

1 **Preadaptation of pandemic GII.4 noroviruses in hidden virus reservoirs years before emergence**

2 **Authors:** Christopher Ruis^{1*}, Lisa C. Lindesmith², Michael L. Mallory², Paul D. Brewer-Jensen²,
3 Josephine M. Bryant¹, Veronica Costantini³, Christopher Monit¹, Jan Vinjé³, Ralph S. Baric², Richard
4 A. Goldstein^{1*#}, Judith Breuer^{1,4#}

5 **Affiliations:**

6 ¹ Division of Infection and Immunity, University College London, London, WC1E 6BT, UK.

7 ² Department of Epidemiology, University of North Carolina, Chapel Hill, USA.

8 ³ Division of Viral Diseases, Centers for Disease Control and Prevention, Atlanta, GA, USA.

9 ⁴ Department of Microbiology, Virology and Infection Control, Great Ormond Street Hospital for
10 Children, London, UK.

11 *Correspondence to: Richard A. Goldstein

12 Cruciform Building, Gower Street, University College London, WC1E 6BT

13 +44 20 3108 2130

14 r.goldstein@ucl.ac.uk

15
16 Christopher Ruis

17 Cruciform Building, Gower Street, University College London, WC1E 6BT

18 +44 20 3108 2130

19 christopher.ruis.10@ucl.ac.uk

20
21 #These authors contributed equally to this work.
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37

38 **Abstract**

39
40 The control of pandemic pathogens depends on early prediction of pandemic variants and, more generally,
41 understanding origins of such variants and factors that drive their global spread. This is especially
42 important for GII.4 norovirus, where vaccines under development offer promise to prevent hundreds of
43 millions of annual gastroenteritis cases. Previous studies have suggested that new GII.4 pandemic viruses
44 evolve from previous pandemic variants through substitutions in the antigenic region of the VP1 protein
45 that enable evasion of host population immunity, leading to global spread. In contrast, we show here that
46 the acquisition of new genetic and antigenic characteristics is not the proximal driver of new pandemics.
47 Instead, pandemic GII.4 viruses circulate undetected for years before causing a new pandemic, during
48 which time they diversify and spread over wide geographical areas. Serological data demonstrate that by
49 2003, some nine years before it emerged as a new pandemic, the ancestral 2012 pandemic strain had
50 already acquired the antigenic characteristics that allowed it to evade prevailing population immunity
51 against the previous 2009 pandemic variant. These results provide strong evidence that viral genetic
52 changes are necessary but not sufficient for GII.4 pandemic spread. Instead, we suggest that it is changes
53 in host population immunity that enable pandemic spread of an antigenically-preadapted GII.4 variant.
54 These results indicate that predicting future GII.4 pandemic variants will require surveillance of currently
55 unsampled reservoir populations. Furthermore, a broadly acting GII.4 vaccine will be critical to prevent
56 future pandemics.

57

58 **Significance**

59 Norovirus pandemics and their associated public health and economic costs could be prevented by
60 effective vaccines. However, vaccine development and distribution will require identification of the
61 sources and drivers of new pandemics. We here use phylogenetics and serological experiments to develop
62 and test a new hypothesis of pandemic norovirus emergence. We find that pandemic noroviruses pre-
63 adapt, diversify and spread worldwide years prior to emergence, strongly indicating that genetic changes
64 are necessary but not sufficient to drive a new pandemic. We instead suggest that changes in population
65 immunity enable pandemic emergence of a pre-adapted low-level variant. These findings indicate that
66 prediction of new pandemics will require surveillance of under-sampled virus reservoirs and that
67 norovirus vaccines will need to elicit broad immunity.

68

69 **Introduction**

70 Noroviruses are the leading cause of acute gastroenteritis in humans worldwide, causing an estimated 684
71 million gastroenteritis episodes, 200,000 deaths and \$65 billion of health and societal costs annually (Kirk
72 et al. 2015; Pires et al. 2015; Bartsch et al. 2016). While more than 30 norovirus genotypes have been
73 described based on sequence variation in the VP1 capsid protein, the GII.4 genotype is responsible for the
74 majority of human cases and outbreaks and has caused six major pandemics since the mid-1990s, each
75 associated with a distinct pandemic variant: US95/96, Farmington Hills 2002, Hunter 2004, Den Haag
76 2006, New Orleans 2009 and Sydney 2012 (where the year denotes the year of onset of the respective
77 pandemic) (Kroneman et al. 2008; Siebenga et al. 2009; White 2014; Vinje 2015; Motoya et al. 2017;
78 Parra et al. 2017), as well as other geographically-limited outbreaks caused by epidemic variants
79 (Siebenga et al. 2009; Eden et al. 2014).

80

81 Vaccines currently under development offer promise to mitigate the global economic and health impact
82 of new GII.4 pandemics (Lindesmith et al. 2015). However, effective vaccine design and distribution
83 depend on understanding the sources from which new pandemic variants emerge and the factors that drive
84 their global circulation especially if, like influenza, vaccine updates are necessary. It has been proposed
85 that new pandemic GII.4 viruses generally evolve from one of the preceding pandemic variants (Siebenga
86 et al. 2007; White 2014) through acquisition of substitutions in the capsid VP1 protein that alter

87 antigenicity and enable evasion of host population immunity (Lindesmith et al. 2008; Lindesmith et al.
88 2012; Debbink et al. 2013; Lindesmith et al. 2013; Eden et al. 2014). The four most recent pandemic
89 variants caused outbreaks up to five years prior to pandemic emergence (Sdiri-Loulizi et al. 2009;
90 Siebenga et al. 2009; Eden et al. 2014; van Beek et al. 2018), leading to suggestions that both New Orleans
91 2009 and Sydney 2012 variants circulated at low levels until the acquisition of the VP1 substitutions
92 necessary for pandemic emergence (Eden et al. 2014; White 2014). Recombination has also been proposed
93 to play a role in pandemic emergence (Eden et al. 2013), although precisely how is unknown.

94

95 Here, we combine phylogenetic and serological analyses to formulate and test a new hypothesis for GII.4
96 pandemic emergence. We demonstrate that pandemic GII.4 variants arise years before pandemic spread
97 and diversify over wide geographical areas over the years prior to their emergence. In depth analysis of
98 sequence data shows that genetic substitutions and recombination events that may be important for
99 pandemic emergence are acquired years before such emergence occurs. Serological assays incorporating
100 reconstructed ancestral strains of the Sydney 2012 pandemic variant demonstrate that key antigenic
101 characteristics required for emergence had already been acquired by 2003, nine years prior to pandemic
102 spread. Together, our results show that viral genetic changes (substitutions and/or recombination events)
103 are necessary but not sufficient for pandemic spread and indicate a role for changes in host immunity in
104 the emergence of new variants.

105

106 **Results and Discussion**

107

108 To understand the origins and spread of norovirus pandemics, we reconstructed the temporal history of
109 each genomic region of GII.4. This indicates that GII.4 was present for at least 50 years prior to the first
110 documented pandemic (Figures 1, S1, Table S1), approximately 20 years earlier than previous estimates
111 (Bok et al. 2009) due to inclusion of an additional early sequence that diverges from the root of the
112 phylogeny.

