Capturing noroviruses circulating in the population: sewage surveillance in Guangdong, China (2013–2018)

J. Lu,[1,2,3*] L. Fang,[2] J.J. Peng,[1,2,3] L.L. Zeng,[1,2] H.F. Lin,[1,2] Q.L. Xiong,[1,2,3] L.N. Yi,[1,2] T. Song,[2] J.F. He,[2] L. M. Sun,[2] C.W. Ke,[2] H. Li,[2] H.Y. Zheng[2*]

Affiliations:

[1] Guangdong Provincial Institution of Public Health, Guangdong Provincial Center for Disease Control and Prevention, Guangzhou, China

[2] Guangdong Provincial Center for Disease Control and Prevention, Guangzhou, China

[3] School of Public Health, Southern Medical University, Guangzhou, China

***Correspondent to:**

Huanying Zheng, email: zhenghy.gdcdc@gmail.com, Guangdong Provincial Center for Disease Control and Prevention, China;

Or

Jing Lu, email: Jimlu0331@163.com, Guangdong Provincial Institution of Public Health, China;

Abstract

Noroviruses (NoVs) are the leading cause of acute gastroenteritis outbreaks. The specific geographical distribution and expanding diversity of NoVs has posed a challenge to NoV surveillance and interventions. Here, we describe the long-term dynamic correlation between NoV distribution in sewage and in the local population by using high-throughput sequencing and operational taxonomic unit (OTU) analysis. The NoV viral loads in sewage were closely associated with the number of NoV outbreaks in the population. Compared with the viral distributions in outbreaks, the dominance of the newly emerged variants, such as GII.P17-GII.17 and GII.P16-GII.2, could be detected two months ahead in sewage. In addition, the dynamics of pre-epidemic variants, which were rarely detected from clinics, could be captured through sewage surveillance, thus improving our understanding of the viral origin and evolution. Our data highlight that the high-throughput environmental screening should become a critical part of the response to infectious diseases.

Introduction

Currently, emerging infectious diseases present one of the greatest public health challenges. Human noroviruses (NoVs) are recognized as the leading cause of acute gastroenteritis (AGE) for all ages (*1*). Viral infections are estimated to be associated with 18% of diarrheal diseases and 212,000 deaths annually (*2*). Due to their high contagiousness, most cluster infections or outbreaks of NoVs have been reported in semienclosed settings such as hospitals, childcare centers, schools, and universities, leading to public panic regarding disease transmission (*3*).

Based on the sequence of the VP1 gene, NoVs can be classified into seven genogroups (GI–GVII) and further subdivided into more than 40 genotypes (*4*). Viruses from the GI, GII, and GIV genogroups are known to infect humans. Currently, there is no drug or licensed vaccine for NoV infections. Disease control has mainly relied on clinical surveillance, which informs us of the emergence and prevalence of different genotypes and their potential risks of causing global outbreaks. Since 1995, GII.4 has been identified as the predominant genotype responsible for 65-80% of NoV infections worldwide (*5*, *6*). The novel GII.4 variant has emerged every 2-5 years and presented amino acid changes in five antigenic sites (*7*) associated with viral escape from herd immunity (*8*, *9*). This epochal evolution pattern is regarded as the principal mechanism adopted by GII.4 to sustain its predominance and has led to successive pandemics in the global population (*10–12*). However, in the winter seasons of 2014-2015 and 2016-2017, novel variants of some genotypes, such as GII.P17-GII.17 and GII.P16-GII.2, emerged. These NoV genotypes, which were rarely identified previously, have replaced GII.4 and caused increasing AGE outbreaks in Asia-Pacific regions. The emergence of these novel variants was also observed in European countries, with a rising prevalence in 2015-2016 (*13*).

The temporal and spatial variances in NoV genetic distribution pose a significant challenge to NoV vaccine design and for early warning of NoV outbreaks, for which more extensive and timely surveillance is warranted. In this study, we developed a high-throughput method to capture NoVs in sewage. With this method, the dynamics of NoV distribution in Guangzhou sewage were investigated from 2013 to 2018. The corresponding epidemiological data of NoV outbreaks were subtracted from the clinical surveillance network to evaluate how sewage data could improve our knowledge of NoV circulation in the population, especially for early detection of novel NoV variants.

Results

Norovirus genotype distributions in AGE outbreaks, 2013-2018

The Guangdong provincial surveillance network was launched in 2013 for monitoring AGE outbreaks in 21 prefectural cities (*14*). In total, 327 NoV-associated outbreaks were reported between January 2013 and December 2018. The majority of these outbreaks (284 of 327, 87%) were caused by the GII genogroup, in which GII.4, GII.17, GII.2, and GII.3 were the predominant genotypes, causing 24, 75, 143 and 42 outbreaks, respectively. As previously reported, a novel variant of GII.17 (GII.P17-GII.17) (*14*) and a recombinant variant GII.2 (GII.P16-GII.2) (*15*) emerged in Guangdong in 2015 and 2017, respectively. These novel variants caused a significant increase (>4-fold) in AGE outbreaks and became predominant in the next 2 or 3 years (Fig. 1A). More importantly, the novel variant outbreaks occurred almost simultaneously in multiple cities, and the epidemic peaked in the following 1-2 months. In this context, an optimized surveillance system that could provide timely and extensive molecular surveillance data is warranted.

Molecular surveillance of norovirus in raw sewage in Guangzhou, 2013-18

Guangzhou is the capital city of Guangdong Province and one of the most populous regions in China. Between 2013-2018, Guangzhou reported the largest number of NoV outbreaks (72 of 327, 22%, Fig. 1B) and was regarded as an epidemiological hub for NoV transmission (*16*). Sewage surveillance in Guangzhou was previously established for the global eradication of poliovirus and was also applied to investigate the potential origins of enterovirus outbreaks (Lu et al., 2020). To explore the dynamic correlations of NoV in sewage and in clinical outbreaks, a total of 72 sewage samples were collected from the Guangzhou wastewater treatment plant (WWTP) each month from January 2013 to December 2018. GI and GII NoVs were detected in 91.7% (66/72) and 100% (72/72) of the samples by RT-PCR. The relative virus copies in sewage samples correlated well with the number of NoV outbreaks in the population, with the most abundant GI and GII NoVs detected at the end of winter and in early spring (November to April of the next year). GI NoV was detected with relatively high viral loads in sewage but rarely identified in clinical outbreaks, suggesting that most GI infections may be asymptomatic and that these viruses circulated undetected in the local population. Although correlation was observed between the viral loads in sewage and the risks of outbreaks in the community, sewage surveillance still provided limited information for early warning of NoV outbreaks. For instance, the largest NoV outbreak was reported in the 2016-17 epidemic season, but the contemporary GII viral copies in sewage were similar to those in the other epidemic seasons. Therefore, the outbreak of NoV was not only driven by the viral prevalence in the community but could also be triggered by the emergence and

transmission of a novel genetic variant for which herd immunity was lacking. Therefore, establishing a method to capture the genetic diversity of NoV in sewage is critical for optimizing surveillance.

Setup and verification of a high-throughput method

It was expected that a sewage sample would contain different NoV genotypes and even genetic variants from a specific genotype. Capturing the distribution of multiple genotypes and variants by the conventional Sanger sequencing/cloning method was a great challenge. In this study, we used amplicon sequencing and calculated the abundances of different genotypes or variants via an operational taxonomic unit (OTU) clustering method, which is commonly used to define the taxonomic units of bacterial species in 16S rRNA sequencing. It was expected that the percentage of different genotypes or variants would be proportional to the corresponding read number in OTU clusters (Fig.2A, details in Material and Methods). To verify the accuracy of this method, we constructed four plasmids, including the genome fragments of the GII.2, GII.3, GII.4, and GII.17 genotypes. The copy numbers of these plasmids were determined through mass measurements. Serial mixtures (groups 1-3) were obtained by adding the GII.2, GII.3, GII.4, and GII.17 plasmids at ratios of 1:1:1:1, 1:3:6:9, and 1:10:100:1000 (Fig. 2B). The total amount of plasmid mixture for each group was ~0.02 ng (approximately $6x10^6$ copies). These samples were first PCR amplified, sequenced with high-throughput sequencing, and analyzed as described in the Material and Methods section. As shown in Fig. 3, the sequencing results illustrated the original genotype distribution well when different genotypes were mixed at a similar ratio. When the genotype ratio was over 1:10, PCR amplification bias likely occurred, and the genotype with a lower rate may have been overrepresented (Fig. 2B). In group 3, the ratio calculated from the number of reads in each genotype cluster was 1:6:64:246, which was basically at the same magnitude in which the plasmids were mixed (1:10:100:1000). The plasmid with the lowest copies ($6x10^3$ copies) finally had 1629 clustered reads. These data suggested that the combination of high-throughput sequencing and the OTU clustering method could roughly identify the percentages of different genotypes in sewage without losing sensitivity for detecting low-proportion genotypes.

