The recombinant origin of emerging human norovirus GI.4/2008: intra-genotypic exchange of the capsid P2 domain

Tommy Tsan-Yuk Lam,1,2,3 Huachen Zhu,2,3 David K. Smith,2,3 Yi Guan,2,3 Edward C. Holmes4,5 and Oliver G. Pybus1

Correspondence Tommy Tsan-Yuk Lam tylam.tommy@gmail.com

1Department of Zoology, University of Oxford, Oxford, UK
2International Institute of Infection and Immunity, Shantou University Medical College, Shantou, Guangdong, PR China
3State Key Laboratory of Emerging Infectious Diseases, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR
4Center for Infectious Disease Dynamics, Department of Biology, The Pennsylvania State University, University Park, PA, USA
5Fogarty International Center, National Institutes of Health, Bethesda, MD, USA

GI.4 noroviruses are a major cause of acute gastroenteritis in humans. A new variant of GI.4, the 2008 variant, has recently increased its prevalence on a global scale. A previous study of this variant in Japan suggested that it might be of recombinant origin, with a breakpoint at the ORF1–ORF2 junction. Here, examination of the evolutionary origin of the 2008 variant based on a larger sample of worldwide GI.4 norovirus sequences revealed a more complex pattern of recombination between the 2006a- and 2006b-like variants of genotype GI.4, involving the P2 antigenic domain. Double (termed ‘2008i’) and triple (termed ‘2008ii’) recombinant forms of 2008 variants were identified. This study highlights the possible importance of intra-genotypic recombination over antigenic regions in driving norovirus evolution, and is suggestive of a process analogous to the antigenic shift of influenza A virus by reassortment.

Noroviruses, a major aetiological agent of gastroenteritis in humans (Atmar & Estes, 2006), are classified into five genogroups (GI–GV), which are in turn divided into a number of genotypes (Zheng et al., 2006). Surveillance since the 1990s has confirmed that genotype 4 of the GII genogroup (denoted GI.4) dominates most worldwide outbreaks of norovirus-related gastroenteritis (Jin et al., 2008; Kroneman et al., 2008; Nayak et al., 2009; Park et al., 2010; Siebenga et al., 2009).

Over the past 20 years GI.4 norovirus has evolved a series of genetic variants, some of which persisted and replaced the previously circulating variants (Reuter et al., 2008; Siebenga et al., 2010). The continuous evolution of the viral capsid VP1 protein has been proposed as a key mechanism by which new antigenic variants are generated. This is supported by observations of temporal changes in histoblood group antigen-binding characteristics, antigenic relatedness and genome composition among GI.4 norovirus variants (Lindesmith et al., 2008).

From 2006 to 2007 onwards the GI.4 ‘2006b variant’ dominated GI.4 norovirus outbreaks worldwide (Chan-It et al., 2011; Chung et al., 2010; Eden et al., 2010; Jin et al., 2008; Kittigul et al., 2010). However, the ‘2008 variant’ has been reported increasingly frequently since its first isolation in 2008 (Belliot et al., 2010; Motomura et al., 2010) and, at the time of writing, has been identified in 12 countries on four continents (Belliot et al., 2010; Eden et al., 2010; Han et al., 2011; Mans et al., 2010; Motomura et al., 2010; Pang et al., 2010; Schenk et al., 2010). Previous phylogenetic analysis showed that the 2008 variant emerged from the Hunter-2006a ancestral lineage, with which it shares 88 % nucleotide similarity (Eden et al., 2010). However, the phylogenetic position of the 2008 variant differed according to the genome region investigated (Belliot et al., 2010; Eden et al., 2010; Han et al., 2011; Mans et al., 2010; Motomura et al., 2010), with a recombination breakpoint identified at the ORF1–ORF2 junction of the Japanese 2008 variants (Motomura et al., 2010).