113

114 The phylogenetic relationships between GII.4 RdRp, VP1 and VP2 sequences are incompatible with the
115 hypothesis that new pandemic/epidemic viruses evolve from previous pandemic/epidemic variants
116 (Figures 1, S1, Table S2). Instead, the deep phylogenetic nodes suggest that GII.4 variants diverge from
117 one another long before emerging to spread pandemically. The long tree branch lengths indicate ongoing
118 undetected circulation of pre-pandemic variants at low level during the period between divergence and
119 pandemic emergence. For example, Den Haag 2006 and New Orleans 2009 were circulating as unique
120 independent lineages by 1997 (VP1, 95% highest probability density (HPD) 1995-2000) and 2004 (VP1,
121 95% HPD 2002-2005), respectively. This undetected persistence over long time periods means that
122 multiple future pandemic/epidemic variants co-circulated simultaneously. For example, at least six
123 unsampled lineages co-circulated in the year 2000, four of which gave rise to five subsequent pandemics
124 (Figures 1, Table S2). In support of our findings, we identified and verified previously sequenced pre-
125 pandemic Farmington Hills 2002, Hunter 2004, New Orleans 2009 and Sydney 2012 and pre-epidemic
126 Osaka 2007 sequences (Figure S2, Tables S3, S4). These pre-pandemic sequences sit closer to the root of
127 the clade than sequences collected during the pandemic/epidemic (Figure S2) with placements in the tree
128 in agreement with their sampling dates, supporting our inferred ancestor and divergence dates. Pre-
129 pandemic circulation of Hunter 2004 (Sdiri-Loulizi et al. 2009), Den Haag 2006 (Siebenga et al. 2009;
130 van Beek et al. 2018), New Orleans 2009 and Sydney 2012 (Eden et al. 2014; van Beek et al. 2018) has
131 also been reported. The presence of multiple long future pandemic lineages that are only rarely sampled
132 suggests that highly diverse viral populations are circulating within reservoirs that are not included in
133 current surveillance.

134

135 We next used datasets of VP1 sequences to reconstruct a more detailed temporal history of the two most
136 recent pandemic variants, New Orleans 2009 and Sydney 2012 (Figures 2, S3). Similar to other pandemic
137 GII.4 variants (Siebenga et al. 2010), New Orleans 2009 and Sydney 2012 VP1 regions underwent a large
138 increase in relative genetic diversity coinciding with their pandemic emergence in 2009 and 2012,
139 respectively (Figure S3). However, each variant had already diversified into many lineages *prior*
140 to pandemic emergence, at least 67 (95% HPD 41-100) and 88 (95% HPD 59-113) lineages for New Orleans
141 2009 and Sydney 2012, respectively (Figures 2A-B, S3). Each of the other pandemic GII.4 variants also
142 exhibit this pre-pandemic divergence (Figure 1). Therefore, pandemic variants not only arise long before
143 the pandemic is observed, but also undergo extensive diversification into multiple related lineages that
144 circulate at low levels for years preceding pandemic emergence.
145

146 To determine the extent of circulation prior to pandemic emergence, we reconstructed the spatiotemporal
147 history of New Orleans 2009 and Sydney 2012, which demonstrated entry into each continent prior to
148 pandemic emergence (Figure 2). Specifically, New Orleans 2009 likely entered Africa, Asia and Oceania
149 more than two years prior to pandemic emergence while Sydney 2012 was likely introduced into Africa,
150 Europe and Oceania more than four years before pandemic emergence (Figure 2C-D). Following their
151 introduction, these data suggest sustained intra- and inter-continental circulation of New Orleans 2009
152 and Sydney 2012 both before (at lower levels) and after (at higher levels) pandemic emergence (Figure
153 2E-F).
154

155 The pandemic variant common ancestor date occurring years prior to pandemic emergence indicates either
156 that the important characteristics for pandemic spread were acquired years before such spread occurred
157 or that such changes were acquired convergently following diversification into multiple lineages. The
158 extent of diversification prior to pandemic emergence argues strongly against the latter scenario. Not only
159 would important changes have to occur in a large number of individual lineages located on multiple
160 continents, but these changes would have to occur approximately simultaneously after a delay of multiple
161 years. We therefore hypothesized that the key characteristics for pandemic spread are acquired by the
162 variant common ancestor years prior to pandemic emergence. To test this hypothesis, we assayed the
163 antigenic properties of two reconstructed Sydney 2012 ancestors: Sydney^{Anc}_{All}, the common ancestor of all
164 sequences genotyped as Sydney 2012 (estimated date: late 2003, 95% HPD early 2000-early 2007) and
165 Sydney^{Anc}_{Pand}, the common ancestor of all pandemic Sydney 2012 viruses (estimated date: late 2008, 95%
166 HPD late 2006-early 2010) (Figure 3A). Sydney^{Anc}_{All} and Sydney^{Anc}_{Pand} exhibit similar or greater resistance
167 to anti-New Orleans 2009 human polyclonal sera (Figure 3B), mouse polyclonal sera (Figure 3C) and
168 mouse monoclonal antibodies (mAbs, Figure S4) compared with a reference Sydney 2012 virus
169 (Sydney^{Ref}) collected during the pandemic. These data indicate that substitutions in VP1 acquired prior to
170 Sydney^{Anc}_{All} provided resistance to the anti-New Orleans 2009 antibody response at least nine years prior
171 to the onset of the Sydney 2012 pandemic and six years prior to the pandemic emergence of New Orleans
172 2009.
173

174 We identified six substitutions in the antigenic P2 domain that could have driven this antigenic change
175 (Figure 3D), including sites within (sites 294, 396 and 413) and structurally close to (site 359) known
176 epitopes (Lindesmith et al. 2012). An additional eight amino acid substitutions occurred in VP1 between
177 Sydney^{Anc}_{All} and Sydney^{Anc}_{Pand} (Figure 3A), of which sites 310 and 368 remain highly conserved within
178 Sydney 2012 (Figure S5), indicating that their acquisition by Sydney^{Anc}_{Pand} may have been important for
179 its subsequent emergence as a new pandemic. Site 368 was previously demonstrated to alter recognition
180 of mAbs raised against New Orleans 2009 (Debbink et al. 2013) while site 310 is located within the NERK
181 motif that regulates particle breathing and antibody access to epitopes (Lindesmith et al. 2014). The
182 acquisition of these substitutions by 2008 may have been important to enable pandemic emergence years
183 later, by further altering antigenicity or by increasing transmissibility, receptor binding, particle stability
184 or other properties.
185

186 We next examined the potential influence of other genomic regions on the pandemic emergence of Sydney
187 2012. It is unlikely that the nonstructural polyprotein drove Sydney 2012 pandemic spread, as this variant
188 co-circulated with the nonstructural polyprotein from the unrelated GII.Pe, GII.P4 and (more recently)
189 GII.P16 genotypes (Figure S6) (Wong et al. 2013; Ruis et al. 2017). Substitutions occurred within VP2
190 leading to Sydney^{Anc}_{All} (site 148) and Sydney^{Anc}_{Pand} (sites 158, 205) and remained highly conserved within
191 the Sydney 2012 clade (Figures 3E, S5). Without better structural or functional characterization of the
192 VP2 protein, interpreting the contribution of these changes is difficult. However, the Sydney^{Anc}_{All} and
193 Sydney^{Anc}_{Pand} VP2 proteins occurred in mid 2004 (95% HPD mid 1999-early 2009) and early 2008 (95%
194 HPD mid 2005-early 2010), respectively (Figure 3E). Therefore, as with VP1, if these substitutions were
195 important for pandemic emergence, they were acquired years earlier.

196

197 A similar process occurred for the other four pandemic GII.4 variants to have emerged since 2002, where
198 in each case, key nonsynonymous substitutions in the nonstructural polyprotein, VP1 and VP2 that
199 characterize the new pandemic variant occurred along the branch leading to the respective common
200 ancestor long before these variants spread pandemically (Tables S5, S6). The fact that the VP1
201 substitutions included sites within known blockade epitopes (Lindesmith et al. 2012) provides further
202 support for the hypothesis that the antigenic changes required for pandemic emergence were acquired
203 years before pandemic emergence actually occurred.

204

205 While recombination has previously been suggested to drive GII.4 pandemics (Eden et al. 2013), we find
206 that each recombination event in which a variant acquired a new nonstructural polyprotein or VP2
207 occurred years prior to pandemic emergence (Table S7, Supplementary text). These results indicate that
208 recombination events are not the proximate drivers of new norovirus pandemics.

