JVI Accepts, published online ahead of print on 23 July 2014
J. Virol. doi:10.1128/JVI.01333-14

Deep sequencing of norovirus genomes defines evolutionary patterns in an urban tropical setting

Matthew Cotten[a], Velislava Petrova[a], My Vu Tra Phan[b], Maia A. Rabaa[b,c], Simon J. Watson[a], Swee Hoe Ong[a], Paul Kellam[a,d]# and Stephen Baker[b,e,f]

The Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, UK[a]; Hospital for Tropical Diseases, Wellcome Trust Major Overseas Programme, Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam[b]; Centre for Immunity, Infection and Evolution, University of Edinburgh, Edinburgh, UK[c]; Division of Infection & Immunity, University College London, London, UK[d]; The London School of Hygiene and Tropical Medicine, London, UK[e]; Centre for Tropical Medicine, University of Oxford, Oxford, UK[f];

#Address correspondence to: Professor Paul Kellam, pk5@sanger.ac.uk

Running head: Norovirus genomics in Vietnam

Abstract word count: 165

Text word count: 4809

Key words:

Genomics, deep sequencing, norovirus, diarrhea, evolution

ABSTRACT

Norovirus is a highly transmissible infectious agent that causes epidemic gastroenteritis in susceptible children and adults. Norovirus infections can be severe and can be initiated from an exceptionally small number of viral particles. Detailed genome sequence data are useful for tracking norovirus transmission and evolution. To address this need we have developed a whole-genome deep sequencing method that generates entire genome sequences from small

amounts of clinical specimens. This novel approach employs an algorithm for reverse transcription and PCR amplification primer design using all publically-available norovirus sequence data. Deep sequencing and *de novo* assembly were used to generate norovirus genomes from a large set of diarrheal patients attending 3 hospitals in Ho Chi Minh City, Vietnam, over a 2.5-year period. Positive selection analysis and direct examination of protein changes in the virus over time identified codons in the regions encoding proteins VP1, p48 (NS1-2) and p22 (NS4) under positive selection and expands the known targets of norovirus evolutionary pressure.

Importance

The high transmissibility and rapid evolution rate of norovirus, combined with short-lived host immune responses, are thought to be responsible for the virus causing a majority of pediatric viral diarrhea cases. The evolutionary patterns of this RNA virus have only been described in detail for a portion of the virus genome and never from a detailed urban tropical setting. We have developed robust deep sequencing methods for generating complete genome sequences directly from small amounts of patient fecal material. We use this method to provide a detailed sequence description of the noroviruses circulating in three Ho Chi Minh City hospitals over a 2.5-year period. The study identified patterns of virus change in known sites of host immune response and identified three additional regions of the virus genome under selection that were not previously recognized. In addition, the methods described here provide a robust full-genome sequencing platform for community-based virus surveillance.

INTRODUCTION

Norovirus is a non-enveloped positive-sense single-stranded RNA virus of approximately 7.5-7.7 kb in length (reviewed in reference (1)). The viral genome is organized into three (or four in the case of murine norovirus (2)) open reading frames (ORFs), encoding several structural and non-structural proteins. The ORF1 encodes for a large polyprotein,

which is proteolytically cleaved into 6 non-structural proteins, including N-terminal p48 protein (NS1-2), an NTPase protein (NS3), a 3A-like p22 protein (NS4), a viral protein-genome linked VpG protein (NS5), a 3C-like protease 3CLpro protein (NS6) and an RNA-dependent RNA polymerase RdRp (NS7). Note that the nomenclature for the NS proteins is currently in flux and both existing names have been included (3). The ORF2 overlaps ORF1 by a short region and encodes the major capsid protein VP1, comprising an S (shell) domain connecting the two P (protruding) sub-domains, P1 and P2 with the P2 domain binding to histo-blood group antigens (HBGAs) on target host cells. ORF3, located at the 3' end of the genome, encodes the minor capsid protein VP2.

Norovirus comprises one of the genera in the Caliciviridae family of viruses, and can be further classified into different genogroups (reviewed in (1)). Noroviruses are known to cause diseases in humans (genogroup GI, GII and GIV) and a number of other mammals including porcine (GII), ovine/bovine (GIII), canine (GIV) and murine (MNV, forming a distinct GV) viruses(4-12).

In humans, norovirus is a highly infectious pathogen that causes a severe gastrointestinal disease in susceptible individual after the ingestion of an exceptionally small number of viral particles. The virus is so infectious that the probability of symptomatic disease from a single norovirus virion has been estimated to be as high as 0.5 (13). A determination of the infectious dose required to infect 50% of subjects (ID_{50}) estimates this as 1,000-3,000 virus genome equivalents(14). A typical norovirus infection can result in profuse volumes of feces and vomitus containing 10^6-10^9 stable, non-enveloped virions per ml of excreta, creating almost infinite opportunities for onward transmission and additional infections. An inability to culture human noroviruses in a laboratory prevents the testing of inactivation and disinfection methods and further complicates control efforts. These issues highlight some of the difficulties in eliminating infectious norovirus from food supplies and the environment, and indicate the need for the development of intelligent approaches to prevent norovirus transmission and infection.

An effective approach to controlling norovirus may be to understand how norovirus evades the human immune system and use this information to develop novel therapeutic options. Norovirus infection in a "healthy" individual is typically short and self-limiting, which results in transient or long-lived immunity (15, 16). No approved drugs exist to block virus replication. Accordingly, public health measures to identify and eliminate sources of infection or behavior leading to virus spread are warranted (17, 18). The utility of viral sequencing to track norovirus in transmission studies has been explored with fragments of the viral genome (19-22). As a consequence of the speed of disease onset and high transmissibility, the number nucleotide and amino acid sequence changes within a local outbreak may be rare so the sequencing of larger genomic fragments should provide greater resolution for defining transmission patterns.

The natural duration and specificity of immune responses to norovirus are difficult to measure due to the lack of a cell culture system for norovirus neutralization studies and the inability to grow a defined virus for such trials (reviewed in (16, 23)). The duration of norovirus immunity may be limited due to the short period of a typical infection and a corresponding short exposure to viral antigens. Periodic population-level replacement of norovirus lineages with viruses with surface residues under positive selection is evidence of immune response-driven antigenic change and suggests that these immune responses are of sufficient strength to drive viral evolution(24-26). Immune studies have identified blockade epitopes in VP1, the major capsid protein. These epitopes are important for interacting with histo-blood group antigens (HBGAs) on target host cells; high titers of antibodies that block virus-like particles binding to HBGAs correlate with protection to norovirus challenge (27-29).

Diarrheal diseases are a serious health problem, especially in developing countries when combined with nutritional problems, co-infection with other pathogens, crowding and limited access to health care. It is clear that norovirus and rotavirus are frequently associated with diarrhea in this setting (30) and it is essential to closely follow the local evolution of norovirus. We describe here a method for deep sequencing the approximately 7,500

nucleotide norovirus RNA genome directly from patient material and used this method to provide a detailed description of genome and community-wide norovirus evolution.

MATERIALS AND METHODS

Primer design. Primers were designed using Python algorithms to identify highly conserved primer targets in the appropriate genome locations. Briefly the algorithm takes as input all complete genome sequence of human norovirus available in GenBank (January 2012, 260 GII.4 entries, 5 GI entries, total sequence 1.9x 10^6 nt). A counting method was employed to identify all highly conserved primer-like sequences with a G+C percentage between 30 and 75 %, a calculated melting temperature between 55 and 59 °C, and with no single nucleotide comprising greater than 40 % of the sequence. The norovirus genome was divided into three overlapping 2.5-3 kb amplicons and the highest frequency primer site in the first and last 800 nt of each amplicon were selected. Finally the primers were used in a virtual PCR reaction to determining the binding behavior of the primer set all available full norovirus genomes (see Figure 1). Primer details are summarized in Table 1.

Sample collection. Stool samples were obtained as part of a larger study examining causes of pediatric diarrhea in subjects presenting to Children's Hospital 1, Children's Hospital 2 and The Hospital for Tropical Diseases, Ho Chi Minh City (HCMC), Vietnam (30, 31). Additional samples came from an *ad hoc* enrollment of children admitted to Children's Hospital 2 with potentially hospital-acquired norovirus diarrhea or prolonged norovirus incubation. In the *ad hoc* collection, pediatric patients were admitted to the hospital due to diseases other than diarrheal diseases and had no diarrhea when presenting at the hospital. These patients eventually developed diarrhea after 48 hours of hospital admission and diarrhea lasted for at least 3 days after the onset. Ethical approval was granted from the institutional ethical review boards and the University of Oxford Tropical Research Ethics Committee (OxTREC No. 0109).