Norovirus genotype distributions in sewage, 2013-2018

We next applied the above method to investigate NoV distribution in Guangzhou sewage from 2013-2018. Approximately 1,000,000 reads (range 112,358–2,947,207) were generated from each amplicon (Table S2). Genotype distributions in the GI and GII genogroups were determined separately by measuring the number of reads in each genotype cluster. Genotypes with a percentage less than 0.1% of total reads were removed. In total, 9 GI genotypes and 15 GII genotypes were identified in sewage

127 water from 2013-2018 (Fig. 3A&3C). For the GI genotype, GI.2, GI.8 and GI.9 were found to be
128 dominant in sewage samples (percentages were over 30%) but with different temporal distributions
129 (Fig. 3A & 3B). For the GI genogroup, GI.2 was continuously circulating in the population from 2013
130 through 2018 (Fig. 3A), and its predominance was significant in the 2017-2018 winter season (Fig. 3B).
131 In contrast, the GI.8 and GI.9 genotypes were dominant in 2013 but were rarely detected in 2017 and
132 2018 (Fig. 3A).
133
134 GII NoV was the dominant genogroup causing AGE outbreaks. The temporal pattern of GII genotypes
135 in sewage was well captured. GII.2, GII.4, and GII.17 were the major circulating genotypes in sewage,
136 and other genotypes were detected with lower prevalence (<25%, Fig. 3C). More specifically, GII.4
137 was dominant (> 50%) in 2013 and several months of 2014 but detected at lower level in the winter
138 season (January to April) from 2015 through 2018. GII.17 was occasionally identified as a dominant
139 genotype in September 2013, and increasing detection of this genotype was observed in the winter
140 season of 2014 (Fig. 3D). The percentage of the GII.17 genotype was over 50% in nearly all months
141 from November 2014 through April 2016, consistent with its dominance in the outbreaks (Fig. 1). GII.2
142 was detected at a relatively low level before mid-2016 but has been increasingly identified since
143 October 2016. Interestingly, the high prevalence of GII.4 in sewage was only observed in months
144 before epidemic seasons (mainly July to December), and this dominance was quickly taken over by
145 GII.17 or GII.2 in all the epidemic seasons from 2014-2018 (Fig. 3D). These data suggested that
146 compared to GII.4, the newly emerged variants of GII.17 and GII.2 likely have higher transmission
147 efficiency when etiological factors are favorable.
148
149 **Phylogenetic analyses**
150 The emergence and spread of new NoV variants always lead to an increase in NoV outbreaks in the
151 population. One priority of sewage surveillance is to identify the early emergence of novel genetic
152 variants and characterize their dynamic distributions. In this study, we set the sequence similarity
153 threshold at 97% in the clustering method to distinguish the different genetic variants (see Material and
154 Methods for details). Through this method, we obtained a consensus sequence of a genetic cluster and
155 the number of reads in it, which was proportional to the abundance in sewage. To define these genetic
156 clusters, we constructed maximum-likelihood (ML) trees for the GII.2, GII.17 and GII.4 genotypes by
157 combining all consensus sequences recovered from sewage and representative variant sequences from
158 public databases (Fig.S1&S2&S3). The 350-bp amplicon sequenced in this study represented one of
159 the most divergent parts of the viral genome and is widely used for NoV genotyping. For GII.4, nearly

all variants identified after 2013 could be classified into the GII.4_Sydeny2012 lineage (Fig. 4A, top panel). The GII.4 sequences from sewage between 2013 and 2015 were near the root of the GII.4_Sydeny2012 lineage. Notably, two distinct subclusters of GII.4_Sydeny2012 emerged in August 2016 (Fig. 4A). The GII.4_Syd12_Var1 strains started to be detected in sewage in August 2016 and were more prevalent than other GII.4 variants in the following seasons (Fig. 4A, lower panel). In addition to the emergence of GII.4_Syd12_Var1, GII.4-associated outbreaks increased in winter 2016-17 compared to the low epidemic level observed in 2014 and 2015 (Fig. 4A). Phylogenetic analysis showed that the GII.4_Syd12_Var1 strains were closely related to the GII.Pe-GII.4_Sydeny2012 identified in global through the NoroNet network(*19*). The contemporary increasing epidemic observed at the end of 2016 highlighted that further characterizing the antigenicity of GII.4_Syd12_Var1 strains and their prevalence in other regions are warranted to evaluate their epidemic risk. The GII.4_Syd12_Var2 strains were closely related to the GII.P16-GII.4_Sydeny2012 strains recently reported in Europe and America (*20*–*22*). GII.P16-GII.4_Sydeny2012 is associated with the emergence of the novel recombinant strains GII.P16-GII.2 detected at the end of 2016 (*23*, *24*).

The GII.17 variant was first identified in sewage samples in September 2013, and this variant was closely related to the pre-epidemic lineage (Kawaski323-like virus) (*25*). The epidemic variant of GII.17 (Kawaski308-like virus) was first detected in sewage in September 2014, which was two months before the occurrence of AGE outbreaks. Notably, in the sewage samples, we found a GII.17 genetic variant (GII.17 variant1) that was close to the GII.Pe-GII.17 viruses recently reported in Hong Kong (KT589391_CUHK-NS-682_HK_2015-06) and southern Brazil (*26*) (Fig. 4B). This variant represented a distinct lineage originating from prototype strain JN699043_GF_1978, and it had not been identified from the outbreaks but was frequently detected in sewage between 2016 and 2018, indicative of its hidden circulation in the local community.

For GII.2, four genetic clusters were observed based on the surveillance data, and most of the GII.2 strains in sewage fell into the Var1 and Var2 clades. The Var1 clade, including the dominant strain in sewage in 2016-2018, represented the epidemic lineage GII.P16-GII.2. This lineage was regarded as a recombinant of GII.2 and GII.P16-GII.4 and caused outbreaks in multiple regions in 2016 (*23*, *24*, *27*). The Var2 clade included GII.2 strains mainly circulating in sewage between 2013 and 2015. Consistent with this, the closely related reference sequences for Var2 were identified in Asian-Pacific countries

between 2012 and 2015 (Fig. 4C upper panel). The dynamic distribution showed that the percentage of Var2 fluctuated at a relatively low level (<25%, Fig. 4C lower panel) in sewage in 2013-2015, and this variant rarely caused outbreaks in the community. In contrast, Var1, representing the 2016 epidemic strains, was first detected in sewage in September 2016, and its percentage steadily increased in the following three months. The first occurrence of GII.P16-GII.2 in outbreak surveillance was in November 2016, which was two months after its dominance in the sewage sample.

Discussion

Currently, multiple NoV genotypes are cocirculating worldwide with continuous and rapid changes in NoV genetic diversity (*13*). In contrast to the global predominance of GII.4 observed in the past decade, the distribution of NoV genotypes is now varied among countries, indicating that spatial and temporal differences exist. Due to this remarkable change, regional surveillance data are critical for early warning and vaccine design, similar to the actions taken in response to seasonal influenza viruses. To date, extensive and timely clinical surveillance is difficult to achieve in most regions. In this study, we prove that high-throughput sequencing combined with the OTU clustering method could be successfully applied to investigate the dynamic distribution of NoVs in sewage. By comparison with long-term surveillance data from NoV outbreaks, our results highlight that sewage surveillance serves as an efficient tool to capture the NoV distribution in the specific population.

