Title: Population Genomics of GII.4 Noroviruses Reveal Complex Diversification and New Antigenic Sites Involved in the Emergence of Pandemic Strains

Running Title: Diversification and Emergence of GII.4 Noroviruses

Authors: Kentaro Tohma*, Cara J. Lepore*, Yamei Gao, Lauren A. Ford-Siltz, Gabriel I. Parra*#

Affiliation: Division of Viral Products, Food and Drug Administration, Silver Spring, MD, United States of America

*These authors contributed equally to this manuscript

Word count in Abstract: 245 (max 250)

Word count in Importance: 110 (max 150)

Word count in Main Text: 5698

 # Corresponding author: Dr. Gabriel I. Parra

10903 New Hampshire Avenue

Building 52/72, Room 1376

Silver Spring, MD 20993

Phone: +1-240-4026935

Fax: +1-301-5951070

Email: gabriel.parra@fda.hhs.gov

Abstract

GII.4 noroviruses are a major cause of acute gastroenteritis. Their dominance has been partially explained by the continuous emergence of antigenically distinct variants. To gain insights on the mechanisms of viral emergence and population dynamics of GII.4 noroviruses, we performed large-scale genomics, structural, and mutational analyses of the viral capsid protein (VP1). GII.4 noroviruses exhibited a periodic replacement of predominant variants with accumulation of amino acid substitutions. Genomic analyses revealed (i) a large number (87%) of conserved residues; (ii) variable residues that map on the previously determined antigenic sites; and (iii) variable residues that map outside of the antigenic sites. Residues from the third pattern formed motifs on the surface of VP1, which suggested extensions of previously predicted and new uncharacterized antigenic sites. The role of two motifs (C and G) in the antigenic make-up of the GII.4 capsid protein was confirmed with monoclonal antibodies and carbohydrate blocking assays. Amino acid profiles from antigenic sites (A, C, D, E, and G) correlated with the circulation patterns of GII.4 variants, with two of them (C and G) containing residues (352, 357, 378) linked with the emergence of new GII.4 variants. Notably, the emergence of each variant was followed by a stochastic diversification with minimal changes at the antigenic sites that did not progress towards the next variant. This study provides a methodological framework for antigenic characterization of viruses, and expands our understanding of the dynamics of GII.4 noroviruses that could facilitate the design of cross-reactive vaccines.

45 **Importance**

46

47 Noroviruses are an important cause of viral gastroenteritis around the world. An obstacle
48 delaying the development of norovirus vaccines is an inadequate understanding of the role of
49 norovirus diversity in immunity. Using a population genomics approach, we identified new
50 residues on the viral capsid protein (VP1) from GII.4 noroviruses, the predominant genotype,
51 that appear to be involved in the emergence and antigenic topology of GII.4 variants. Careful
52 monitoring of the substitutions in those residues involved in the diversification and emergence of
53 new viruses could help in the early detection of future novel variants with pandemic potential.
54 Therefore, this novel information on the antigenic diversification could facilitate GII.4 norovirus
55 vaccine design.

56

Introduction

Noroviruses are a leading cause of acute gastroenteritis affecting all ages worldwide (1). They are second only to rotavirus, although this is changing in places where rotavirus vaccination is effective (2, 3). It is estimated that norovirus is responsible for approximately 685 million infections and 200,000 deaths worldwide, with a primary public health concern in children, the elderly, and immunocompromised individuals (4-6). Norovirus outbreaks usually occur during the winter season in enclosed settings such as schools, hospitals, military facilities, and cruise ships. Because norovirus is highly contagious, outbreaks can be hard to control.

The norovirus genome is a positive-sense, single-stranded RNA molecule that is organized into three open reading frames (ORFs). ORF1 encodes for a polyprotein that is co-translationally cleaved by the viral protease into six non-structural (NS) proteins required for replication. ORF2 encodes the major capsid protein (VP1), and ORF3 the minor capsid protein (VP2). The norovirus capsid consists of 180 copies of VP1, arranged in a T=3 icosahedral symmetry. X-ray crystallography of the VP1 revealed two structural domains: the shell (S) and protruding (P). The S domain is relatively conserved and forms the core of the capsid, while the P domain is more variable and extends to the exterior of the capsid protein (7, 8). The P domain interacts with host attachment factors, namely human histo-blood group antigen (HBGA) carbohydrates, which could facilitate infection. Antibody-mediated blocking of the VP1:HBGA interaction correlates with protection against norovirus disease (9, 10). Expression of VP1 results in the self-assembly of virus-like particles (VLPs) that are structurally and antigenically similar to native virions (7, 11, 12). Given the lack of a traditional cell culture system for human norovirus, experimentally

developed VLPs have been an important tool to study norovirus immune responses and vaccine design.

Norovirus is highly diverse, with at least seven genogroups (GI-GVII) and over 40 genotypes defined based on differences in their VP1 sequences (13). While over 30 different genotypes from GI, GII, and GIV can infect humans, noroviruses from the GII.4 genotype are responsible for at least 70% of infections worldwide (14). Since the mid-1990s, six major norovirus GII.4 pandemics have been recorded worldwide and were associated with the following variants: Grimsby 1995 (or US95_96), Farmington Hills 2002, Hunter 2004, Den Haag 2006b, New Orleans 2009, and Sydney 2012 (15-17). The predominance of GII.4 viruses has been linked to the chronological emergence of variants in the human population, with new variants emerging around the time that the previous declines (15). The emergence of these variants has been correlated with changes on five different variable antigenic sites (A-E) that map on the surface of the P domain; thus, new viruses can evade the human immune responses elicited to previously-circulating variants (11, 18-20). Using the recently developed cell culture system for human noroviruses (21), two of these sites have been confirmed to be involved in virus neutralization (22). Studies have shown that antibodies that map to these sites can block the interaction of VP1 with carbohydrates from the HBGA; however, the antigenic sites from several monoclonal antibodies (mAbs) with blocking activity, raised against GII.4 viruses, have not been determined (11, 12, 19, 23-25). The evolving nature of GII.4 noroviruses could challenge the development of cross-protective vaccines against noroviruses; therefore, a better understanding of the mechanisms responsible for the antigenic diversification will facilitate vaccine design.

In this study, we adopted a large-scale genomics approach to identify sites that play a role in GII.4 evolution and antigenic diversification. We (re)defined the sites involved in the antigenic make-up of GII.4 pandemic variants, and found that intra-variant diversification exhibited a stochastic pattern of evolution. Importantly, we identified three sites (amino acids 352, 357, and 378) implicated with the emergence of predominant GII.4 variants, and that could help in the early detection of the next pandemic variant.

Results

GII.4 Inter-variant Evolution is Characterized by the Accumulation of Substitutions in the P domain

In order to investigate the evolutionary pattern of GII.4 strains, we calculated the genetic differences of 1601 nearly-full-length (≥1560 nt) VP1 sequences from GII.4 strains collected from 1974 to 2016 (Table S1). The phylogenetic tree of these sequences showed the presence of at least 11 different GII.4 variants emerging since 1995 (Fig. 1a). As shown previously (15, 26), GII.4 strains presented a chronological replacement of variants, with several unassigned intermediate strains (Fig. S1). Genetic analyses revealed an accumulation of substitutions in both nucleotide (Fig. 1b) and amino acid (Fig. 1c) sequences (coefficient of determination [R^2] of linear regression = 0.87 and 0.78, respectively). This pattern of accumulation of mutations was observed in all the subdomains of VP1 (Shell, P1, and P2; Fig. S2); however, higher slopes were noted in P2, where the five variable antigenic sites (A-E) are located, thus suggesting their role in the evolution and antigenic diversification of GII.4 noroviruses.

126

127 *Identification of New Antigenic Sites of GII.4 Noroviruses*

128

129 To pinpoint the role of each amino acid within the P domain in the evolution of GII.4, we
130 calculated the Shannon Entropy to measure the residue diversity at the inter-variant and intra-
131 variant level. Because Grimsby-like viruses were the first recorded to cause large outbreaks
132 worldwide, entropy values were calculated with strains (1572 VP1 sequences) detected from
133 1995 to 2016. While minimal diversity was detected at the intra-variant level (Fig. S3), the inter-
134 variant level revealed three substitution patterns: (i) a large number (87%, 285/325) of highly
135 conserved residues, which include conserved antigenic site F (27), and likely maintain the
136 structural integrity of the P domain; (ii) a small number (5%, 17/325) of variable residues that
137 map on the previously determined antigenic sites: A-E (19); and (iii) 23 variable residues that
138 map outside of those antigenic sites (Fig. 2a). Among the 23 non-antigenic variable sites from
139 the third pattern, two residues (residues 228 and 255) were surrounded by conserved residues on
140 the surface of VP1, and fourteen residues formed clusters (motifs) on the surface that could
141 represent new, or extensions of previously predicted, uncharacterized antigenic sites (Fig. 2b).
142 One motif comprises residues 339, 340, 341, 375, 376, 377, and 378. Because this motif included
143 two residues (340 and 376) previously described as one of the five major antigenic sites (20), we
144 extended the number of residues forming this antigenic site C (Fig. 2b). Although residue 375
145 presented low variability, the site was included as part of the antigenic site as it mapped on the
146 surface of the molecule and could potentially play a role in antibody recognition. Antigenic site
147 D was also extended to include two additional residues (396 and 397), as they clustered with
148 original residues (393, 394, and 395) on the surface (19). The new motif, denominated G,

149 included residues 352, 355, 356, 357, 359, and 364; and the last motif, denominated H, included
150 the residues 309 and 310. A summary of the residues from each of the motifs/antigenic sites is
151 shown in Fig. 2b. Profiling of the temporal frequency of the amino acid sequence pattern
152 (mutational pattern) of the previously characterized antigenic site A, the expanded antigenic site
153 C, and the new antigenic site G (confirmed in this study) indicated that their mutational patterns
154 correlated well with the fluctuation of GII.4 variants in the human population, suggesting a major
155 role in the emergence of variants (Fig. 3 and Fig. S4a). Correlation between the antigenic site
156 mutational patterns and GII.4 variants was assessed using adjusted Rand Index, and antigenic
157 site G was the one presenting the best correlation with the GII.4 variant circulation pattern, while
158 showing low sequence variation (Fig. 3b and c). Of note is that the mutational pattern from the
159 old antigenic site C did not correlate with the circulation of variants in nature (Fig. S4a and b),
160 suggesting that our population-guided antigenic site characterization provides a better resolution
161 of the antigenic profiling of GII.4 noroviruses. Neither the previously defined putative epitope
162 (motif) B (19) nor the new variable motif H correlated with GII.4 variant circulation (Fig. 3c and
163 Fig. S4a). Of note is that motif B presented differences for Farmington Hills 2002 and Hunter
164 2004 strains, while the motif H showed differences for Apeldoorn 2007, New Orleans 2009, and
165 Sydney 2012 (Fig. S4a). Mutations on antigenic sites D and E have been shown to alter both
166 antigenicity and binding to HBGA carbohydrates (19, 24, 28), while showing modest correlation
167 with GII.4 variant emergence (Fig. 3c). Thus, antigenic sites D and E might be under different
168 evolutionary pressures than those only involved in the antigenic characteristics of the virus. Of
169 note, the mutational pattern from expanded antigenic site D correlates much better than that from
170 the original antigenic site D (Fig. S4a and b), and this improvement of correlation is consistent
171 with the recent addition of residue 396 to this antigenic site (29). On the contrary, the mutational

pattern from the expanded antigenic site E showed lower correlation than that from the original antigenic site (Fig. S4a and b). This could be due to the expanded site (inclusion of residue 414) showing variation within the variants, Den Haag 2006b and Sydney 2012, making the profiles less consistent with the inter-variant diversity. Since our dataset presents sampling bias, i.e. over 50% of the sequences in the dataset were collected during 2010-2016, the Shannon entropy analysis was re-conducted using a randomly-subsampled dataset that includes a maximum of 50 sequences per variant (Fig. S5a). This dataset sensitively reflected the variation of viruses circulating before 2010, and showed 13 additional variable sites mapping outside the newly defined variable motifs/antigenic sites. Only three of them were located on the surface of the VP1; one of them (residue 250) did not cluster with any variable residues, and other two (residues 300 and 329) mapped close to the motif/antigenic site G (Fig. S5b). These two residues could be a part of the antigenic site G, however, both presented minor variation over decades suggesting a subtle impact on variant emergence (Fig. S5c).

