appeared as the primary cause of outbreaks in some Asian countries and replaced the Sydney_2012 variant [14]. However, among children, the genotype GII.3 commonly causes irregular NoV infection [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 the NoV infection [17, 6]. The complete genome sequences of NoV GII isolates from different continental regions are available in the public genome repositories. The high prevalence rate of NoV GII along with the recent emergence of novel strains provoked us to examine this genogroup complete genome sequences to understand their genetic composition, distinction and extent of possible genetic admixture. Besides, the role of natural selection and recombination analyses were performed to understand the possible role of these events to shape the genetic structure of NoV GII. Understanding these genomic features across the worldwide NoV GII isolates may implicate for devising effective vaccine designing against the NoV infection.

2 Methodology

2.1 Genome sequences retrieval

Complete genome sequences data of human NoV GII isolates were obtained from the Virus Pathogen Resource database hosted by NCBI [18]. Several NoV genome sequences are deposited in public databases with no genotype information. The genotype information for such sequences was obtained from NoV automated online genotyping tool (version 2.0) [19]. The sequences submitted with unknown location, host, and sampling time information were excluded. Finally, a dataset comprised of 822 sequences of NoV GII was generated **(Table S1)**.

2.2 Multiple sequence alignment and parsimony-informative sites identification

The multiple sequence alignment (MSA) was performed using Clustal omega3 [20]. Mega-X was used to extract parsimony informative (PI) sites from alignment. Total 4069 PI sites were acquired from aligned data.

2.3 Linkage analysis

The LIAN v3.5 tool was employed to examine the null hypothesis and the linkage equilibrium within NoV GII genomes data [21]. This program calculates the standardized index of association ($I^S A$) to quantify

the haplotype-wide linkage derived from the dataset. In addition, the |D'| and r^2 were computed via DnaSpv6.0 [22] to measure the linkage disequilibrium (LD). The |D'| represents the absolute value of the difference between observed and the expected haplotype frequency in absence of LD. The variance of the allele frequency between observed and expected haplotype is represented by r^2 [23].

2.4 Population structure analysis

The genetic structure of NoV GII was analyzed by 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. The STRUCTURE operated 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 pursued with 100,000, burn-in length, followed by 100,000 MCMC iteration with default parameters values (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 sub-clusters) was determined by the formula depicted by the Evanno ΔK approach using the STRUCTURE HARVESTER resource [26, 27]. The plot of K vs ΔK was used for K_{opt}. The value of K_{opt} validated by various combination of burn-ins and burn-ins length i.e. ranges from 50,000-50,000, 70,000-70,000 and 100,000-100,000.

2.5 F-statistics and PCA analyses

The NoV GII genetic composition estimates were additionally corroborated by F-statistics known as fixation indices (F_{ST}) calculation and principal component analysis (PCA). The Fst was calculated by analysis of molecular variance (AMOVA) implemented in ARLEQUINv3.11 with 1,000 permutations [28]. The AMOVA calculates the partitioning variance at different levels of population subdivision and yield Fst. The PCA was acquired via PLINKv1.9 [29] and the output results visualized with the R built-in function "prcomp".

2.6 Phylogenetic analysis

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

2.7 Recombination analysis

The aligned complete genome dataset was used for the identification of potential recombination events by following the seven different methods implemented in RDP4 package. These methods are RDP [33], GENECONV [34] BOOTSCAN [35], MaxChi [36], CHIMAERA [37], SiSCAN [38], and 3SEQ [39]. A sample predicted to be recombinant by at least three of the above-mentioned methods with a p-value of 0.00001 was considered, to prevent the false positive recombination calls.

2.8 Demography estimation of NoV GII

The fluctuation in the effective population size of NoV GII with respect to time was inferred for available isolates genomic data using the Bayesian skyline model [40], via BEAST2 [41]. The selection of the best-fit nucleotide substitution model was achieved via the 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. The 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 the convergence. The convergence of the MCMC log output files and effective sample size (ESS) > 200 was analyzed via the Tracerv1.7 program [43].

2.9 Natural selection analysis

A dataset of 538 NoV GII sequences was prepared for natural selection analyses. The potential recombinant samples were excluded from the analyses to avoid inferential biases. Total 8 datasets were generated, according to 8 protein coding sequences, i.e. p48, Ntpase, p22, vpg, protease, RdRp, ORF2, and ORF3. The accuracy of selection pressure calculation mainly depends upon the quality of MSA. Therefore, the quality of the MUSCLE generated MSA was checked by the GUIDANCE server. The Guidance server checks the unreliable alignment region within MSA by following a confidence threshold i.e. score of ∼1 [44]. All 8 datasets were separately analyzed using different ML-based methods with the default value of 0.1. These methods are Single Likelihood Ancestor Counting (SLAC) [45], Internal Fixed Effect Likelihood (IFEL) [46], and Fixed Effect Likelihood (FEL) [47] accessible through Datamonkey web-server in the HYPHY package [48, 49]. These three methods identify the 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 at the Datamonkey server. The episodic positive selection signatures were detected via MEME (Mixed Effects Model of Evolution) method available at the datamonkey server. Episodic positive selection affects a few lineages even in a condition when majorities of the lineages undergoing purifying selection [50].

3. Results

3.1 Linkage analysis

Prior to the genetic composition assessment via the STRUCTURE program, the loci linkage pattern needs to be evaluated. LD is the nonrandom association of alleles at different polymorphic sites. In the case of free recombination, the value of $I^S A$ calculated via 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) that indicates a signal of linkage equilibrium and weak LD. To confirm further the existence of low LD, the plots of |D| and r^2 were computed by DnaSP v5. The D is the function of LD measurement. The average value of |D| and r^2 were

found to be 0.8206 and 0.0522 respectively. This inferred that the loci are weakly linked and the STRUCTURE program usage is therefore appropriate for NoV isolates dataset.