209

210 Together, our results indicate that pandemic GII.4 variants arise, diversify and spread widely years before
211 they emerge to cause a pandemic. If a new pandemic was triggered by antigenic change or another viral
212 characteristic, a single lineage would rapidly increase in prevalence, as is observed for influenza A H3N2
213 (Bedford et al. 2015). Our data instead strongly support a scenario where the key antigenic and other
214 changes are acquired through substitutions and/or recombination events years before pandemic
215 emergence. Therefore the pandemic event is not proximally driven by genetic changes in any of the
216 genomic regions altering antigenicity, receptor binding or another property.

217

218 These results raise the question of what drives a variant that has been circulating widely and cryptically
219 for years to suddenly increase in frequency, dominate outbreaks worldwide and rapidly replace the
220 preceding pandemic variant. It is unlikely that stochastic factors would enable widely spread lineages
221 from a single variant within an individual genotype to emerge simultaneously. We therefore suggest that
222 norovirus pandemics are driven by a change in host factors. Given the importance of herd immunity in
223 variant emergence (Lindesmith et al. 2012; Debbink et al. 2013; Lindesmith et al. 2013), we hypothesize
224 that a shift in host population immunity (potentially driven by growing immunity against the previous
225 pandemic variant) opens a population-wide immunological niche into which the multiple circulating, but
226 hidden, lineages of the new pandemic variant can expand, having acquired the necessary antigenic
227 characteristics to do so years before. Therefore viral genetic changes are necessary but not sufficient for
228 pandemic emergence. Instead, a shift in host immunity combines with antigenic preadaptation to drive a
229 new pandemic. Vaccination against norovirus has previously been demonstrated to alter the serological
230 blockade repertoire (Lindesmith et al. 2015), supporting the notion that infection of a large number of
231 individuals can alter population-level immunity. Alternatively, variant emergence could be delayed by
232 widespread heterotypic immunity that decays faster than homotypic immunity raised against the preceding
233 pandemic variant. A natural corollary is that future pandemic GII.4 variants are continuing to circulate
234 and diversify undetected within the reservoir until such time as changes in host immunity should favor
235 the emergence of a new pandemic variant.

236

237 These results also raise the question of where pre-pandemic variants circulate over the years prior to
238 emergence. Both immunocompromised patients and animals have been mooted as a potential source of
239 pandemic GII.4 variants (Kari Debbink et al. 2014; Karst and Baric 2015). While it is possible that either
240 of these hosts could be the source of the ancestral variant, both seem unlikely to be the source of the
241 diversifying pre-pandemic lineage. It is unlikely that multiple pandemic lineages could emerge
242 simultaneously from a single immunocompromised host at pandemic onset. It is also difficult to explain
243 how multiple immunocompromised patients could form the inter-connected intercontinental transmission
244 network required for this pre-diversification, supporting recent suggestions that immunocompromised
245 patients are an unlikely reservoir (Eden et al. 2017). Additionally, while GII.4 viruses have occasionally
246 been detected in stool samples from cows, pigs and dogs (Mattison et al. 2007; Summa et al. 2012),
247 concurrent emergence of multiple lineages would require multiple zoonotic transmissions and no such
248 transmissions have been observed (Wilhelm et al. 2015). A more parsimonious explanation for our
249 findings is that pandemic GII.4 variants circulate within the community and are not detected by current
250 surveillance efforts that largely target outbreaks, predominantly in hospital and institutional settings (Inns
251 et al. 2017). More extensive co-circulation of viral lineages has been noted in influenza A H1N1 and
252 influenza B compared with influenza A H3N2 and has been correlated with slower rates of antigenic drift
253 and a lower average age of infection (Bedford et al. 2015; Vijaykrishna et al. 2015). While noroviruses
254 are prevalent in individuals of all age groups, the infection rate is highest in young children (Lopman et
255 al. 2016; O'Brien et al. 2016). In this regard, of the 16 identified pre-pandemic samples for which this
256 information is available, 15 are either from children or were sampled in a nursery or primary school
257 (Figure S2, Table S3). The ability of the ancestral Sydney 2012 VLPs to evade anti-New Orleans 2009
258 polyclonal sera raised in mice that have not previously been exposed to norovirus (Figure 3C) suggests
259 that these viruses would have been able to evade immunity raised in young children. Thus, while
260 continued strain monitoring, as captured by norovirus outbreak surveillance networks such as CaliciNet
261 (Vega et al. 2011) and NoroNet (van Beek et al. 2018), has significant value for vaccine development,
262 additional efforts should focus on identifying potential reservoirs from which future pandemic norovirus
263 variants could emerge, including symptomatic and asymptomatic children (Rouhani et al. 2016) in
264 healthcare and community settings. The antigenic preadaptation of future pandemic variants suggests that
265 such a surveillance system combined with antigenic testing could efficiently predict currently low-level
266 variants that might become pandemic in the future. It will also be key to understand how these variants
267 interact with prevailing population immunity.

268

269 Our results also have important implications for current efforts to develop norovirus vaccines. Should the
270 new pandemic variant emerge from the preceding variant, a vaccine targeting the current variant may
271 prevent the next pandemic. However, under our hypothesis, a vaccine that boosts immunity to the current
272 variant may actually hasten emergence of the next pandemic. It is therefore essential that norovirus
273 vaccines provide broad immunity against GII.4 viruses (Lindesmith et al. 2015).

274

275 **Materials and Methods**

276

277 *Reconstruction of the temporal history of GII.4 norovirus*

278

279 We collected all norovirus sequences available on GenBank as of 30/10/2015 containing the complete
280 RNA-dependent RNA polymerase (RdRp), VP1 and VP2 genome regions. Each sequence was genotyped
281 using the norovirus genotyping tool (Kroneman et al. 2011) and the 871 sequences containing the GII.4
282 VP1 were retained. Due to inter-genotype recombination events close to the ORF1-ORF2 boundary (Eden
283 et al. 2013), the dataset contains sequences with the GII.P1, GII.P4, GII.P12 and GII.Pe RdRps.

284

285 Due to the presence of recombination close to ORF boundaries, we ran all analyses on the RdRp, VP1
286 and VP2 separately. Each genome region was aligned at the amino acid level using MUSCLE (Edgar
287 2004). We screened for the presence of intra-genic recombination in the RdRp, VP1 and VP2 separately
288 using the Single Breakpoint (SBP) method implemented in HyPhy (Pond et al. 2006), identifying 19
289 sequences as potential recombinants in VP1 (Table S8). These sequences clustered differently with strong
290 support on either side of the putative breakpoint in phylogenetic trees reconstructed with RAxML v8.1
291 (Stamatakis 2014) with the GTR model and gamma rate heterogeneity with four gamma classes. SBP was
292 then run on the alignment again following removal of these samples to ensure the recombination signal
293 had been removed. A summary of the number of remaining sequences from each GII.4 variant and each
294 RdRp genotype is shown in Table S9. The variant names used here are those returned by the norovirus
295 genotyping tool (Kroneman et al. 2011).

296
297 Methods employing sequence data and sampling dates to infer divergence times are valid only if there is
298 a temporal evolutionary signal in the dataset (Rambaut et al. 2016). To assess whether each of our datasets
299 exhibits a temporal evolutionary signal, we reconstructed a maximum likelihood tree using RAxML, as
300 above. We identified the best-fitting root position using TempEst v1.5 (Rambaut et al. 2016) and
301 calculated the R² correlation between root-to-tip distance and sampling date (Figure S7). There was a
302 significant temporal evolutionary signal within each dataset ($p < 0.001$).

