Selection of a phylogenetically informative region of the norovirus genome for outbreak linkage

Linda Verhoef · Kelly P. Williams · Annelies Kroneman · Bruno Sobral · Wilfrid van Pelt · Marion Koopmans · on behalf the FBVE network

Received: 25 May 2011 / Accepted: 12 September 2011 / Published online: 30 September 2011 © The Author(s) 2011. This article is published with open access at Springerlink.com

Abstract The recognition of a common source norovirus outbreak is supported by finding identical norovirus sequences in patients. Norovirus sequencing has been established in many (national) public health laboratories and academic centers, but often partial and different genome sequences are used. Therefore, agreement on a target sequence of sufficient diversity to resolve links between outbreaks is crucial. Although harmonization of laboratory methods is one of the keystone activities of networks that have the aim to identify common source norovirus outbreaks, this has proven difficult to accomplish, particularly in the international context. Here, we aimed at providing a method enabling identification of the genomic region informative of a common source norovirus outbreak by bio-informatic tools. The data set of 502 unique full length capsid gene sequences available from the public domain, combined with epidemiological data including linkage information was used to build over 3,000 maximum likelihood (ML) trees for different sequence lengths and regions. All ML trees were evaluated for robustness and specificity of clustering of known linked norovirus outbreaks against the background diversity of strains. Great differences were seen in the robustness of commonly used PCR targets for cluster detection. The capsid gene region spanning nucleotides 900–1,400 was identified as the region optimally substituting for the full length capsid region. Reliability of this approach depends on the quality of the background data set, and we recommend periodic reassessment of this growing data set. The approach may be applicable to multiple sequence-based data sets of other pathogens.

Keywords Norovirus · Nucleotide sequence data · Surveillance · Outbreaks · Outbreak linkage

Introduction

Cluster detection based on identification of related nucleotide (nt) sequences of pathogens in patients is an important tool to support outbreak investigations in modern day public health and clinical laboratories. The added value of these approaches is particularly clear in detection of links that are difficult to unravel through classical epidemiological investigations, for instance in diffuse food-borne outbreaks involving several countries. Viral contamination of food can occur through infected food handlers, or the use of sewage contaminated water during cultivation, production, and processing of foods. Unfortunately, the standard legally required quality-control criteria for food are not adequate for detection or exclusion of viral contamination [1, 2]. Given the globalization of the food market, a single
batch of food is often consumed in several countries simul-
taneously [3], and may consequently cause international
market. However, such decisions depend on the timely
detection of viruses in food and patients, and correct inter-
pretation of sequence comparison between countries. A
prerequisite for comparing sequences internationally is the
presence of a sufficiently large up-to-date background data
set describing viruses endemic in the community, which is
collected according to standardized detection and typing
methods. Such data may only be available for a select
number of countries [4, 5].

The need for standardized molecular surveillance is a
general problem increasingly recognized for food- and
waterborne diseases. Several electronic surveillance sys-
tems for norovirus outbreaks have recently been initiated in
the US (Calicinet), Canada (Vironet), Australia, and New
Zealand (Norovirus Surveillance Network) and globally
(Noronet) [6]. The Foodborne Viruses in Europe (FBVE)
network was one of the pioneers collecting both laboratory
and epidemiological data on norovirus outbreaks in Europe
since 1999 [7]. Noroviruses (NoVs) are among the most
prevalent causative agents of community acquired viral
gastroenteritis [8, 9]. NoVs constitute a genetically diverse,
single-stranded, positive-sense RNA virus genus within the
family Caliciviridae and contain a ~7.5 kb genome with
three open reading frames (ORF1-3). ORF1 encodes a
polyprotein comprising all non-structural proteins, which is
autocatalytically cleaved to produce the non-structural
proteins. ORF2 encodes the major structural capsid protein
containing the major antigenic and receptor binding sites of
the virus; the different domains of ORF2 are the N-terminal
domain, the shell domain, and the protruding domain
subdivided into P1 and P2. The norovirus particle is built
from 180 copies of the capsid protein (90 dimers). ORF3
encodes protein the function of which is unknown but may
involve regulation of the expression and stability of VP1
[10]. Currently five norovirus genogroups have been
described which can be subdivided into at least 40 geno-
types based on their amino acid capsid sequence [11, 12].