3.2 Genetic composition analyses

3.2.1 Clustering analysis via 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 section). The K_{opt} of 2 was detected in the plot of *K* vs *ΔK* (Figure 1A). This unveiled the basic stratification of all the NoV isolates samples into two subpopulations. Additionally, the AMOVA test clued marked genetic distinction, i.e. Fst=0.53293 (P-value = 0000), between the two sub-population genetic components. The cluster-1 (C-1) acquired at K_{opt} 2 comprises all the NoV genotypes samples except the GII.4. The GII.4 samples comprised a separate subpopulation (C-2) (Figure 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 distinction and origin.

Additional analyses were pursued to further investigate the genetic stratification in each of the major genetic components acquired in the above analyses. 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 (Figure 2A). The K_{opt} 3 reveals the diversification of C-1 into further three subpopulation/lineages represented with C-1.1, C-1.2, and C-1.3 (Figure 2B). The C-1.1 consists of genotype GII.2 strains. The UK strains from GII.2 genotype were noticed to be admixed having significant membership scores ranges from 0.500- 0.434 for the clusters C-1.2, and C-1.3. The 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 genotypes samples. At K_{opt} = 4, the C-1.3 further stratified into two sub-clusters (i.e. C-1.3a and C-1.3b) (Figure 2B). The GII.3 samples constitute a C-1.3a, while the samples from genotypes GII.5, GII.6, GII.7, GII.12, GII.13, and GII.26 are clumped into C-1.3b. Likewise, the K_{opt} of 5 unveiled the sub-clustering of the GII.2 genotype, i.e. formerly comprises the C-1.1 cluster at K_{opt} 3, further stratify into two lineages i.e. C-1.1a and C-1.1b at K_{opt} = 5 (Figure 2B). The overall clustering pattern of samples obtained at K_{opt} of 3, 4, and 5 did not reveal any geography-based genetic distinction among NoV GII isolates and the sub-population's genetic components and stratification of samples mainly based on the genotype identity.

3.2.2 Genetic stratification of GII.4 samples

The GII.4 samples, initially split at *K*=2, additionally stratified during subsequent Bayesian clustering analysis (Figure 2). The samples of genotype GII.4 were stratified into two significant major and five minor lineages (Figure 2C). The C-2.1 cluster corresponds the GII.4-sydeny_2012, GII.4-New Orleans_2009, and GII.4-Apeldoorn_2007 strains. Whereas, the C-2.2 cluster corresponds to the strains of Den Haag_2006b. In the case of *K*=5, the C-2.1 stratified into three lineages, i.e. C-2.1a, C-2.1b, and C-2.1c, while the C-2.2 split into two sub-lineages, i.e. C-2.2a and C-2.2b (Figure 2D). The C-2.1a subpopulation

holds the Sydney_2012 strains and the C-2.1b comprised of Sydney_2012 and New Orleans_2009 GII.4 samples. Whereas, the C-2.1c cluster holds the GII.4 Sydney_2012 samples. The GII.4-Den Haag_2006b samples initially clustered in C-2.2, however, additionally stratified into two lineages represented as C-2.2a and C-2.2b (**Table S2**).

3.2.3 PCA analysis

PCA was pursued to further validate 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 (Figure 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 genotypes clustered closely. The stratification pattern observed in the PCA plot is in agreement with the STRUCTURE findings.

3.2.4 Phylogenetic analyses

The ML and NJ-based methods produced congruent tree topologies. Phylogenetic tree result obtained with NJ-based method was examined according to the clustering pattern acquired via STRUCTURE. The NoV GII samples grouped into eight independent clades in the NJ tree (Figure 4A) which corresponds to the eight clusters (C-1.1a, C-1.1b, C-1.2, C-1.3a, C-1.3b, C-2.2a, C-2.2b, C-2.2c, C-2.1a), acquired during STRUCTURE analyses. Some admixed strains were observed in the phylogenetic tree cladding represented with the C-1.b*, C-1.2b*, C-1.3b*, and C-2.1a* clusters. The ML phylogenetic tree revealed that cluster C-1.1 further stratified into five minor clades **(**Figure 4B**)**. Contrary to phylogenetic tree stratification, the STRUCTURE failed to delineate the GII.2 into additional lineages and identified only the main two subpopulations within the genotype GII.2 samples (Figure 2A **& B**). The ML-based tree inferred 13 total clades with high bootstrap >90 support (Figure 4B).

3.3 Recombination pattern and distinction

Recombination analyses were performed to validate the admixed samples observed in STRUCTURE and phylogenetic tree analyses. A total of 40 recombinant strains from 822 sequences of NoV GII were identified with threshold of $p < 0.00001$ (**Table S3)**. The STRUCTURE and RDP4 results were found congruent in the case of many admixed and recombinant strains with few exceptions. Different recombination breakpoints were observed in the GII.4 and non-GII.4 genotypes. In non-GII.4 genotypes majority of recombination breakpoints were detected at the junction of ORF2 and ORF3, while in GII.4 genotypes recombination breakpoints were mostly detected in the ORF1 region. Few strains were indicated to have multiple recombination breakpoints. For instance, the sample MH218571.1 was observed to have three recombination events. The RDP4 also described this as 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 (i.e. MG745991.1) from genotype GII.2 undergone intra-genotype recombination have major (i.e. MG746023.1), and minor parents (i.e. MG745990.1) both from the GII.2 genotype. While a Japanese strain, i.e. LC209439.1 from genotype GII.2

undergone inter-genotypic recombination event and originated from the GII.2 major parent (i.e. LC209463.1) and a minor parent (i.e. KJ196283.1) from the GII.4 genotype.