Generation of amplified cDNA for deep sequencing. For RNA extraction, 140 µl of each stool specimen was subjected to automated extraction into a final 50 µl elution using the MagNA Pure 96 automated extraction machine according to manufacturer's instruction (Roche). Reverse transcription was performed as previously described(32). Briefly, a primer mix was prepared separately for each amplicon: the reverse primers for the amplicon were

pooled in an equimolar ratio and water up to 7 µl primer mix (7.6 pmol of each primer; 0.38pmol/ µl per reaction). Extracted norovirus RNA was diluted in 1:10 in water; 5 µl of this dilution was added to the primer mix and heated for 5 min at 65 °C and immediately transferred to an ice block for 1 min. An enzyme mix was then added to each reaction and mixed by pipetting. Each 20 µl reaction contained 4µl 5x First Strand buffer (250 mM Tris-HCl (pH 8.3), 375mM KCl, 15 mM $MgCl_2$), 1µl 0.1 M DTT, 1µl 10mM dNTPs, 1 µl RNase Inhibitor (Promega, 40U/µl), 1 µl SuperScript III RT (Life Technologies, 200U/µl). Reverse transcription was performed at 50°C for 60 min, followed by 70 °C for 15 min.

PCR amplification. Amplification was performed with primer mix solutions prepared for each amplicon. Primer mix (per 25 µl reaction): the forward and reverse primers from each amplicon were pooled together in 1.5:1 ratio (1.9 pmol of each forward primer and 1.26 pmol of each reverse primer; 0.08 pmol/µl and 0.05 pmol/ µl respectively). A 5 µl aliquot of the reverse transcription reaction for the each amplicon was used as a template for the PCR step. The following thermal cycling conditions were used: enzyme activation: 98°C for 30 sec; cycling (35 cycles): 98°C for 10 sec, 53°C for 30 sec, 72°C for 3.0 min; final extension: 72°C for 10 min; hold: 4°C.

Sequencing and genome assembly. Pooled amplicons for each sample (approximately 1.2 µg) were individually indexed and subjected to sequencing with Illumina MiSeq (33, 34) to generate approximately 300,000 reads of 149 nt per sample (median value 302904 reads). All reads were processed using QUASR (35) to remove sequencing adapters and index sequences and to trim primer sequences present within a fixed distance from the 5' or 3' end of a read. Reads were then trimmed from the 3' end to reach a minimum median Phred quality score of 35 and reads below 125 nt in length were removed. After primer trimming and quality control for each sample, *de novo* assembly using SPAdes (36) was used to generate full norovirus genomes. Intact open reading frames (ORFs) were checked using Python scripts as a measure of correct genome assembly.

Recombination detection. The 119 complete genome of all GII noroviruses from this study and from global data (retrieved from GenBank) were manually aligned using Se-AL v2.0 (http://tree.bio.ed.ac.uk/software/seal/). Only full-length sequences with information on sample date and location were included in this analysis. Potential presence of recombination in these complete sequences was screened using the Recombination Detection Program version 4 (RDP4) software (37). RDP, GENECONV, 3SEQ and MAXCHI methods were employed for primary screening and the BOOTSCAN and SISCAN methods for automatic checking of the recombination signals, as described previously (38). The automask xfunction in RDP4 was selected for optimal recombination detection, *i.e.* one representative strain within each group of similar sequences was examined during the primary/exploratory search for recombination signals while the remaining sequences within groups of sequences with high similarity were automatically masked. Using this method, masked sequences were examined for the presence of recombination if the program detected a recombination signal in the representative unmasked sequence. Each test of recombination used a 400 nt sliding window, and any recombination signals with significant p-values for ≥3 test parameters were considered as potential recombination events. A further analysis on these potential recombinants, comparing tree topologies with likelihood (Shimodaira-Hidetoshi, SH test) was employed to determine which of the testing strains were likely to be true recombinants and which were not. All intra-ORF recombinant strains (GenBank accession numbers EU921388, AB541275, GU991355 and AB541254) were excluded from the estimation of selection rates.

Phylogenetic analysis. An alignment of non-recombinant sequences, including all full genomes determined in this analysis and global background sequences obtained from GenBank, was utilized to reconstruct evolutionary relationships among norovirus sequences. A phylogenetic tree was inferred using aligned nucleotide sequences, employing a maximum likelihood (ML) method in RaxML (39), under the GTR+Γ model of substitution, determined to be the best fit model to our data using jModelTest version 2.1.1 (40). Tree topology was

assessed through bootstrapping with 1,000 pseudo-replicates. The resulting phylogenetic tree was visualized and edited in FigTree v1.4.0 (http://tree.bio.ed.ac.uk/software/figtree/).

Evolutionary rate estimations. Evolutionary rates were estimated using a Bayesian Markov chain Monte Carlo (BMCMC) method implemented in BEAST version 1.7.2 (41). A relaxed uncorrelated lognormal molecular clock was employed to account for lineage-specific rates, and a GMRF Bayesian skyride coalescent (42) was used to model the population dynamics. The relevant substitution models for each alignment were selected using jModelTest version 2.1.1 (40). The mean evolutionary rate and the 95% upper and lower highest posterior density (HPD) intervals were inferred from the posterior tree distribution generated from the BMCMC runs using Tracer version 1.6 (http://tree.bio.ed.ac.uk/software/tracer/).

Positive selection analysis. To determine evolutionary patterns of norovirus, selection analyses were performed on the regions encoding VP1, VP2, and ORF1-encoded p48 (NS1-2) and p22 (NS4) proteins. Norovirus codons under selective pressure were first determined using the Mixed Effects Model of Evolution (MEME; p-value < 0.05) (43) and Fast Unconstrained Bayesian AppRoximation (FUBAR; posterior probability >0.9) (44) implemented through the Datamonkey web browser (45). Codons that were found to be under positive selection by either method were inspected at the sequence alignment, and those with no evidence of polymorphisms were considered a false positive and discarded.

Ancestral sequences were reconstructed from the sequence alignment and inferred phylogeny using the joint likelihood method implemented in Hyphy (46) under a GTR+Γ model of evolution.

The GenBank accession numbers for all new norovirus sequences reported here are reported in Table 2. Also listed are the samples collection dates, the genetic cluster (see Figure 2 and Table 4) and the European Nucleotide Archive accession numbers for the raw sequence data. In addition, 89 GII.4 genomes from the same HCMC study were also publically available in GenBank, with the following accession numbers: Cluster 1: KC409244,

KC409245, KC409246,KC409257, KC409258, KC409259, KC409260, KC409261, KC409262, KC409264, KC409265, KC409266, KC409267, KC409268, KC409269, KC409270, KC409271, KC409272, KC409273, KC409274, KC409275, KC409276, KC409277, KC409279, KC409280, KC409281, KC409282, KC409283, KC409284, KC409285, KC409286, KC409287, KC409288, KC409289, KC409290, KC409291, KC409293, KC409294, KC409295, KC409296, KC409297, KC409298, KC409304, KC409305, KC409306, KC409307, KC409308, KC409309, KC409310, KC409312, KC409313, KC409314, KC409315, KC409318, KC175360, KC175365, KC175366, KC175371, KC175373, KC175381, KC175388, KC175389, KC175390, KC175391, KC175392, KC175393, KC175394, KC175395, KC175396, KC175406, KC175407, KC175408, KC175409, KC175410

Cluster 3: KC409256, KC409263, KC409278

Cluster 4: KC409240, KC409241, KC409242, KC409243, KC409299, KC409301, KC409302, KC409303, KC175384, KC175385, KC175386, KC175387.

RESULTS

Norovirus sequencing strategy. A novel general strategy for designing PCR primers was developed that would permit the production of complete norovirus genome sequences. Deep sequencing of RNA virus genomes requires reverse transcription (RT) of viral RNA and the amplification of the resulting cDNA that encompass the entire viral genome. Python algorithms were used to process all available norovirus full genome data (265 full genomes,, January 2012) and to select primer target sequences suitable for whole genome amplification. Briefly the algorithm processes the norovirus sequence data into primer-sized sequences trimmed to a calculated melting temperature. The frequency of each sequence in the entire set is calculated, with high frequency sequences correlating with conserved sites across the viral genome. The norovirus genome was divided into three overlapping amplicons, potential primers were mapped to a reference genome and the highest frequency sequences mapping within the terminal 800 nt of each amplicon were identified. Reverse complements of the primers mapping to the 3' end were prepared. A virtual PCR was performed to examine the potential function of the primers across all known full norovirus genomes, the output of such an analysis is shown in Figure 1 (left panel) with blue markers indicating the position of the each primer and grey bars indicating the expected PCR product. The actual function of the primer set is demonstrated in Figure 1 (right panel) with each lane showing the PCR products from 14 samples, present by amplicon. Each reverse transcription reaction contained the 2 (or 3 for amplicon 3) reverse primers for each amplicon 1, 2 or 3, and each PCR reaction contained the 2 (or 3 for amplicon 3) forward and reverse primers for amplicon 1, 2, or 3. Within these samples, sample 7 failed, however the remaining 13 samples provided sufficient material for deep sequencing.

A summary of the predicted performance of the norovirus primer set on all available norovirus genomes is included in Table 1. All full length Norovirus GII genomes (TaxonID 142786, length 7000-8000 nt, 517 entries) or all Norovirus genomes (TaxonID 122929, length 7000-8000 nt, 753 entries) were retrieved from GenBank. The target sequence for each

primer was examined in these genome sets and the percentage of the genomes with a perfect match to the target sequence or with a functional match (0-3 mismatches) to the target sequence was reported. For the Norovirus GII genomes, the primers have a perfect match to 79% of the genomes, and a functional match (up to 3 mismatches) for 97% of the genomes. For the complete set of Norovirus genomes (this includes all GI, GII and all animal noroviruses), the primers have a perfect match to 65% of the genomes, and a functional match (0-3 mismatches) for 82% of the genomes. These values and the details of the analysis as well as the GC content and calculated Tm for all primers are listed in Table 1.