To better understand the role of recombination in the evolutionary origin of the 2008 variant, partial and complete genome sequences of GI.4 noroviruses from
GenBank were examined. After alignment using MUSCLE v3.8 (Edgar, 2004), a maximum-likelihood (ML) phylogeny of GII.4 ORF2 sequences (1620 nt; n=600) was estimated using PhyML v3.0 (Guindon & Gascuel, 2003). Topological robustness was assessed by performing ML bootstrap analysis (1000 pseudoreplicates; using PhyML) and Bayesian phylogeny sampling (5000 samplings, each at every 2000 states; using MrBayes 3.1.2; Ronquist & Huelsenbeck, 2003). This revealed a distinct monophyletic lineage of variant strains (n=35; Fig. 1) that was previously

![Fig. 1. ML phylogeny of ORF2 of GII.4 norovirus (n=600). (a) ML tree of ORF2 of GII.4. Variant lineages are compressed for clarity. They are Japan01 (Kaiso/030556/2003/JP), Grimsby (Grimsby), Cairo (Cairo2/2006/EGY), Hunter (2004/NL), 2006a (Yerseke3B/2006/NL), Henry (Houston/TCH186/2002/US), Farmington-Hills (Langen1061/2002/DE), Asia03 (Sakai/04-179/2005/JP), 2008 (JB-15/KOR/2008, Osaka1/2008/JP, Orange/NBW-001P/2008/AU), 2006b (DenHaag89/2006/NL) and 2008b (Hokkaido4/2008/JP). Representative strains (in parentheses), mostly from Siebenga et al. (2009) are indicated by black circles. Bootstrap support and clade posterior probabilities are shown on the left and right side of the slashes. (b) The lineage of 2008 variants extracted from the larger tree. Putative 2008i and 2008ii variants are indicated in orange and yellow, respectively. Virus strains for which complete genome sequences are available for analysis are indicated with asterisks. A 2008-derived recombinant is indicated by a hash. Branch lengths are in unit of substitutions per site.](image-url)
defined as ‘2008 variant’ (Belliot et al., 2010; Mans et al., 2010), and was also known as ‘2008a variant’ according to Motomura et al. (2010). Other known GII.4 variants are also identified in Fig. 1 (colour bars).

Systematic screening for recombination was undertaken using the suite of statistical tests implemented in RDP v3.41 (Martin & Rybicki, 2000). Sequences of 2008 variant viruses (1620 nt; n = 35; Supplementary Table S1, available in JGV Online) consistently exhibited a mosaic fragment about 300–500 nt long in the ORF2 P2 domain (statistically significant in $\geq$2 RDP tests, after Bonferroni correction). Results suggest that the 2006b and 2006a variants were the ancestors of the small and large ORF2 fragments, respectively. Next, we studied the complete genomes of 2008 variants (7509 bp; n = 11; Supplementary Table S1) that revealed similar breakpoints in the P2 region. However, all 2008 variants (except Orange/NSW001P/2008/AU, denoted ‘NSW001P’) also exhibited an additional breakpoint near the ORF1–ORF2 junction (statistically significant in $\geq$4 RDP tests, after Bonferroni correction). Further analysis using similarity plots in Simplot v3.5.1 (Lole et al., 1999) gave concordant results (Supplementary Fig. S1a, b, available in JGV Online).

Using the best recombination model selected by the phylogenetic GARD method (Kosakovsky Pond et al., 2006), the recombination breakpoints were estimated to be at nucleotide positions 5398, 5866 (5932 for NSW001P) and 6477 (Supplementary Fig. S1a, b; right axis).