3.4 Phylodynamics of NoV GII

The GII NoV isolates genome sequences data deposited in the public databases almost from two decades. The BSP plot analysis, pursued in current study, inferred predominantly 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 **(**Figure 5**).**

3.5 Episodic positive selection signatures across NoV genomes

The signature of episodic positive selection was found in all the coding genes of NoV. The MEME method identified a total of 72 codons that possibly evolved under significant episodic diversifying selection. (**Table S4**). Most of these codons are found in the VP1 and the RdRp coding genes. The VP2 and Ntpase have 25 and 11 codons underlie selection respectively. The protease gene hold 6 sites with evidence of episodic positive selection despite that the coding region is comparatively short.

3.5.1 Footprints of pervasive positive selection

The analysis conducted via FEL, IFEL, and SLAC identified limited signals of pervasive positive selection in case of non-structural proteins, i.e. p48, Ntpase, p22, vpg and RdRp (Table 1). However, in the case of VP1 and VP2 structural protein coding genes, many codons appeared under the influence of pervasive positive selection (Table 1). Noticeably, the large number of codons is 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.

Table 1

Pervasive positive selection sites across NoV genome via FEL, IFEL and SLAC methods with default p-value 0.1. * represents the selection signal at respective codon site in each method.

S.NO	Codon	FEL	IFEL	SLAC
NS1/2				
1	81	*	*	*
2	89	*	*	
3	9	*		*
4	150		*	
5	305	*	*	*
NS4/P22				
6	66	*	*	
NS5/vpg				
7	8	*	*	
NS7/RdRP				
8	497			*
VP1				
9	6		*	
10	22		*	
11	61		*	
12	297		*	
13	302	*		
VP2				
14	81	*	*	*
15	109	*	*	*
16	145	*	*	*
17	151	*	*	*
18	165	*	*	*
19	171	*		*

S.NO	Codon	FEL	IFEL	SLAC
20	176	*		*
21	182	*		*
22	229	*		*
23	242	*	*	*
24	269	*		*

Table 2
Total number of negative selection codons in each gene detected by 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

3.5.2 Differential evolutionary pressure across NoVs genotypes

Analyses were pursued to assess whether the NoV GII population clusters and genotypes have undergone differential or homogeneous natural selection signatures. This actually highlights the differences in the antigenicity and dispersal pattern of the pathogen. The analysis unveiled differential positive selection signatures across the structural and non-structural proteins of NoV GII. In the VP1 protein, there were 19 distinct sites detected with episodic positive selection feature specifically in the GII.4 strains (**Table S5**). Likewise, in the case of VP2 protein, the 14 distinct codons undergone episodic positive selection in the GII.4 genotypes only (**Table S5**).

Among the nonstructural proteins, i.e. NS1/2, NS3, NS4, NS5, NS6, and NS7, evidence of differential episodic positive selection was observed as well among different genotypes. Total 13 codons underlie positive natural selection signature in the NS1/2(p48) gene. Among these, the 6 sites were positive selected in the GII.4 genotype specifically, while the rest of the 7 codons underline selected among the non-GII.4 genotypes samples (**Table S5**). The codon-44 of NS1/2 codes for Serine (S) in the UK isolates

of genotype GII.3, which is substituted with amino acid phenylalanine (F) in Asian GII.3 strains. Similarly, in case of NS3 (Ntpase) gene, there were 2 codons specifically under positive selection among the isolates of GII.4 genotypes. While, among the rest of all non-GII.4 genotypes, 11 distinct sites underline positive selection pressure (Table S5). The Histidine-224 of the Ntapse substituted with lysine (K) in GII.6, GII.7, and GII.14 genotypes, while in the case of GII.17 genotype, the Histidine-224 specifically substituted with the Glutamine (Q) (**Table S5**). The selection signature was observed across the seven codons in NS4 (P22) gene, differentially selected among the GII.4, GII.2, and GII.3 samples (**Table S5**). In case of NS5 (Vpg) protein, only codon-127, coding for asparagine (N) is under the influence of episodic positive selection, which is mutated to Histidine (H) in GII.17 genotype samples. Likewise, in the case of NS6 (protease) protein, only one codon is under the episodic positive selection in GII.4 samples, however, in the case of non-GII4 genotypes, there were 7 different codons observed under episodic positive selection. Similarly, the NS7 (RdRp) protein coding gene was also observed as a target of episodic diversifying selection and 17 codons are specifically selected in the genotype GII.4. Besides, different codons of the NS7 coding gene are underlined selection in GII.2, GII.3, and GII.17 genotypes samples as well (**Table S5**). Overall, major differential natural selection features were observed between the GII.4 and non-GII.4 genotypes. Besides, marked differential selection signatures were also observed among non-GII.4 samples, including the GII.2, GII.3 and G.II.17 genotypes.

4. Discussion

The fast evolutionary rate, selection pressure, and recombination act as a prodigious evolutionary forces to intensify the genetic diversity in the norovirus [50]. Owing to smaller genome sizes, higher mutation rates, short generation time, and large population sizes, the RNA viruses are suitable models to study evolution under the conceptual perception of population genetics. Prior studies are reported by focusing only on specific NoV genotypes or part of the genome from a specific continental region [1, 51]. The current study pursued a genome-wide comprehensive analysis of NoV GII isolates from different continental regions to gain a better understanding of their genetic structure, recombination events, and natural selection pattern.