A summary of the performance of the norovirus primer set for amplifying and sequencing 188 fecal-derived RNA samples is presented in Table 3. PCR success was defined as obtaining the three amplicon-specific RT-PCR products of the predicted size with sufficient yield for sequencing library preparation. The overall RT-PCR success rate was 78.2% (147 out of the 188 tested clinical samples). The most common genotype globally (GII.4), had the highest PCR success rate (93.7%, 74 of 79 samples), followed by GII.6 (88%, 7 of 8 samples), GII.13 (83%, 5 of 6 samples), and GII.3 (77%, 26 of 34 samples). Much lower amplification efficiency was observed for genogroup I (GI) strains, with successful PCR genome amplification in only two of 10 tested samples. The high success with GII with respect to GI strains (especially GII.4) was predicable given that GII.4 genomes dominate sequences in public databases. Future primer sets could be reiteratively designed using targeted and revised genome data sets.

Norovirus diversity in HCMC. Using the developed whole genome sequencing technique, 112 novel GII norovirus genomic sequences were generated. In addition, 89 GII.4 genomes from the same HCMC study were also publically available in GenBank; these were included in the following analysis for a total of 201 complete genomes, with collection dates between April 2009 and December 2011. A phylogenetic analysis of the 201 genomes defined eight genotypes of GII norovirus using maximum likelihood methods (Figure 2). Consistent with previous characterization of norovirus infections in HCMC (30), and global norovirus

patterns, the most prevalent genotype GII.4 found in this study belonged to the GII.4 Den Haag genotype lineage (Figure 2, Clusters 1, 2 and 3), which is most genetically similar to the GII.4 Minerva_2006b partial sequence and Taiwan_2006 (GenBank JN400601). Phylogenetically, the GII.4 strains in Cluster 4 (Figure 2A) were most closely related to GII.4 New_Orleans_2010 (GenBank JN595867), while GII strains in Cluster 5 were classified as GII.P21_GII.3 and most closely related to the strain NV_Pune_2007 (GenBank EU921389). A small number of strains belonged to genotype GII.Pg_GII.12 (Cluster 6), while viruses of the GII.P7_GII.6 genotype fell into the 2 distinct lineages, Clusters 7 and 8 (Figure 2). Our genotype assignment using phylogenetic reconstruction was consistent with the genotype designation generated by the RIVM algorithm (47) (Table 4). Additionally, the relative frequency of each genotype observed in the full genome set was similar to the frequencies determined by My *et al.* (31) from a larger set of HCMC samples using ORF1 and 2 fragments (Table 4), indicating that the generation of full genome sequences was not strongly influenced by genotype-based selection biases. Viruses of the GII.4 Den Haag and GII.P21/GII.3 genotypes, (Cluster 1 and 5), were identified over two sampling periods, from 2009-2010 and later in 2011, while other virus genotypes were detected only in the first sampling period.

The temporal occurrence of sampled noroviruses is shown in Figure 3, with samples stratified by genotype cluster. The three GII.P4/GII.4(2006) genotypes (Clusters 1,2, and 3) were present in the first half of 2009, with the GII.P4/GII.4(2010) genotype (Cluster 4, grey) first appearing at the end of 2009. There was a pause in sampling in the first half of 2011, followed by sampling in the second half of 2011. Reduced diversity was observed in 2011, with only Clusters 3 and 5 sequenced from these samples. Changes in sampling protocols between 2010 and 2011 preclude inference on how this reduced diversity may relate to norovirus epidemiology and evolution. However, the identification of clusters of phylogenetically-related viruses undergoing *in situ* evolution in this region over the observation period allowed an examination of evolutionary processes that may allow the continued transmission and maintenance of viral lineages in the presence of population immune responses. Characterizing

317 such changes in the norovirus population may provide important clues about how the virus
318 evades host immunity.

319 **Evolutionary rates within each cluster.** A sufficient number of genomes were
320 available from Clusters 1, 4 and 5 for well-supported evolutionary rate estimations (Table 5).
321 Mean evolutionary rates of 6.15, 5.73 and 5.34 (x 10^{-3} substitutions per site per year) were
322 estimated from the full genomes for Cluster 1, 4 and 5. Figure 4 plots the rates for the GII.4
323 Cluster 1 viruses by the region of the genome used for each calculation.

324 ORF-specific rates estimated for the three genetic clusters show that the ORF1 regions
325 exhibited a reduced rate compared to those of the ORF2 (VP1) regions. For all three clusters
326 the ORF1 and ORF2 (VP1) regions showed modestly reduced rates relative to the full
327 genome, while the ORF3 (VP2) substitution rates for both Cluster 1 (8.99 x 10^{-3} substitutions
328 per site per year) and Cluster 5 viruses (7.38 x 10^{-3} substitutions per site per year) were higher
329 than those of the whole genome. The overlapping confidence intervals for these estimations
330 make these conclusions less secure. The amount of signal available for the Cluster 4 ORF3
331 was not sufficient to yield a reliable rate estimate.

332 The norovirus ORF1 encodes a large polyprotein containing the viral polymerase,
333 protease and several essential replicase components. Evolutionary rates were estimated
334 separately for these individual coding regions of the Cluster 1 ORF1 (Table 5, Figure 4). The
335 region encoding p22 (NS4) showed the highest levels of change (6.60 and 8.21 x 10^{-3}
336 substitutions per site per year, Figure 4), greater than the whole genome rates for Cluster 1
337 (6.15 x 10^{-3} substitutions per site per year). The enzymes (NTPase (NS3), protease and RdRp
338 (NS7)) and VP1 show modestly lower substitution rates than observed across the whole
339 genome.

340 **Amino acid changes in norovirus proteins.** The evolutionary patterns of four
341 norovirus-encoded proteins with the higher evolutionary rates were examined (VP1, VP2 and
342 p48 (NS1-2) and p22 (NS4)). An alignment of protein sequences ordered by time was used to
343 detect sustained versus sporadic changes in the protein relative to a reconstructed ancestral

sequence. Information about the biochemical properties of the protein was gathered from published literature. Positive selection analysis was performed with the mixed effects model of evolution (MEME) (43) or fast unconstrained Bayesian approximation (FUBAR) (44).

Cluster 1 VP1 showed changes in multiple patients relative to the ancestral sequence: Q106R, S174P, N298D (in blockade epitope A), G340E, G393S (in blockade epitope D) (Figure 5). Additional substitutions were seen at lower frequency suggesting evolution during the course of transmission through HCMC. Position 298 in the blockade epitope A was found to be positively selected using FUBAR, while both FUBAR and MEME identified position 106 within the shell domain (Figure 5) to be under positive selection (Table 6).

An alignment of VP2 protein sequences ordered by time was used to detect sustained versus sporadic changes in the protein relative to the ancestral sequence. Several changes, including T139M/A, I144V/T, and Y169H occurred in multiple HCMC Cluster 1 viruses with a much higher frequency of changes in the internal region of the protein (Figure 6). It was previously noted that changes in this region of VP2 (VP1 interacting domain (VP_ID) were associated with changes in VP1(48). Both MEME and FUBAR identified VP2 codon 144 (marked with red asterisk in Figure 6) as being under positive selection.

The region encoding p22 (NS4) from the Cluster 1 viruses showed elevated evolutionary rates relative to the full genome (Table 5, Figure 4). Analysis of all encoded p22 (NS4) molecules from Cluster 1 (Figure 7) showed amino acid changes relative to the ancestral sequence. Substitutions were observed in multiple isolates suggesting neutral or positive selection (I29V, E46D, N77S, R82K, T86S and D174V). Analysis of all encoded p48 (NS1-2) molecules from Cluster 1 (Figure 8) showed amino acid changes appearing in multiple isolates suggesting neutral consequences with no constraints to limit change or positive selection (D7V, N15D, R55K, V79T or A, and S184P. Both MEME and FUBAR identified p48 (NS1-2) codon 79 to be under positive selection (Table 6).

DISCUSSION

Our work outlines a strategy for full genome deep sequencing of norovirus directly from fecal specimens and we have applied the strategy to characterize norovirus samples collected across a clinical spectrum of pediatric norovirus infections in HCMC, Vietnam. An essential component of the methods is a primer design algorithm that takes as input all available sequence data for a virus and quickly provides a set of functional primers. The flexible design of the primer design algorithm avoids a cumbersome alignment step in the process and facilitates regular updates with new sequence data. This is essential to avoid perpetuating a bias in the sequence data whereby sequences are only obtained if primers have functioned, and primers are designed on antiquated data sets. The method showed a high success rate of full genome sequencing of GII noroviruses, especially GII.4, which was predictable given that GII.4 genomes dominated the sequence data set used to design the primers. Future primer sets will be designed using more targeted and updated genomes sets and including more sequence data from other genogroups.