The current international or regional NoV surveillance networks mostly focus on AGE outbreaks or sporadic infections from clinics. As a member of CaliciNet (*3*), the Guangdong provincial surveillance system followed a similar scheme to report NoV-associated AGE outbreaks in 21 prefectural cities. Great efforts have been made toward sequencing and molecular typing to capture genotype distributions, especially for the emerging novel genetic variants in the local population (*14*, *15*). Several limitations have been found for the current surveillance system. The sporadic infections reported by sentinel hospitals are often biased, toward the detection of NoV genotypes associated with a relatively severe symptom. Thus, the method always oversamples specific groups of the population, such as the young or old age groups and immunocompromised individuals, who tend to have more severe infections and prolonged symptoms. Outbreak surveillance mostly identifies the genotypes that may already be efficiently transmitted in the population, and it is difficult to detect those variants before they are established in the local population. In particular, NoV genotypes such as GII.4 and GII.17 can have sudden transitions in their antigenicity or infectivity through a few amino acid mutations in their structural proteins (*28–30*). Before the transitions, the variants may have limited infectivity and circulate in the population at a relatively low level. These variants were challenging to identify during the early stage based on the conventional clinical surveillance system. For example, the pre-epidemic GII.P17_GII.17 strains had already disseminated in Guangdong in the early winter of 2013-2014 (Fig. 4B) without any associated AGE outbreaks reported. Several amino acid substitutions in the antigenic region of VP1 emerged almost simultaneously in the pre-epidemic GII.p17_GII.17 strain, resulting in the emergence of GII.p17_GII.17 epidemic strains at the end of 2014 (Fig.4B). This lineage replaced GII.4 as the predominant genotype and caused AGE outbreaks in Asian-Pacific

232 regions in 2014-2015 (*31–33*). A similar pattern was also observed for GII.P16_GII.2, which was
233 almost simultaneously identified in Germany and multiple areas in China at the end of 2016 (*15*, *34*)
234 and led to an exponential increase in NoV outbreaks (Fig. 4C). From this perspective, the time window
235 for our response to the epidemic is too short when we rely on the current clinical surveillance system.
236
237 In this study, we discuss the possibility of capturing the NoV distribution in the local population
238 through sewage surveillance in a hub city. Before this study, several groups tried to use Sanger
239 sequencing/cloning or NGS to investigate NoV genotype diversity and temporal dynamics in sewage
240 (*35–37*). However, a high-throughput method for calculating the relative abundance of different
241 genotypes and genetic variants in sewage has not been established. The central question regarding the
242 dynamic relationship between NoVs in sewage and the human population has still not been adequately
243 addressed. In this study, the OTU method, which has previously been used for bacterial typing and
244 distribution counting, was successfully applied for measuring the NoV distribution in sewage. The
245 study results support sewage surveillance as an efficient way to capture the dynamic distribution of
246 NoVs in the population nearly in real time (high-throughput sequencing and analyzing could be
247 completed in 24 hours). First, the correlation between the relative viral loads in sewage and the number
248 of outbreaks in the community suggests that the viral RNA concentration in sewage is an indicator of
249 outbreak risks. Second, the dynamics of the novel genetic variants in sewage could be used as a
250 warning of outbreaks in the population. For example, the epidemic strain GII.P17-GII.17 was dominant
251 in the sewage sample in September 2014, and the increasing prevalence of this variant in October and
252 November highlights its outbreak risk in the following epidemic season (Fig. 4B). This pattern was
253 repeatedly observed two years later, during the outbreak of GII.P16-GII.2 (Fig. 4C). More importantly,
254 via the clustering method, we could reveal the evolutionary history of these genetic variants, which was
255 difficult to fully capture by clinical surveillance. As shown in Fig. 4B, the pre-epidemic GII.P17-GII.17
256 variant was rarely detected in clinics, possibly due to asymptomatic infections. In contrast, this variant
257 was dominant in sewage in September 2013 and May 2014, with percentages reaching 50% and 75%,
258 respectively. The establishment of a pre-epidemic strain in the local population provides a foundation
259 for generating the epidemic GII.P17-GII.17 variant several months later. Similarly, the emergence and
260 prevalence of GII.P16-GII.4-like strains identified in sewage in the middle of 2016 support the
261 following recombination identified between GII.P16-GII.4 and GII.2 viruses.
262
263 There remain some limitations to the present study. First, specific PCR amplification is required before
264 high-throughput sequencing, and the PCR bias would lead to low-proportion (<10%) samples being

overrepresented (Fig. 2B). However, we need to balance the sensitivity and precision of this high-throughput method. The RT-PCR results showed a relatively low copy of NoVs in the raw sewage (90% samples less than $3.1x10^5$ copies/L), and direct metagenomic sequencing could obtain only a few fragments of the most abundant genotype(s). Therefore, PCR amplification is required to analyze the distribution of multiple genotypes in sewage, especially during the nonepidemic season. Second, the modified primer set G2SKF and G2SKR was used in this study to produce an ~350-bp fragment covering part of the VP1 region. From the sequencing data, we could not determine the RdRp genotype of NoV in sewage. Currently, it remains challenging to find a general primer set for amplifying the RdRp and VP1 fragments simultaneously, especially because this primer set should have a similar amplification efficiency among different NoV genotypes. Additionally, limited by the sequence length of the Illumina sequencing platform, the long fragment could not be sequenced without fragmentation.

For viral infectious diseases, the surveillance system is designed to answer a set of questions that are central to disease mitigation and control. These questions include what the virus is, whether it is genetically different, and how the kinetics of disease transmission are in different populations (*38*, *39*). Sometimes, clinical surveillance provides a short time window for us to answer these questions and respond in an informed manner before the occurrence of large outbreaks or exponential growth of the epidemic. The global pandemic of COVID-19 is another example of how extensive clinical surveillance is needed for early warning of infectious disease epidemics. Recent sewage surveillance performed in multiple regions indicated that SARS-CoV-2 RNA could be detected in sewage before the first COVID-19 cases were reported by local clinical surveillance, and SARS-CoV-2 RNA concentration in sewage served as a leading indicator of community infection (*40*, *41*). These data again highlight the importance of sewage surveillance in the early warning of emerging infectious diseases. With the significant advancement of sequencing technologies, viral disease surveillance has finally reached a point where high-throughput screening of the associated environmental samples, including sewage, air, and soil, can become a critical part of the response to infectious diseases.

Material and Methods

Norovirus outbreak surveillance system

The Guangdong provincial AGE outbreak surveillance network has been established since 2013 for monitoring AGE outbreaks in all 21 cities in Guangdong, China. An outbreak with a cluster of at least 20 AGE cases (meeting the Kaplan criterion) within 3 days must be reported to Guangdong Provincial Center for Disease Control and Prevention. Samples from each outbreak were first tested for NoV (Norovirus RT-PCR Kit; Shanghai ZJ Bio-Tech Co., Ltd., China) and for intestinal bacteria at the local Centers for Disease Control and Prevention. The NoV-positive specimens were delivered to the Guangdong Provincial Center for Disease Control and Prevention for further genotyping.

Sewage surveillance

Raw sewage samples were collected monthly from January 2013 to December 2018 from the primary sedimentation tanks at the Liede wastewater treatment plant (WWTP) in Guangzhou city, China. This WWTP serves a population of approximately 2,150,000. A four-liter sample was obtained from the inlets of the primary sedimentation tanks on a routine basis each month. The sample was immediately transported to the laboratory, and sample treatment was started within 2 h after the samples arrived at the laboratory. Viruses in the sewage samples were concentrated using an improved negative-charge filter membrane absorption and sonication method, as previously described (*18*). Briefly, the collected sewage samples were centrifuged at 3,000 rpm for 30 min. Then, $MgCl_2$ was added to the supernatant at a final concentration of 0.05 M, and the pH was adjusted to 3.5 to 4.0 with HCl. The samples were then slowly passed through a negatively charged membrane filter (mixed cellulose ester membrane filter; Advantec Co. Ltd., Japan) under gentle positive pressure. To elute the viruses, the filter with adsorbed viruses was cut into pieces and sonicated for 1 min in 10 ml of a 3% beef extract solution (pH 9.6), followed by centrifugation at 1,940 x g for 30 min to yield 8 ml of eluents. Finally, the eluents were passed through a 0.22 μm syringe filter to remove bacteria and fungi.

NGS and bioinformatic analysis

The nucleic acids were extracted from 140 μl of concentrated sewage with a QIAamp Viral RNA Minikit (Qiagen, USA). RT-PCR was performed with Superscript III (Thermo Fisher, USA) by using random primers. The cDNA was then used in the first round of PCR amplification with the primer set described in Table S1. After 27 cycles of amplification, 2.5 μl of the first PCR products were used as a template in the second-round PCR. After 35 cycles of amplification, the positive products (~354 nucleotides) were subjected to Illumina MiSeq sequencing (PE250), and approximately 1,000,000

reads were generated for each PCR product. The datasets were quality controlled (QC) using fastp (*42*) to trim artificial sequences (adapters), to cut low-quality bases (quality scores <□30) and to remove short reads (<□50□bp). Paired-end reads were assembled and PCR primers were trimmed by using cutadapt. Then, the sequences were clustered into OTUs with the following clustering criteria: 1) all pairs of OTU sequences should have <97% pairwise sequence identity; 2) an OTU sequence should be the most abundant within a 97% neighborhood. These criteria have been tested to distinguish the different genetic variants of the NoV genotype, including GII.4, GII.17, and GII.2. After clustering, genotype assignment was performed using blastn (*43*) by alignment with the local NoV nucleotide database. The abundance of different OTUs and genotypes was calculated with the relative number of raw reads using R scripts. All sequencing metadata generated in this study have been submitted to the National Genomics Data Center (https://bigd.big.ac.cn/) with submission number subCRA003702.