The genetic structure analyses in the current study identified no geographical based distinction among the NoV GII isolates. Due to high mobility and traveling in the modern world, the NoV GII isolates might have disseminated worldwide, and hence no continental-specific distinction remained among the GII isolates. The genetic structure analysis unveiled the genotype GII.4 samples distinction from the rest of NoV GII genotypes **(**Figure 1). The stratification of entire NoV GII samples into two main subpopulations was also strengthened by the branching pattern of a phylogenetic tree and PCA analysis **(**Figure 3 & 4**).** This is somehow contrary to the findings of Kobayshi et al. (2016), where the NoV samples were stratified into three main populations based on OFR1, and the genotype GII.4 samples were reported to cluster with GII.15 and GII.20 genotypes [52]. The complete genome sequences based analyses pursued in the current study unveiled a clear distinction of GII.4 compare to the rest of the GII genotypes, including the GII.15 and GII.20. Moreover, additional analyses of GII.4 isolates sequences suggested extra clustering at K=2

and 5 (Figure 2C). At K=5, the GII.4 sydney_2012 variants stratified into three lineages. Such stratification pattern of the Sydney_2012 variant is also reported earlier based on ORF2 gene sequences [53].

The current study identified admixture strains using the admixture model/linkage model implemented in the STRUCTURE program. The admixture model fails to take into account the physical relation between loci, and the proportion of admixed strains may sometimes be under or over-estimated. Therefore, to optimize the membership scores given to the admixed strains linkage correlated model was applied that report for potential linkage. The 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. Few of these admixed strains are reported to be globally prevalent such as GII.Pb/GII.3, GII.Pb/GII.13, and GII.Pg/GII.12 [54]. Recombination among the NoV strains occurs at high frequency and acts as a major driving force of viral evolution. Recombination allows the virus to increase its genetic fitness, evolve, 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 C-1.2b and C-1.3b clusters. Likewise, the $|D'|$, r^2 & I^S_A statistics inferred poor linkage evidence for norovirus GII isolates in the current study and indirectly justifying the role of recombination to shape the Norovirus GII isolates evolution.

The steady BSP plot generated on the basis of complete genome markers, speculate predominantly a stable effective population size for the NoV GII isolates originated from the Human host (Figure 5). The sharp decrease in effective population size of NoV in 2003 might be caused by the introduction of GII.4 as a new variant. In 2002-03, a marked increase in NoV infection was reported in England and other countries due to the emergence of GII.4 Farmington Hills and b4s6 variants [56]. The BSP plot also inferred a rapid increase in the effective population size during 2009-10. This might be accompanied by the large outbreaks and epidemicity of 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 fells sharply in 2015 that may likely correspond to the gain of host immunity against the dominant NoV variants infection.

The substantial signals of episodic diversifying selection were observed across all the proteins, including both the structural and non-structural proteins. However, limited pervasive positive selection signals identified for NoV GII samples at the VP1 and VPG genes. Although Xingguang et al., 2021, formerly reported no episodic positive selection signals detection for the genotype GII.2 isolates and speculated the genetic drift as a possible mechanism for NoV GII.2 evolution [59]. However, in the current study, significant positive selection signatures were identified for the GII.2 strains (Table S4, S5). This speculates the selection pressure as a possible driven force accompanied with the GII.2 evolution. Several other studies also reported small numbers 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 the host cell and considered to be a key site for immune recognition and receptor binding. Therefore, this protein might possibly be a potential target for vaccine development [61]. We observed several sites under the positive selection in both the P1 and P2 as well as the Shell domain of VP1 protein. The mutation at positions 282 to 395 of VP1 (Table S5) is a part of its P2 domain and this region reported to play an important role in

the interaction with the human blood group antigen (HBGA) [62]. The S domain is mainly conserved across different genotypes and mapping antigenic sites across this domain are mostly cross-reactive [63]. Besides positive selection, a large number of sites were under the influence of negative selection and signifying a scenario of purifying selection. In general, positive selection sites may be responsible for the immune pressure leading to an escape mutation, and negative selection sites may prevent deterioration of antigenic function and structures [64]. The sites under positive selection could provide markers for vaccine designing. The negatively selected sites identified in NoV GII genes may worthy to identify the highly conserved regions useful to implement new diagnostic protocols [65]. A marked differentiation was observed in the positive selection signatures pattern in the GII.4 samples compare to the rest of the GII genotypes, which might have shaped the differential genetic composition of the GII.4 genotype, as identified in the current analyses.

5. Conclusion

The complete genome-based population genetic analyses, pursued in the current study, unveiled significant distinctions of the GII.4 genotypes compare to the rest of the NoV GII genotypes. This differential genetic composition of GII.4 might be raised due to its specific positive selection signatures as observed in current study. The genetic stratification of GII.4 samples speculate about the emergence of additional GII.4 lineages. The analyses identified no continental-specific genetic composition of the NoV GII samples. Besides, the analyses of the current study speculating the recombination and selection pressure as major factors driving the genetic diversification and emergence of new lineages in the NoV GII strains. The findings of the current study may implicate for planning effective strategies to combat the NoV GII mediated infection.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors read and agreed to publish the study.

Availability of supplementary data

The supplementary data relevant to this study is provided.

Competing interests

The authors declare that they have no competing interests.