The estimated breakpoints thus define three non-recombinant genomic regions of the 2008 variant: region 1 (1–5397 nt), region 2 (5398–5898 nt plus 6478–7509 nt) and region 3 (5899–6477 nt) (Fig. 2). Separate phylogenies of these three regions were estimated using both ML (in PhyML) and Bayesian (in MrBayes) methods (Fig. 2). The GTR+$I_4$ substitution model was used and 1000 bootstrap replicates were computed. The MrBayes analysis was performed using $10^7$ steps, sampled every 2000. In the phylogeny of region 1 (mainly ORF1; Fig. 2 left), NSW001P clustered with the 2006a variants, while the other 2008 variant viruses (n = 10; Supplementary Table S1) clustered with the Hunter variant. For region 2, all 2008 variants clustered with 2006a variants (Fig. 2 centre), whereas in region 3 they clustered with the 2006b variants (Fig. 2 right).

Most groupings are supported by high bootstrap scores and Bayesian posterior probabilities (Fig. 2). It is noteworthy that the 2008 variant group diverged from a node [indicated by clade support values 97/1 (53/.84) in the rightmost tree of Fig. 2] that is genetically distant (0.129
Phylogenetic-based analyses of recombination may be affected by convergent evolution leading to similar sets of amino acid substitutions in independent lineages. To examine such effects, we aligned the consensus sequences of different GII.4 variants, and removed the codon positions that are non-synonymous across variant lineages from the alignment. ML and Bayesian phylogenies were then estimated from these synonymous-site sequences (4710, 1250 and 412 bp in length for the regions 1, 2 and 3, respectively). The resultant tree topologies are largely congruent with regions 1, 2 and 3 phylogenies obtained from the full data (Fig. 2). Clustering of 2008 and 2006b variants in region 3 (Supplementary Fig. S3, available in JGV Online).

There are 35 viral isolates falling inside the 2008 variant lineage in ORF2 phylogeny (Fig. 1). Only 11 of them have complete genomes that were confidently classified into two recombinant forms (Figs 2 and 3): (i) a 2006a/2006b-like double recombinant and (ii) a 2006a/2006b-like/Hunter triple recombinant, which we term ‘2008i’ and ‘2008ii’, respectively. The remaining 24 of them showed a recombinant pattern in the P2 domain similar to that observed in the 11 complete genomes, hence these 24 are, at least, double recombinants like 2008i. Phylogenetic analyses of six of them (those for which sufficient sequence before nucleotide position 5397 is available to distinguish between 2008i and 2008ii; see Supplementary Fig. S2a, available in JGV Online) suggests that three have the 2008i recombination structure (orange without asterisk in Fig. 1b) and three have the 2008ii structure (yellow without asterisk in Fig. 1b). Furthermore, Han et al. (2011) reported five 2008 variants in Korea based on their partial sequences at ORF1–ORF2 junction (GenBank accession numbers: HM635099–HM635103; the available ORF2 region is too short to be included in the ORF2 phylogeny of Fig. 1). Our re-analysis suggests that three of them might have the 2008ii genome structure, one might have the 2008i structure and one (Seoul/0654/2009/KOR) might be a Hunter variant, because it branched from a node closer to the Hunter lineage than to the 2008ii lineage, but could also be derived from the early ancestor of 2008ii (Supplementary Fig. S2a). Complete genomes of these isolates would help to confirm these speculations.

A simple recombination history may explain the emergence of the norovirus GII.4 2008 variants (Fig. 3). An ancestral 2006b-like virus might have recombined (by contributing a fragment of its P2 domain) with an ancestral 2006a virus creating the 2008i recombinant. Some of these continue to circulate (e.g. NSW001P). Others may have subsequently acquired a 5’ genomic region (5397 nt in length) from an ancestral Hunter variant to form the triple recombinant 2008ii. This scenario involves the least number of recombination steps, but conflicts with the observation that the 2008ii lineage is slightly more distant than the 2008i lineage to the most recent common ancestor of the 2008 lineage (Fig. 1b). To resolve the order of recombination with more certainty, full-length genome sequences of early 2008 variants (such as 8483/2008/ZAF and 2405/2008/ZAF) and molecular dating analyses (Lam et al., 2008; Tee et al., 2009) are needed.