303
304 We reconstructed the temporal evolutionary history of each genomic region using the Bayesian Markov
305 chain Monte Carlo approach implemented in BEAST version 2.2.1 (Bouckaert et al. 2014). Analyses were
306 run independently on the RdRp, VP1 and VP2. As the Den Haag 2006 and New Orleans 2009 variants
307 have a greater number of sequences compared with other variants (Table S9), we took three random
308 subsamples of 41 sequences from each of these variants (with 41 chosen to match the number of sequences
309 in the third most numerous variant); results were insensitive to the subsampled dataset (Table S10). Each
310 sequence was labelled with the most accurate collection date possible: the day of collection if available,
311 the month of collection if the day of collection was not available or the year of collection if the month of
312 collection was not available. Each dataset was analyzed using the GTR substitution model with gamma
313 rate heterogeneity and partitioned so codon positions 1 and 2 shared a substitution model and codon
314 position 3 had a different substitution model. We employed both the strict and relaxed lognormal clock
315 models to examine variation in the substitution rate within each dataset. We used a lognormal prior
316 distribution for each dataset with mean 4.3×10^{-3} substitutions/site/year and standard deviation 0.1 for the
317 VP1 (Bok et al. 2009) and with mean 4.32×10^{-3} substitutions/site/year and standard deviation 0.1 for the
318 RdRp (Siebenga et al. 2010). At the time of our analysis, there was no previously published substitution
319 rate for VP2 across the GII.4 clade. We therefore employed the same prior as for the VP1 dataset. We
320 applied a coalescent Bayesian skyline tree prior. Three replicate runs with different starting values were
321 performed for each dataset and clock model and run until convergence, as assessed using Tracer v1.6
322 (Rambaut et al. 2014). The replicate runs were combined with removal of suitable burnin using
323 LogCombiner v2.2.1 and maximum clade credibility trees were obtained using TreeAnnotator v2.2.1. In
324 each case there was strong support to reject the strict clock model in favor of the relaxed lognormal clock
325 model. Therefore the results employing the relaxed lognormal clock model were used for all further
326 analyses.

327
328 We estimated variant divergence dates and recombination dates by combining the posterior distribution
329 of trees from each subsampled dataset into a single posterior distribution. We inferred variant divergence
330 dates by calculating the date of the most recent common ancestor between each pair of variants in each
331 tree in this posterior distribution. To calculate the date of each recombination event, we identified the
332 mean and 95% HPD of the distribution of branch start and end times of the corresponding branch in each
333 tree in the posterior distribution.

334
335 *Identification of pre-pandemic and pre-epidemic GII.4 sequences*

336

337 We defined pre-pandemic/pre-epidemic sequences as those that cluster with a GII.4 variant but were
338 collected prior to the year in which that variant emerged pandemically or epidemically. We genotyped all
339 GII.4 VP1 sequences present on GenBank containing more than 400 nucleotides (Kroneman et al. 2011).
340 We identified 50 sequences with a reported collection date earlier than the year of pandemic/epidemic
341 emergence of the respective variant. We estimated the collection date of each of these sequences using
342 BEAST v2.4.2 (Bouckaert et al. 2014), assuming a uniform prior distribution with minimum 1974.5 and
343 maximum 2015.446575 (the dates of the earliest and latest collected sequences in the main dataset). The
344 95% HPD of the estimated collection date overlapping with the reported collection date was taken as
345 evidence to support, but not confirm, the reported collection date.

346 347 *Reconstruction of the temporal history of New Orleans 2009 and Sydney 2012*

348
349 We compiled expanded datasets from GenBank containing 460 and 533 P2 domain sequences for New
350 Orleans 2009 and Sydney 2012, respectively. Each dataset was aligned and inferred to exhibit a temporal
351 evolutionary signal ($p < 0.001$) as above. We reconstructed the evolutionary dynamics of New Orleans
352 2009 and Sydney 2012 independently using BEAST v2.2.1 (Bouckaert et al. 2014), using the HKY
353 nucleotide substitution model with four gamma classes. We employed a lognormal prior on the
354 substitution rate with mean 6.83×10^{-3} and standard deviation 0.1 to accommodate the mean and 95% HPD
355 of our estimate of the substitution rate across the complete GII.4 VP1 clade. We applied the strict and
356 relaxed lognormal clock models and found strong support (\log_{10} Bayes factor > 100) to reject the strict
357 clock model in favor of the relaxed lognormal clock model in each case. We therefore used the results
358 from the relaxed lognormal clock runs in all further analyses. Bayesian skyline plots were reconstructed
359 using Tracer v1.5. The date at which the Bayesian skyline plot exhibits a large increase in relative genetic
360 diversity was used as the time of pandemic onset. The number of lineages present at the onset of the
361 pandemic was calculated as the number of lineages present at this point in time in each tree in the posterior
362 distribution.

363 364 *Reconstruction of the spatiotemporal history of New Orleans 2009 and Sydney 2012*

365
366 We collected datasets containing all available VP1 sequences on GenBank as of 09/02/2017 from New
367 Orleans 2009 ($n=565$) and Sydney 2012 ($n=708$). Each dataset exhibited temporal signal as above.
368 Examination of the sampling locations showed that there was typically only a small number of sequences
369 from each country (Table S11). We therefore used the continent of collection as the location label. The
370 New Orleans 2009 dataset contained a large number of sequences from Asia and Oceania relative to the
371 other continents, while the Sydney 2012 dataset contained a large number of sequences from Asia relative
372 to the other continents (Table S11). Should sequences from the same continent cluster together within the
373 tree, an excess of sequences from one continent is unlikely to alter estimates of ancestral locations.
374 However, if sequences from the different continents are typically interspersed within the tree, an excess
375 of sequences from one continent could result in artifactual support for this continent being the location of
376 ancestral nodes. The sequences from each continent are interspersed throughout the tree in both New
377 Orleans 2009 and Sydney 2012 (Figure S8). We therefore randomly down-sampled New Orleans 2009
378 sequences from Asia and Oceania and Sydney 2012 sequences from Asia to match the number of
379 sequences from the next most commonly represented continent. We carried out three random subsamples
380 and performed all analyses on each subsampled dataset. All results were insensitive to the subsampled
381 dataset.

382
383 We used discrete phylogeography (Lemey, Philippe et al. 2009) implemented in BEAST v2.4.2 (Bouckaert
384 et al. 2014) to reconstruct the spatiotemporal history of New Orleans 2009 and Sydney 2012. Sequences
385 were labelled with the most accurate collection date possible. We modelled the nucleotide substitution
386 process using the HKY substitution model and gamma rate heterogeneity with four gamma classes. We
387 applied both the strict and relaxed lognormal clock models. In each case there was strong support to reject
388 the strict clock model in favor of the relaxed lognormal clock model (\log_{10} Bayes factor 66-83) and we

389 therefore used the relaxed lognormal clock model for our inferences. However, the results with the strict
390 clock model are qualitatively very similar, indicating that potential over-parameterization due to the large
391 number of branch-specific rates with the relaxed lognormal clock model has not influenced our results.
392 We applied a lognormal prior on the substitution rate with mean 7×10^{-3} for New Orleans 2009 and 6.4×10^{-3}
393 for Sydney 2012 and standard deviation 0.1 in each case, based on the posterior estimates of the variant
394 substitution rates inferred previously. We employed a Bayesian coalescent skyline tree prior for each
395 dataset. We applied a discrete phylogeographic model to describe lineage migrations within each dataset
396 (24), consisting of a symmetric transition matrix for the migration rates and a set of Bayesian stochastic
397 search variable selection (BSSVS) indicator variables. We assumed a Poisson prior for the number of ‘on’
398 BSSVS variables with mean 5 for the New Orleans 2009 datasets and mean 4 for the Sydney 2012
399 datasets. We used an exponential prior with mean 1.0 migration rate per lineage per year for the overall
400 rate of geographical transition and a gamma prior with shape 1.0 and scale 1.0 for each of the relative
401 geographical transition rates. Three replicate runs with different starting parameters were carried out with
402 each dataset and run until convergence, as assessed using Tracer v1.6 (Rambaut et al. 2014). Runs from
403 each subsampled dataset were combined into a single posterior distribution for downstream analyses.