Funding

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

Authors' contribution

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

Acknowledgment

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

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. Journal of virology 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 et al 2020
11. Gaythorpe KAM, Trotter CL, Lopman B, Steele M, Conlan AJK (2018) Norovirus transmission dynamics: a modelling review. Epidemiology Infection 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. Journal of virology 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 N Costa, Teixeira LCP, Portela DM, de Lima ACR, da Silva Bandeira ICG, Júnior R, Siqueira ECS, Resque JAM, da Silva HR, L.D. and 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. Journal of virology 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. Molecular biology evolution 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. Molecular ecology 14:2611–2620
27. Earl DA (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation genetics resources 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 bioinformatics 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. The American journal of human genetics 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. Molecular biology evolution 32(1):268–274
31. Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS (2018) UFBoot2: improving the ultrafast bootstrap approximation. Molecular biology evolution 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 evolution 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 Research Human Retroviruses 21(1):98–102
36. Smith JM (1992) Analyzing the mosaic structure of genes. Journal of molecular evolution 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. Molecular biology evolution 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. Molecular biology evolution 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 research 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. Molecular biology evolution 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 computational 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. Statistical methods in molecular evolution. Springer. 125–181
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. Scientific reports 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. Scientific reports 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. Archives of virology 160(7):1705–1713
56. Bull RA, Tu ET, McIver CJ, Rawlinson WD, White PA (2006) Emergence of a new norovirus genotype II. 4 variant associated with global outbreaks of gastroenteritis. J Clin Microbiol 44(2):327–333

57. 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. Clinical infectious diseases 53(6):568–571
58. Vega E, Vinjé J, 2011. Novel GII. 12 norovirus strain, United States, 2009–2010. *Emerging infectious diseases*, *17*(8), p.1516
59. Li X, Liu H, Magalis BR, Pond SLK, Volz EM (2021) Molecular evolution of human norovirus GII. 2 clusters. *Frontiers in microbiology*, *12*
60. 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
61. 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
62. Hardy MJ, Kuczera G, Coombes PJ (2005) Integrated urban water cycle management: the UrbanCycle model. Water science technology 52(9):1–9
63. 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):67592. .e p.
64. Domingo E (2007) Virus evolution./Fields' Virology. Knipe D (ed), P. Howly
65. Presti AL, Rezza G, Stefanelli P (2020) Selective pressure on SARS-CoV-2 protein coding genes and glycosylation site prediction. Heliyon 6(9):.e05001