Phylogenetic analysis

The phylogenetic analyses were performed by combining the partial VP1 gene sequences generated from sewage samples and the selected reference sequences from the public database (Fig. S1&S2&S3) to classify the different genetic lineages. ML trees were estimated for sequences from GII.4, GII.17, and GII.2 by using IQ-Tree (*44*). Phylogenetic trees were annotated and visualized with ggtree (*45*). The dynamic changes in different genetic lineages were visualized using R scripts.

References

1. R. L. Atmar, M. K. Estes, The epidemiologic and clinical importance of norovirus infection. *Gastroenterol. Clin. North Am.* **35**, 275–290, viii (2006).

2. S. M. Pires, C. L. Fischer-Walker, C. F. Lanata, B. Devleesschauwer, A. J. Hall, M. D. Kirk, A. S. R. Duarte, R. E. Black, F. J. Angulo, Aetiology-Specific Estimates of the Global and Regional Incidence and Mortality of Diarrhoeal Diseases Commonly Transmitted through Food. *PLoS ONE*. **10**, e0142927 (2015).

3. A. J. Hall, M. E. Wikswo, K. Manikonda, V. A. Roberts, J. S. Yoder, L. H. Gould, Acute gastroenteritis surveillance through the National Outbreak Reporting System, United States. *Emerging Infect. Dis.* **19**, 1305–1309 (2013).

4. A. Kroneman, E. Vega, H. Vennema, J. Vinjé, P. A. White, G. Hansman, K. Green, V. Martella, K. Katayama, M. Koopmans, Proposal for a unified norovirus nomenclature and genotyping. *Archives of Virology*. **158**, 2059–2068 (2013).

5. J. J. Siebenga, H. Vennema, D.-P. Zheng, J. Vinjé, B. E. Lee, X.-L. Pang, E. C. M. Ho, W. Lim, A. Choudekar, S. Broor, T. Halperin, N. B. G. Rasool, J. Hewitt, G. E. Greening, M. Jin, Z.-J. Duan, Y. Lucero, M. O'Ryan, M. Hoehne, E. Schreier, R. M. Ratcliff, P. A. White, N. Iritani, G. Reuter, M. Koopmans, Norovirus Illness Is a Global Problem: Emergence and Spread of Norovirus GII.4 Variants, 2001–2007. *J Infect Dis*. **200**, 802–812 (2009).

6. E. Vega, L. Barclay, N. Gregoricus, S. H. Shirley, D. Lee, J. Vinjé, Genotypic and Epidemiologic Trends of Norovirus Outbreaks in the United States, 2009 to 2013. *Journal of Clinical Microbiology*. **52**, 147–155 (2014).

7. K. Tohma, C. J. Lepore, Y. Gao, L. A. Ford-Siltz, G. I. Parra, Population Genomics of GII.4 Noroviruses Reveal Complex Diversification and New Antigenic Sites Involved in the Emergence of Pandemic Strains. *mBio*. **10** (2019), doi:10.1128/mBio.02202-19.

8. G. I. Parra, E. J. Abente, C. Sandoval-Jaime, S. V. Sosnovtsev, K. Bok, K. Y. Green, Multiple Antigenic Sites Are Involved in Blocking the Interaction of GII.4 Norovirus Capsid with ABH Histo-Blood Group Antigens. *Journal of Virology*. **86**, 7414–7426 (2012).

9. K. Debbink, E. F. Donaldson, L. C. Lindesmith, R. S. Baric, Genetic Mapping of a Highly

Variable Norovirus GII.4 Blockade Epitope: Potential Role in Escape from Human Herd Immunity. *Journal of Virology*. **86**, 1214–1226 (2012).

10. L. C. Lindesmith, E. F. Donaldson, A. D. LoBue, J. L. Cannon, D.-P. Zheng, J. Vinje, R. S. Baric, Mechanisms of GII.4 Norovirus Persistence in Human Populations. *PLoS Medicine*. **5**, e31 (2008).

11. K. Debbink, L. C. Lindesmith, E. F. Donaldson, R. S. Baric, Norovirus Immunity and the Great Escape. *PLoS Pathogens*. **8**, e1002921 (2012).

12. J. J. Siebenga, H. Vennema, B. Renckens, E. de Bruin, B. van der Veer, R. J. Siezen, M. Koopmans, Epochal Evolution of GGII.4 Norovirus Capsid Proteins from 1995 to 2006. *Journal of Virology*. **81**, 9932–9941 (2007).

13. J. van Beek, M. de Graaf, H. Al-Hello, D. J. Allen, K. Ambert-Balay, N. Botteldoorn, M. Brytting, J. Buesa, M. Cabrerizo, M. Chan, F. Cloak, I. Di Bartolo, S. Guix, J. Hewitt, N. Iritani, M. Jin, R. Johne, I. Lederer, J. Mans, V. Martella, L. Maunula, G. McAllister, S. Niendorf, H. G. Niesters, A. T. Podkolzin, M. Poljsak-Prijatelj, L. D. Rasmussen, G. Reuter, G. Tuite, A. Kroneman, H. Vennema, M. P. G. Koopmans, Molecular surveillance of norovirus, 2005–16: an epidemiological analysis of data collected from the NoroNet network. *The Lancet Infectious Diseases* (2018), doi:10.1016/S1473-3099(18)30059-8.

14. J. Lu, L. Sun, L. Fang, F. Yang, Y. Mo, J. Lao, H. Zheng, X. Tan, H. Lin, S. Rutherford, L. Guo, C. Ke, L. Hui, Gastroenteritis Outbreaks Caused by Norovirus GII.17, Guangdong Province, China, 2014–2015. *Emerging Infectious Diseases*. **21**, 1240–1242 (2015).

15. J. Lu, L. Fang, L. Sun, H. Zeng, Y. Li, H. Zheng, S. Wu, F. Yang, T. Song, J. Lin, C. Ke, Y. Zhang, J. Vinjé, H. Li, Association of GII.P16-GII.2 Recombinant Norovirus Strain with Increased Norovirus Outbreaks, Guangdong, China, 2016. *Emerging Infectious Diseases*. **23**, 1188–1190 (2017).

16. J. Lu, L. Fang, H. Zheng, J. Lao, F. Yang, L. Sun, J. Xiao, J. Lin, T. Song, T. Ni, J. Raghwani, C. Ke, N. R. Faria, T. A. Bowden, O. G. Pybus, H. Li, The Evolution and Transmission of Epidemic GII.17 Noroviruses. *Journal of Infectious Diseases*. **214**, 556–564 (2016).

17. J. Lu, M. Kang, H. Zeng, Y. Zhong, L. Fang, X. Zheng, L. Liu, L. Yi, H. Lin, J. Peng, C. Li, Y.

399 Zhang, L. Sun, S. Luo, J. Xiao, B. B. O. Munnink, M. P. G. Koopmans, J. Wu, Y. Zhang, Y.
400 Zhang, T. Song, H. Li, H. Zheng, Tracking echovirus eleven outbreaks in Guangdong, China: a
401 metatranscriptomic, phylogenetic, and epidemiological study. *Virus Evol*. **6** (2020),
402 doi:10.1093/ve/veaa029.

403 18. H. Zheng, J. Lu, Y. Zhang, H. Yoshida, X. Guo, L. Liu, H. Li, H. Zeng, L. Fang, Y. Mo, L. Yi, T.
404 Chosa, W. Xu, C. Ke, Prevalence of Nonpolio Enteroviruses in the Sewage of Guangzhou City,
405 China, from 2009 to 2012. *Applied and Environmental Microbiology*. **79**, 7679–7683 (2013).