There are subsequent recombinants derived from 2008 and other GII.4 variants. Recently, Motomura et al. (2010) reported a recombinant group, called 2008b, that is 2008-like in ORF1 and 2006b-like in other genomic regions. Our phylogenetic analysis (Supplementary Fig. S2b) suggests that the origin of the ORF1 of 2008b is most likely 2008ii: the 2008b group diverged before (but quite close to) the 2008ii lineage in a phylogeny of positions 1–5096 nt, and is placed inside the 2008ii in a phylogeny of positions 4264–5096 nt (Supplementary Fig. S2b). There are two other Japanese sporadic recombinant strains, Iwate5/2007/JP (termed 2007b in Motomura et al., 2010), which is a recombinant of 2008b and 2006b (Supplementary Fig. S2c), and Toyama5/2008/JP, which is a recombinant of 2006b and 2008ii (Supplementary Fig. S2d). A South African isolate (4638/2008/ZAF) for which only an ORF2 sequence is available, may be a mosaic of Hunter and 2008i variants, but with a different breakpoint location (at position 5966 nt) to that of the 2008ii strain (Supplementary Fig. S2e).
Both mutation and recombination are important mechanisms in norovirus evolution (Bull & White, 2011; Rohayem et al., 2005). As the entire clade of 2008 variant viruses appear to have a recombinant origin, intra-genotypic recombination may be more important in the emergence of new norovirus GI.4 variants than previously suggested (Lindesmith et al., 2008; Siebenga et al., 2007). By exchanging the P2 domain, which contains numerous receptor-binding and antigenic sites (Donaldson et al., 2010; Lindesmith et al., 2008; Prasad et al., 1999; Siebenga et al., 2007; Tan et al., 2003) between two GI.4 lineages, the virus may be able to move more rapidly across its fitness landscape (Burke, 1997) in the face of high levels of immunity in the host population. Incorporation by recombination of a P2 domain from a minority strain that exhibits epitopes unfamiliar to host immunity may result in a fitter recombinant progeny. This scenario may explain the emergence of the 2008 variant, whose 2006b-like parent may represent a minority circulating lineage that has not been discovered in the past. The putative antigenic change caused by recombination within the P2 domain of norovirus may resemble the antigenic shift caused by reassortment in human influenza virus, whereby gene segments encoding viral surface proteins are replaced by those of other (usually zoonotic) origins and representing novel antigenicity to human. However, such an analogy must remain speculative until a robust way of classifying noroviruses by antigenicity becomes available.

Although the pathogenicity and transmissibility of the 2008 variant remain to be studied experimentally, its novel intra-genotypic recombinant nature and widespread global distribution suggest it should be a surveillance target. Although the prevalence of the 2008 variant among GI.4 infections varies geographically [1% in Canada (Pang et al., 2010), ~8% in France (Belliot et al., 2010), ~22% in Korea (Han et al., 2011), and up to 80% in South Africa (Mans et al., 2010)], its increasing prevalence (Belliot et al., 2010; Motomura et al., 2010) and involvement in other recombinants raise a global public health concern. Despite this, we note that emergence of new GI.4 variants do not always cause dramatic outbreaks, e.g. during the 2009-2010 winter in the USA, when the 2008 variant first appeared there (Yen et al., 2011). In general, recombination of the norovirus antigenic region could have major implications for the design and effectiveness of the norovirus vaccines.

Acknowledgements

T.T.-Y.L. is supported by Newton International Fellowship from Royal Society. H.Z., D.K.S. and Y.G. are supported by Li Ka Shing Foundation and Research Grant Council of Hong Kong SAR (HKU 765809M). O.G.P. is supported by Royal Society, UK. Additional bioinformatic and computational resources were provided by the Computer Centre at The University of Hong Kong (HKU). We thank the two anonymous reviewers for useful suggestions. We also thank W. K. Kwan, Frankie Cheung and Lilian Chan (HKU) for technical support.

References