Results using this method have provided a large set of norovirus genome sequences derived from longitudinal samples from one location At the start of this study, 265 full norovirus genomes were available in GenBank and the current study added an additional 112 genomes. The data allowed estimation of evolutionary rates for several genotypes, for full genomes as well as for sub-genomic regions. The evolutionary pressures and the constraints to avoid change are not expected to be uniform across the virus genome. Selection pressures are likely to vary greatly depending on the function of the encoded proteins, with enzymatic and structural regions more constrained then surface and immune-exposed or spacer regions with less defined functions. ORF-specific substitution rates estimated for the three phylogenetic clusters show that for the ORF1 regions exhibited a reduced evolutionary rate compared to those of the ORF2 (VP1) regions rates.

Previous studies have estimated norovirus GII.4 and GII.3 VP1 capsid regions to evolve at 5.1 to 5.8 x 10^{-3} substitutions per site per year (49) (50) (51), while the GII.4 polymerase

region was estimated to evolve between 4.33 - 8.98 x 10^{-3} substitutions per site per year depending on the data set used (49). Our estimates using the HCMC data are consistent with these published figures. The evolutionary rate determined for GII.4 Cluster 1 was higher than the estimated rates for GII.4 Cluster 4 and the GII.3 Cluster 5 viruses, which is perhaps due to a greater number of Cluster 1 infections/unit time and thus a greater number of replication events. Alternatively, the three virus genotypes might have intrinsically different replication properties, polymerase fidelity, or immune selection pressure that result in the differing rates.

The norovirus sequence data obtained from this study allowed an analysis of the evolutionary patterns of the second viral capsid protein VP2. The high evolutionary rates reported here (Cluster 5: 7.38 x 10^{-3} substitutions per site per year, Cluster 1: 8.99 x 10^{-3} substitutions per site per year) have not been observed previously, as this region is seldom included in previous sequencing projects. The structure of VP2 is not defined although there is evidence that the protein is interior to the VP1 shell and may be important for assembly of the VP1 structure (52). The protein is moderately basic and the C-terminal half of the protein is rich in serine and threonine residues (providing possible phosphorylation sites) and proline residues (perhaps accounting for the inability to define a structure for this protein). Evidence that changes in VP2 accompany changes in VP1 has been presented (48). Recently the murine norovirus (MNV) VP2 has been shown to influence the host immune response to the virus, with MNV1 VP2 interfering with antigen presenting cell (APC) function and MNV3 VP2 promoting the response (53). These observations identify a possible site of virus/host interaction that could be a source of selective pressure. The evolutionary rates for the VP2-encoding regions were found here to be much higher than the well-studied norovirus VP1 region and the higher rates are consistent with less-constrained protein product, stronger selection pressures or both. Positive selection analysis across the VP2 region identified position 144 to be under selection, this position region of the protein was previously found to be involved in interactions of VP2 with VP1(52). A high evolutionary rate in a virus capsid protein suggests a region of the virion experiencing immune selection. Vaccine development

efforts should take this accelerated rate of change into consideration when selecting components for a vaccine.

Humoral immunity to norovirus (at least GII.4) may involve blockade antibodies that bind and block the VP1 residues required for binding to HBGAs (16, 54, 55). The correlation of high titer blockade antibodies with protection from gastroenteritis in challenge studies (29) and the frequent evolution of these sites (blockade epitopes A, D and E) suggest that these amino acid residues may be frequent targets of immune selection(55). Blockade epitope D may be directly involved in HBGA binding (16, 54). Our observation of changes in VP1 position 298 epitope A, position 393 epitope D, and position 412 epitope E support these previous conclusions. Several additional changes were located outside of blockade epitopes (S78G, S174P, G340E, T502N, Figure 5). Further studies should investigate whether these are founder effect changes of neutral consequence, or if they provide an advantage for the virus.

Similar mean evolutionary rates for full genome were found for Cluster 1, 4 and 5, with 95% confidence ranges largely overlapping. One might expect a higher evolutionary rate for GII.4 viruses compared to GII.3 viruses if the 10-fold higher detection frequency than GII.3, viruses directly reflects the community prevalence of these two infections. The similar full genome rates suggests that either the number of active infections is not a large factor in rate, or that the less frequently diagnosed GII.3 infection may be as frequent in the population as GII.4, but does not appear as frequently in clinics.

The ORF1-encoded p22 (NS4) regions showed a higher evolutionary rate compared to the full genome and p48 (NS1-2) codon 79 was found to be under positive selection. The function of p22 (NS4) is not known but the protein has been observed to localize to the Golgi/ER and influence the host secretory pathway with a centrally located MERES motif required for localization (56, 57). The function of p48 (NS1-2) in norovirus infection is also largely unexplored, although the protein is reported to localize to vesicles and has been proposed to influence protein trafficking (58). The evidence that these viral proteins interact with host proteins, combined with higher evolutionary rate or positive selection described here,

suggest that these proteins may be interacting with host restriction factors. Alternatively these regions with higher rates of change could encode proteins with no constraint. Further studies are needed to clarify this.

Extensive work has been done using the Feline Calicivirus (FCV) and MNV models to elucidate the roles and interactions of the non-structural (NS1-7) and structural proteins (VP1 and VP2) in regulating virus replication and infectivity, as comprehensively reviewed (1). However a functional profiling of human norovirus is not yet possible due to the lack of tissue culture and animal models for human norovirus replication. The full genome sequences of human norovirus available from this study provides valuable data on the spectrum of changes in the viral proteins allowed by the virus while awaiting alternative models for functional experiments.

This study has provided a description of norovirus evolution rates across HCMC over a 2.5-year period for full genomes as well as for subgenomic regions of the virus. We reveal for the first time a higher evolutionary rate for three regions of the genome (VP2, p22 (NS4) and p48 (NS1-2)) and provide evidence of positive selection in two coding regions (VP2 and p48 (NS1-2)). We suggest that these regions should be monitored for interactions with the host that might be a source of selective pressure. Finally, we believe this study and the methods we have described will provide a useful template for community-wide studies of full genome evolution for many RNA virus pathogens.

Acknowledgements.

We thank the Sanger Institute Illumina C team for their sequencing support. We are grateful to members of the HCMC study teams for their support and efforts for enrolling patients and sample collections at three collaborative hospitals: Children's Hospital 1, Children's Hospital 2 and Hospital for Tropical Diseases (HCMC, Vietnam). This work was supported by the Wellcome Trust Strategic Award (VIZIONS – Vietnam Initiative on Zoonotic Infections) program and the European Community's Seventh Framework Programme

479 (FP7/2007_2013) under the project EMPERIE, European Community grant agreement number
480 223498.
481

482 **TABLES**

483 **Table 1. Primers used.**

Primer	Sequence	Strand	Position[1]	Tm[2]	GC_ fraction	Norovirus GII Genomes (517)[3] 0 MM[4]	Norovirus GII Genomes (517)[3] 0-3 MM[5]	Norovirus Genomes (753)[6] 0 MM[7]	Norovirus Genomes (753)[6] 0-3 MM[8]
UNP_47	GTGAATGAAGATGGCGTCTAAC	Plus	1	55.52	0.45	98	100	83	84
UNP_45	TCTAACGACGCTTCCGCTG	Plus	17	58.30	0.58	75	96	62	80
UNP_201R	GCAATGGCCACCTCCTCAT	Minus	2808	57.95	0.58	97	100	80	85
UNP_226R	TTGGCCTCCTCCTCTTCACA	Minus	2850	58.21	0.55	92	99	76	82
UNP_339	GGCAAGAAGCACACAGCC	Plus	2660	57.48	0.61	88	92	73	78
UNP_1316	TGGTCCAAGCCACAAGTGG	Plus	2519	58.05	0.58	11	100	13	89
UNP_82	GACCTCTGGGACGAGGTTG	Minus	5150	57.41	0.63	87	96	70	80
UNP_135	CTCCACCAGGGGCTTGTAC	Minus	5271	57.63	0.63	89	94	73	78
UNP_2	GGGAGGGCGATCGCAAT	Plus	5049	57.57	0.65	88	96	72	79
UNP_23	TTGTGAATGAAGATGGCGTCGA	Plus	5079	58.53	0.45	56	100	42	84
UNP_100	GCCAGTCCAGGAGTCCAA	Minus	7447	56.43	0.61	74	97	61	83
UNP_44	GCACGGTTGAGACTGTGC	Minus	7418	57.27	0.61	84	98	69	82
UNP_20	CGAGGGGAGTCACGGGT	Minus	7493	58.34	0.71	86	97	70	83

484

485 Footnotes.

486 1. Primer mapping position in norovirus GII4 norovirus genome JQ613552

487 3.Tm (melting temperature) calculated using a Python script that approximates the Breslauer
488 method(59).

489 3. All GenBank entries (July 2014) for Norovirus GII (TaxonID 142786, length 7000-8000 nt)=
490 517 entries.

491 4. Percentage of Norovirus GII genomes (n=517) showing perfect homology to primer.

492 5. Percentage of Norovirus GII genomes (n=517) showing the target sequence for the primer
493 with up to 3 mismatches.