Figures

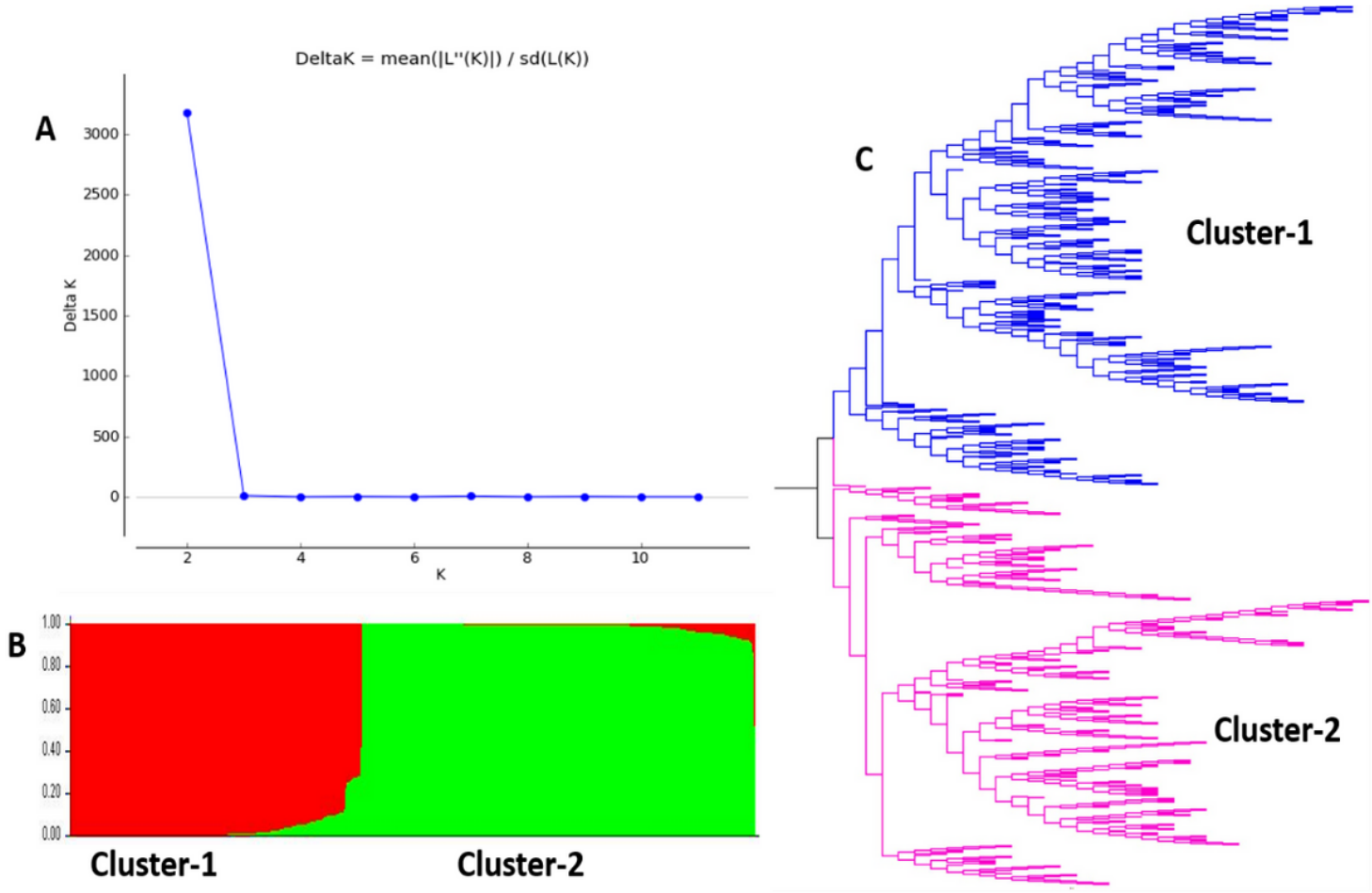

Figure 1

[A] Determination of Kopt for NoV GII: The graph show plot of K versus delta K, which defines the optimum number of cluster Kopt in the NoV GII population. K denotes the number of clusters while delta K illustrates the rate of change of likelihood posterior probability for the given number of sub-cluster K. The plot was executed at high simulation burn-ins (100,000) and burn-ins length (100,000). The plot representing the major peak at K=2 depicts that NoV genetic structure is grouped into two main subpopulations. (B) Estimate of population genetic structure of NoV at Kopt of 2 using admixture model in the STRUCTURE software. C-1 comprised 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 comprised of genotype GII.4. [C] The initial clustering pattern of phylogenetic tree results is congruent to the STRUCTURE result.

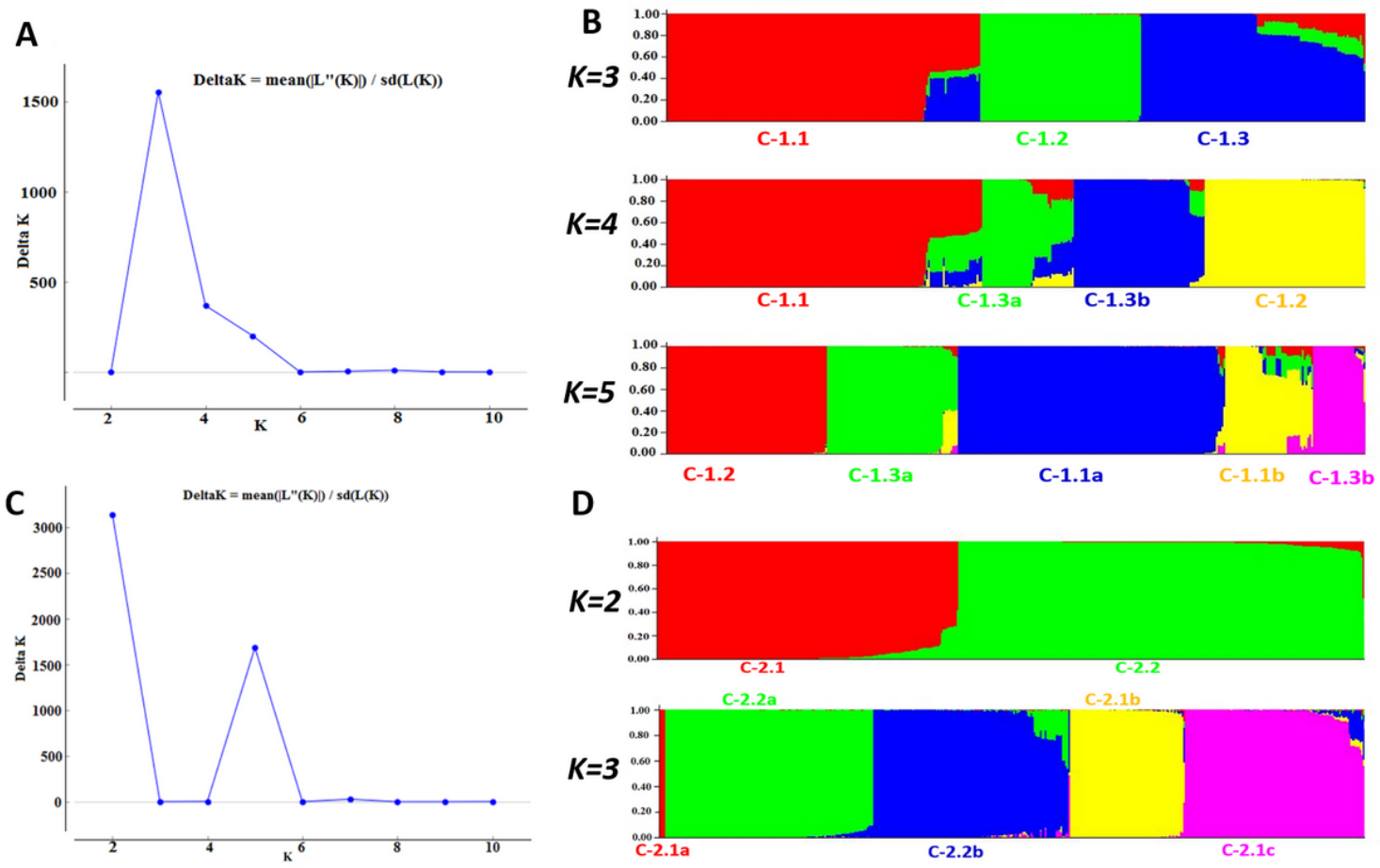

Figure 2

additional genetic structure analysis of C-1 and C-2 using admixture model in the STRUCTURE: (A) plot of K versus delta K shows the optimum number of subpopulations in C-1. The plot depicts a significant peak at K=3 followed by two minor peaks at K=4 and K=5 respectively. [B] Sublevel genetic structure of C-1 (non-GII.4 genotypes) obtained using STRUCTURE program applying 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 genotype 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 i.e. 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 showing the possible clustering of genotype GII.2 into two lineages (C-1.1a and C-1.1b). (C) Genetic structure of genotype GII.4 (cluster 2): At K=2 C-2.1 (GII.4 sydney_2012, New Orleans_2009, and Apeldoorn) and C-2.2 (Den_Haag_2006b). At K=5 C-2.1a (Sydney_2012), C-2.1b (Sydney_2012 and New Orleans), C-2.1c (Sydney_2012), C-2.2a (Den_Haag_2006), C-2.2b (Den_Haag_2006). [D] The plot of K versus delta K represents major peak at K=2 showing significant distinction of C-2 (GII.4) into two subpopulations. The second peak was found at K=5 which shows additional diversification of GII.4 into 5 lineages.