404
405 We identified the date of first import into each continent by calculating either the root date if the continent
406 was inferred to be the root location or the earliest branch midpoint where the downstream node was
407 inferred to be within the continent. We calculated this date within each tree in the posterior distribution.
408 We therefore assume that migration events occurred at the midpoint of the branch. We obtain similar
409 results using the earliest non-root node inferred to be within the continent, which assumes that migration
410 occurs at the end of the branch. We used the program posterior analysis of coalescent trees (PACT) v0.9.4
411 to compute the proportion of lineages present on each continent through time.

412 413 *Reconstruction of ancestral Sydney 2012 viruses and identification of substitutions leading to each GII.4* 414 *variant*

415
416 We collected a dataset containing all 2198 available GII.4 VP1 sequences, including sequences from all
417 of the major GII.4 variants (Table S12). Alignment and phylogenetic reconstruction was carried out as
418 above. We used RAxML to optimize branch lengths within the maximum likelihood phylogenetic tree
419 and ten bootstrap tree topologies using the amino acid alignment and the WAG substitution model with
420 optimized base frequencies. We used multiple tree topologies to assess the influence of tree topology on
421 our inferences. We carried out ancestral reconstruction at the amino acid level with PAML v4.9 (Yang
422 2007) using the WAG substitution model and optimized base frequencies. The ancestral sequence at
423 Sydney^{Anc}_{All} and Sydney^{Anc}_{Pand} was identical with each tree topology and the residue inferred at each site
424 was supported by posterior probability > 0.95. We identified VP1 substitutions by comparing the sequence
425 of the variant root ancestor with the sequence of the immediately upstream node in each tree.

426
427 To identify substitutions with the nonstructural polyprotein and VP2 we collected datasets containing all
428 available GII and GIV nonstructural polyprotein sequences (Table S13) and all available GII.4 VP2
429 sequences (Table S12), respectively. We identified substitutions leading to each GII.4 variant using the
430 same process described for VP1 above.

431 432 *Surrogate neutralization assay (Antibody blockade assay)*

433
434 Sydney^{Anc}_{All} and Sydney^{Anc}_{Pand} VP1 genes were codon optimized for mammalian expression and
435 synthesized by Bio Basic Inc. (Amherst, NY) and VLPs were expressed in baby hamster kidney cells from
436 Venezuelan equine encephalitis virus replicons (11). 0.25 µg/ml VLP was pretreated with decreasing
437 concentrations of antibody/serum for 1 h before addition to pig gastric mucin type III (PGM, Sigma
438 Aldrich, St. Louis, MO) coated plates for 1 h. Bound VLP was detected with anti-GII.4 2012 ref rabbit
439 hyperimmune sera. Percent control binding is defined as the binding with antibody/serum pretreatment
440 compared to the binding without multiplied by 100. All incubations were done at room temperature. The

441 blockade data were fit using sigmoidal dose-response curve analysis of nonlinear data in GraphPad Prism
442 702 and IC50 titers with 95% confidence intervals calculated. Antibodies that did not block 50% of
443 binding at the highest dilution tested were assigned an IC50 of 2 times the assay limit of detection for
444 statistical comparison (11). Anti-New Orleans 2009 polyclonal sera was collected from ten patients
445 naturally infected during a New Orleans 2009 outbreak in a long-term care facility in March 2010.
446 Infection was confirmed by symptoms and norovirus detection in acute stool. The human serum samples
447 were collected from a study that was approved by institutional review board of the Centers for Disease
448 Control and Prevention (CDC Protocol #5051). Mouse anti-GII.4 New Orleans 2009^{Ref} sera were
449 generated in immunized mice as described in (K. Debbink et al. 2014). Mouse monoclonal antibodies to
450 GII.4 New Orleans 2009^{Ref} (Lindesmith et al. 2013) were generated by VLP immunization.

451

452 *Data availability*

453 All alignments, phylogenetic trees and BEAST XML files will be made available on GitHub on
454 acceptance of the manuscript. VLPs are available from the authors.

455

456 **References**

- 457 Bartsch SM, Lopman BA, Ozawa S, Hall AJ, Lee BY. 2016. Global economic burden of norovirus
458 gastroenteritis. *PLoS One* 11.
- 459 Bedford T, Riley S, Barr IG, Broor S, Chadha M, Cox NJ, Daniels RS, Gunasekaran CP, Hurt AC,
460 Kelso A, et al. 2015. Global circulation patterns of seasonal influenza viruses vary with antigenic
461 drift. *Nature* 523:217–220.
- 462 van Beek J, de Graaf M, Al-Hello H, Allen DJ, Ambert-Balay K, Botteldoorn N, Brytting M, Buesa J,
463 Cabrerizo M, Chan M, et al. 2018. Molecular surveillance of norovirus, 2005-16: an
464 epidemiological analysis of data collected from the NoroNet network. *Lancet. Infect. Dis.* 0.
- 465 Bok K, Abente EJ, Realpe-Quintero M, Mitra T, Sosnovtsev S V, Kapikian AZ, Green KY. 2009.
466 Evolutionary dynamics of GII.4 noroviruses over a 34-year period. *J. Virol.* 83:11890–11901.
- 467 Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond
468 AJ. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput.*
469 *Biol.*
- 470 Debbink K, Lindesmith LC, Donaldson EF, Costantini V, Beltramello M, Corti D, Swanstrom J,
471 Lanzavecchia A, Vinje J, Baric RS. 2013. Emergence of New Pandemic GII.4 Sydney Norovirus
472 Strain Correlates With Escape From Herd Immunity. *J. Infect. Dis.* 208:1877–1887.
- 473 Debbink K, Lindesmith LC, Donaldson EF, Swanstrom J, Baric RS. 2014. Chimeric GII.4 Norovirus
474 Virus-Like-Particle-Based Vaccines Induce Broadly Blocking Immune Responses. *J. Virol.*
- 475 Debbink K, Lindesmith LC, Ferris MT, Swanstrom J, Beltramello M, Corti D, Lanzavecchia A, Baric
476 RS. 2014. Within-host evolution results in antigenically distinct GII.4 noroviruses. *J. Virol.*
477 88:7244–7255.
- 478 Eden J-S, Chisholm RH, Bull RA, White PA, Holmes EC, Tanaka MM. 2017. Persistent infections in
479 immunocompromised hosts are rarely sources of new pathogen variants. *Virus Evol.* 3:219–222.
- 480 Eden J-S, Hewitt J, Lim KL, Boni MF, Merif J, Greening G, Ratcliff RM, Holmes EC, Tanaka MM,
481 Rawlinson WD, et al. 2014. The emergence and evolution of the novel epidemic norovirus GII.4
482 variant Sydney 2012. *Virology* 450–451:106–113.
- 483 Eden J-S, Tanaka MM, Boni MF, Rawlinson WD, White PA. 2013. Recombination within the pandemic
484 norovirus GII.4 lineage. *J. Virol.* 87:6270–6282.
- 485 Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput.