494 6. All GenBank entries (July 2014) for Norovirus (TaxonID 122929, length 7000-8000 nt) = 753
495 entries.

496 7. Percentage of Norovirus genomes (n=753) showing perfect homology to primer.

497 5. Percentage of Norovirus GII genomes (n=753) showing the target sequence for the primer
498 with up to 3 mismatches.

499

500

501 Table 2. GenBank and ENA accession numbers, genetic cluster and samples dates.

502

Virus	GenBank_Acc_No[1]	ENA_Acc_No[2]	Cluster[3]	Sample_Date[4]
Hu_GII_10116_2009_VNM	KM198480	ERR212491	1	09-Jul-2009
Hu_GII_10054_2009_VNM	KM198481	ERR225641	1	21-May-2009
Hu_GII_10114_2009_VNM	KM198482	ERR212490	1	09-Jul-2009
Hu_GII_10313_2010_VNM	KM198483	ERR217285	4	22-Feb-2010
Hu_GII_30212_2009_VNM	KM198484	ERR223539	5	06-Oct-2009
Hu_GII_10148_2009_VNM	KM198485	ERR212498	2	11-Aug-2009
Hu_GII_C2H-18_2011_VNM	KM198486	ERR225628	3	30-Aug-2011
Hu_GII_10110_2009_VNM	KM198487	ERR212489	1	06-Jul-2009
Hu_GII_10325_2010_VNM	KM198488	ERR217290	4	26-Feb-2010
Hu_GII_10101_2009_VNM	KM198489	ERR212487	1	25-Jun-2009
Hu_GII_10002_2009_VNM	KM198490	ERR225635	1	04-May-2009
Hu_GII_30351_2009_VNM	KM198491	ERR138007	4	17-Dec-2009
Hu_GII_30448_2010_VNM	KM198492	ERR223547	6	29-Jan-2010
Hu_GII_30468_2010_VNM	KM198493	ERR223549	5	24-Feb-2010
Hu_GII_10247_2009_VNM	KM198494	ERR217278	1	10-Dec-2009
Hu_GII_10193_2009_VNM	KM198495	ERR138002	1	05-Oct-2009
Hu_GII_20419_2010_VNM	KM198496	ERR223554	5	01-Feb-2010
Hu_GII_10236_2009_VNM	KM198497	ERR217280	1	19-Nov-2009
Hu_GII_20088_2009_VNM	KM198498	ERR223553	8	28-Jul-2009
Hu_GII_20118_2009_VNM	KM198499	ERR212481	1	28-Aug-2009
Hu_GII_C2H-20_2011_VNM	KM198500	ERR225629	5	05-Sep-2011
Hu_GII_10173_2009_VNM	KM198501	ERR212503	1	11-Sep-2009
Hu_GII_10136_2009_VNM	KM198502	ERR212495	1	03-Aug-2009
Hu_GII_C2033_2010_VNM	KM198503	ERR212484	6	28-Jun-2010
Hu_GII_20151_2009_VNM	KM198504	ERR212467	1	16-Sep-2009
Hu_GII_20460_2010_VNM	KM198505	ERR223530	5	04-Mar-2010
Hu_GII_10199_2009_VNM	KM198506	ERR217283	1	20-Oct-2009
Hu_GII_C2007_2010_VNM	KM198507	ERR138011	4	02-Apr-2010
Hu_GII_20066_2009_VNM	KM198508	ERR212470	1	14-Jul-2009
Hu_GII_20479_2010_VNM	KM198509	ERR223531	5	16-Mar-2010
Hu_GII_10012_2009_VNM	KM198510	ERR225637	1	07-May-2009
Hu_GII_C2H-24_2011_VNM	KM198511	ERR225631	5	16-Sep-2011

Hu_GII_10062_2009_VNM	KM198512	ERR225642	1	29-May-2009
Hu_GII_20494_2010_VNM	KM198513	ERR223534	1	19-Mar-2010
Hu_GII_10079_2009_VNM	KM198514	ERR212486	1	11-Jun-2009
Hu_GII_C2H-31_2011_VNM	KM198515	ERR217269	3	28-Sep-2011
Hu_GII_10285_2010_VNM	KM198516	ERR217288	4	18-Jan-2010
Hu_GII_30399_2010_VNM	KM198517	ERR138008	1	11-Jan-2010
Hu_GII_10182_2009_VNM	KM198518	ERR212507	1	17-Sep-2009
Hu_GII_30082_2009_VNM	KM198519	ERR223537	8	23-Jun-2009
Hu_GII_10158_2009_VNM	KM198520	ERR212499	1	21-Aug-2009
Hu_GII_10176_2009_VNM	KM198521	ERR212504	1	14-Sep-2009
Hu_GII_20150_2009_VNM	KM198522	ERR212466	1	15-Sep-2009
Hu_GII_10204_2009_VNM	KM198523	ERR217287	1	29-Oct-2009
Hu_GII_10034_2009_VNM	KM198524	ERR225638	1	15-May-2009
Hu_GII_10163_2009_VNM	KM198525	ERR212501	1	28-Aug-2009
Hu_GII_10075_2009_VNM	KM198526	ERR225644	1	09-Jun-2009
Hu_GII_10074_2009_VNM	KM198527	ERR225643	1	09-Jun-2009
Hu_GII_C2H-25_2011_VNM	KM198528	ERR225632	5	20-Sep-2011
Hu_GII_C2H-27_2011_VNM	KM198529	ERR225633	5	21-Sep-2011
Hu_GII_20486_2010_VNM	KM198530	ERR223532	7	18-Mar-2010
Hu_GII_30116_2009_VNM	KM198531	ERR223538	7	09-Jul-2009
Hu_GII_10108_2009_VNM	KM198532	ERR212488	1	02-Jul-2009
Hu_GII_30241_2009_VNM	KM198533	ERR223540	1	26-Oct-2009
Hu_GII_30443_2010_VNM	KM198534	ERR223546	8	28-Jan-2010
Hu_GII_20092_2009_VNM	KM198535	ERR212474	1	31-Jul-2009
Hu_GII_20079_2009_VNM	KM198536	ERR212473	1	24-Jul-2009
Hu_GII_10137_2009_VNM	KM198537	ERR212496	1	04-Aug-2009
Hu_GII_10051_2009_VNM	KM198538	ERR225640	1	21-May-2009
Hu_GII_C2H-36_2011_VNM	KM198539	ERR217266	3	25-Oct-2011
Hu_GII_20145_2009_VNM	KM198540	ERR212464	1	11-Sep-2009
Hu_GII_20188_2009_VNM	KM198541	ERR138004	1	07-Oct-2009
Hu_GII_20094_2009_VNM	KM198542	ERR212476	2	03-Aug-2009
Hu_GII_20357_2009_VNM	KM198543	ERR138005	4	30-Dec-2009
Hu_GII_C2418_2010_VNM	KM198544	ERR138012	4	01-Nov-2010
Hu_GII_10195_2009_VNM	KM198545	ERR217284	1	13-Oct-2009
Hu_GII_20067_2009_VNM	KM198546	ERR212471	1	15-Jul-2009

Hu_GII_C2H-47_2011_VNM	KM198547	ERR217275	5	03-Nov-2011
Hu_GII_20107_2009_VNM	KM198548	ERR212477	1	20-Aug-2009
Hu_GII_30473_2010_VNM	KM198549	ERR223550	7	01-Mar-2010
Hu_GII_10078_2009_VNM	KM198550	ERR225645	1	10-Jun-2009
Hu_GII_20108_2009_VNM	KM198551	ERR212478	1	20-Aug-2009
Hu_GII_20154_2009_VNM	KM198552	ERR212469	1	16-Sep-2009
Hu_GII_30381_2010_VNM	KM198553	ERR223544	5	04-Jan-2010
Hu_GII_C2H-48_2011_VNM	KM198554	ERR217271	5	04-Nov-2011
Hu_GII_C2035_2010_VNM	KM198555	ERR138006	4	28-Jun-2010
Hu_GII_10127_2009_VNM	KM198556	ERR212492	1	17-Jul-2009
Hu_GII_10194_2009_VNM	KM198557	ERR217286	1	13-Oct-2009
Hu_GII_30257_2009_VNM	KM198558	ERR223541	1	30-Oct-2009
Hu_GII_10129_2009_VNM	KM198559	ERR212493	1	20-Jul-2009
Hu_GII_C2H-50_2011_VNM	KM198560	ERR217268	3	22-Nov-2011
Hu_GII_30303_2009_VNM	KM198561	ERR223542	5	23-Nov-2009
Hu_GII_C2H-55_2011_VNM	KM198562	ERR212509	3	25-Nov-2011
Hu_GII_C2365_2010_VNM	KM198563	ERR212485	5	15-Sep-2010
Hu_GII_C2H-44_2011_VNM	KM198564	ERR217276	3	31-Oct-2011
Hu_GII_10169_2009_VNM	KM198565	ERR212502	1	04-Sep-2009
Hu_GII_20093_2009_VNM	KM198566	ERR212475	1	03-Aug-2009
Hu_GII_10255_2009_VNM	KM198567	ERR217273	1	15-Dec-2009
Hu_GII_C2H-62_2011_VNM	KM198568	ERR217267	3	14-Dec-2011
Hu_GII_10235_2009_VNM	KM198569	ERR217274	1	19-Nov-2009
Hu_GII_20146_2009_VNM	KM198570	ERR212465	1	11-Sep-2009
Hu_GII_20123_2009_VNM	KM198571	ERR212479	1	01-Sep-2009
Hu_GII_20370_2010_VNM	KM198572	ERR212461	5	06-Jan-2010
Hu_GII_C2H-45_2011_VNM	KM198573	ERR217277	5	02-Nov-2011
Hu_GII_10183_2009_VNM	KM198574	ERR217289	1	22-Sep-2009
Hu_GII_20069_2009_VNM	KM198575	ERR212472	1	16-Jul-2009
Hu_GII_C2H-43_2011_VNM	KM198576	ERR217281	3	14-Oct-2011
Hu_GII_10145_2009_VNM	KM198577	ERR212497	1	07-Aug-2009
Hu_GII_C2H-52_2011_VNM	KM198578	ERR212508	3	24-Nov-2011
Hu_GII_10160_2009_VNM	KM198579	ERR212500	1	26-Aug-2009
Hu_GII_10223_2009_VNM	KM198580	ERR217272	1	06-Nov-2009
Hu_GII_10003_2009_VNM	KM198581	ERR225636	1	05-May-2009