The mutational patterns from three motifs (A, C, and G) correlated well with the fluctuation of GII.4 variants (Fig. 3c). The motif/antigenic site A was previously confirmed as an antigenic site (23), while expanded/new motifs (C and G) were not yet confirmed experimentally. To confirm the role of these two motifs in the antigenic make-up of the GII.4 capsid, we replaced residues of the VP1 from a Farmington Hills 2002 variant (MD2004 strain; accession number: DQ658413) with those of a Sydney 2012 variant (RockvilleD1 strain; accession number: KY424328; Fig. 4) and vice versa, and produced the corresponding VLPs. We tested the mutant and wild-type VLPs against guinea pig hyper-immune sera and mouse mAbs (two uncharacterized mAbs: B11 and B12 raised against Farmington Hills 2002 variant (MD2004 strain) (11), and four mAbs: 1C10,

6E6, 17A5 and 18G12 newly developed against Sydney 2012 variant (RockvilleD1 strain)). We found that mutations at the new motifs C and G abrogated binding of mAbs B11, B12, 6E6, 18G12, and mAbs 1C10 and 17A5, respectively (Fig. 4a). Reconstitution of the motif C in the Sydney 2012 wild-type strain can be achieved by reverting those sites to Farmington Hills 2002 wild-type strain (2012: C2004 VLPs) (Fig. 4a, left panel). Of note, when mutations were introduced at residues 340 and 376, which were regarded as the original antigenic site C, no differences in binding were observed with mAb B11 and B12 (11). While changes at residues 377 and 378 reduced binding to those mAbs, an additional mutation at residue 340 was required for complete antigenic site-depletion (Fig. S6). Likewise, reconstitution of the motifs C and G in the Farmington Hills 2002 wild-type strain was achieved by reverting those sites to Sydney 2012 wild-type strain (2004: C2012 and 2004: G2012 VLPs) (Fig. 4a, right panel).

The impact of mutations on the motifs/antigenic sites C and G against the immune response was evaluated as blockade potential (a surrogate of norovirus neutralization) using HBGA-blocking assays. Polyclonal sera against Farmington Hills 2002 wild-type VLP showed high blocking activity against homologous wild-type VLP (low EC_{50} value), but reduced blocking against the Sydney 2012 wild-type VLP (high EC_{50} value) (Fig. 4b and c). Mutations on antigenic sites C and G on those wild-type VLPs altered the blockade potential of polyclonal sera. Thus, when antigenic site C from Farmington Hills 2002 wild-type strain was transplanted into Sydney 2012 wild-type VLP, polyclonal sera raised against Farmington Hills 2002 wild-type strain gained blocking potential against the Sydney 2012 mutant (2012: C2004) VLP, but minimally with the mutant 2012: G2004. Notably, none of the Farmington Hills 2002 mutant VLPs (2004: C2012 and 2004: G2012) lost the blocking ability when using sera raised against the Farmington Hills

strain. Similarly, polyclonal sera raised against Sydney 2012 wild-type VLPs blocked the homologous VLPs, but not the Farmington Hills 2002 wild-type VLPs. Transplanting of antigenic site C and G into Farmington Hills 2002 VLPs (2004: C2012 and 2004: G2012) recovered its blocking ability of sera raised against Sydney 2012 wild-type strain. Polyclonal sera raised against Sydney 2012 wild-type VLPs did not lose blocking activity against antigenic site C mutant VLP (2012: C2004), however, a slight lost in the blocking activity was detected for antigenic site G mutant (2012: G2004) as compared with wild-type VLP. In summary, both antigenic sites (C and G) showed distinctive roles as blockade sites, suggesting their potential as protective antigenic sites.

Intra-variant Evolution is Driven by Stochastic Processes

In contrast to the accumulation of nucleotides and amino acids detected at the inter-variant level (Fig. 1b and 1c), there were limited accumulation of nucleotide (data not shown) and amino acid substitutions within variants (intra-variants level) (Fig. 5a). Despite this limited accumulation of substitutions, each of the pandemic variant presented diversity in their sequences (average of 4.8 amino acid substitutions). Interestingly, this diversity was detected at most of the major antigenic sites and variants (Fig. 5b and Fig. S7). Thus, while each variant presented a major amino acid combination for each antigenic site, most of the GII.4 variants presented other minor amino acid patterns on those antigenic sites. Moreover, the analysis of intra-variant diversification showed that their evolution was stochastic in time and location (Fig. 6), in contrast to the temporally clustered inter-variant evolution of GII.4 strains (Fig. 1). While some variants (e.g. Hunter 2004, Den Haag 2006b, New Orleans 2009, and Sydney 2012) presented diversity in their major

antigenic sites after 3-4 years of circulation, which could suggest that immune pressure acts at the intra-variant levels (Fig. 5), the number of strains (and sequences) is limited and do not represent dominant strains. Two other important observations could be made while analyzing the mutational pattern of each of the antigenic sites at the intra-variant level: (i) major differences at the amino acid sequence were detected in early strains for New Orleans 2009 and Sydney 2012 variants (Fig. 5 and Fig. S7), and those early strains did not present the major amino acid combination for any of the four major antigenic sites (A, C, D, and E; Fig. S7); and (ii) none of the preceding strains evolved towards (or presented) the amino acid motif from future strains. This, together with the phylogenetic analyses, suggests that each pandemic variant presented a different origin and did not follow a trunk-like linear evolution such as that seen in H3N2 influenza viruses (30).

Differences in the inter- and intra-variant evolutionary pattern were also confirmed by the Bayesian Markov Chain Monte Carlo (MCMC) analysis. Substitution rates from seven variants with > 50 sequences were calculated and summarized in Table 1. The rate of inter-variant (overall) GII.4 strains was reported elsewhere (15). Intra-variant substitution rates ranged from $1.57 – 4.64 \times 10^{-3}$ substitutions/site/year, all of them lower than the inter-variant (overall) substitution rate (5.4×10^{-3} substitutions/site/year) (15). Among the variants, Sydney 2012 variants had the higher substitution rate (Table 1).

Diversifying Pressure Drives Emergence of Pandemic GII.4 Noroviruses

263 Because different factors seemed to drive the inter-variant (overall) and intra-variant evolution,
264 we performed selection analysis using the Mixed Effect Model of Evolution (MEME) method
265 and looked for evidence of site-by-site episodic diversifying pressure on the VP1 along the
266 branches of its evolutionary tree. To analyze the overall evolutionary process, we randomly
267 subsampled sequences from the original dataset that included a maximum of 30 strains per
268 variant. During the overall evolution, we found 9 positively selected sites ($p < 0.05$ and empirical
269 Bayes factor > 100) on the P2 subdomain on the surface (codon sites: 327, 335, 352, 355, 357,
270 366, 368, 375, and 378) distributed on 21 branches of the phylogenetic tree (Fig. 7a). Branches
271 connecting discrete GII.4 variants presented sites under episodic diversification (codon sites:
272 352, 355, 357, 368, 378) that mapped on the antigenic sites. The mutational pattern of these sites
273 correlated well with GII.4 variant emergence, with higher adjusted Rand Index than any
274 antigenic sites (Fig. 7b and c). This suggests that residues on the antigenic sites experienced
275 episodic diversifying pressure during inter-variant evolution. When analyzing only each variant,
276 most of the diversifying pressure was found on the branches connecting to the tips rather than on
277 the internal branches (Fig. S8), suggesting that non-synonymous substitutions were deleterious
278 and rarely fixed in the population during the intra-variant evolution. The more comprehensive
279 dataset presented by New Orleans and Sydney variants may have included many viruses with
280 deleterious mutations that would not persist (as indicated by higher number of non-synonymous
281 substitutions on branches connecting to the tips), but would account for an artificial higher
282 substitution rate as compared with the other variants (Table 1). Together, this shows that the
283 diversifying pressure has driven the inter-variant, but not the intra-variant, evolution of GII.4
284 strains.
285

Discussion

GII.4 noroviruses are the most common cause of norovirus infections worldwide. Although other norovirus genotypes have predominated in specific locations and time, the global dominance of GII.4 has been recorded for almost three decades (16, 31). The persistence and dominance of GII.4 over all other norovirus genotypes has been explained by the chronological emergence of variants, which enables the virus to evade the immunity acquired to previously-circulating GII.4 variants, a process similar to H3N2 influenza viruses (26, 28, 30). Since the mid-1990s, over 10 different variants have been reported, with six of them associated with large outbreaks worldwide. The overall evolutionary pattern of GII.4 viruses presents a strong linear accumulation of amino acid substitutions during inter-variant diversification (15), with most substitutions occurring in the P2 subdomain (28). Antigenic differences among variants have been largely attributed to highly variable residues that map on the surface of the P domain, leading to the identification of five (A-E) motifs that are part of GII.4-specific antigenic sites (18-20). While the binding site was characterized for different GII.4 specific mAbs, the same studies have reported numerous GII.4 specific mAbs whose binding sites have not been determined (11, 12, 19, 23-25). Applying a population genomics approach we found new (or expanded) motifs on the surface of the capsid that presented mutational patterns that correlates with the circulation of GII.4 variants. The role of antigenicity of two of those motifs (antigenic sites C and G) were confirmed with (previously) uncharacterized HBGA-blockade mAbs (B11 and B12), newly developed mAbs (1C10, 6E6, 17A5, and 18G12), and polyclonal sera from guinea pigs immunized with wild-type VLPs. The antigenic site G presented the strongest correlation with the emergence and circulation of new GII.4 variants, while being more

conserved when compared with other large antigenic sites (e.g. A, C, or E). This indicates that newly discovered antigenic site G plays a pivotal role in shaping the GII.4 noroviruses into new variants with pandemic potential. The antigenic site C is close to the previously-defined antigenic site A, but substitutions on antigenic site A did not affect the binding of those mAbs (11). Notably, competition analyses among different mAbs showed that mAbs B11 and B12 partially blocked interactions with mAbs mapping to antigenic site A (11), suggesting that this new motif is part of an antigenic site that involves more than one epitope (at least antigenic sites A and C). Recently, Koromyslova and colleagues have shown that the footprint of an antibody that neutralized human noroviruses mapped to residues from antigenic sites C and D (22). This demonstrates that same epitope could be shared by the different antigenic sites and their interaction could result in differences in the evolutionary pattern presented. In addition, differences on the HBGA-blocking ability between antigenic sites (C and G) and between variants (Farmington Hills 2002 and Sydney 2012) further highlight the complexity of the antigenic topology of GII.4 noroviruses.

Large-scale analysis reduces biases on the role of individual amino acid substitution on the emergence of the new GII.4 variants. Recently, it was suggested that mutations in the capsid protein of Sydney strains circulating in 2015 resulted in antigenically different viruses from those circulating in 2012 (32). Our large-scale intra-variant analyses show that (i) the strain selected as Sydney 2012 for the antigenic study (32) was not representative of the predominant virus, and (ii) strains with similar antigenic site sequences as those regarded as “GII.4 2015” by Lindesmith et al. (32) have been circulating since 2010 in the human population as part of the overall population of the Sydney 2012 variant (Fig. S7). Similarly, the NERK motif, which was

suggested to occlude conserved GII.4 antigenic sites and effect the antibody blockade potency (33, 34), was shown to be highly conserved among GII.4 variants. Previous studies pointed out a mutation on this motif in Sydney 2012 variant, however, our large-scale analysis showed that the New Orleans variant was the only one showing -at the population level- any mutation at this motif (Fig. S9). Thus, the profiling of mutational patterns we implemented, which included the use of a large number of sequences, could provide a better understanding on the role of individual mutations on the circulation and predominance of the pandemic GII.4 variants. Immunological analyses that include multiple different viruses from each of the pandemic variants are needed to better delineate the meaning of minor mutations on the antigenic differences among the variants.