- 486 Nucleic Acids Res.
- 487 Inns T, Harris J, Vivancos R, Iturriza-Gomara M, O'Brien S. 2017. Community-based surveillance of
488 norovirus disease: a systematic review. *BMC Infect. Dis.* 17:657.
- 489 Karst SM, Baric RS. 2015. What is the reservoir of emergent human norovirus strains? *J. Virol.*
490 89:5756–5759.
- 491 Kirk MD, Pires SM, Black RE, Caipo M, Crump JA, Devleeschauwer B, Döpfer D, Fazil A, Fischer-
492 Walker CL, Hald T, et al. 2015. World Health Organization Estimates of the Global and Regional
493 Disease Burden of 22 Foodborne Bacterial, Protozoal, and Viral Diseases, 2010: A Data Synthesis.
494 *PLoS Med.* 12.
- 495 Kroneman A, Vennema H, Deforche K, Avoort H, Peñaranda S, Oberste MS, Vinjé J, Koopmans M.
496 2011. An automated genotyping tool for enteroviruses and noroviruses. *J. Clin. Virol.*
- 497 Kroneman A, Verhoef L, Harris J, Vennema H, Duizer E, Van Duynhoven Y, Gray J, Iturriza M,
498 Böttiger B, Falkenhorst G, et al. 2008. Analysis of integrated virological and epidemiological
499 reports of norovirus outbreaks collected within the Foodborne Viruses in Europe network from 1
500 July 2001 to 30 June 2006. *J. Clin. Microbiol.* 46:2959–2965.
- 501 Lemey, Philippe, Rambaut, Andrew, Drummond, Alexei J., Suchard, Marc A. 2009. Bayesian
502 phylogeography finds its roots. *PLoS Comput. Biol.*
- 503 Lindesmith LC, Beltramello M, Donaldson EF, Corti D, Swanstrom J, Debbink K, Lanzavecchia A,
504 Baric RS. 2012. Immunogenetic mechanisms driving norovirus GII.4 antigenic variation. *PLoS*
505 *Pathog.* 8.
- 506 Lindesmith LC, Costantini V, Swanstrom J, Debbink K, Donaldson EF, Vinjé J, Baric RS. 2013.
507 Emergence of a norovirus GII.4 strain correlates with changes in evolving blockade epitopes. *J.*
508 *Virol.* 87:2803–2813.
- 509 Lindesmith LC, Donaldson EF, Beltramello M, Pintus S, Corti D, Swanstrom J, Debbink K, Jones TA,
510 Lanzavecchia A, Baric RS. 2014. Particle Conformation Regulates Antibody Access to a
511 Conserved GII.4 Norovirus Blockade Epitope. *J. Virol.* 88:8826–8842.
- 512 Lindesmith LC, Donaldson EF, Lobue AD, Cannon JL, Zheng D-P, Vinje J, Baric RS. 2008.
513 Mechanisms of GII.4 norovirus persistence in human populations. *PLoS Med.* 5:e31.
- 514 Lindesmith LC, Ferris MT, Mullan CW, Ferreira J, Debbink K, Swanstrom J, Richardson C, Goodwin
515 RR, Baehner F, Mendelman PM, et al. 2015. Broad Blockade Antibody Responses in Human
516 Volunteers after Immunization with a Multivalent Norovirus VLP Candidate Vaccine:
517 Immunological Analyses from a Phase I Clinical Trial. *PLoS Med.* 12.
- 518 Lopman BA, Steele D, Kirkwood CD, Parashar UD. 2016. The Vast and Varied Global Burden of
519 Norovirus: Prospects for Prevention and Control. *PLoS Med.* 13.
- 520 Mattison K, Shukla A, Cook A, Pollari F, Friendship R, Kelton D, Bidawid S, Farber JM. 2007. Human
521 noroviruses in swine and cattle. *Emerg. Infect. Dis.* 13:1184–1188.
- 522 Motoya T, Nagasawa K, Matsushima Y, Nagata N, Ryo A, Sekizuka T, Yamashita A, Kuroda M,
523 Morita Y, Suzuki Y, et al. 2017. Molecular evolution of the VP1 gene in human norovirus GII.4
524 variants in 1974-2015. *Front. Microbiol.*
- 525 O'Brien SJ, Donaldson AL, Iturriza-Gomara M, Tam CC. 2016. Age-Specific Incidence Rates for
526 Norovirus in the Community and Presenting to Primary Healthcare Facilities in the United
527 Kingdom. *J. Infect. Dis.* 213:S15–S18.
- 528 Parra GI, Squires RB, Karangwa CK, Johnson JA, Lepore CJ, Sosnovtsev S V., Green KY. 2017. Static
529 and Evolving Norovirus Genotypes: Implications for Epidemiology and Immunity. *PLoS Pathog.*
530 13.

- 531 Pires SM, Fischer-Walker CL, Lanata CF, Devleeschauwer B, Hall AJ, Kirk MD, Duarte ASR, Black
532 RE, Angulo FJ. 2015. Aetiology-specific estimates of the global and regional incidence and
533 mortality of diarrhoeal diseases commonly transmitted through food. *PLoS One* 10.
- 534 Pond SLK, Posada D, Gravenor MB, Woelk CH, Frost SDW. 2006. Automated phylogenetic detection
535 of recombination using a genetic algorithm. *Mol. Biol. Evol.*
- 536 Rambaut A, Lam TT, Max Carvalho L, Pybus OG. 2016. Exploring the temporal structure of
537 heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.*
- 538 Rambaut A, Surchard MA, Xie D, Drummond AJ. 2014. Tracer v1.6. Available from
539 <http://beast.bio.ed.ac.uk/Tracer>.
- 540 Rouhani S, Peñataro Yori P, Paredes Olortegui M, Siguas Salas M, Rengifo Trigos D, Mondal D,
541 Bodhidatta L, Platts-Mills J, Samie A, Kabir F, et al. 2016. Norovirus infection and acquired
542 immunity in 8 countries: Results from the MAL-ED study. *Clin. Infect. Dis.* 62:1210–1217.
- 543 Ruis C, Roy S, Brown JR, Allen DJ, Goldstein RA, Breuer J. 2017. The emerging GII.P16-GII.4
544 Sydney 2012 norovirus lineage is circulating worldwide, arose by late-2014 and contains
545 polymerase changes that may increase virus transmission. *PLoS One* 12.
- 546 Sdiri-Loulizi K, Ambert-Balay K, Gharbi-Khelifi H, Sakly N, Hassine M, Chouchane S, Guediche MN,
547 Pothier P, Aouni M. 2009. Molecular epidemiology of norovirus gastroenteritis investigated using
548 samples collected from children in Tunisia during a four-year period: Detection of the norovirus
549 variant GGII.4 hunter as early as January 2003. *J. Clin. Microbiol.* 47:421–429.
- 550 Siebenga JJ, Lemey P, Pond SLK, Rambaut A, Vennema H, Koopmans M. 2010. Phylodynamic
551 reconstruction reveals norovirus GII.4 epidemic expansions and their molecular determinants.
552 *PLoS Pathog.* 6:1–13.
- 553 Siebenga JJ, Vennema H, Renckens B, de Bruin E, van der Veer B, Siezen RJ, Koopmans M. 2007.
554 Epochal evolution of GGII.4 norovirus capsid proteins from 1995 to 2006. *J. Virol.* 81:9932–9941.
- 555 Siebenga JJ, Vennema H, Zheng D-P, Vinjé J, Lee BE, Pang X-L, Ho ECM, Lim W, Choudekar A,
556 Broor S, et al. 2009. Norovirus illness is a global problem: emergence and spread of norovirus
557 GII.4 variants, 2001-2007. *J. Infect. Dis.* 200:802–812.
- 558 Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large
559 phylogenies. *Bioinformatics.*
- 560 Summa M, von Bonsdorff CH, Maunula L. 2012. Pet dogs-A transmission route for human noroviruses?
561 *J. Clin. Virol.* 53:244–247.
- 562 Vega E, Barclay L, Gregoricus N, Williams K, Lee D, Vinjé J. 2011. Novel surveillance network for
563 norovirus gastroenteritis outbreaks, United States. *Emerg. Infect. Dis.* 17:1389–1395.
- 564 Vijaykrishna D, Holmes EC, Joseph U, Fourment M, Su YCF, Halpin R, Lee RTC, Deng Y-M, Gunalan
565 V, Lin X, et al. 2015. The contrasting phylodynamics of human influenza B viruses. *Elife*
566 4:e05055.
- 567 Vinje J. 2015. Advances in Laboratory Methods for Detection and Typing of Norovirus. *J Clin*
568 *Microbiol* 53:373–381.
- 569 White PA. 2014. Evolution of norovirus. *Clin. Microbiol. Infect.* 20:741–745.
- 570 Wilhelm B, Waddell L, Greig J, Rajić A, Houde A, McEwen SA. 2015. A scoping review of the
571 evidence for public health risks of three emerging potentially zoonotic viruses: Hepatitis E virus,
572 norovirus, and rotavirus. *Prev. Vet. Med.* 119:61–79.
- 573 Wong THN, Dearlove BL, Hedge J, Giess AP, Piazza P, Trebes A, Paul J, Smit E, Smith EG, Sutton JK,
574 et al. 2013. Whole genome sequencing and de novo assembly identifies Sydney-like variant
575 noroviruses and recombinants during the winter 2012/2013 outbreak in England. *Virol. J.* 10:335.