406 19. J. van Beek, M. de Graaf, H. Al-Hello, D. J. Allen, K. Ambert-Balay, N. Botteldoorn, M.
407 Brytting, J. Buesa, M. Cabrerizo, M. Chan, F. Cloak, I. D. Bartolo, S. Guix, J. Hewitt, N. Iritani,
408 M. Jin, R. Johne, I. Lederer, J. Mans, V. Martella, L. Maunula, G. McAllister, S. Niendorf, H. G.
409 Niesters, A. T. Podkolzin, M. Poljsak-Prijatelj, L. D. Rasmussen, G. Reuter, G. Tuite, A.
410 Kroneman, H. Vennema, M. P. G. Koopmans, Molecular surveillance of norovirus, 2005–16: an
411 epidemiological analysis of data collected from the NoroNet network. *The Lancet Infectious*
412 *Diseases*. **18**, 545–553 (2018).

413 20. L. Barclay, J. L. Cannon, M. E. Wikswo, A. R. Phillips, H. Browne, A. M. Montmayeur, R. L.
414 Tatusov, R. M. Burke, A. J. Hall, J. Vinjé, Emerging Novel GII.P16 Noroviruses Associated with
415 Multiple Capsid Genotypes. *Viruses*. **11** (2019), doi:10.3390/v11060535.

416 21. M. Bidalot, L. Théry, J. Kaplon, A. De Rougemont, K. Ambert-Balay, Emergence of new
417 recombinant noroviruses GII.p16-GII.4 and GII.p16-GII.2, France, winter 2016 to 2017. *Euro*
418 *Surveill*. **22** (2017), doi:10.2807/1560-7917.ES.2017.22.15.30508.

419 22. M. C. Medici, F. Tummolo, V. Martella, F. De Conto, M. C. Arcangeletti, F. Pinardi, F. Ferraglia,
420 C. Chezzi, A. Calderaro, Emergence of novel recombinant GII.P16_GII.2 and GII. P16_GII.4
421 Sydney 2012 norovirus strains in Italy, winter 2016/2017. *New Microbiol*. **41**, 71–72 (2018).

422 23. J. Lu, L. Fang, L. Sun, H. Zeng, Y. Li, H. Zheng, S. Wu, F. Yang, T. Song, J. Lin, C. Ke, Y.
423 Zhang, J. Vinjé, H. Li, Association of GII.P16-GII.2 Recombinant Norovirus Strain with
424 Increased Norovirus Outbreaks, Guangdong, China, 2016. *Emerging Infectious Diseases*. **23**,
425 1188–1190 (2017).

426 24. S. Niendorf, S. Jacobsen, M. Faber, A. M. Eis-Hubinger, J. Hofmann, O. Zimmermann, M.
427 Hohne, C. T. Bock, Steep rise in norovirus cases and emergence of a new recombinant strain

GII.P16-GII.2, Germany, winter 2016. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*. **22** (2017), doi:10.2807/1560-7917.ES.2017.22.4.30447.

25. Y. Matsushima, M. Ishikawa, T. Shimizu, A. Komane, S. Kasuo, M. Shinohara, K. Nagasawa, H. Kimura, A. Ryo, N. Okabe, K. Haga, Y. H. Doan, K. Katayama, H. Shimizu, Genetic analyses of GII.17 norovirus strains in diarrheal disease outbreaks from December 2014 to March 2015 in Japan reveal a novel polymerase sequence and amino acid substitutions in the capsid region. *Euro Surveill*. **20** (2015), doi:10.2807/1560-7917.es2015.20.26.21173.

26. T. M. Fumian, Norovirus Recombinant Strains Isolated from Gastroenteritis Outbreaks in Southern Brazil, 2004–2011. *PLOS ONE*, 12 (2016).

27. K. Kwok, S. Niendorf, N. Lee, T.-N. Hung, L.-Y. Chan, S. Jacobsen, E. A. S. Nelson, T. F. Leung, R. W. M. Lai, P. K. S. Chan, M. C. W. Chan, Increased Detection of Emergent Recombinant Norovirus GII.P16-GII.2 Strains in Young Adults, Hong Kong, China, 2016–2017. *Emerging Infectious Diseases*. **23**, 1852–1855 (2017).

28. L. C. Lindesmith, J. F. Kocher, E. F. Donaldson, K. Debbink, M. L. Mallory, E. W. Swann, P. D. Brewer-Jensen, R. S. Baric, Emergence of Novel Human Norovirus GII.17 Strains Correlates With Changes in Blockade Antibody Epitopes. *J Infect Dis*. **216**, 1227–1234 (2017).

29. G. I. Parra, R. B. Squires, C. K. Karangwa, J. A. Johnson, C. J. Lepore, S. V. Sosnovtsev, K. Y. Green, Static and Evolving Norovirus Genotypes: Implications for Epidemiology and Immunity. *PLOS Pathogens*. **13**, e1006136 (2017).

30. Y. Qian, M. Song, X. Jiang, M. Xia, J. Meller, M. Tan, Y. Chen, X. Li, Z. Rao, Structural Adaptations of Norovirus GII.17/13/21 Lineage through Two Distinct Evolutionary Paths. *Journal of Virology*. **93** (2018), doi:10.1128/JVI.01655-18.

31. M. C. W. Chan, N. Lee, T.-N. Hung, K. Kwok, K. Cheung, E. K. Y. Tin, R. W. M. Lai, E. A. S. Nelson, T. F. Leung, P. K. S. Chan, Rapid emergence and predominance of a broadly recognizing and fast-evolving norovirus GII.17 variant in late 2014. *Nature Communications*. **6**, 10061 (2015).

32. J. Lu, L. Sun, L. Fang, F. Yang, Y. Mo, J. Lao, H. Zheng, X. Tan, H. Lin, S. Rutherford, L. Guo,

C. Ke, L. Hui, Gastroenteritis Outbreaks Caused by Norovirus GII.17, Guangdong Province, China, 2014–2015. *Emerging Infectious Diseases*. **21**, 1240–1242 (2015).

33. Y. Matsushima, M. Ishikawa, T. Shimizu, A. Komane, S. Kasuo, M. Shinohara, K. Nagasawa, H. Kimura, A. Ryo, N. Okabe, K. Haga, Y. Doan, K. Katayama, H. Shimizu, Genetic analyses of GII.17 norovirus strains in diarrheal disease outbreaks from December 2014 to March 2015 in Japan reveal a novel polymerase sequence and amino acid substitutions in the capsid region. *Eurosurveillance*. **20**, 21173 (2015).

34. Y. Ao, X. Cong, M. Jin, X. Sun, X. Wei, J. Wang, Q. Zhang, J. Song, J. Yu, J. Cui, J. Qi, M. Tan, Z. Duan, Genetic analysis of re-emerging GII.P16–GII.2 noroviruses in 2016– 2017 in China, 33 (2018).

35. B. Prevost, F. S. Lucas, K. Ambert-Balay, P. Pothier, L. Moulin, S. Wurtzer, Deciphering the Diversities of Astroviruses and Noroviruses in Wastewater Treatment Plant Effluents by a High-Throughput Sequencing Method. *Appl. Environ. Microbiol*. **81**, 7215–7222 (2015).

36. J. H. Lun, J. Hewitt, A. Sitabkhan, J.-S. Eden, D. Enosi Tuipulotu, N. E. Netzler, L. Morrell, J. Merif, R. Jones, B. Huang, D. Warrilow, K.-A. Ressler, M. J. Ferson, D. E. Dwyer, J. Kok, W. D. Rawlinson, D. Deere, N. D. Crosbie, P. A. White, Emerging recombinant noroviruses identified by clinical and waste water screening. *Emerging Microbes & Infections.* **7** (2018), doi:10.1038/s41426-018-0047-8.

37. B. Prevost, F. S. Lucas, A. Goncalves, F. Richard, L. Moulin, S. Wurtzer, Large scale survey of enteric viruses in river and waste water underlines the health status of the local population. *Environment International*. **79**, 42–50 (2015).

38. N. D. Grubaugh, J. T. Ladner, P. Lemey, O. G. Pybus, A. Rambaut, E. C. Holmes, K. G. Andersen, Tracking virus outbreaks in the twenty-first century. *Nature Microbiology*. **4**, 10–19 (2019).

39. J. T. Ladner, N. D. Grubaugh, O. G. Pybus, K. G. Andersen, Precision epidemiology for infectious disease control. *Nature Medicine*. **25**, 206–211 (2019).

40. J. Peccia, A. Zulli, D. E. Brackney, N. D. Grubaugh, E. H. Kaplan, A. Casanovas-Massana, A. I. Ko, A. A. Malik, D. Wang, M. Wang, J. L. Warren, D. M. Weinberger, S. B. Omer, *medRxiv*, in

press, doi:10.1101/2020.05.19.20105999.