Hu_GII_30192_2010_VNM	KM198582	ERR138003	1	21-Sep-2009
Hu_GII_20493_2010_VNM	KM198583	ERR223533	5	19-Mar-2010
Hu_GII_10131_2009_VNM	KM198584	ERR212494	1	22-Jul-2009
Hu_GII_10238_2009_VNM	KM198585	ERR217282	1	19-Nov-2009
Hu_GII_30400_2010_VNM	KM198586	ERR223545	5	11-Jan-2010
Hu_GII_20153_2009_VNM	KM198587	ERR212468	1	16-Sep-2009
Hu_GII_10037_2009_VNM	KM198588	ERR225639	1	15-May-2009
Hu_GII_20144_2009_VNM	KM198589	ERR212463	1	10-Sep-2009
Hu_GII_C2H-39_2011_VNM	KM198590	ERR217270	5	24-Oct-2011
Hu_GII_20122_2009_VNM	KM198591	ERR212480	1	31-Aug-2009

503 Footnotes.

504 1. GenBank accession number, accessible at http://www.ncbi.nlm.nih.gov/nuccore/

505 2. European Nucleotide Archive accession number, accessible at http://www.ebi.ac.uk/ena/

506 3. Genetic cluster as defined in Figure 2.

507 4. Date of sample collection.

508

509 **Table 3. PCR and genome sequencing success by norovirus genotypes.**

Genotype[1]	Samples	Amplicon1[2]	Amplicon2	Amplicon3	Genomes[3]	(%) Successful
GII.4	60	55	55	57	55	92
GII (non GII.4)	58	48	45	53	45	74
GI	10	7	4	4	2	20
GII.2	5	4	1	5	2	40
GII.3	34	26	27	27	26	77
GII.6	8	8	7	8	7	88
GII.7	2	2	0	2	0	0
GII.9	1	1	0	1	1	100
GII.12	2	1	2	2	1	50
GII.13	6	5	6	6	5	83

510 Footnotes.

511 1. Sample genotype previously determined (30, 31).

512 2. Successful reverse transcription/ PCR amplification of sufficient DNA (ca. 0.4 µg) for

513 Illumina library preparation.

514 3. Yield of >95% full genome.

515

516 Table 4. **Summary of phylogenetic clusters identified in this study.**

Phylogenetic cluster[1]	Closest genome[2]	Genotype by RIVM algorithm[3]	Number of genomes	Frequency in 201 genomes
1	NV_GII_VNM_2009_KC175360 NV_GII_VNM_2009_KC175395	GII.P4.DH06b_GII.4.DH06b	140	69.65 (67.6)[4]
2	NV_GII_VNM_2009_KC175402	GII.P4.DH06b_GII.4.DH06b	2	
3	NV_GII4_TW_2007_JN400615 NV_GII4_Ehime_2007_AB541241	GII.P4.DH06b_GII.4.DH06b	12	
4	NV_GII_VNM_2010_KC175383	GII.P4.NO09_GII.4.NO09	20	9.95 (9.5%)
5	NV_Pune_2007_EU921389	GII.P21_GII.3	19	9.45 (10.2%)
6	NV_Pune_2007_EU921389 NV_GII2_12_Wahroonga_2009_JQ613568	GII.Pg_GII.12	2	0.99 (0.6%)
7	NV_GII_Gifu_1999_AB084071 (<50%)	GII.P7_GII.6	3	1.49 (2.5%)
8	NV_GII_Gifu_1999_AB084071 (<50%)	GII.P7_GII.6	3	1.49 (2.5%)

517 Footnotes

518 1. Phylogenetic classification (see Figure 2A).

519 2. Based on number of reads mapped.

520 3. Based on the RIVM algorithm(47).

521 4. Values in parentheses are the genotype frequency values determined by My *et al.* (31).

522

523 **Table 5. Evolutionary rates**

Sequence sets	Genomic region	Mean rate[1]	95% HPD[1]	Substitution Model
Cluster 1: GII.P4 Den Haag 2006b_GII.4 Den Haag 2006b	Whole genome	6.15	5.39 - 6.86	SRD06
	ORF1	5.94	5.04 - 6.94	SRD06
	ORF2	5.69	4.54 - 6.90	SRD06
	ORF3	8.99	6.59 - 11.6	SRD06
	p48 (NS1-2)	6.60	4.83 - 8.47	GTR+G
	NTPase (NS3)	5.41	4.04 - 6.89	GTR+G
	p22 (NS4)	8.21	5.48 - 11.11	GTR+G
	VPg (NS5)	5.95	3.27 - 8.94	HKY+G
	3CLpro (NS6)	6.03	3.57 - 8.61	GTR+G
	RdRp (NS7)	4.74	3.50 - 6.10	GTR+G
Cluster 4: GII.P4 New Orleans 2009_GII.4 New Orleans 2009	Whole genome	5.73	3.74 - 7.81	GTR+G
	ORF1	4.03	1.77 - 6.33	HKY+G
	ORF2	5.60	0.68 - 9.82	HKY+G
	ORF3[1]			
Cluster 5: GII.P21_GII.3	Whole genome	5.34	4.06 - 6.82	SRD06
	ORF1	4.81	3.45 - 6.17	SRD06
	ORF2	5.99	3.75 - 8.39	SRD06
	ORF3	7.38	2.06 - 13.9	SRD06

Footnotes.

1. Evolutionary rates were measured as $x10^{-3}$ substitutions per site per year. The mean evolutionary rate ($x10^{-3}$ substitutions per site per year) and the 95% upper and lower highest posterior density (HPD) were determined as described in the Materials and Methods section.

2. There was insufficient signal for the algorithms to return a reliable evolutionary rate for ORF3 region of sequences from GII.4 Cluster 4.

532 **Table 6. Positive selection analysis**

Codon position[1]	FUBAR[2]	MEME[3]
ORF1 (p48 (NS1-2)) 79	0.991	0.037
ORF2 (VP1) 106	0.993	0.014
ORF2 (VP1) 298	0.984	>0.05
ORF3 (VP2) 144[4]	0.983	0.043

533 Footnotes.

534 1. Codons under positive selection in 140 Cluster 1 norovirus genomes sequenced in this

535 study as detected using FUBAR(44) and MEME(43).

536 2.The values of the posterior probability (FUBAR) or the *p*-value (MEME) are indicated.

537 3. Analysis for ORF3 covered the first 247 of the protein's 268 codons.

538
539

540

541

542

FIGURE LEGENDS

Figure 1. Primer design and function for full genome deep sequencing amplification. **Left panel:** Virtual PCR showing the mapping of the designed primers to a norovirus GII.4 genome (GenBank JQ613552). Colored circles indicate the position of the each primer and number of mismatches, grey bars indicate the predicted size of the PCR products. A schematic of the open reading frame organization of the virus is shown at the tope of the panel. **Right panel:** the PCR products from 14 samples for individual primer pairs for Amplicon 1, 2 and 3. Sample 7 failed to amplify, **c**: water control, **m**: size marker. The size of relevant marker bands (in kilobasepairs) is indicated to the right of the image.

Figure 2. Maximum likelihood phylogenetic tree of all HCMC GII genomes from this study (112) and 89 GII.4 genomes from the same HCMC cohort that were sequenced separately and made publically available in GenBank, plus selected global reference genomes. The eight phylogenetic clusters of norovirus identified in this study are marked with colored bars. Bootstrap support ≥ 0.85 at key nodes is indicated with asterisks. The tree is midpoint rooted for purposes of clarity, and all horizontal branch lengths are drawn to a scale of nucleotide substitutions per site.

Figure 3. The temporal appearance of the HCMC norovirus GII genotypes during the study period. Genomes were stratified by genotype (from Figure 2), color-coded and plotted by date of sample isolation.