Although harmonization of laboratory methods was one
of the keystone activities of the FBVE network, this has
proven difficult to accomplish, as no consensus could be
reached among laboratories with respect to the genomic
target region. For monitoring trends at the level of geno-
types, resolution of sequence-based typing does not need to
be very high, and this can be achieved by sequencing a
relatively conserved genomic fragment [6]. However, the
identification of epidemiologically linked patients requires
sequence typing at a much higher resolution level.

Koopmans et al. [13] illustrated that increasing deletions in
a fragment leads to misclassification, since the number of
sequences clustered as different decrease. As sequencing
may be done by local public health laboratories with lim-
ited resources, for the time being the option of full genome
sequencing in all participating laboratories is excluded. A
more recent study made a start toward a scientific basis for
harmonization by comparing two standardized ORF2
genotyping protocols in a small set of pre-selected noro-
virus strains [14]. However, with the rapid development of
technology of both sequencing methods as well as com-
putational tools [15], systematic analysis of large databases
with both sequences and epidemiological data like those in
the FBVE database is now possible. Here, we provide a
bioinformatic approach for identifying the most informa-
tive region of NoV for outbreak investigations including
outbreak linkage, to guide future laboratory efforts in
harmonizing typing methods.

Methods

Data set

Sequence selection

We compiled 573 norovirus capsid gene sequences with
background epidemiological data, as available from the
public domain on April 1, 2010, representing the diversity
of norovirus strains detected since 1999. Sequences were
collected through the FBVE network (N = 163) and
Genbank (N = 410). In all, these 573 sequences consisted
of 502 unique sequences. Accession numbers and back-
ground information of sequences used are provided in
electronic supplementary material 1.

Sequence classification

Genogroups, genotypes, and variants were assigned
according to capsid-based phylogenetic clustering (http://
www.rivm.nl/mpf/norovirus/typingtool) [16]. Variants are
emerging lineages of the GI.4 genotype displacing the
resident viruses, and which are, with some exceptions,
found to descend from their chronologic predecessor [17].
A total of 72 sequences (53 unique) were defined as clus-
ters of linked sequences, on the basis of available epi-
demiological and molecular information reported to FBVE,
and published work [18–21]. Two clusters (Table 1) rep-
resented known common source internationally dispersed
outbreaks for which multiple strains of patients had been
detected in different countries, i.e., ‘event 1’ with two II.1
sequences [22], and ‘event 2’ with nine II.4-2006b
sequences [18, 20]. Nine additional clusters consisting of
sequences detected in single patients over a prolonged period of time. These were added to the data set as linked sequences that may represent the diversity within outbreaks (Table 1, shedders 1–9). The nine shedder clusters consisted of 42 strains, i.e., II.3 (shedders 4, 6, 9), II.4-2006a (shedders 2, 5, 8), II.4-2006b (shedder 7), and II.4-2004 (shedder 1, 3) [19, 21]. We further refer to these 11 clusters as ‘outbreak events.’

Phylogenetic analysis

Full alignment

Full capsid nt sequences \((n = 502)\) were translated into amino acids (AA) and protein sequences were aligned using the default mode in MUSCLE version 3.6 [23]. Alignments for each genogroup were prepared separately, then merged as profiles into the full alignment. The AA in the full capsid alignment were converted back to the corresponding codon triplets. Tulane virus (EU391643) was included as outgroup, which is likely to be the type species of the new genus Recovirus of the Caliciviridae closely related to norovirus [24].

Subalignments

A script was written in Perl (www.perl.com) to prepare alignments of simulated PCR products of sequences of different lengths and targets alignments, i.e., subalignments, which were sliding windows of varying size (100, 200, 250, 300, 400, and 500 nt). To standardize window-taking from the gapped full alignment, a reference sequence was arbitrarily selected (a II.4 genotype strain; AB220921, 1,620 nt excluding the stop codon). Subalignments were taken to include the given number of nt from this reference sequence. The program was written to start in the middle position of the aligned reference sequence (for window-100: \((1,620 – 100)/2 = 760\), then repositioned to both ends (for window-100: 1, and 1,520), then repositioned to the two middles (for window-100: 760/2 = 380, and 380 + 760 = 1,140) and so on, so that the windows would be evenly distributed over the capsid gene.