41. W. Randazzo, P. Truchado, E. Cuevas-Ferrando, P. Simón, A. Allende, G. Sánchez, SARS-CoV-2 RNA in wastewater anticipated COVID-19 occurrence in a low prevalence area. *Water Research*. **181**, 115942 (2020).

42. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. **34**, i884–i890 (2018).

43. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, BLAST+: architecture and applications. *BMC Bioinformatics*. **10**, 421 (2009).

44. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol*. **32**, 268–274 (2015).

45. G. Yu, D. K. Smith, H. Zhu, Y. Guan, T. T.-Y. Lam, ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*. **8**, 28–36 (2017).

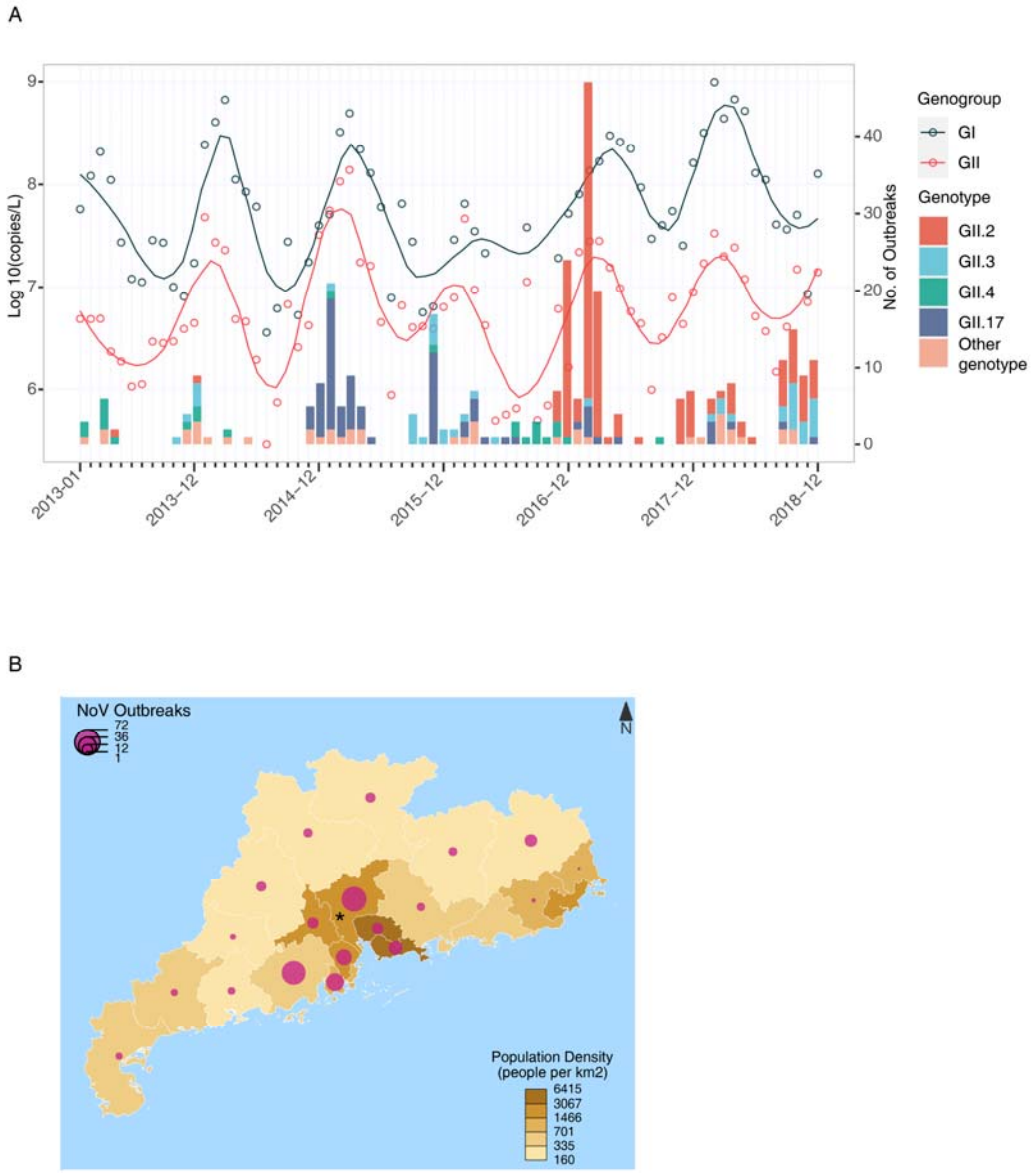
46. S. Kojima, T. Kageyama, S. Fukushi, F. B. Hoshino, M. Shinohara, K. Uchida, K. Natori, N. Takeda, K. Katayama, Genogroup-specific PCR primers for detection of Norwalk-like viruses. *Journal of virological methods*. **100**, 107–14 (2002).

Acknowledgement: This study was supported by The Key Research and Development Program of Guangdong Province (2019B111103001) and Science and Technology Planning Project of Guangzhou (201804010030).

Competing interests: All authors declare no competing interests.

Author contributions: JL designed the study. JL and HYZ had full access to all data in the study, and take responsibility for the integrity and accuracy of the data. JL, LF, JJP, LLZ, HFL and QLX contributed to NGS sequencing and the sequences analysis. JL, JJP, LLZ and QLX contributed to literature search writing of the manuscript. All authors contributed to data acquisition, data interpretation, and all reviewed and approved the final version of the manuscript.

513 **Figures and Figure legends**



514

515 **Fig. 1. Norovirus outbreaks in Guangdong Province, China, January 2013- December 2018.** (A)
516 Temporal distribution of viral genotypes in norovirus outbreaks. The relative numbers of viral copies of
517 GI and GII norovirus in Guangzhou sewage were plotted each month. The samples with viral loads
518 below the detection threshold were set to a Ct-value of 40 ($<1x10^5$ copies/L) for convenience. (B)
519 Spatial distribution of norovirus outbreaks in 21 prefectural cities in Guangdong, China. Guangzhou,
520 the hub of the norovirus epidemic, is marked with a star.