Improvements on the understanding of viral dynamics and the correlation between antigenic and genetic changes in influenza viruses have facilitated the selection of virus strains to be included in upcoming seasonal vaccines (35-37). To better understand viral dynamics of GII.4 noroviruses, we performed selection analyses including all variants reported for over four decades. Episodic diversifying (positive) selection was observed in five residues (352, 355, 357, 368, and 378) during inter-variant evolution (Fig. 7a); these residues are part of antigenic sites A, C, and G. Three of these residues (352, 357, and 378) showed positive selection on major branches of the GII.4 tree, indicating a role in the emergence of new pandemic noroviruses. In these three sites, strains that emerged and predominated from 1995 to 2006 (Grimsby 1995, Farmington Hills 2002, Hunter 2004, and Yerseke 2006a) presented the motif SHG; the strain Den Haag 2006b that emerged and predominated from 2006 to 2009 presented the motif YPH; the strains that have predominated since 2009 (Apeldoorn 2007, New Orleans 2009, and Sydney

2012) presented the motif YDN. Notably, while these three residues have been understudied pertaining to the evolution of GII.4 noroviruses (26, 28, 38), they seem to play a major role in the antigenic topology and emergence of new GII.4 variants. Of note, residue 368 presented episodic diversification that distinguished the Sydney 2012 strains from Apeldoorn 2007 and New Orleans 2009. Changes on this residue were shown to be involved in the antigenic diversification of the Sydney 2012 strain (39), thus supporting our *in silico* observation. Initial studies suggested that most positively-selected sites for GII.4 noroviruses were located in the S domain (28, 38, 40); however, our analyses, with an expanded number of sequences and variants, showed that the emergence of new variants stemmed from the positive selection of residues mapping on the surface of the P domain. Pinpointing the changes required for emergence of a new antigenic variant of seasonal influenza or the emergence of a pandemic virus is the "holy grail" for controlling viruses undergoing constant change. While prediction of viral emergence requires a wholistic approach that includes studies of the virus, environment, and the host (41-43), careful monitoring of the substitutions in residues involved in the diversification and emergence of new GII.4 viruses (i.e. residues 352, 357, and 378) could help in the early detection of future novel variants with pandemic potential.

The GII.4 inter-variant diversification, likely driven by the immune status of the population, is correlated with the accumulation of amino acid substitutions in the major capsid protein. In contrast, our analyses suggest that the diversification at intra-variant level is much restricted, with amino acid substitutions occurring without indication of diversifying selection. Minimal substitutions of the amino acids that mapped on the major antigenic sites were observed over the predominance of each variant, and most seemed to follow a stochastic process. The latter could

be a result of multiple pressures on the virus, including but not limited to individual virus-host interactions and dispersion. For two variants (New Orleans 2009 and Sydney 2012) a larger, more comprehensive dataset was available, and pre-pandemic strains have been reported (44-47). Interestingly, although these pre-pandemic strains cluster within their respective variants, they present multiple differences (mostly mapping at the antigenic sites), as compared to the virus population that later established a worldwide dominance. Together, these results suggest that emergence of GII.4 variants could occur in steps: (1) a pre-pandemic stage characterized by acquisition of mutations that facilitate viral emergence and episodic diversification (exemplified here by residues 352, 357, 368, and 378), followed by (2) a short period (1-2 years) of adaptability (with different antigenic motifs) that precedes the pandemic phase, and finally (3) the pandemic phase, where the virus is dominant and only explores a narrow space of sequence diversity. A similar pattern has been observed for H3N2 influenza viruses (48) or rotaviruses (49), in which viruses that circulate at low levels could predominate in the upcoming season without major changes in their genetic background. This order of events could have also occurred during the recent emergence and predominance of GII.17 noroviruses in many Asian countries. During 2013-2014, a “new” GII.17 (variant C) was detected circulating in different countries (Japan, Hong Kong, China), but in the next epidemic season, this variant was ultimately replaced by the variant D that spread worldwide and predominated from 2014-2016 (15, 50-52). In the same context, we found six GII.4 strains that did not cluster with any GII.4 variant (Fig. S1), and could represent strains that did not adapt well to the human population. These strains presented different sequence patterns in their antigenic sites, and therefore could constitute strains in the pre-pandemic stage that explored the sequence space but failed to thrive to the next evolutionary level.

401

402 Our findings suggest two different mechanisms behind the evolutionary dynamics of the major
403 capsid protein from GII.4 noroviruses: (1) a steady pandemic phase governed by stochastic
404 processes, which is preceded by (2) substitutions that arise from positive selection. Our study
405 also provides a methodological framework that could facilitate the characterization of variable
406 antigenic sites that play a relevant role in the emergence of new viruses in the human population.
407 Studies that include large and long time-scale datasets of full-length genomes would help to
408 determine the factors involved in norovirus predominance and persistence in the human
409 population.

410

411 **Materials and Methods**

412

413 *1. Data mining and sequence analyses*

414

415 A total of 1601 full-length (1623 nt) and nearly full-length (≥1560 nt) VP1 sequences of the
416 GII.4 genotype, in which sequences from immunocompromised patients and environmental
417 samples were removed, were downloaded from GenBank (accessed July 2017) (Table S1). The
418 sequences spanned 42 years from 1974 to 2016. Sequences were aligned using ClustalW as
419 implemented in MEGA v7 (53), and visually inspected to confirm proper alignment.
420 Nomenclature for GII.4 variants was adopted as previously indicated (15), and variant datasets
421 were parsed following such information. Sequence analyses were performed using total dataset
422 except where indicated. Entropy analysis and profiling of mutational patterns only included
423 strains (1572 VP1 sequences) collected from 1995 to 2016 as Grimsby-like viruses were the first

424 recorded to cause large outbreaks worldwide (31), and included the following 11 variants in
425 approximate order of emergence: Grimsby 1995 (or US95_96), Farmington Hills 2002, Lanzhou
426 2002, Sakai 2003 (or Asia 2003), Hunter 2004, Yerseke 2006a, Den Haag 2006b, Osaka 2007,
427 Apeldoorn 2007, New Orleans 2009, and Sydney 2012. Intra-variant and inter-variant Shannon
428 entropy was calculated using the Shannon Entropy-One tool as implemented in Los Alamos
429 National Laboratory (www.hiv.lanl.gov). Entropy values for each position were plotted in
430 GraphPad Prism v7. The structural model of GII.4 norovirus P domain dimer (Protein Data Base
431 [PDB] accession number: 2OBS) was rendered using UCSF Chimera (version 1.11.2) (54).
432 Profiling of mutational patterns of motifs/antigenic sites was performed using R 3.4.2 (55).
433 Amino acid sequences of each motif/antigenic site was profiled by year and its mutational
434 patterns were plotted in a composite bar graph showing the number of strains with each pattern
435 as a fraction (%) of a whole (total number of strains). The correlation of the mutational patterns
436 and variant distribution was assessed using adjusted Rand Index, known as a clustering analysis
437 method, which evaluated the degree of the matches between mutational patterns and variant
438 classification. Adjusted Rand Index was calculated using R and *mclust* package (56). To account
439 for the sampling bias from the original dataset, we repeated entropy and structural analysis using
440 randomly subsampled dataset that included a maximum 50 strains/variant (n = 474) from the
441 original GII.4 dataset (1572 sequences).
442
443 *2. Diversifying selection analysis*
444
445 Maximum likelihood (ML) phylogenetic trees of VP1-enconding nucleotide sequences for all
446 variants (inter-variant analyses) and each variant (intra-variant analyses) were constructed using

447 the PhyML (57). The best substitution models were selected based on the lowest Akaike
448 Information Criterion (AICc) for each dataset using jModelTest v2 (58, 59). Larger datasets (i.e.
449 Sydney 2012 and New Orleans 2009 variants) tended to favor a Generalized time-reversible
450 (GTR) substitution model, while variants with fewer reported sequences favored a Tamura-Nei
451 93 (TN93) model. Diversifying selection on the VP1-encoding sequence through its ML
452 phylogenetic tree was analyzed by using MEME methods (60, 61). We aimed to detect codon
453 sites under positive selection (i.e. more non-synonymous substitutions than synonymous
454 substitutions) during the evolution and focused on the sites on or near the major antigenic sites of
455 GII.4. Significant positive selection was indicated by $p < 0.05$. The branches under diversifying
456 positive selection were explored by using empirical Bayes factor >100 in MEME. To reduce the
457 data size for the inter-variant analyses of positive selection, we randomly subsampled a
458 maximum 30 strains/variant (n = 308) from the original GII.4 dataset (1601 sequences) as
459 indicated previously.
460
461 *3. Bayesian analyses of nucleotide substitution rate*
462
463 Using the VP1 sequences and respective collection years from each strain, temporal phylogenetic
464 analysis was calculated using Bayesian Markov Chain Monte Carlo (MCMC) methodology in
465 BEAST v1.8.3 (62). The best substitution models were selected based on the lowest Akaike
466 Information Criterion (AICc) as mentioned above. The clock models (strict or relaxed lognormal
467 clock) and tree priors (constant population size, exponential growth, or skyline) were tested, and
468 best models were selected based on the model selection procedure using AIC through MCMC.
469 The MCMC runs were performed until all the parameters reached convergence. MCMC runs

were analyzed using Tracer v1.6 (http://tree.bio.ed.ac.uk/software/tracer/). The initial 10% of the logs from the MCMC run were removed before summarizing the mean and the 95% highest posterior density interval of the substitution rates.

4. Site-directed Mutagenesis and VLPs production

The VP1-encoding sequences from a GII.4 Farmington Hills 2002 (MD2004-3) and Sydney 2012 (RockvilleD1) strain were ligated into pFastBac1 vectors using SalI and NotI restriction sites (11, 63). Site-directed mutagenesis of pFastBac-MD2004-3 and pFastBac-RockvilleD1 was performed using mutation-specific forward and reverse primers, followed by purification with illustra MicroSpin G-50 Columns (GE Healthcare, Buckinghamshire, United Kingdom). Parental DNA was digested using the Dpn1 enzyme (New England BioLabs, Massachusetts, USA). VLPs presenting multiple mutations were developed by cloning a chemically synthesized P domain into a pFastBac1 plasmid containing the S domain using PspXI and NotI restriction sites (11). Each pFastBac construct was transformed via electroporation into ElectroMAX DH10B Cells (Thermo Fisher Scientific, California, USA), and grown on LB plates with ampicillin overnight at 37°C. Selected colonies were used to extract plasmid DNA (QIAprep Spin Miniprep Kit; Qiagen, Hilden, Germany). Introduction of mutations were confirmed by Sanger sequencing. VLPs were produced using the Bac-to-Bac Baculovirus Expression System (Invitrogen, California, USA) and purified through a cesium chloride gradient as previously described (63). Expression of VP1 protein was confirmed by western blot, and VLP integrity was confirmed by electron microscopy.

493 *5. Immunoassays*

494

495 Mutants and wild-type norovirus VLPs were analyzed for reactivity to monoclonal antibodies by
496 enzyme-linked immunosorbent assay (ELISA) as described previously (11). The mAbs B11 and
497 B12 were obtained from mice immunized with GII.4 Farmington Hills 2002 variant (MD2004-3
498 strain) VLP (11), and generously provided by Dr. Kim Y. Green (National Institutes of Health,
499 USA). The GII.4 Sydney 2012 variant specific-mAbs (1C10, 6E6, 17A5, and 18G12) were
500 developed from mice immunized with RockvilleD1 strain VLP (GenScript, NJ, USA). HBGA-
501 blocking assays were performed using mutant and wild-type VLPs, HBGA molecules derived
502 from human saliva, and polyclonal antibodies. The polyclonal antibodies were obtained from
503 guinea pigs and mice immunized with GII.4 Farmington Hills 2002 (MD2004-3) and Sydney
504 2012 (RockvilleD1 strain) VLPs (11, 63). Human saliva was collected from a healthy adult
505 volunteer. Saliva sample was boiled at 100^{o}C for 10 mins immediately after the collection, and
506 centrifuged at 13,000 rpm for 5min. The saliva supernatant was collected and used for HBGA-
507 binding and -blocking assays. Briefly, serial dilution of guinea pig polyclonal sera were mixed
508 with wild-type or mutant VLPs and incubated on the saliva-coated plate for 1h at 37^{o}C. Plates
509 were washed four times to remove unbound (i.e. blocked by guinea pig polyclonal sera) VLPs.
510 Pooled sera from mice immunized with Farmington Hills 2002 VLP and Sydney 2012 VLP was
511 used to detect the VLPs attached to the plate. Goat anti-mouse IgG conjugated with horseradish
512 peroxidase and 2,2'-azino-bis(3-ethylbenzothiazoline-6-sulphonic acid) (ABTS) substrate
513 (SeraCare, MA, USA) were used to develop the blue-green color on the VLP-attached plate. The
514 degree of blocking was evaluated using optical density (OD) at 405 nm and EC_{50} of sera dilution.
515 The EC_{50} was calculated from the OD curve using GraphPad Prism v7.