576 Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*

577

578 **Acknowledgments:** The authors would like to thank Victoria Madden of Microscopy Services
579 Laboratory, Department of Pathology and Laboratory Medicine, University of North Carolina-Chapel
580 Hill for expert technical support. The authors are grateful to Oliver G. Pybus of Department of Zoology,
581 University of Oxford for critical review of the manuscript. This work was supported by the Wellcome
582 Trust (grant number 203268/Z/16/Z to J.B), a studentship from UCL CoMPLEX (to C.R), the National
583 Institute for Health Research UCL-UCLH Biomedical Resource Centre, the Medical Research Council
584 (grant number U117573805 to R.A.G), the National Institutes of Health, Allergy and Infectious
585 Diseases (grant number U19 AI109761 to R.S.B), the EU FP7 PATHSEEK grant and a Medical
586 Research Council studentship. The funders had no role in study design, data collection and
587 interpretation, or the decision to submit the work for publication. The findings and conclusions in this
588 article are those of the authors and do not necessarily represent the official position of the Centers for
589 Disease Control and Prevention.

590

591

592

593

594

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

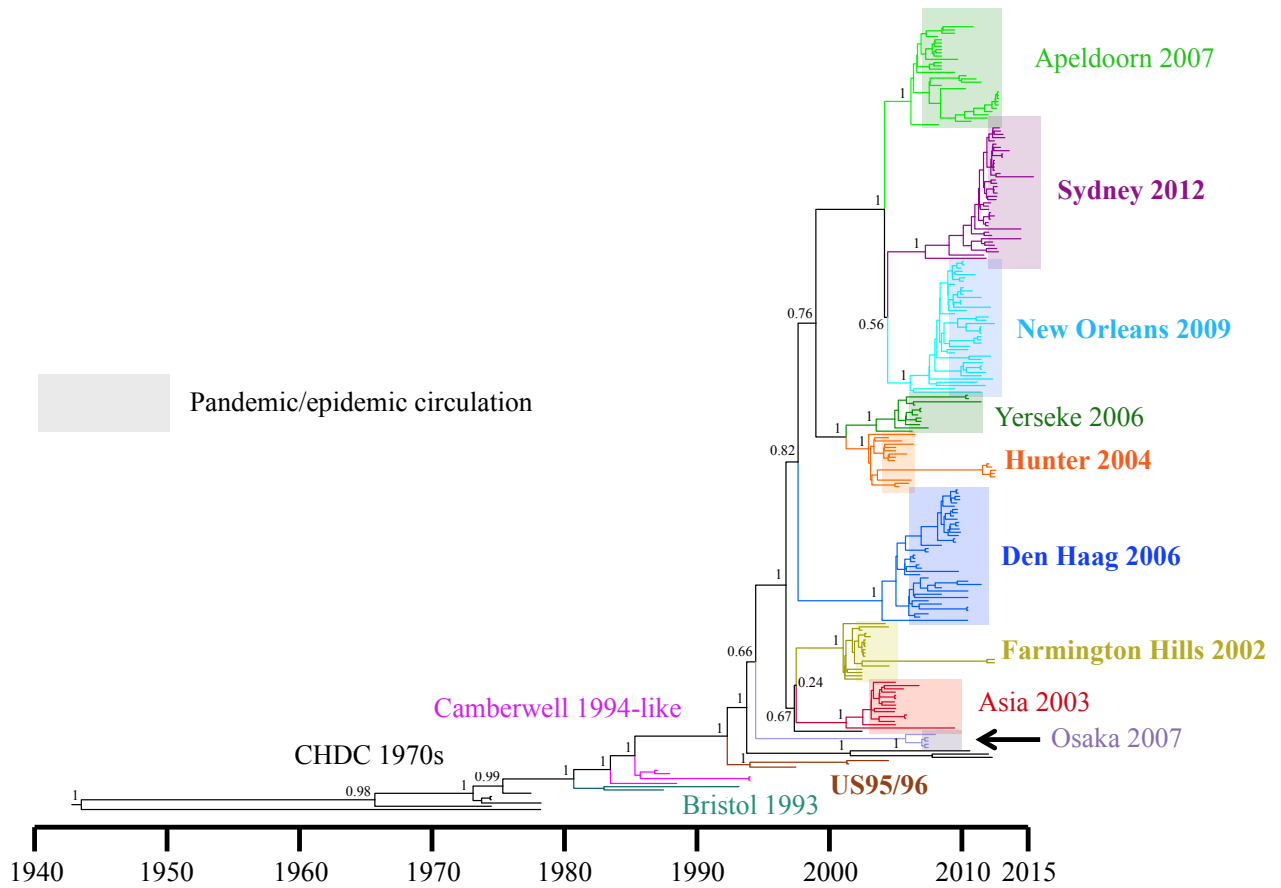
622

623

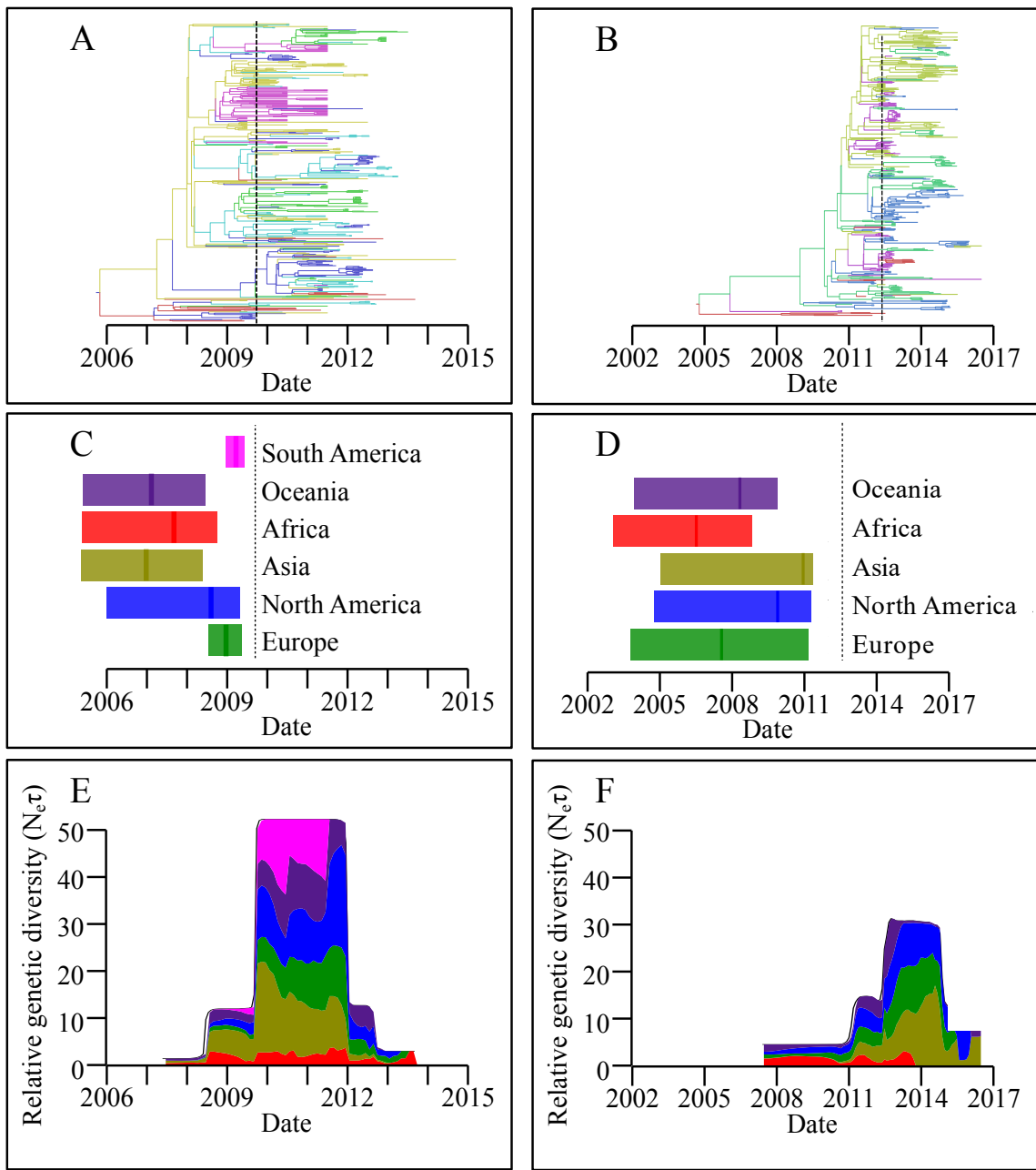
624

625

626

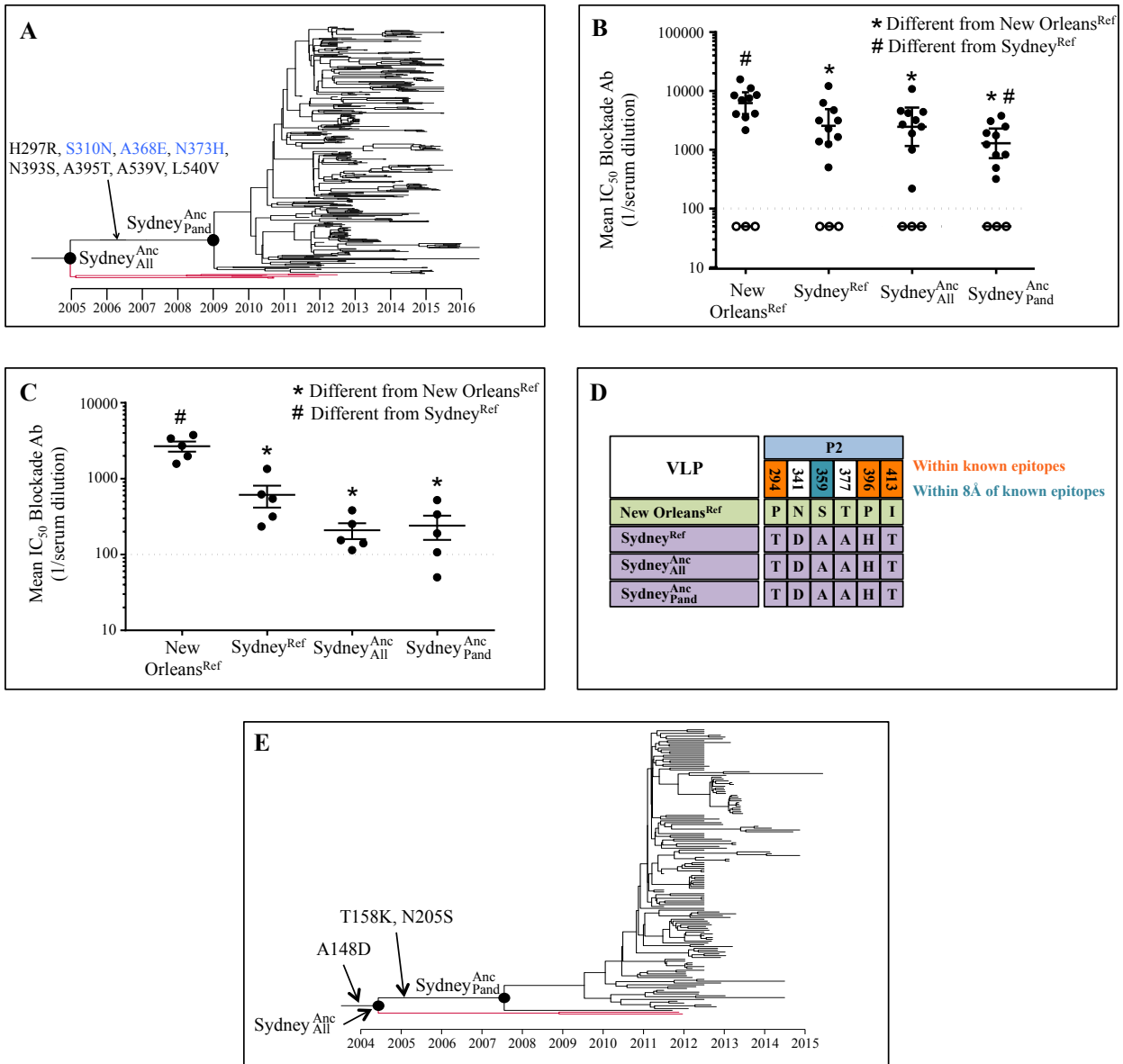


629
630 **Figure 1.** Temporal maximum clade credibility (MCC) tree of GII.4 VP1 sequences from major pandemic
631 and epidemic variants. Variants diverge from all other sampled variants years before their emergence as
632 a pandemic or epidemic (represented by the shaded area). Long branches throughout the tree indicate a
633 high level of unsampled diversity through time. Posterior supports are shown on trunk nodes.
634



635
 636
 637
 638
 639
 640
 641
 642
 643
 644
 645