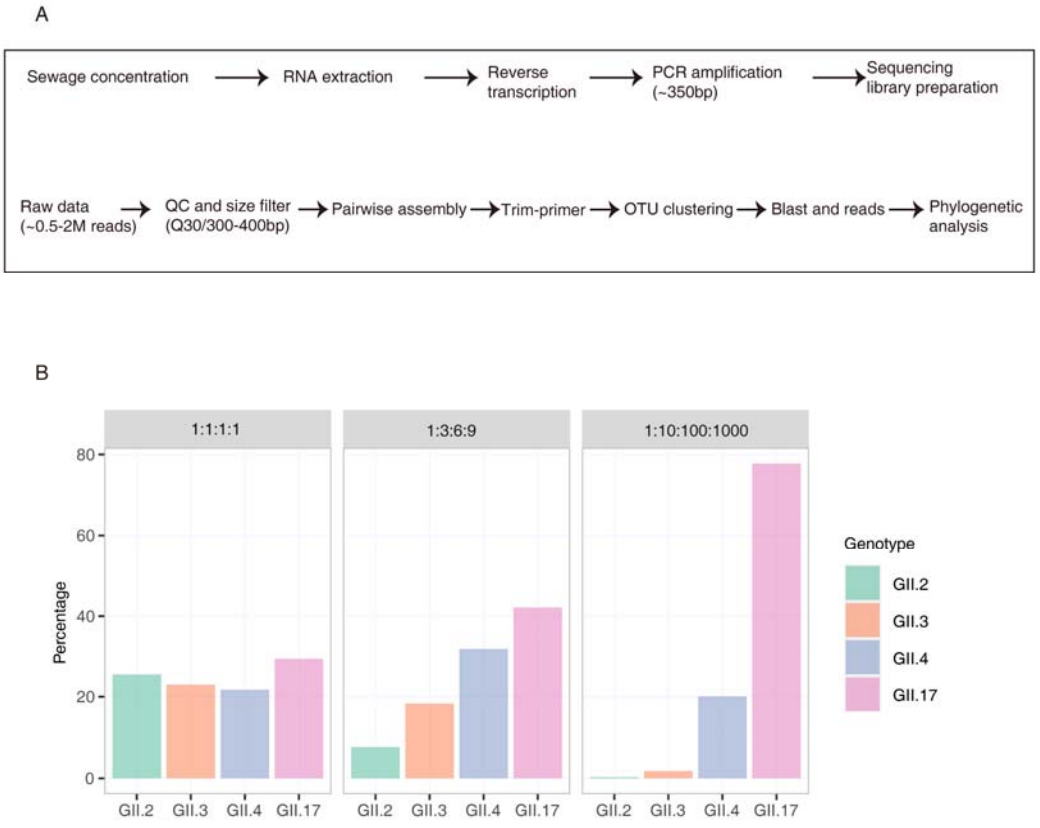
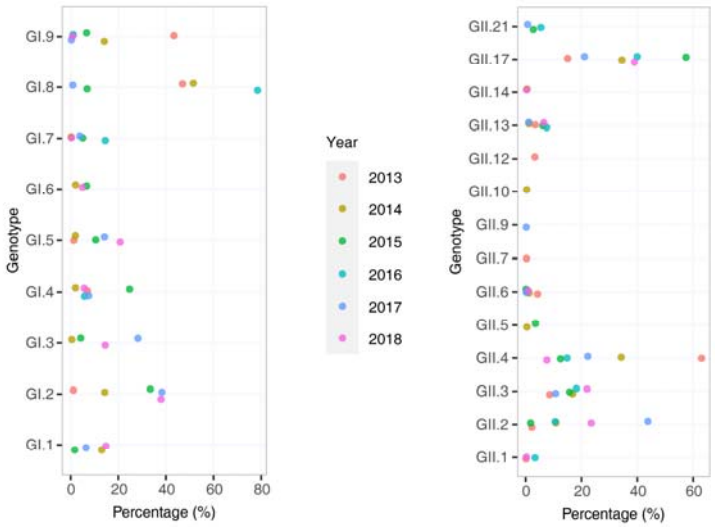
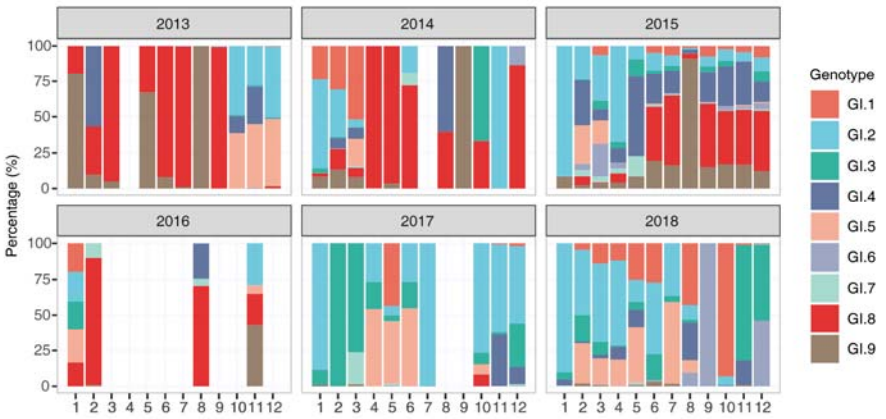


Fig. 2. A high-throughput method for detecting NoVs in sewage. (A) workflow for calculating genotype distribution in sewage. (B) Method verification by using GII.2, GII.3, GII.4, and GII.17 plasmid constructs mixed in different ratios.