Figure 4. Summary of evolutionary rates inferred for the genomic regions GII.4 Cluster 1. Evolutionary rates were estimated as described in the Methods section, mean values are indicated by colored circles with the 95% credible interval (CI) indicated. The region of the norovirus genome used for calculation is labelled, and the two regions with rates elevated relative to full genome are marked in red.

Figure 5. Changes in the GII.4 Cluster 1 VP1 protein. The protein sequences were aligned, and amino acid differences from the reconstructed ancestral sequence of Cluster 1 were determined and marked with vertical colored bars, with the new amino acid residue color-coded as shown in the bottom panel; gray bars indicate a gap in the query sequence. The sequences were ordered by sample date with earliest samples at the bottom of the graph. Functional domains of the VP1 protein are indicated at the top of the graph and include the shell domain the protruding 1 (P1) and protruding 2 (P2) domains. The locations of the blockade epitopes A-E are also indicated (B_A - B_E). A histogram in the second panel indicates total changes at each position. The protein changes occurring in more than four

samples are annotated with the parental amino acid, the position and the new amino acid. Codons found to be under positive selection by MEME or FUBAR are indicated with a red asterisk.

Figure 6. Changes in the Cluster 1 minor capsid protein VP2 over time. Protein changes were analyzed and depicted as described for Figure 5. Functional domains of the VP2 protein including the VP1 interacting region (VP1-ID) are marked at the top of the graph. A histogram in the second panel indicates total changes at each position. Codons found to be under positive selection by MEME or FUBAR are indicated with a red asterisk.

Figure 7. Changes in the Cluster 1 p22 proteins over time. Protein changes were analyzed and depicted as described for Figure 5. Functional domains of the p22 protein including the MERES domain are marked at the top of the graph. A histogram in the second panel indicates total changes at each position.

Figure 8. Changes in the Cluster 1 p48 proteins over time. Protein changes were analyzed and depicted as described for Figure 5. Functional domains of the p48 protein including the transmembrane domain are marked at the top of the graph. A histogram in the second panel indicates total changes at each position. Codons found to be under positive selection by MEME or FUBAR are indicated with a red asterisk.

REFERENCES

1. **Thorne LG, Goodfellow IG.** 2014. Norovirus gene expression and replication. J Gen Virol **95:**278-291.
2. **McFadden N, Bailey D, Carrara G, Benson A, Chaudhry Y, Shortland A, Heeney J, Yarovinsky F, Simmonds P, Macdonald A, Goodfellow I.** 2011. Norovirus regulation of the innate immune response and apoptosis occurs via the product of the alternative open reading frame 4. PLoS pathogens **7:**e1002413.
3. **Sosnovtsev SV, Belliot G, Chang KO, Prikhodko VG, Thackray LB, Wobus CE, Karst SM, Virgin HW, Green KY.** 2006. Cleavage map and proteolytic processing of the murine norovirus nonstructural polyprotein in infected cells. Journal of virology **80:**7816-7831.
4. **Wolf S, Williamson W, Hewitt J, Lin S, Rivera-Aban M, Ball A, Scholes P, Savill M, Greening GE.** 2009. Molecular detection of norovirus in sheep and pigs in New Zealand farms. Veterinary microbiology **133:**184-189.

5. **Wang QH, Han MG, Cheetham S, Souza M, Funk JA, Saif LJ.** 2005. Porcine noroviruses related to human noroviruses. Emerging infectious diseases **11:**1874-1881.

6. **Wang QH, Chang KO, Han MG, Sreevatsan S, Saif LJ.** 2006. Development of a new microwell hybridization assay and an internal control RNA for the detection of porcine noroviruses and sapoviruses by reverse transcription-PCR. J Virol Methods **132:**135-145.

7. **Sugieda M, Nakajima S.** 2002. Viruses detected in the caecum contents of healthy pigs representing a new genetic cluster in genogroup II of the genus "Norwalk-like viruses". Virus research **87:**165-172.

8. **Oliver SL, Dastjerdi AM, Wong S, El-Attar L, Gallimore C, Brown DW, Green J, Bridger JC.** 2003. Molecular characterization of bovine enteric caliciviruses: a distinct third genogroup of noroviruses (Norwalk-like viruses) unlikely to be of risk to humans. Journal of virology **77:**2789-2798.

9. **Martella V, Lorusso E, Decaro N, Elia G, Radogna A, D'Abramo M, Desario C, Cavalli A, Corrente M, Camero M, Germinario CA, Banyai K, Di Martino B, Marsilio F, Carmichael LE, Buonavoglia C.** 2008. Detection and molecular characterization of a canine norovirus. Emerging infectious diseases **14:**1306-1308.

10. **Martella V, Campolo M, Lorusso E, Cavicchio P, Camero M, Bellacicco AL, Decaro N, Elia G, Greco G, Corrente M, Desario C, Arista S, Banyai K, Koopmans M, Buonavoglia C.** 2007. Norovirus in captive lion cub (Panthera leo). Emerging infectious diseases **13:**1071-1073.

11. **Liu BL, Lambden PR, Gunther H, Otto P, Elschner M, Clarke IN.** 1999. Molecular characterization of a bovine enteric calicivirus: relationship to the Norwalk-like viruses. Journal of virology **73:**819-825.

12. **Hsu CC, Riley LK, Livingston RS.** 2007. Molecular characterization of three novel murine noroviruses. Virus Genes **34:**147-155.

13. **Teunis PF, Moe CL, Liu P, Miller SE, Lindesmith L, Baric RS, Le Pendu J, Calderon RL.** 2008. Norwalk virus: how infectious is it? J Med Virol **80:**1468-1476.

14. **Atmar RL, Opekun AR, Gilger MA, Estes MK, Crawford SE, Neill FH, Ramani S, Hill H, Ferreira J, Graham DY.** 2013. Determination of the 50% Human Infectious Dose for Norwalk Virus. The Journal of infectious diseases.

15. **Simmons K, Gambhir M, Leon J, Lopman B.** 2013. Duration of immunity to norovirus gastroenteritis. Emerging infectious diseases **19:**1260-1267.

16. **Debbink K, Lindesmith LC, Donaldson EF, Baric RS.** 2012. Norovirus immunity and the great escape. PLoS pathogens **8:**e1002921.

17. **Sukhrie FH, Teunis P, Vennema H, Copra C, Thijs Beersma MF, Bogerman J, Koopmans M.** 2012. Nosocomial transmission of norovirus is mainly caused by symptomatic

651 cases. Clinical infectious diseases : an official publication of the Infectious Diseases Society of
652 America **54:**931-937.
653 18. **Siebenga JJ, Beersma MF, Vennema H, van Biezen P, Hartwig NJ, Koopmans M.**
654 2008. High prevalence of prolonged norovirus shedding and illness among hospitalized
655 patients: a model for in vivo molecular evolution. The Journal of infectious diseases **198:**994-
656 1001.
657 19. **Verhoef L, Williams KP, Kroneman A, Sobral B, van Pelt W, Koopmans M.** 2012.
658 Selection of a phylogenetically informative region of the norovirus genome for outbreak
659 linkage. Virus Genes **44:**8-18.
660 20. **Sukhrie FH, Teunis P, Vennema H, Bogerman J, van Marm S, Beersma MF,**
661 **Koopmans M.** 2013. P2 domain profiles and shedding dynamics in prospectively monitored
662 norovirus outbreaks. Journal of clinical virology : the official publication of the Pan American
663 Society for Clinical Virology **56:**286-292.
664 21. **Bull RA, Eden JS, Luciani F, McElroy K, Rawlinson WD, White PA.** 2012.
665 Contribution of intra- and interhost dynamics to norovirus evolution. Journal of virology
666 **86:**3219-3229.
667 22. **Kundu S, Lockwood J, Depledge DP, Chaudhry Y, Aston A, Rao K, Hartley JC,**
668 **Goodfellow I, Breuer J.** 2013. Next-generation whole genome sequencing identifies the
669 direction of norovirus transmission in linked patients. Clinical infectious diseases : an official
670 publication of the Infectious Diseases Society of America **57:**407-414.
671 23. **Lindesmith L, Moe C, Lependu J, Frelinger JA, Treanor J, Baric RS.** 2005. Cellular
672 and humoral immunity following Snow Mountain virus challenge. Journal of virology **79:**2900-
673 2909.
674 24. **Cannon JL, Lindesmith LC, Donaldson EF, Saxe L, Baric RS, Vinje J.** 2009. Herd
675 immunity to GII.4 noroviruses is supported by outbreak patient sera. Journal of virology
676 **83:**5363-5374.
677 25. **Donaldson EF, Lindesmith LC, Lobue AD, Baric RS.** 2008. Norovirus pathogenesis:
678 mechanisms of persistence and immune evasion in human populations. Immunol Rev
679 **225:**190-211.
680 26. **Lindesmith LC, Beltramello M, Donaldson EF, Corti D, Swanstrom J, Debbink K,**
681 **Lanzavecchia A, Baric RS.** 2012. Immunogenetic mechanisms driving norovirus GII.4
682 antigenic variation. PLoS pathogens **8:**e1002705.
683 27. **LoBue AD, Lindesmith L, Yount B, Harrington PR, Thompson JM, Johnston RE,**
684 **Moe CL, Baric RS.** 2006. Multivalent norovirus vaccines induce strong mucosal and systemic
685 blocking antibodies against multiple strains. Vaccine **24:**5220-5234.