Maximum likelihood (ML) tree building

A script was written to run ML tree building for the full alignment as well as for all sub-alignments in RAxML 7.0.4 [25]. ML tree building was done using the substitution model GTRGAMMA [25–27], partitioning the 3rd codon position from the other two codon positions, as mutations in this position rarely result in amino acid changes.

Tree comparison

ML tree scoring

All ML trees were evaluated for their ability to cluster each group of a type (these types being genotypes, variants, and outbreak events) in isolation from other groups of the same type. To do so, a script was written to calculate a clade impurity score for each group type in each tree, using the bipartitions analysis in Phylip version 3.69 (http://en.bio-soft.net/tree/Phylip.html), as follows. For each group, the smallest clade containing all sequences belonging to the group was identified; this can be called the ‘minimal differentiating clade.’ The number of invading clades was counted, i.e., subclades within the minimal differentiating clade that consist only of sequences not belonging to the group. This invader count was normalized by dividing by the maximum possible invader count for the group, i.e., the number of leaves in the tree that are not part of the group. Such normalized invader counts were averaged for all groups of the same type, yielding a clade impurity score between 0 and 1 for the tree, with a score of 0 for the optimal case of a pure clade where no groups have any invading clades.

Tree scores comparison

The clade impurity score for each sub-alignment tree was plotted against its mid-position of the subalignment in the capsid gene, and for all window sizes in order to identify the optimal area of the capsid gene with the lowest impurity scores.

Validation

Bootstrap analysis

In order to identify the optimal typing region and to evaluate fragment length in the optimal typing region, non-parametric bootstrap values were calculated for the full capsid gene tree as well as the subalignment trees for each of six window sizes for several center positions within the identified area in the tree-score comparison step. In addition, bootstrap analysis were performed for ML trees based on the alignments which would be obtained when using genotyping protocols for genomic regions currently commonly applied [14, 28–31] for region C, D, E, and the P2 domain. This analysis step thus resulted in bootstrap values for 43 ML trees. One hundred runs of exhaustive bootstrap analysis were performed using RAxML 7.0.4 (i.e., in script: option -f i).
Clade support values

For each tree, all clade-supporting bootstrap values were identified supporting the minimal differentiating clades of single genotypes (or variants, or outbreak events). A bootstrap value \( C_{70} \) was considered well performing [32]. Results were ranked into six categories, based on the clade-supporting bootstrap values and the number of invading sequences (Table 1).

Specificity analysis

Specificity was considered ‘able to cluster strains from a specific genotype (or variant, or outbreak event) together as a pure clade separated from others in the ML tree.’ To validate phylogenetic trees constructed from fragments of various sizes part of the complete capsid nt sequences, the specificity was calculated as the percentage of genotypes (or variants or outbreak events) meeting criterion 1 (i.e., pure clades with bootstrap value \( \geq 70 \)) as well as for criteria 1 and 2 together (i.e., pure clades irrespective of the bootstrap values), with 80% being considered adequate specificity.

Results

Phylogenetic analysis

Full alignment ML tree

The alignment covered a total of 1,957 positions, with the longest sequence containing 1,677 nt. The full alignment
resulted in a ML tree capable of clustering all genotypes, 10/12 variants and all outbreak events grouped together but separately from other genotypes, variants, and outbreak events (Fig. 1, of which detailed tree information in Nexus format is provided in electronic supplementary material 2).

Subalignment ML trees

In a preliminary run of 780 ML trees (130 for each window size), we found the window-100 performed poorly for segregation of variants and outbreak events, while window-250 clearly performed better (data not shown), confirming previous observations [13]. In the second tree building run, leaving out the window-100, and running extra analysis for window-250, a total of 2,295 additional trees were built for evenly distributed windows of 200–500 nt, including a higher frequency for window-250, for a total of 3,075 ML trees for analysis: 130 window-100; 513 window-200; 959 window-250; 513 window-300; 513 window-400; and 513 window-500 trees. Thus, average window spacing was 11.7 nt for window-100, 1.5 nt for window-250, and 2.2–2.8 nt for the others. This is very high coverage of the 7,976 possible windows; the remainder would have added little information and they were not processed.