A



B

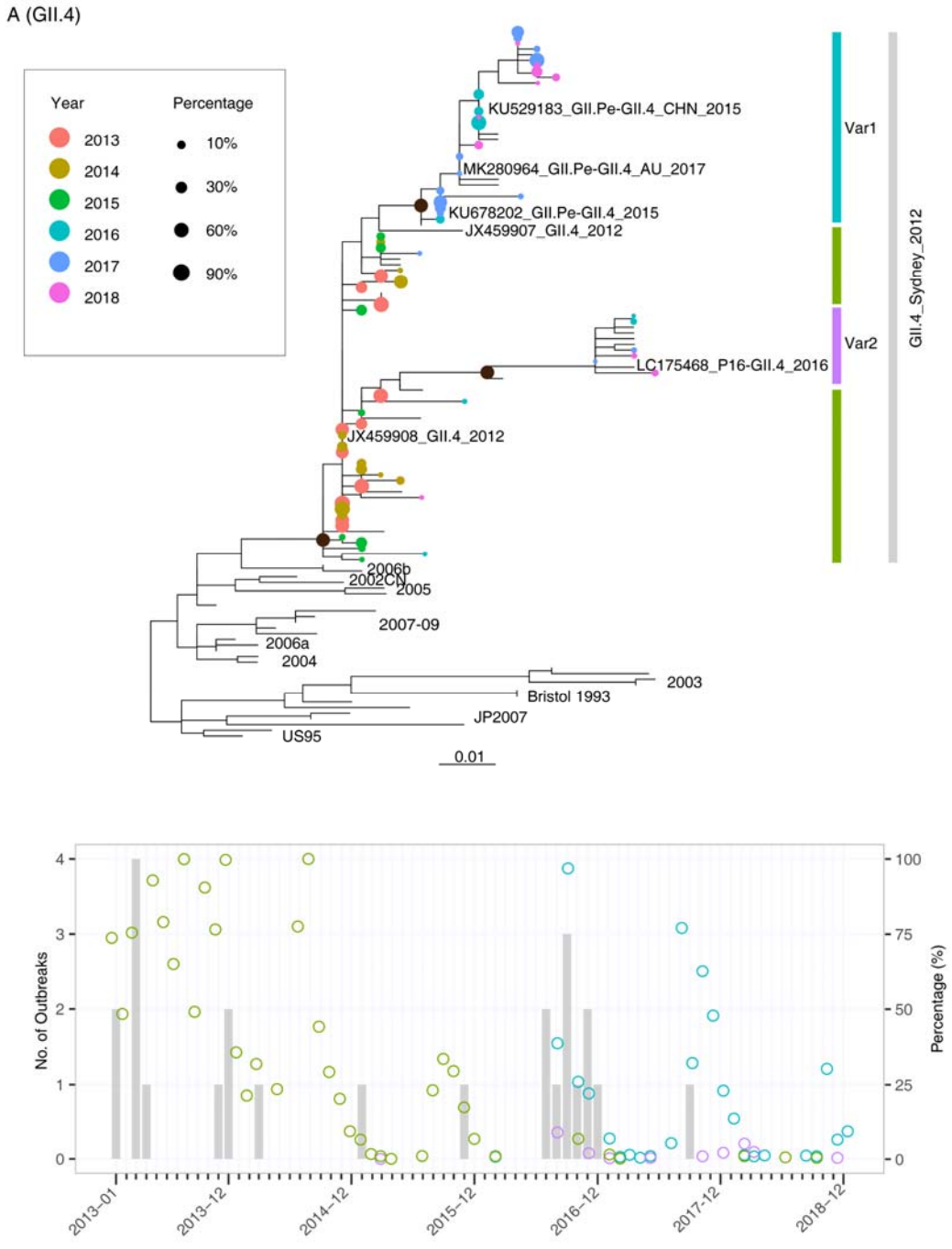


525

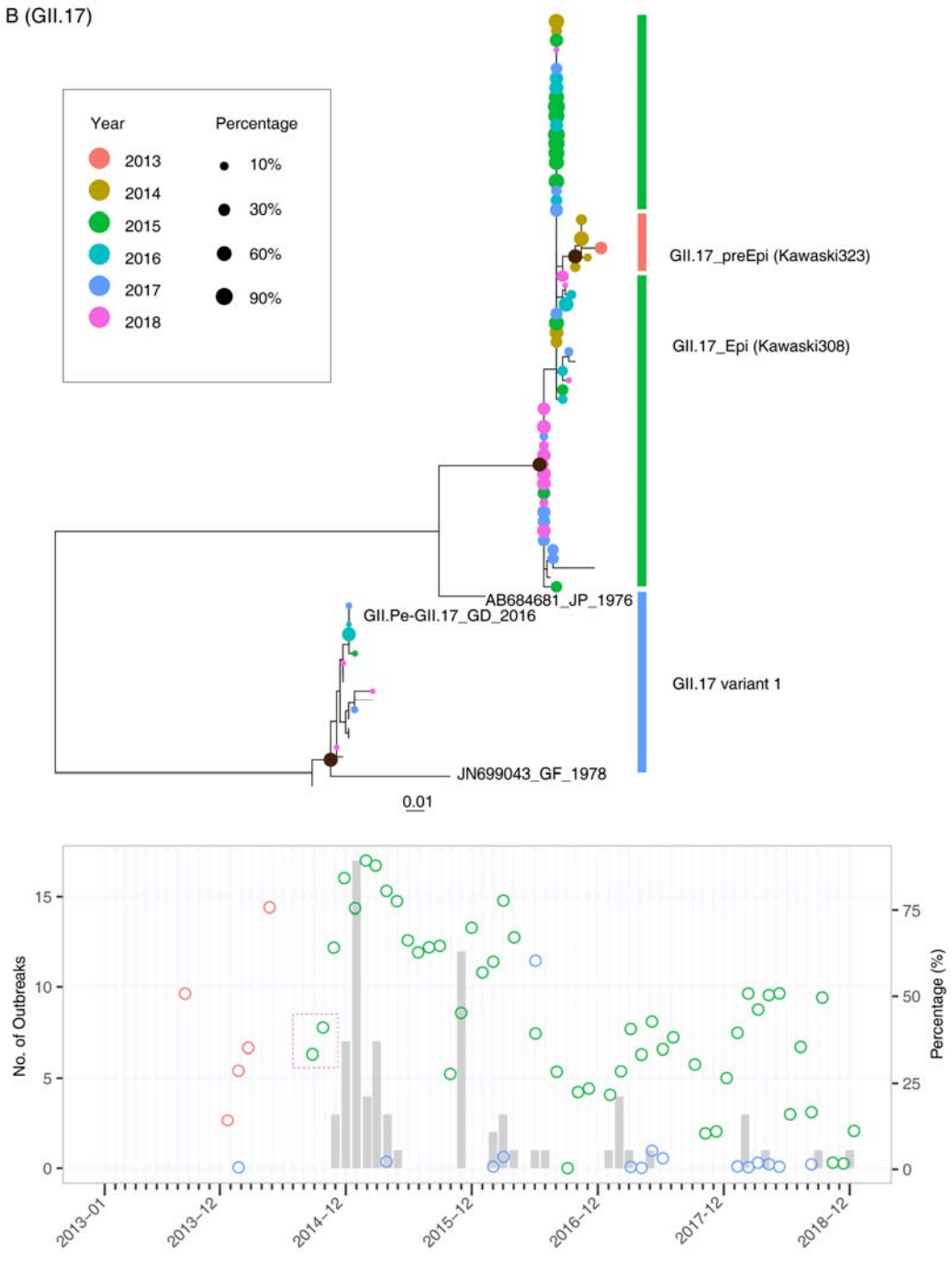
C

526

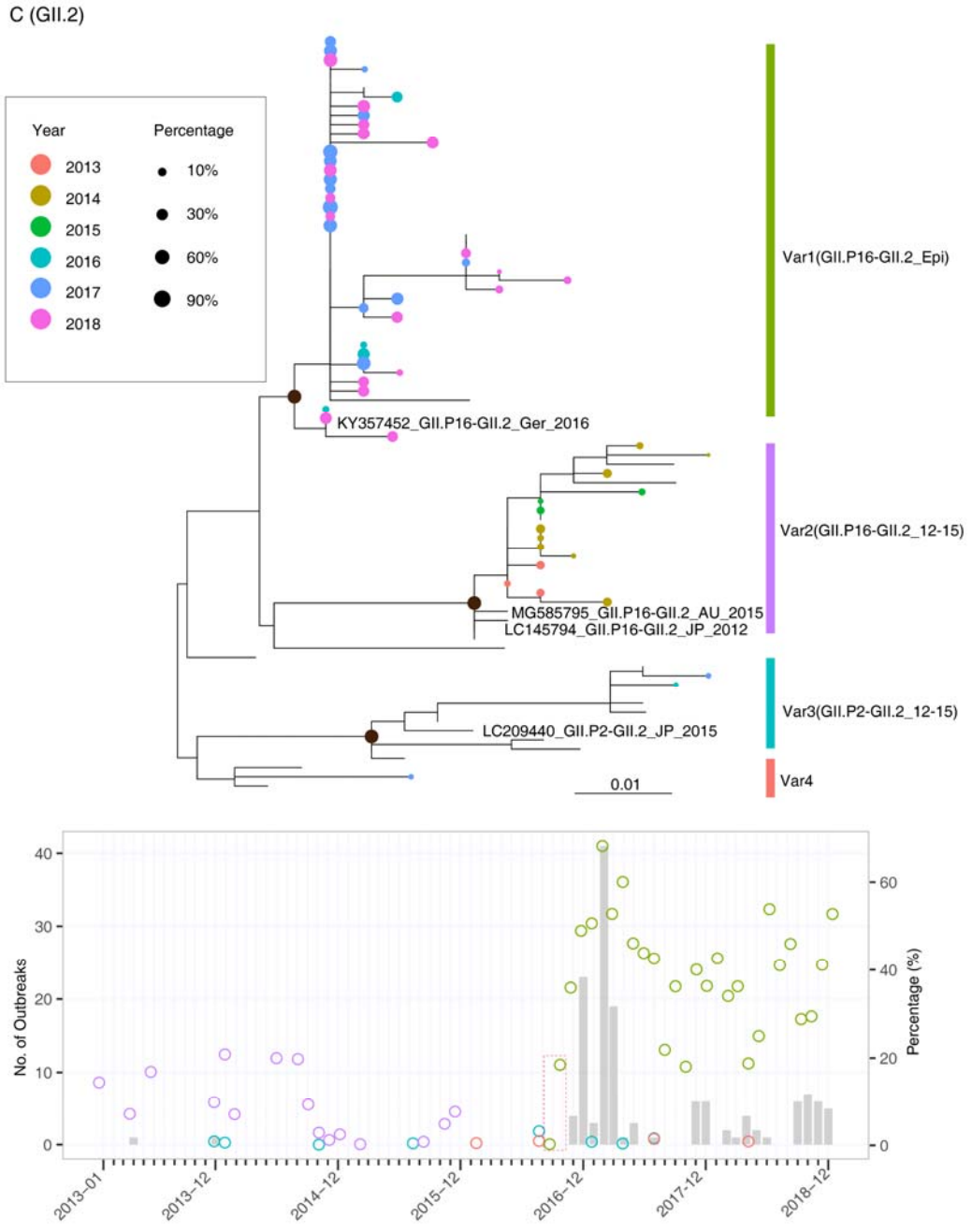
527 **Fig. 3. Distributions of norovirus genotypes in sewage samples.** Annual distribution (A) and
528 monthly distribution (B) of different GI genotypes in sewage samples, January 2013–December 2018.
529 Annual distribution (C) and monthly distribution (D) of different GII genotypes, January
530 2013–December 2018.



531
532



533



534

535 **Fig. 4. Phylogenetic analysis of different genetic variants and their dynamic changes in sewage**
536 **from January 2013 to December 2018.** Upper panel: maximum-likelihood trees were constructed by
537 using the consensus sequences generated from OTU clusters of GII.4 (A), GII.17 (B), and GII.2 (C).
538 The tip points are colored according to the year in which the samples were collected, and the size of the
539 points was adjusted according to their abundances in sewage samples. Lower panel: The percentages of
540 different genetic variants in sewage are plotted with circles, which are colored according to the clusters
541 defined in the phylogenetic trees. The early detection of novel genetic variants in sewage was
542 highlighted with a red box.

543 **Supplementary Materials**

Table S1 The primer set used in the PCR amplification

	Primer*	Seq (5 ' to 3 ')	Orientation
	COG1F	CGYTGGATGCGNTTYCATGA	Forward
GI	G1SKF	CTGCCCGAATTYGTAAATGA	Forward
	G1SKR	CCAACCCARCCATTRTACA	Reverse
	COG2F	CARGARBCNATGTTYAGRTGGATGAG	Forward
GII	G2-SKF	TGGGAGGGCGATCGCAA	Forward
	G2SKR	CCWCCNGCATRHCCRTTRTACAT	Reverse

544 *Primers are set according to the previous publication with a few modifications (*46*).

545

546 **Table S2 Reads numbers of NoV collected from sewage sample in 2013-2018**

547

548 **Fig. S1. Maximum Likelihood tree of GII.4.** The accession numbers of reference sequences and the
549 name of sequences recovered from different sewage samples have been noted. The tree was annotated
550 with the same scheme in Fig. 4A.

551

552 **Fig. S2. Maximum Likelihood tree of GII.17.** The accession numbers of reference sequences and
553 the name of sequences recovered from different sewage samples have been noted. The tree was
554 annotated with the same scheme in Fig. 4B.

555

556 **Fig. S3. Maximum Likelihood tree of GII.2.** The accession numbers of reference sequences and the
557 name of sequences recovered from different sewage samples have been noted. The tree was annotated
558 with the same scheme in Fig. 4C.

559