516

517 **Acknowledgments**

518 We thank Dr. Steve Rubin for the critical reading of the manuscript. Financial support for this
519 work was provided by the Food and Drug Administration intramural funds [Program Number
520 Z01 BK 04012-01 LHV to G.I.P]. K.T and C.J.L are recipients of a CBER/FDA-sponsored Oak
521 Ridge Institute for Science and Education (ORISE) fellowship.

522

523

524 **References**

525

526 1. Ahmed SM, Lopman BA, Levy K. 2013. A systematic review and meta-analysis of the
527 global seasonality of norovirus. PLoS One 8:e75922.
528 2. Payne DC, Vinje J, Szilagyi PG, Edwards KM, Staat MA, Weinberg GA, Hall CB,
529 Chappell J, Bernstein DI, Curns AT, Wikswo M, Shirley SH, Hall AJ, Lopman B,
530 Parashar UD. 2013. Norovirus and medically attended gastroenteritis in U.S. children. N
531 Engl J Med 368:1121-30.
532 3. Koo HL, Neill FH, Estes MK, Munoz FM, Cameron A, DuPont HL, Atmar RL. 2013.
533 Noroviruses: The Most Common Pediatric Viral Enteric Pathogen at a Large University
534 Hospital After Introduction of Rotavirus Vaccination. J Pediatric Infect Dis Soc 2:57-60.
535 4. Kirk MD, Pires SM, Black RE, Caipo M, Crump JA, Devleesschauwer B, Dopfer D,
536 Fazil A, Fischer-Walker CL, Hald T, Hall AJ, Keddy KH, Lake RJ, Lanata CF,
537 Torgerson PR, Havelaar AH, Angulo FJ. 2015. World Health Organization Estimates of
538 the Global and Regional Disease Burden of 22 Foodborne Bacterial, Protozoal, and Viral
539 Diseases, 2010: A Data Synthesis. PLoS Med 12:e1001921.
540 5. Patel MM, Widdowson MA, Glass RI, Akazawa K, Vinje J, Parashar UD. 2008.
541 Systematic literature review of role of noroviruses in sporadic gastroenteritis. Emerg
542 Infect Dis 14:1224-31.
543 6. Pires SM, Fischer-Walker CL, Lanata CF, Devleesschauwer B, Hall AJ, Kirk MD,
544 Duarte AS, Black RE, Angulo FJ. 2015. Aetiology-Specific Estimates of the Global and
545 Regional Incidence and Mortality of Diarrhoeal Diseases Commonly Transmitted
546 through Food. PLoS One 10:e0142927.

547 7. Green KY. 2013. Caliciviridae: The Noroviruses, p 582-608. *In* Knipe DM, Howley PM
548 (ed), Fields Virology, 6th ed. Lippincott, Williams & Wilkins, Philadelphia, Pa.
549 8. Prasad BV, Hardy ME, Dokland T, Bella J, Rossmann MG, Estes MK. 1999. X-ray
550 crystallographic structure of the Norwalk virus capsid. Science 286:287-90.
551 9. Reeck A, Kavanagh O, Estes MK, Opekun AR, Gilger MA, Graham DY, Atmar RL.
552 2010. Serological correlate of protection against norovirus-induced gastroenteritis. J
553 Infect Dis 202:1212-8.
554 10. Czako R, Atmar RL, Opekun AR, Gilger MA, Graham DY, Estes MK. 2012. Serum
555 hemagglutination inhibition activity correlates with protection from gastroenteritis in
556 persons infected with Norwalk virus. Clin Vaccine Immunol 19:284-7.
557 11. Parra GI, Abente EJ, Sandoval-Jaime C, Sosnovtsev SV, Bok K, Green KY. 2012.
558 Multiple antigenic sites are involved in blocking the interaction of GII.4 norovirus capsid
559 with ABH histo-blood group antigens. J Virol 86:7414-26.
560 12. Alvarado G, Ettayebi K, Atmar RL, Bombardi RG, Kose N, Estes MK, Crowe JE, Jr.
561 2018. Human Monoclonal Antibodies That Neutralize Pandemic GII.4 Noroviruses.
562 Gastroenterology doi:10.1053/j.gastro.2018.08.039.
563 13. Vinje J. 2015. Advances in laboratory methods for detection and typing of norovirus. J
564 Clin Microbiol 53:373-81.
565 14. Hoa Tran TN, Trainor E, Nakagomi T, Cunliffe NA, Nakagomi O. 2013. Molecular
566 epidemiology of noroviruses associated with acute sporadic gastroenteritis in children:
567 global distribution of genogroups, genotypes and GII.4 variants. J Clin Virol 56:185-93.

15. 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 Pathog 13:e1006136.

16. Siebenga JJ, Vennema H, Zheng DP, Vinje J, Lee BE, Pang XL, Ho EC, Lim W, Choudekar A, Broor S, Halperin T, Rasool NB, Hewitt J, Greening GE, Jin M, Duan ZJ, Lucero Y, O'Ryan M, Hoehne M, Schreier E, Ratcliff RM, White PA, Iritani N, Reuter G, Koopmans M. 2009. Norovirus illness is a global problem: emergence and spread of norovirus GII.4 variants, 2001-2007. J Infect Dis 200:802-12.

17. van Beek J, Ambert-Balay K, Botteldoorn N, Eden JS, Fonager J, Hewitt J, Iritani N, Kroneman A, Vennema H, Vinje J, White PA, Koopmans M, NoroNet. 2013. Indications for worldwide increased norovirus activity associated with emergence of a new variant of genotype II.4, late 2012. Euro Surveill 18:8-9.

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

19. Lindesmith LC, Beltramello M, Donaldson EF, Corti D, Swanstrom J, Debbink K, Lanzavecchia A, Baric RS. 2012. Immunogenetic mechanisms driving norovirus GII.4 antigenic variation. PLoS Pathog 8:e1002705.

20. Garaicoechea L, Aguilar A, Parra GI, Bok M, Sosnovtsev SV, Canziani G, Green KY, Bok K, Parreno V. 2015. Llama nanoantibodies with therapeutic potential against human norovirus diarrhea. PLoS One 10:e0133665.

21. Ettayebi K, Crawford SE, Murakami K, Broughman JR, Karandikar U, Tenge VR, Neill FH, Blutt SE, Zeng XL, Qu L, Kou B, Opekun AR, Burrin D, Graham DY, Ramani S,

590 Atmar RL, Estes MK. 2016. Replication of human noroviruses in stem cell-derived
591 human enteroids. Science 353:1387-1393.
592 22. Koromyslova AD, Morozov VA, Hefele L, Hansman GS. 2018. Human Norovirus
593 Neutralized by a Monoclonal Antibody Targeting the HBGA Pocket. J Virol
594 doi:10.1128/JVI.02174-18.
595 23. Debbink K, Donaldson EF, Lindesmith LC, Baric RS. 2012. Genetic mapping of a highly
596 variable norovirus GII.4 blockade epitope: potential role in escape from human herd
597 immunity. J Virol 86:1214-26.
598 24. Lindesmith LC, Debbink K, Swanstrom J, Vinje J, Costantini V, Baric RS, Donaldson
599 EF. 2012. Monoclonal antibody-based antigenic mapping of norovirus GII.4-2002. J
600 Virol 86:873-83.
601 25. Lindesmith LC, Costantini V, Swanstrom J, Debbink K, Donaldson EF, Vinje J, Baric
602 RS. 2013. Emergence of a norovirus GII.4 strain correlates with changes in evolving
603 blockade epitopes. J Virol 87:2803-13.
604 26. Siebenga JJ, Vennema H, Renckens B, de Bruin E, van der Veer B, Siezen RJ,
605 Koopmans M. 2007. Epochal evolution of GGII.4 norovirus capsid proteins from 1995 to
606 2006. J Virol 81:9932-41.
607 27. Lindesmith LC, Mallory ML, Debbink K, Donaldson EF, Brewer-Jensen PD, Swann EW,
608 Sheahan TP, Graham RL, Beltramello M, Corti D, Lanzavecchia A, Baric RS. 2018.
609 Conformational Occlusion of Blockade Antibody Epitopes, a Novel Mechanism of GII.4
610 Human Norovirus Immune Evasion. mSphere 3: e00518-17.

611 28. Lindesmith LC, Donaldson EF, Lobue AD, Cannon JL, Zheng DP, Vinje J, Baric RS.
612 2008. Mechanisms of GII.4 norovirus persistence in human populations. PLoS Med
613 5:e31.
614 29. Lindesmith LC, Brewer-Jensen PD, Mallory ML, Yount B, Collins MH, Debbink K,
615 Graham RL, Baric RS. 2019. Human Norovirus Epitope D Plasticity Allows Escape from
616 Antibody Immunity without Loss of Capacity for Binding Cellular Ligands. J Virol
617 93:e01813-18.
618 30. Koelle K, Cobey S, Grenfell B, Pascual M. 2006. Epochal evolution shapes the
619 phylodynamics of interpandemic influenza A (H3N2) in humans. Science 314:1898-903.
620 31. Vinje J, Altena SA, Koopmans MP. 1997. The incidence and genetic variability of small
621 round-structured viruses in outbreaks of gastroenteritis in The Netherlands. J Infect Dis
622 176:1374-8.
623 32. Lindesmith LC, Brewer-Jensen PD, Mallory ML, Debbink K, Swann EW, Vinje J, Baric
624 RS. 2018. Antigenic Characterization of a Novel Recombinant GII.P16-GII.4 Sydney
625 Norovirus Strain With Minor Sequence Variation Leading to Antibody Escape. J Infect
626 Dis 217:1145-1152.
627 33. Lindesmith LC, Donaldson EF, Beltramello M, Pintus S, Corti D, Swanstrom J, Debbink
628 K, Jones TA, Lanzavecchia A, Baric RS. 2014. Particle conformation regulates antibody
629 access to a conserved GII.4 norovirus blockade epitope. J Virol 88:8826-42.
630 34. Tamminen K, Malm M, Vesikari T, Blazevic V. 2019. Immunological Cross-Reactivity
631 of an Ancestral and the Most Recent Pandemic Norovirus GII.4 Variant. Viruses 11:E91.

632 35. Neher RA, Bedford T, Daniels RS, Russell CA, Shraiman BI. 2016. Prediction,
633 dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. Proc
634 Natl Acad Sci U S A 113:E1701-9.
635 36. Morris DH, Gostic KM, Pompei S, Bedford T, Luksza M, Neher RA, Grenfell BT, Lassig
636 M, McCauley JW. 2018. Predictive Modeling of Influenza Shows the Promise of Applied
637 Evolutionary Biology. Trends Microbiol 26:102-118.
638 37. Klingen TR, Reimering S, Guzman CA, McHardy AC. 2018. In Silico Vaccine Strain
639 Prediction for Human Influenza Viruses. Trends Microbiol 26:119-131.
640 38. Siebenga JJ, Lemey P, Kosakovsky Pond SL, Rambaut A, Vennema H, Koopmans M.
641 2010. Phylodynamic reconstruction reveals norovirus GII.4 epidemic expansions and
642 their molecular determinants. PLoS Pathog 6:e1000884.
643 39. Debbink K, Lindesmith LC, Donaldson EF, Costantini V, Beltramello M, Corti D,
644 Swanstrom J, Lanzavecchia A, Vinje J, Baric RS. 2013. Emergence of new pandemic
645 GII.4 Sydney norovirus strain correlates with escape from herd immunity. J Infect Dis
646 208:1877-87.
647 40. Bok K, Abente EJ, Realpe-Quintero M, Mitra T, Sosnovtsev SV, Kapikian AZ, Green
648 KY. 2009. Evolutionary dynamics of GII.4 noroviruses over a 34-year period. J Virol
649 83:11890-901.
650 41. Holmes EC, Rambaut A, Andersen KG. 2018. Pandemics: spend on surveillance, not
651 prediction. Nature 558:180-182.
652 42. Holmes EC. 2013. What can we predict about viral evolution and emergence? Curr Opin
653 Virol 3:180-4.