686 28. **Lindesmith LC, Donaldson E, Leon J, Moe CL, Frelinger JA, Johnston RE, Weber**
687 **DJ, Baric RS.** 2010. Heterotypic humoral and cellular immune responses following Norwalk
688 virus infection. Journal of virology **84:**1800-1815.
689 29. **Reeck A, Kavanagh O, Estes MK, Opekun AR, Gilger MA, Graham DY, Atmar RL.**
690 2010. Serological correlate of protection against norovirus-induced gastroenteritis. The Journal
691 of infectious diseases **202:**1212-1218.
692 30. **My PV, Thompson C, Phuc HL, Tuyet PT, Vinh H, Hoang NV, Minh PV, Vinh NT,**
693 **Thuy CT, Nga TT, Hau NT, Campbell J, Chinh NT, Thuong TC, Tuan HM, Farrar J, Baker**
694 **S.** 2013. Endemic norovirus infections in children, Ho Chi Minh City, Vietnam, 2009-2010.
695 Emerging infectious diseases **19:**977-980.
696 31. **Tra My PV, Lam HM, Thompson CN, Phuc HL, Tuyet PT, Vinh H, Hoang NV, Minh**
697 **P, Vinh NT, Thuy CT, Nga TT, Hau NT, Chinh NT, Thuong TC, Tuan HM, Campbell JI,**
698 **Clements AC, Farrar J, Boni MF, Baker S.** 2013. The dynamics of GII.4 Norovirus in Ho Chi
699 Minh City, Vietnam. Infect Genet Evol **18:**335-343.
700 32. **Cotten M, Lam TT, Watson SJ, Palser AL, Petrova V, Grant P, Pybus OG, Rambaut**
701 **A, Guan Y, Pillay D, Kellam P, Nastouli E.** 2013. Full-genome deep sequencing and
702 phylogenetic analysis of novel human betacoronavirus. Emerging infectious diseases **19:**736-
703 742B.
704 33. **Quail MA, Otto TD, Gu Y, Harris SR, Skelly TF, McQuillan JA, Swerdlow HP, Oyola**
705 **SO.** 2012. Optimal enzymes for amplifying sequencing libraries. Nature methods **9:**10-11.
706 34. **Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H,**
707 **Turner DJ.** 2008. A large genome center's improvements to the Illumina sequencing system.
708 Nature methods **5:**1005-1010.
709 35. **Watson SJ, Welkers MR, Depledge DP, Coulter E, Breuer JM, de Jong MD, Kellam**
710 **P.** 2013. Viral population analysis and minority-variant detection using short read next-
711 generation sequencing. Philos Trans R Soc Lond B Biol Sci **368:**20120205.
712 36. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM,**
713 **Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G,**
714 **Alekseyev MA, Pevzner PA.** 2012. SPAdes: a new genome assembly algorithm and its
715 applications to single-cell sequencing. J Comput Biol **19:**455-477.
716 37. **Martin D, Rybicki E.** 2000. RDP: detection of recombination amongst aligned
717 sequences. Bioinformatics **16:**562-563.
718 38. **Dubot-Peres A, Vongphrachanh P, Denny J, Phetsouvanh R, Linthavong S,**
719 **Sengkeopraseuth B, Khasing A, Xaythideth V, Moore CE, Vongsouvath M, Castonguay-**
720 **Vanier J, Sibounheuang B, Taojaikong T, Chanthongthip A, de Lamballerie X, Newton**

721 **PN.** 2013. An epidemic of dengue-1 in a remote village in rural Laos. PLoS Negl Trop Dis
722 **7:**e2360.
723 39. **Stamatakis A.** 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic
724 analyses with thousands of taxa and mixed models. Bioinformatics **22:**2688-2690.
725 40. **Darriba D, Taboada GL, Doallo R, Posada D.** 2012. jModelTest 2: more models, new
726 heuristics and parallel computing. Nature methods **9:**772.
727 41. **Drummond AJ, Suchard MA, Xie D, Rambaut A.** 2012. Bayesian phylogenetics with
728 BEAUti and the BEAST 1.7. Mol Biol Evol **29:**1969-1973.
729 42. **Minin VN, Bloomquist EW, Suchard MA.** 2008. Smooth skyride through a rough
730 skyline: Bayesian coalescent-based inference of population dynamics. Mol Biol Evol **25:**1459-
731 1471.
732 43. **Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL.**
733 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet
734 **8:**e1002764.
735 44. **Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL,**
736 **Scheffler K.** 2013. FUBAR: a fast, unconstrained bayesian approximation for inferring
737 selection. Mol Biol Evol **30:**1196-1205.
738 45. **Delport W, Poon AF, Frost SD, Kosakovsky Pond SL.** 2010. Datamonkey 2010: a
739 suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics **26:**2455-2457.
740 46. **Pond SLK, Frost SD, Muse SV.** 2005. HyPhy: hypothesis testing using phylogenies.
741 BIOINFORMATICS **21:**676-679.
742 47. **Kroneman A, Vennema H, Deforche K, v d Avoort H, Penaranda S, Oberste MS,**
743 **Vinje J, Koopmans M.** 2011. An automated genotyping tool for enteroviruses and
744 noroviruses. Journal of clinical virology : the official publication of the Pan American Society for
745 Clinical Virology **51:**121-125.
746 48. **Chan MC, Lee N, Ho WS, Law CO, Lau TC, Tsui SK, Sung JJ.** 2012. Covariation of
747 major and minor viral capsid proteins in norovirus genogroup II genotype 4 strains. Journal of
748 virology **86:**1227-1232.
749 49. **Siebenga JJ, Lemey P, Kosakovsky Pond SL, Rambaut A, Vennema H, Koopmans**
750 **M.** 2010. Phylodynamic reconstruction reveals norovirus GII.4 epidemic expansions and their
751 molecular determinants. PLoS pathogens **6:**e1000884.
752 50. **Bok K, Abente EJ, Realpe-Quintero M, Mitra T, Sosnovtsev SV, Kapikian AZ,**
753 **Green KY.** 2009. Evolutionary dynamics of GII.4 noroviruses over a 34-year period. Journal of
754 virology **83:**11890-11901.

755 51. **Boon D, Mahar JE, Abente EJ, Kirkwood CD, Purcell RH, Kapikian AZ, Green KY,**
756 **Bok K.** 2011. Comparative evolution of GII.3 and GII.4 norovirus over a 31-year period.
757 Journal of virology **85:**8656-8666.
758 52. **Vongpunsawad S, Venkataram Prasad BV, Estes MK.** 2013. Norwalk Virus Minor
759 Capsid Protein VP2 Associates within the VP1 Shell Domain. Journal of virology **87:**4818-
760 4825.
761 53. **Zhu S, Regev D, Watanabe M, Hickman D, Moussatche N, Jesus DM, Kahan SM,**
762 **Napthine S, Brierley I, Hunter RN, 3rd, Devabhaktuni D, Jones MK, Karst SM.** 2013.
763 Identification of immune and viral correlates of norovirus protective immunity through
764 comparative study of intra-cluster norovirus strains. PLoS pathogens **9:**e1003592.
765 54. **de Rougemont A, Ruvoen-Clouet N, Simon B, Estienney M, Elie-Caille C, Aho S,**
766 **Pothier P, Le Pendu J, Boireau W, Belliot G.** 2011. Qualitative and quantitative analysis of
767 the binding of GII.4 norovirus variants onto human blood group antigens. Journal of virology
768 **85:**4057-4070.
769 55. **Debbink K, Donaldson EF, Lindesmith LC, Baric RS.** 2012. Genetic mapping of a
770 highly variable norovirus GII.4 blockade epitope: potential role in escape from human herd
771 immunity. Journal of virology **86:**1214-1226.
772 56. **Sharp TM, Crawford SE, Ajami NJ, Neill FH, Atmar RL, Katayama K, Utama B,**
773 **Estes MK.** 2012. Secretory pathway antagonism by calicivirus homologues of Norwalk virus
774 nonstructural protein p22 is restricted to noroviruses. Virol J **9:**181.
775 57. **Sharp TM, Guix S, Katayama K, Crawford SE, Estes MK.** 2010. Inhibition of cellular
776 protein secretion by norwalk virus nonstructural protein p22 requires a mimic of an
777 endoplasmic reticulum export signal. PloS one **5:**e13130.
778 58. **Ettayebi K, Hardy ME.** 2003. Norwalk virus nonstructural protein p48 forms a complex
779 with the SNARE regulator VAP-A and prevents cell surface expression of vesicular stomatitis
780 virus G protein. Journal of virology **77:**11790-11797.
781 59. **Breslauer KJ, Frank R, Blocker H, Marky LA.** 1986. Predicting DNA duplex stability
782 from the base sequence. Proc Natl Acad Sci U S A **83:**3746-3750.

783
784