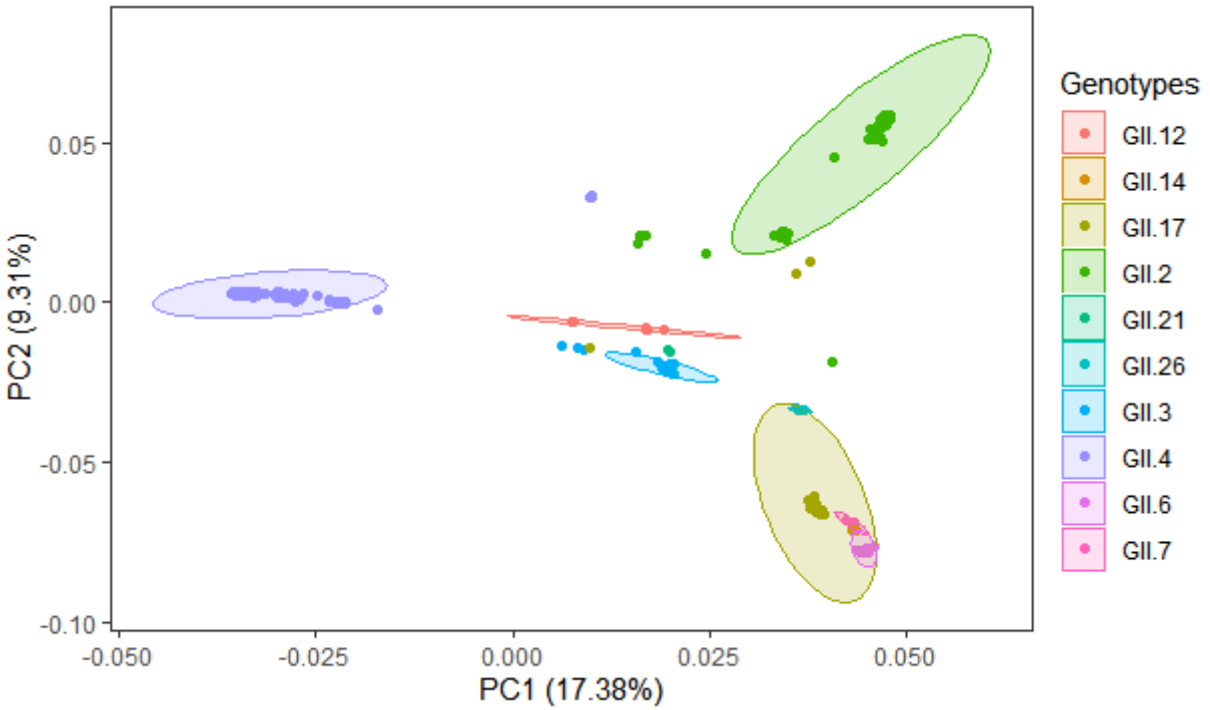

Figure 3

The two-dimensional PCA analysis of NoV GII samples.