654 43. Lipsitch M, Barclay W, Raman R, Russell CJ, Belser JA, Cobey S, Kasson PM, Lloyd-
655 Smith JO, Maurer-Stroh S, Riley S, Beauchemin CA, Bedford T, Friedrich TC, Handel
656 A, Herfst S, Murcia PR, Roche B, Wilke CO, Russell CA. 2016. Viral factors in
657 influenza pandemic risk assessment. Elife 5:e18491.
658 44. Giammanco GM, De Grazia S, Terio V, Lanave G, Catella C, Bonura F, Saporito L,
659 Medici MC, Tummolo F, Calderaro A, Banyai K, Hansman G, Martella V. 2014.
660 Analysis of early strains of the norovirus pandemic variant GII.4 Sydney 2012 identifies
661 mutations in adaptive sites of the capsid protein. Virology 450-451:355-8.
662 45. Eden JS, Hewitt J, Lim KL, Boni MF, Merif J, Greening G, Ratcliff RM, Holmes EC,
663 Tanaka MM, Rawlinson WD, White PA. 2014. The emergence and evolution of the
664 novel epidemic norovirus GII.4 variant Sydney 2012. Virology 450-451:106-13.
665 46. Hasing ME, Lee BE, Preiksaitis JK, Tellier R, Honish L, Senthilselvan A, Pang XL.
666 2013. Emergence of a new norovirus GII.4 variant and changes in the historical biennial
667 pattern of norovirus outbreak activity in Alberta, Canada, from 2008 to 2013. J Clin
668 Microbiol 51:2204-11.
669 47. Eden JS, Bull RA, Tu E, McIver CJ, Lyon MJ, Marshall JA, Smith DW, Musto J,
670 Rawlinson WD, White PA. 2010. Norovirus GII.4 variant 2006b caused epidemics of
671 acute gastroenteritis in Australia during 2007 and 2008. J Clin Virol 49:265-71.
672 48. Wolf YI, Viboud C, Holmes EC, Koonin EV, Lipman DJ. 2006. Long intervals of stasis
673 punctuated by bursts of positive selection in the seasonal evolution of influenza A virus.
674 Biol Direct 1:34.
675 49. Parra GI. 2009. Seasonal shifts of group A rotavirus strains as a possible mechanism of
676 persistence in the human population. J Med Virol 81:568-71.

677 50. Chan MC, Lee N, Hung TN, Kwok K, Cheung K, Tin EK, Lai RW, Nelson EA, Leung
678 TF, Chan PK. 2015. Rapid emergence and predominance of a broadly recognizing and
679 fast-evolving norovirus GII.17 variant in late 2014. Nat Commun 6:10061.
680 51. Lu J, Fang L, Zheng H, Lao J, Yang F, Sun L, Xiao J, Lin J, Song T, Ni T, Raghwani J,
681 Ke C, Faria NR, Bowden TA, Pybus OG, Li H. 2016. The Evolution and Transmission of
682 Epidemic GII.17 Noroviruses. J Infect Dis 214:556-64.
683 52. Matsushima Y, Ishikawa M, Shimizu T, Komane A, Kasuo S, Shinohara M, Nagasawa
684 K, Kimura H, Ryo A, Okabe N, Haga K, Doan YH, Katayama K, Shimizu H. 2015.
685 Genetic analyses of GII.17 norovirus strains in diarrheal disease outbreaks from
686 December 2014 to March 2015 in Japan reveal a novel polymerase sequence and amino
687 acid substitutions in the capsid region. Euro Surveill 20:pii=21173.
688 53. Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics
689 Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol 33:1870-4.
690 54. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE.
691 2004. UCSF Chimera--a visualization system for exploratory research and analysis. J
692 Comput Chem 25:1605-12.
693 55. R Core Team. 2017. R: A language and environment for statistical computing. R
694 Foundation for Statistical Computing, Vienna, Austria.
695 56. Scrucca L, Fop M, Murphy TB, Raftery AE. 2016. mclust 5: Clustering, Classification
696 and Density Estimation Using Gaussian Finite Mixture Models. R J 8:289-317.
697 57. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New
698 algorithms and methods to estimate maximum-likelihood phylogenies: assessing the
699 performance of PhyML 3.0. Syst Biol 59:307-21.

700 58. Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new
701 heuristics and parallel computing. Nat Methods 9:772.
702 59. Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large
703 phylogenies by maximum likelihood. Syst Biol 52:696-704.
704 60. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012.
705 Detecting individual sites subject to episodic diversifying selection. PLoS Genet
706 8:e1002764.
707 61. Delport W, Poon AF, Frost SD, Kosakovsky Pond SL. 2010. Datamonkey 2010: a suite
708 of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26:2455-7.
709 62. Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with
710 BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969-73.
711 63. Parra GI, Green KY. 2014. Sequential gastroenteritis episodes caused by 2 norovirus
712 genotypes. Emerg Infect Dis 20:1016-8.
713

Figure Legends

Figure 1: Evolution of the major capsid protein (VP1) from GII.4 noroviruses results in the accumulation of substitutions and the periodic emergence of variants. (a) Maximum likelihood tree of GII.4 noroviruses showing the circulation of different variants overtime. Branches are colored based on variant determination from the online Norovirus Typing Tool (https://www.rivm.nl/mpf/typingtool/norovirus/). Black branches in the tree represent sequences that did not cluster into a variant that circulated between 1995 and 2016. A subset of 308 sequences, from a total of 1601, were used for tree reconstruction as indicated in Materials and Methods section. Node support values calculated by the approximate likelihood-ratio test were shown on the major branches. Graphical representation of nucleotide (b) and amino acid (c) pairwise differences of each sequence in the dataset as compared to the earliest strain, a GII.4 strain collected in 1974 (AB303922). A total of 1601 nearly full-length ORF2 GII.4 sequences were included for the pairwise analyses. Black solid line indicates the linear regression line, with dotted lines showing the 95% confidence interval of the best-fit line.

Figure 2: Conservation analyses redefined antigenic sites of the major capsid protein (VP1) from GII.4 noroviruses. (a) Shannon entropy was calculated to quantify amino acid variation for each site in the VP1. Analyses were calculated with strains (1572 VP1 sequences) detected from 1995 to 2016. Data from the P domain is included here (amino acids 216-540). Residues were grouped depending on the degree of variability into: conserved sites (Shannon entropy value ≤ 0.3), sites mapping on antigenic sites (A-E), and variable sites that map outside of antigenic sites (left-side dot plot). Based on structural analyses, 14 variable residues that mapped

737 outside antigenic sites were clustered as part of novel motifs (potential antigenic sites) or
738 extension of previously defined antigenic sites (right-side dot plot). (b) Residues forming these
739 novel or expanded motifs/antigenic sites are colored accordingly on the surface of the GII.4
740 major capsid protein.
741
742 **Figure 3: Predominant sequence patterns for proposed antigenic sites correlate with GII.4**
743 **variant circulation in the human population. (a)** Amino acids from new and expanded
744 motifs/antigenic sites were tabulated using 1572 sequences of GII.4 norovirus that circulated
745 from 1995 to 2016. GII.4 variant yearly distribution was tabulated using the same sequence
746 database. Colors of the bars for the profiling graphs correspond to the predominant sequence
747 pattern presented in that antigenic site for each GII.4 variant. Colors and variant assignment
748 follows those from Fig. 1. Patterns of the bars represent minor variations of the sequences in the
749 motifs/antigenic sites. Motifs/Antigenic sites A, C, and G appear to follow the pattern of GII.4
750 variant distribution over time, implying the potential role of these sites in the emergence of new
751 pandemic GII.4 variants. (b) Number of sequence patterns of each antigenic site and (c) the
752 correlation between variant classification and mutational patterns of each motif/antigenic site
753 were calculated. The degree of correlation was assessed by adjusted Rand Index, in which higher
754 index indicates higher correlation between variant distribution and mutational pattern of the
755 motif/antigenic site.
756
757 **Figure 4: Mutational analyses on the major capsid protein (VP1) of GII.4 noroviruses**
758 **confirmed newly proposed antigenic sites C and G**. (a) Immunoassay performed with virus-
759 like particles (VLPs) and mAbs raised against the Farmington Hills 2002 strain (MD2004-3

760 [DQ658413]; WT 2004 [(11)]) (left) and the Sydney 2012 strain (RockvilleD1 [KY424328]; WT
761 2012) (right). Mean and standard deviation (SD) were calculated from duplicate wells. (b)
762 HBGA-blocking assays were performed with hyperimmune (polyclonal) sera raised against a
763 Farmington Hills strain (MD2004-3; WT 2004) and a Sydney strain (RockvilleD1; WT 2012)
764 and wild-type and mutant VLPs. Experiments were performed using sera from two guinea pigs
765 for each strain, and mean and SD were calculated in duplicate wells. (c) The half-maximal
766 binding (EC_{50}) values (sera dilution) of each wild-type and mutant VLPs was calculated from the
767 OD values from the blocking assays.
768
769 **Figure 5: Intra-variant diversity of GII.4 noroviruses reveals minimal accumulation of**
770 **mutations.** (a) Amino acid pairwise differences among viruses from each of the pandemic GII.4
771 variants as compared to the earliest strain from each given variant. (b) Intra-variant mutational
772 pattern of each antigenic site for each pandemic GII.4 variant was calculated as indicated in Fig.
773 3.
774
775 **Figure 6: Intra-variant diversification of GII.4 noroviruses is governed by stochastic**
776 **events.**
777 Maximum likelihood trees of two major GII.4 variants, Den Haag 2006b and New Orleans 2009,
778 showing the collection year (a and c) or country (b and d) for each of their strains. Phylogenetic
779 clustering of the strains did not present any pattern indicating randomness of the intra-variant
780 evolution in time and space.
781

782 **Figure 7: Diversifying pressure on GII.4 evolution. (**a) Maximum likelihood tree of all GII.4
783 variants indicates the branches under possible diversifying selection. The branches under
784 diversifying selection (empirical Bayes factor > 100 and $p < 0.05$) were explored using Mixed
785 Effect Model of Evolution method and codon positions that are located on the capsid surface are
786 represented by black branches. Diversifying selection at codon positions 352, 355, 357, 368, and
787 378 appeared on the inter-variant branches, suggesting their significant role in the emergence of
788 new GII.4 variants. (b) Predominant amino acids present in each of the sites under diversifying
789 selection for each GII.4 variant that emerged since 1995. (c) The mutational patterns and
790 adjusted Rand Index for the sites under diversifying selection were calculated as indicated in Fig.
791 3.
792

793 **Supplemental Material**

794

795 **Supplementary Figure 1: Unassigned strains which could not be clustered with any**
796 **variants in the phylogenetic tree.** (a) Six strains which could not be assigned and located
797 outside the variant clusters were indicated on the Maximum likelihood tree of GII.4 noroviruses.
798 Branches are colored based on variant determination as in Fig. 1. The table shows amino acid
799 sequences of the variable motifs/antigenic sites of those unassigned strains. The color in the table
800 corresponds to the predominant pattern presented in that antigenic site for each GII.4 variant.

801

802 **Supplementary Figure 2: Pairwise differences of GII.4 VP1 sequence database indicated by**
803 **the structural subdomains.** (a) The structural model of norovirus VP1 (PDB accession number:
804 1IHM) was rendered using UCSF Chimera (version 1.11.2). (b-d) Pairwise differences were
805 calculated and plotted as in Fig. 1b, except that sequences spanning each of the individual
806 structural subdomains of VP1 were used for the analyses. The structural subdomains of norovirus
807 VP1 are defined as the following: P2 (b), amino acids 281-415; P1 (c), N-terminal amino acids
808 216-280, C-terminal amino acids 416-540; and Shell (d), amino acids 1-215 .

809

810 **Supplementary Figure 3: Conservation analyses of the major capsid protein (VP1) from**
811 **GII.4 noroviruses.** Shannon entropy values were calculated to quantify amino acid variation for
812 each site in the VP1. The top panel presents Shannon entropy values for the P domain for all
813 GII.4 sequences available in public databases. The bottom panels present Shannon entropy
814 values for the P domain of each of the seven major GII.4 variants that emerged since 1995. Sites
815 mapping at the variable motifs/antigenic sites (A-E, G, and H) are indicated by different colors.