Figure 2. GII.4 variants New Orleans 2009 and Sydney 2012 diversified and spread widely prior to pandemic emergence. (A and B) Spatiotemporally resolved MCC trees for New Orleans 2009 (A) and Sydney 2012 (B) with each branch coloured by inferred location, as in panels C and D. (C and D) Summary of continent import dates for New Orleans 2009 (C) and Sydney 2012 (D); the vertical line is the median import date and the shaded area the 95% HPD. The dashed vertical black line is the inferred date of pandemic emergence. (E and F) Summary of the spatiotemporal distribution of lineages from New Orleans 2009 (E) and Sydney 2012 (F). The proportion of lineages on each continent through time is plotted as a stacked area plot, scaled to the estimated relative genetic diversity.



646

647

648

649

650

651

652

653

654

655

656

657

658

659

660

661

662

663

664

665

Figure 3. Sydney 2012 could resist anti-New Orleans 2009 immunity by 2003. (A) Temporally resolved Sydney 2012 tree with Sydney^{Anc}_{All} and Sydney^{Anc}_{Pand} labelled. The lineages that diverged between Sydney^{Anc}_{All} and Sydney^{Anc}_{Pand} (shown in red) did not persist in the population. Nonsynonymous substitutions that occurred leading to Sydney^{Anc}_{Pand} are labelled. Substitutions labelled in blue remained highly conserved in Sydney 2012 (Figure S5). (B) Blockade of Sydney^{Anc}_{All}, Sydney^{Anc}_{Pand}, New Orleans^{Ref} and Sydney^{Ref} interaction with pig gastric mucin (PGM) by polyclonal sera from patients infected with New Orleans 2009 (closed circles) or healthy blood donors (open circles), did not block VLPs at the assay limit of detection. Bars are geometric mean values with 95% confidence intervals. Dashed line is assay limit of detection. Statistical significance calculated using the Wilcoxon test. (C) As in B but using polyclonal sera collected from mice exposed to New Orleans 2009. (D) Six amino acid sites in the antigenic VP1 P2 subdomain exhibit a different residue in all three Sydney 2012 VLPs compared with New Orleans 2009. (E) Temporally resolved Sydney 2012 VP2 phylogeny with Sydney^{Anc}_{All} and Sydney^{Anc}_{Pand} labelled. Nonsynonymous substitutions leading to Sydney^{Anc}_{All} and Sydney^{Anc}_{Pand} are labelled.