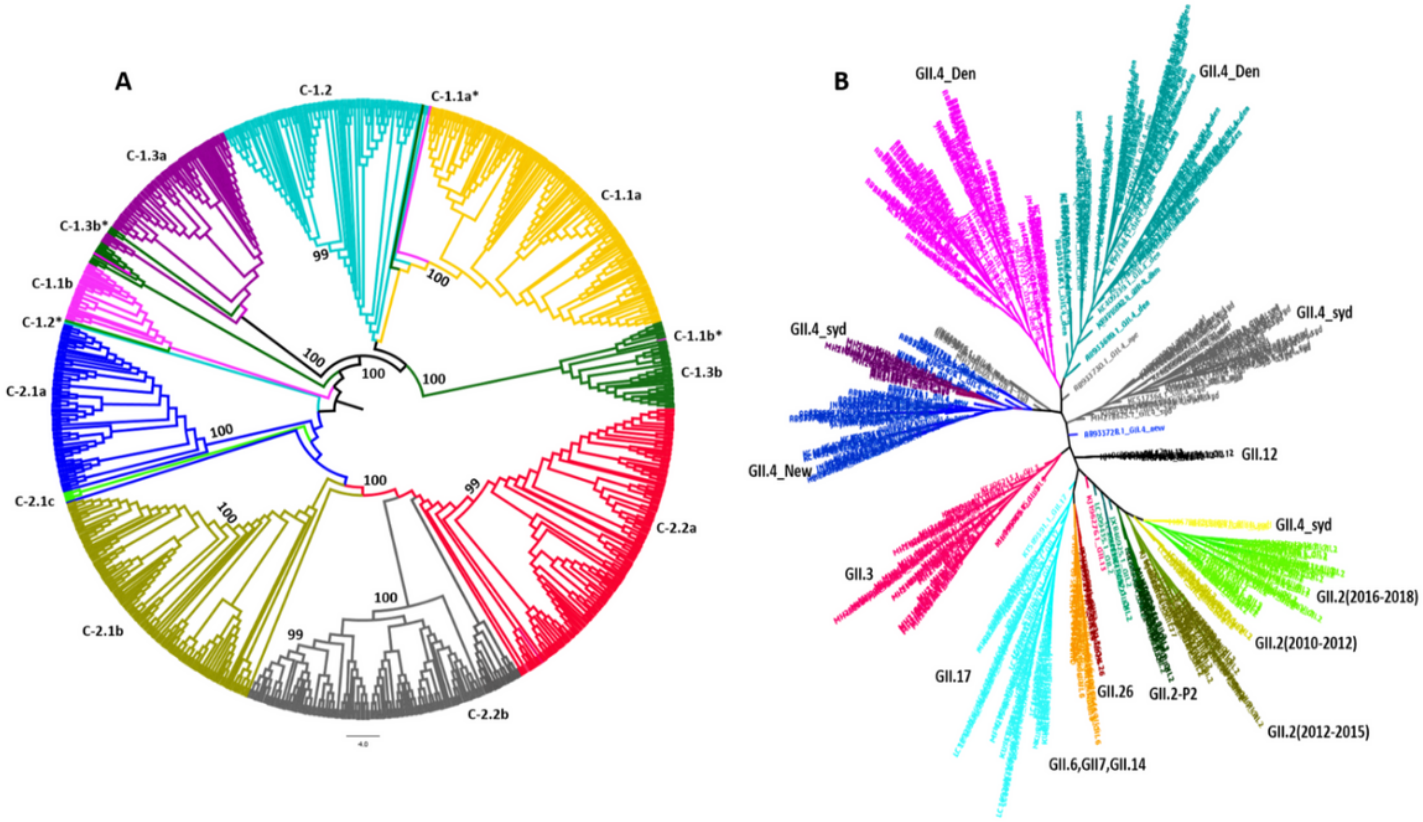

Figure 4

Phylogenetic tree analysis of the whole genome of NoV GII isolates: A) NJ-based tree of 822 strains were constructed using Mega-X. Ten clades were observed in the phylogenetic tree, which is congruent with the clustering pattern observed by 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), C-2.2b (Den Haag_2006b). The branches of recombinant /admixed strains are represented with an asterisk (*). B) Maximum likelihood tree of NoV GII.2. All major clades of NoV are colored and labeled.

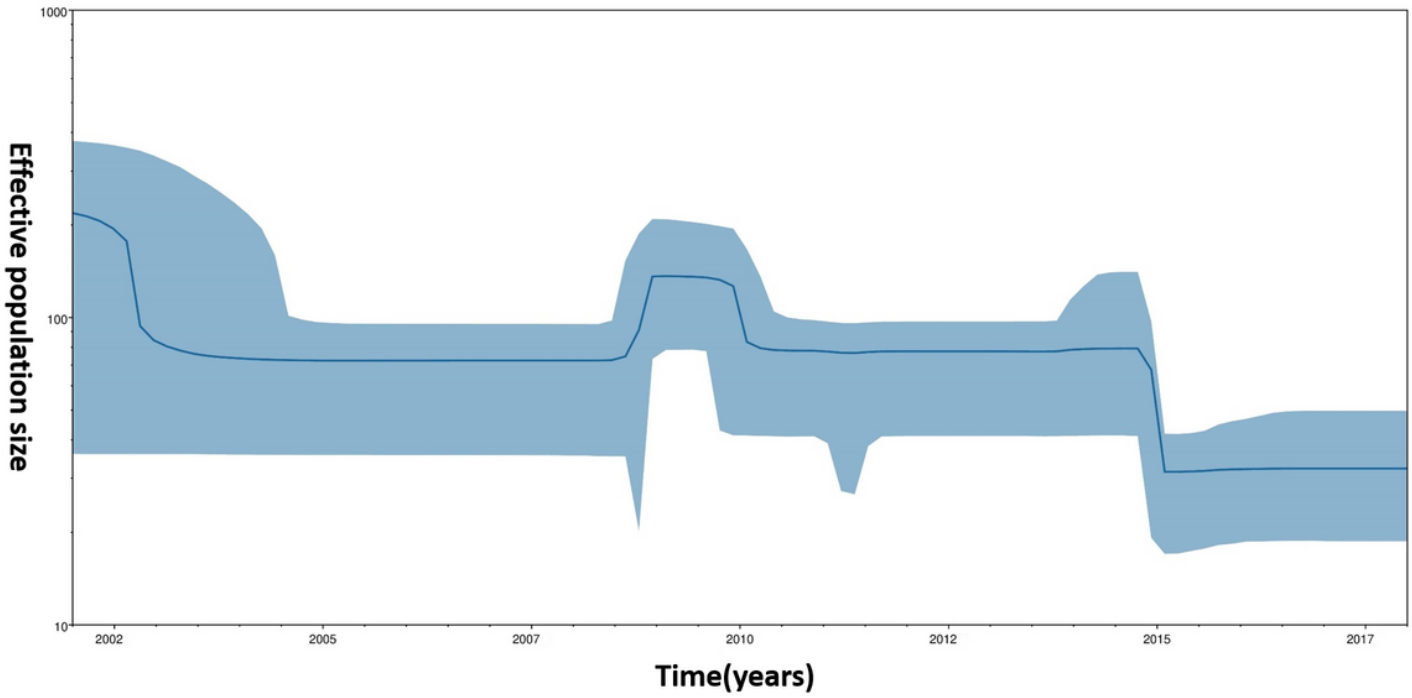

Figure 5

Bayesian skyline plot of Human NoV GII.2. The Y-axis represents effective population size, while the X-axis denotes time in years. The solid black line indicates the mean posterior value and the blue shaded area represents 95% HPD intervals.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1NoVGII.xlsx
- TableS2NoVGII.xlsx
- TableS3NoVGII.xlsx
- TableS4NoVGII.xlsx
- TableS5NoVGII.xlsx