816

817 **Supplementary Figure 4: Amino acid mutational patterns comparing the new and old**
818 **variable motifs/antigenic sites.** (a) Mutational patterns of all the variable motifs/antigenic sites
819 (A-E, G, and H) and the previously defined original antigenic sites C (amino acids 340, 376), D
820 (amino acids 393-395), and E (amino acids 407, 411-413). GII.4 variant distribution was plotted
821 as described in Fig. 3. Amino acids from each new and expanded motifs/antigenic sites were
822 tabulated using 1572 sequences of GII.4 norovirus that circulated from 1995 to 2016. Colors of
823 the bars for the profiling graphs correspond to the predominant sequence pattern presented in that
824 motif/antigenic site for each GII.4 variant. Patterns of the bars represent minor variations of
825 sequences in the motifs/antigenic sites. The amino acid sequence patterns of each motif/antigenic
826 site were listed in the legend of each bar graph. (b) Adjusted Rand Index shows higher
827 correlation of mutational pattern of expanded antigenic site C and D compared with that of the
828 original antigenic site C and D, respectively. Mutational patterns of expanded antigenic site E
829 was less correlated with variant distribution as compared with that of the original antigenic site
830 E.

831

832 **Supplementary Figure 5: Conservation analyses of the major capsid protein (VP1) from a**
833 **randomly subsampled dataset.** (a) To account for sampling bias, Shannon entropy was re-
834 calculated to quantify amino acid variation for each site in the VP1 using randomly subsampled
835 dataset (maximum 50 strains per variant; 474 VP1 sequences detected from 1995 to 2016). Data
836 from the P domain is included here (amino acids 216-540). Residues were grouped into
837 conserved sites (Shannon entropy value $\leq$ 0.3), sites mapping on newly defined variable
838 motifs/antigenic sites (A-E, G, and H), and variable sites that mapped outside of the

motifs/antigenic sites as in Fig. 2. (b) Based on structural analyses, four residues (250, 255, 300, and 329) were mapped on the surface of the VP1 protein outside of the motifs/antigenic sites. (c) Minor variation of the sequence patterns of those four sites suggested subtle impact of these residues on the variant emergence.

Supplementary Figure 6: Mutagenesis analyses for antigenic site C mapping. Differences among a Farmington Hills strain (MD2004-3 [DQ658413]), and a Sydney strain (RockvilleD1 [KY424328]) are shown for those residues from antigenic site C. Immunoassay performed with previously uncharacterized mAbs (B11 and B12) raised against the Farmington Hills strain (MD2004-3). Data from one mAb (B12) is shown in the context of different mutant VLPs from the Farmington Hills strain (MD2004-3). Mutation at residue 340 does not affect binding while progressive reduction of binding is detected when multiple substitutions are introduced in the VLPs.

Supplementary Figure 7: Intra-variant mutational patterns of all the variable motifs/antigenic sites (A-E, G, and H) with legend. (a) Amino acid pairwise differences among viruses from each of the GII.4 variants as compared to the earliest strain from each given variant. (b) Amino acids from each new and expanded motifs/antigenic sites were tabulated for each variant from GII.4 norovirus. Colors of the bars for the profiling graphs correspond to the predominant sequence pattern presented in that motif/antigenic site for each GII.4 variant. Patterns of the bars represent minor variations of sequences in the motifs/antigenic sites. The mutational patterns of each motif/antigenic site were listed in the legend of each bar graph.

862 **Supplementary Figure 8: Diversifying pressure on GII.4 intra-variant evolution.** Maximum
863 likelihood trees of major GII.4 variants showing the diversifying pressure during the evolution.
864 The branches under diversifying selection were explored using Mixed Effect Model of Evolution
865 method (empirical Bayes factor > 100 and $p < 0.05$) and indicated in red.
866
867 **Supplementary Figure 9: Sequence pattern of NERK motif**. GII.4 variant distribution was
868 plotted as described in Fig. 3. Amino acids from NERK motif was tabulated using 1572
869 sequences of GII.4 norovirus that circulated from 1995 to 2016. Colors of the bars for the graph
870 correspond to the predominant sequence pattern presented in the NERK motif for each GII.4
871 variant.
872
873 **Supplementary Table 1: Dataset used in this study.**
874

Table 1: GII.4 variants rate of evolution

Variant	N	Detection	Mean substitution rate (subs/site/year x 10^{-3})
Grimsby 1995	71	1995 – 2002	1.57 (1.05 – 2.11)
Farmington Hills 2002	58	2002 – 2004	3.66 (2.76 – 4.63)
Hunter 2004	55	2002 – 2007	1.60 (1.05 – 2.20)
Den Haag 2006b	284	2006 – 2015	1.77 (1.50 – 2.07)
Apeldoorn 2007	55	2007 – 2011	2.76 (1.82 – 3.87)
New Orleans 2009	399	2008 – 2014	3.88 (3.41 – 4.37)
Sydney 2012	526	2010 – 2016	4.64 (4.16 – 5.18)

Figure 1

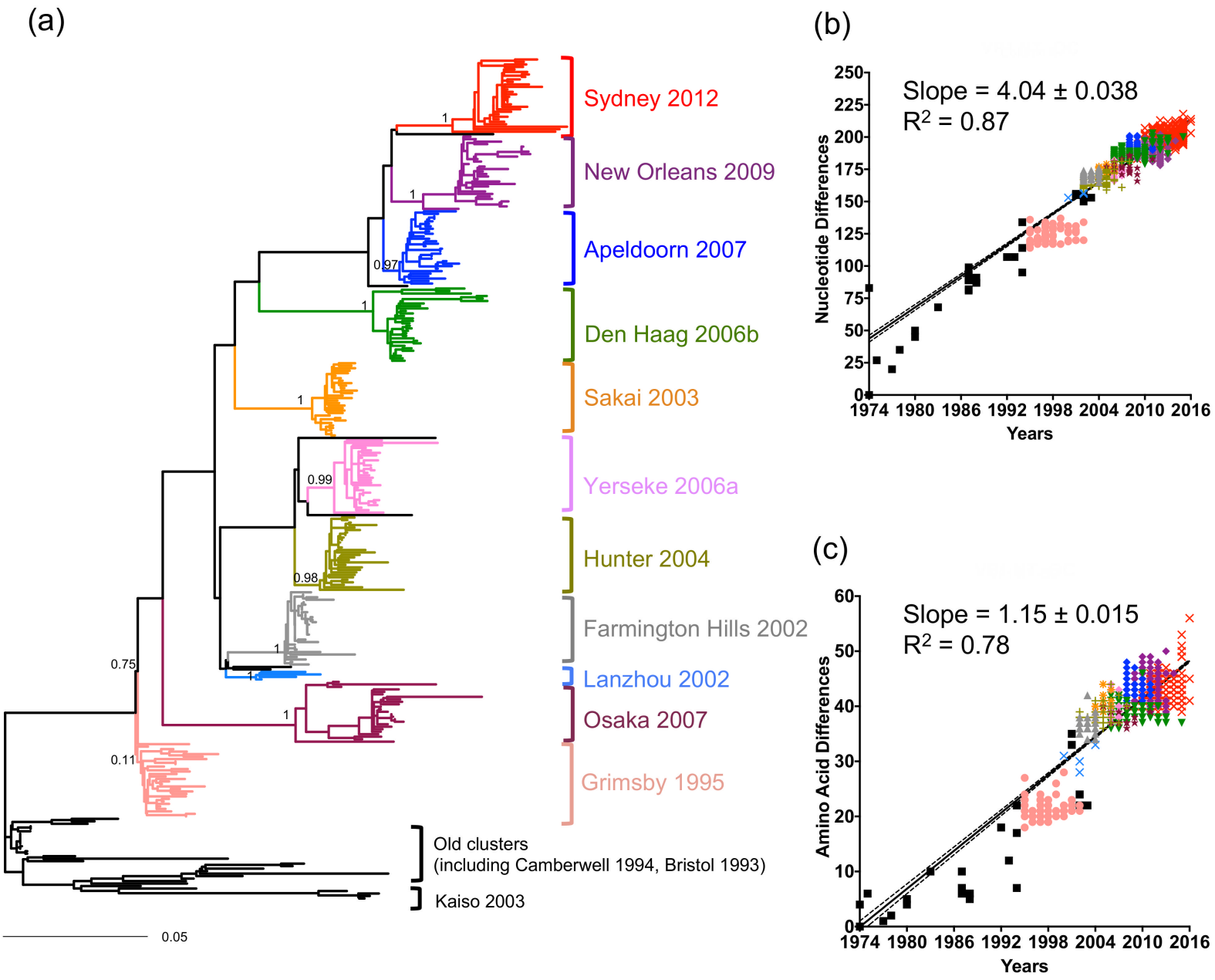
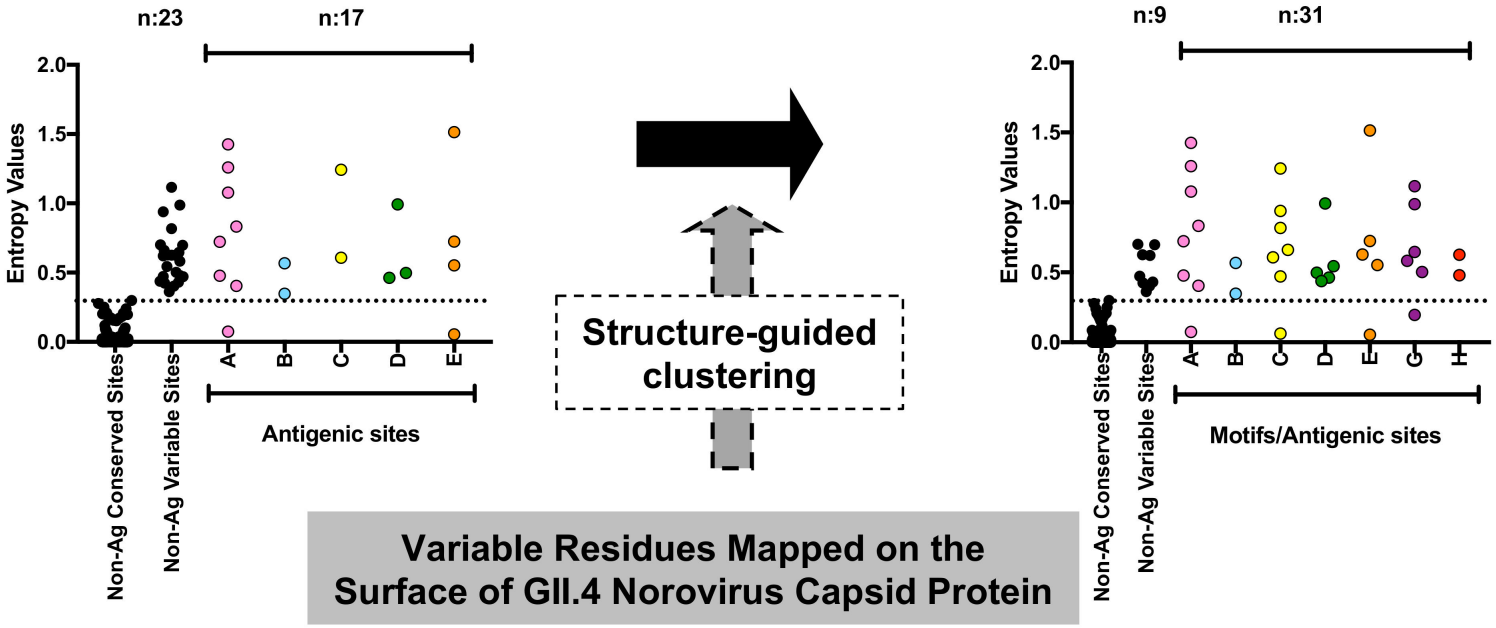
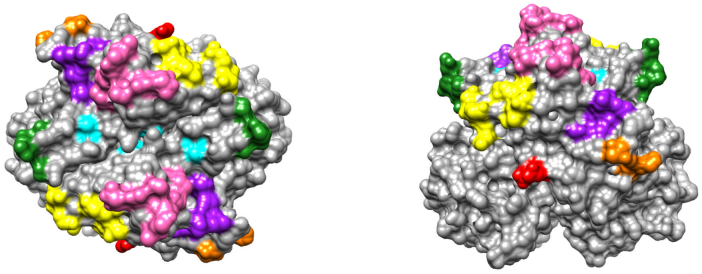


Figure 2

(a)



(b)



Motif/Antigenic site A*	Motif B	Motif/Antigenic site C†
294 295 296 297 298 368 372 373	333 389	339 340 341 375 376 377 378

Motif/Antigenic site D*	Motif/Antigenic site E*	Motif/Antigenic site G†	Motif H
393 394 395 396 397	407 411 412 413 414	352 355 356 357 359 364	309 310

*Confirmed as antigenic sites by mAbs
†Confirmed as antigenic sites by mAbs in this study

Figure 3

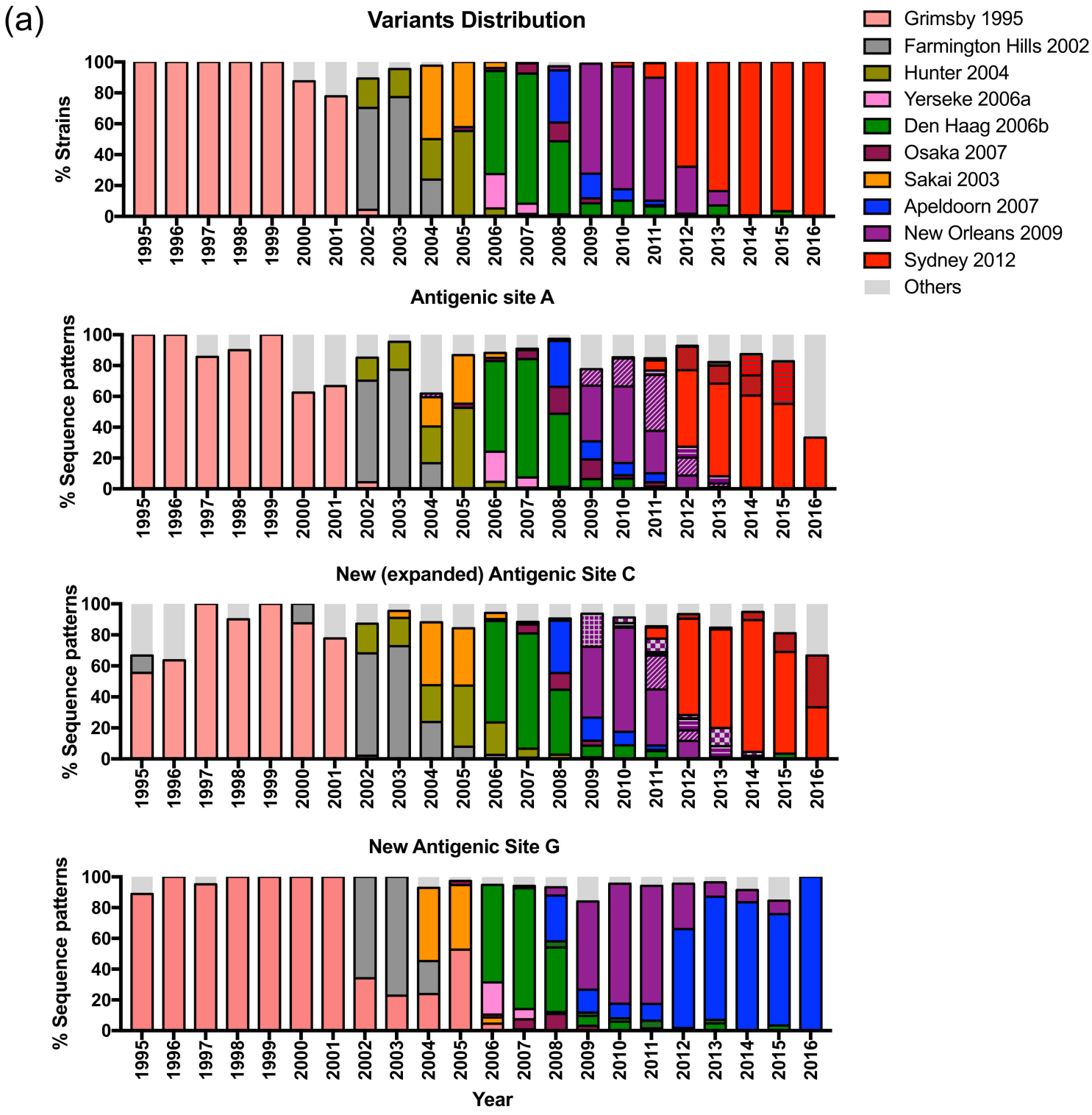
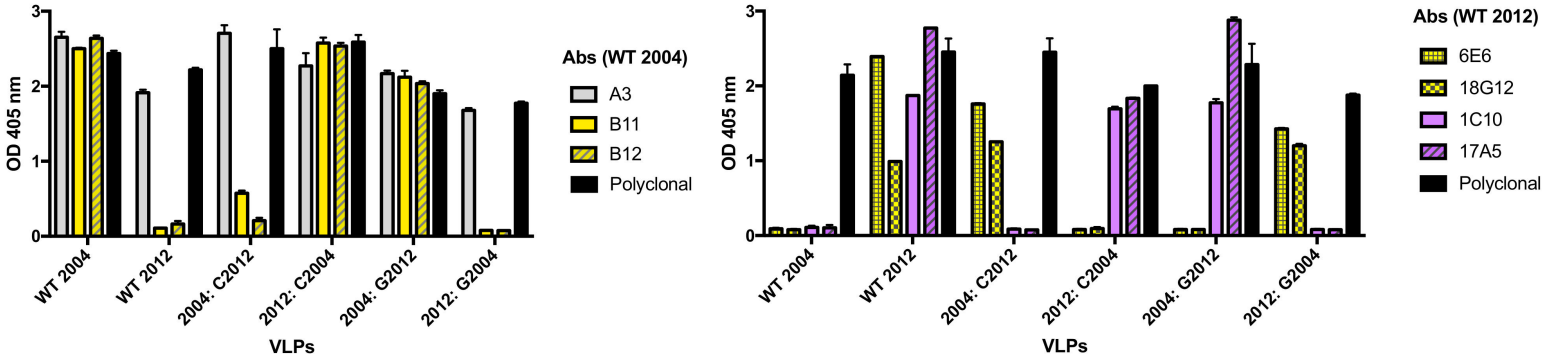
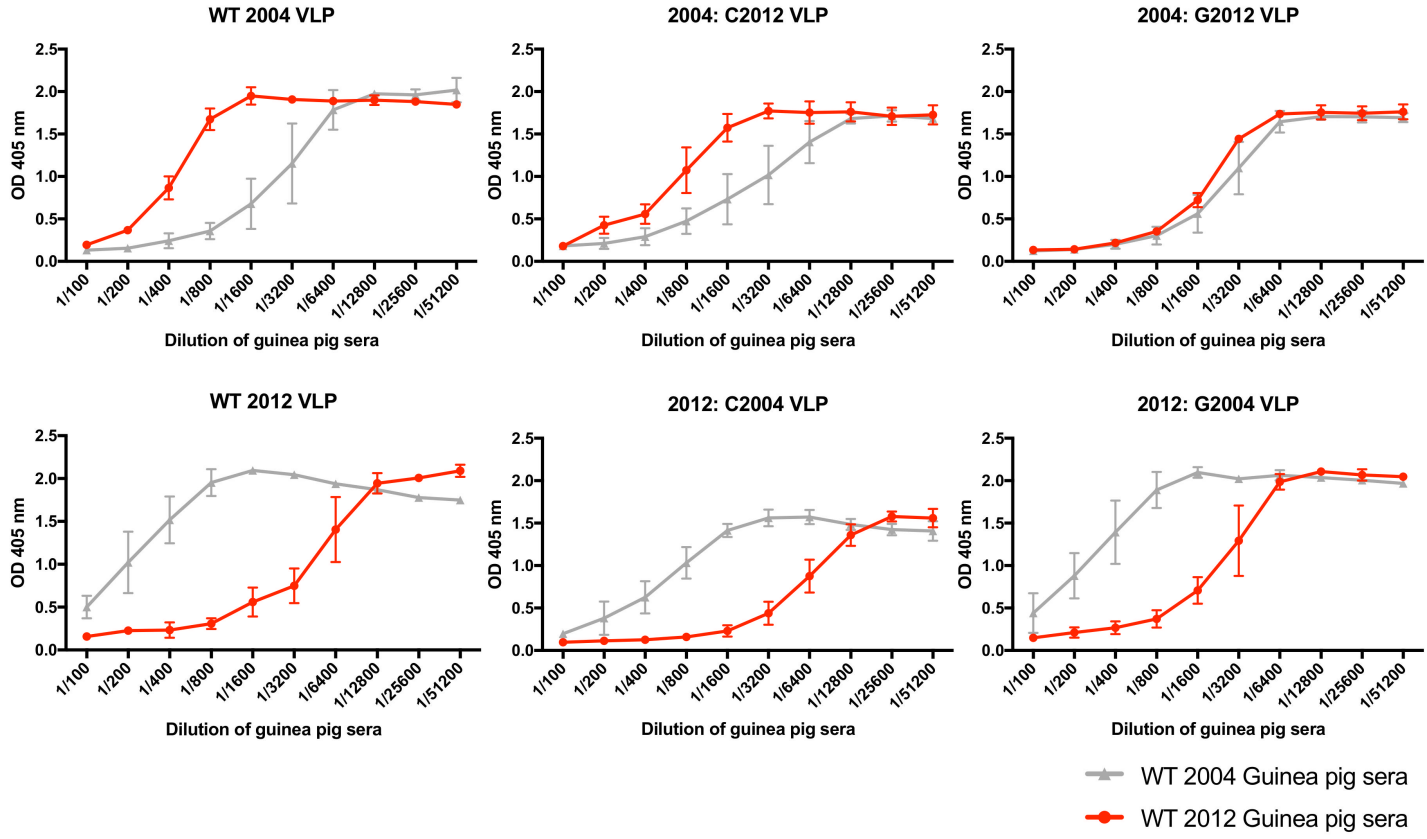


Figure 4

(a)



(b)



(c)

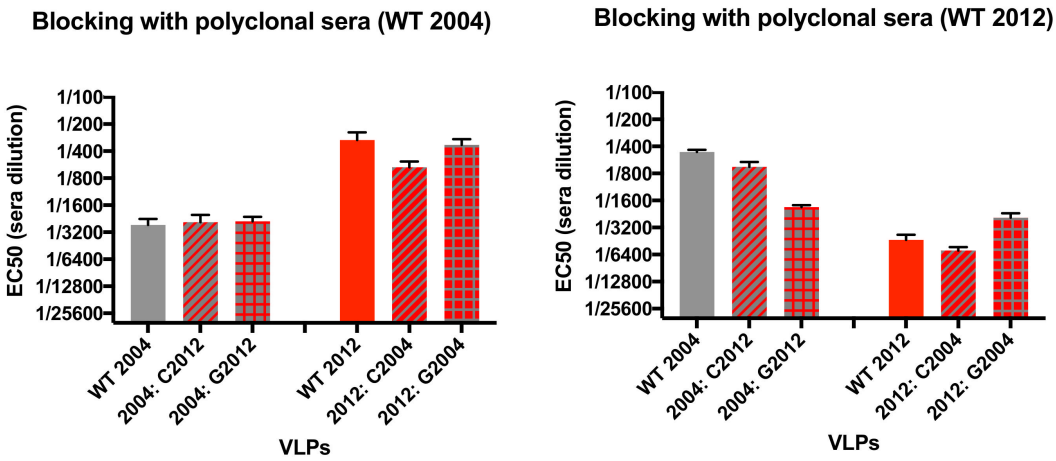


Figure 5

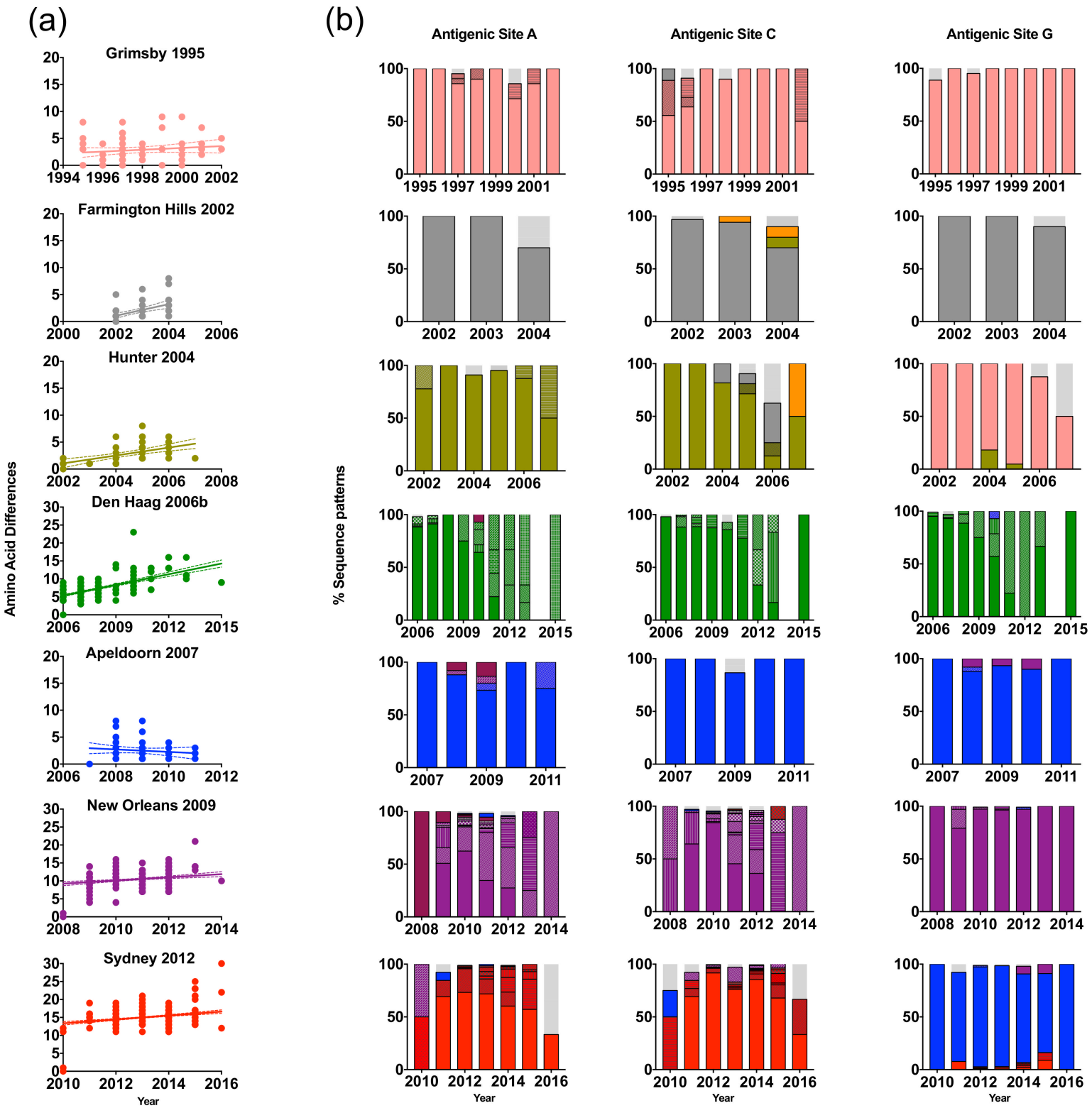
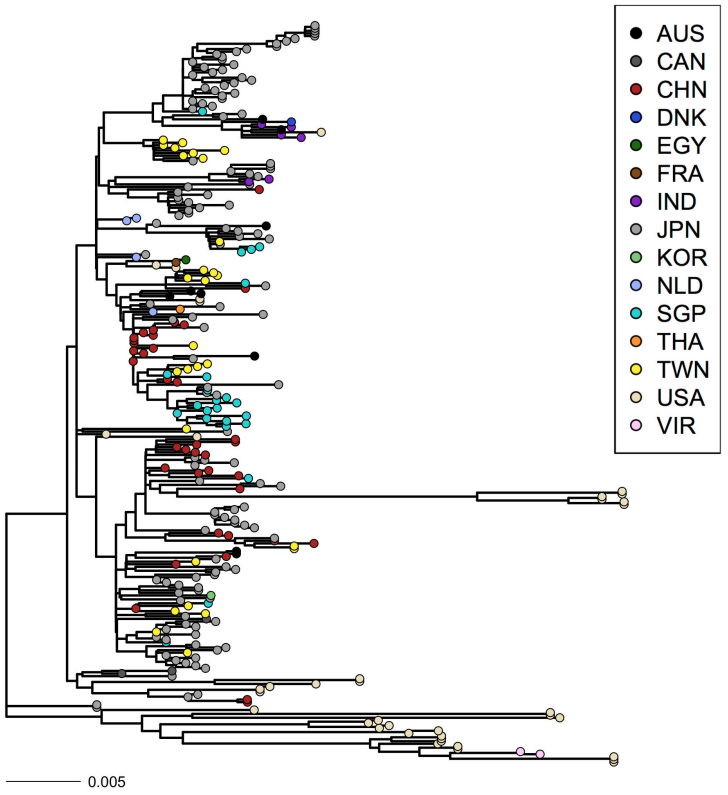


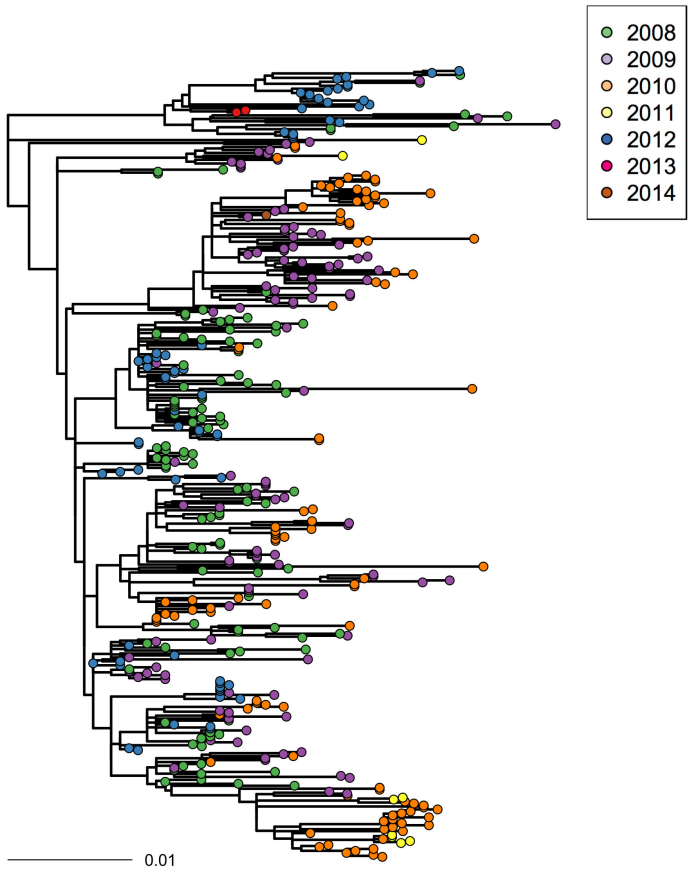
Figure 6

(a) Den Haag 2006b (Year)

(b) Den Haag 2006b (Country)



(c) New Orleans 2009 (Year)



(d) New Orleans 2009 (Country)

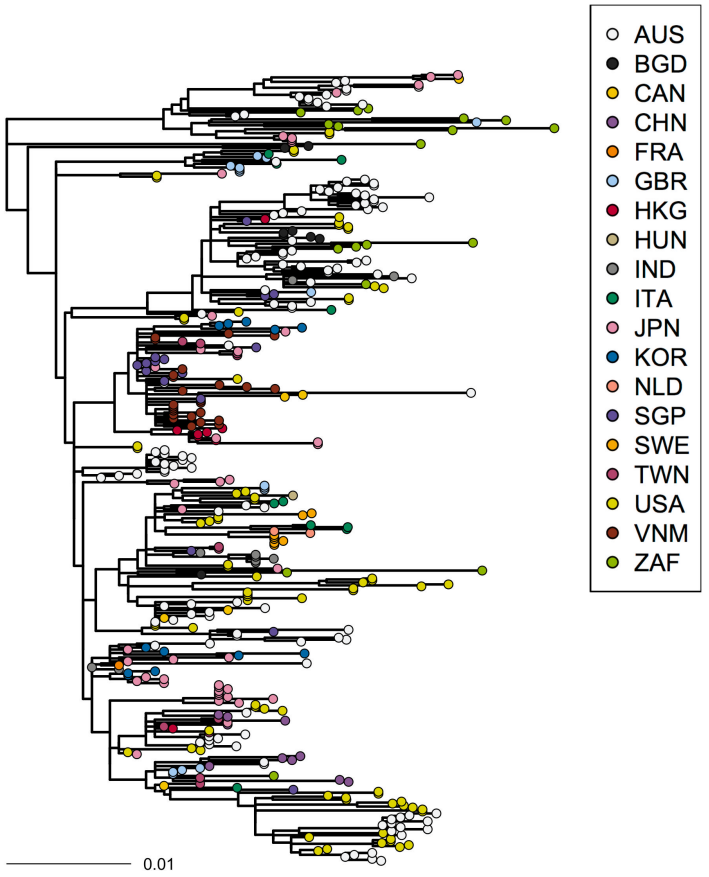
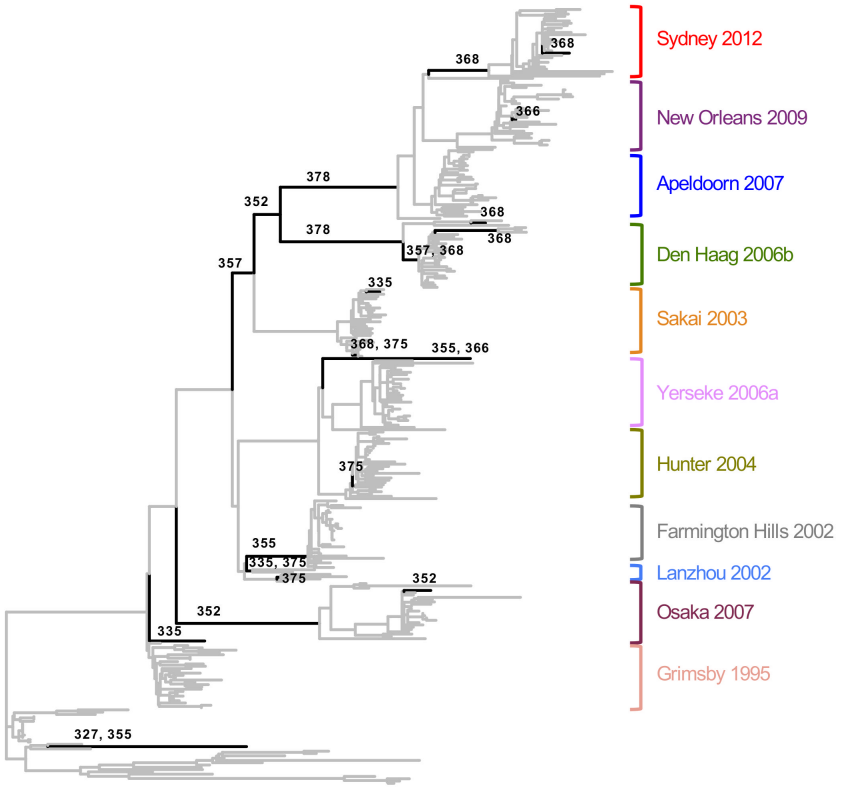


Figure 7

(a)



(b)

Variant	VP1 Amino Acid Site				
	352	355	357	368	378
Sydney 2012	Y	S	D	E	N
New Orleans 2009	Y	S	D	A	N
Apeldoorn 2007	Y	S	D	A	N
Den Haag 2006b	Y	S	P	S	H
Sakai 2003	S	S	D	A	G
Yerseke 2006a	S	S	H	S	G
Hunter 2004	S	S	H	S	G
Farmington Hills 2002	S	D	H	N	G
Osaka 2007	L	S	D	A	G
Grimsby 1995	S	S	H	T	G

(c)