Tree comparison

Maximum likelihood trees were scored for the impurity of clades containing known clusters indicating the ability to discriminate genotypes, variants, or outbreak events. The full alignment tree performed very well, showing 100% segregation of strains belonging to distinct genotypes and outbreak events (score 0), and two invading clades for two variants (score 0.000,759). Scores for subalignment trees were plotted in Fig. 2. The top panel shows that for resolving genotypes all regions across the gene perform well, even with the smallest (100-nt) windows. However, for resolving variants or outbreak events, smaller windows perform distinctly worse, as do particular regions of the capsid gene. The gene region with center positions between 900 and 1,150, encoding the P2 portion of the capsid protein, showed lowest impurity scores for variants and outbreak events (Fig. 2) and was therefore considered promising for resolving clusters.

Validation

Bootstrap analysis and clade support values

Within the promising region, for six center positions of each of the six window sizes bootstrap analysis was performed. The region centered on nt position 1,150 appeared optimal and was selected for further analysis. Table 1 shows 13 bootstrap results, with six colors representing levels of the ability for a region to separate genotypes, variants, and outbreaks. The full alignment tree (column 1 in Table 1) clearly shows best results for all levels of pure clades with bootstrap values ≥70 (criterion 1) and pure clades irrespective of the bootstrap values (criteria 1 and 2). With respect to the subalignments the window of 500 nt, i.e., positions 900–1,400, is best approaching the full alignment.
Table 1 Bootstrap values, as derived from 100 runs of ML trees built in RAxML 7.0.4 [25], for clades without invading sequences or with a maximum of one invading sequence, and for different levels of resolution, i.e., genogroups, genotypes, variants, and outbreak events and bootstraps values were calculated for different fragment lengths within their optimal genomic region, and for several target regions in commonly applied genotyping protocols (color online).

Different colors indicate different levels for mismatching of clades coined to a genotype, variant or outbreak event, as follows:

- Pure clade with bootstrap value > 70;
- Pure clade with bootstrap value < 70
- Impure clade with one invasion sequence and with bootstrap > 70.
- Impure clade with one invasion sequence and with bootstrap < 70

Clade is split in groups, or impure including 2 or more invasion sequences from other genotypes (or variants, or outbreak events), i.e. others of same level; clade support value unresolved

Clade is split in groups, or impure including 2 or more invasion sequences from other genogroups (or genotypes, or variants), i.e. others of a higher level; clade support value unresolved.
Table 1 continued

| Event 1 [II.4] | 100% | 99% | 98% | 97% | 96% | 95% | 94% | 93% | 92% | 91% | 90% | 89% | 88% | 87% | 86% | 85% | 84% | 83% | 82% | 81% | 80% |
| Event 2 (2005) | 94% | 93% | 92% | 91% | 90% | 89% | 88% | 87% | 86% | 85% | 84% | 83% | 82% | 81% | 80% | 79% | 78% | 77% | 76% | 75% | 74% | 73% | 72% | 71% | 70% | 69% | 68% | 67% | 66% | 65% | 64% | 63% | 62% | 61% | 60% | 59% | 58% | 57% | 56% | 55% | 54% | 53% | 52% | 51% | 50% | 49% | 48% | 47% | 46% | 45% | 44% | 43% | 42% | 41% | 40% | 39% | 38% | 37% | 36% | 35% | 34% | 33% | 32% | 31% | 30% | 29% | 28% | 27% | 26% | 25% | 24% | 23% | 22% | 21% | 20% | 19% | 18% | 17% | 16% | 15% | 14% | 13% | 12% | 11% | 10% | 9% | 8% | 7% | 6% | 5% | 4% | 3% | 2% | 1% | 0% |

Specificity

Table 2 shows specificity for each of 13 genomic regions in recognizing genogroups, genotypes, variants, and outbreak events as groups separated from other sequences in their ML tree. The full capsid ML tree clearly shows optimal performance, with 100% specificity for typing of genogroups, genotypes and outbreak events, and 83% specificity for typing of variants. The window sizes with the center being the 1,150th nt approach this optimal performance, with window-500 showing best performance for genogroups, genotypes, variants, and outbreak events simultaneously. Window-400 can be considered the minimum fragment length still able to recognize outbreak events included in this analysis as pure clades with adequate specificity (i.e., >80%) if low bootstrap values are considered acceptable. The specificity for variants in windows 500 on the basis of criteria 1 and 2 can be increased to 83% if recognition of variants is based on a shorter fragment within this region, thereby obtaining the same specificity as the full capsid ML tree.

The P2 domain is the best performing PCR currently available for simultaneous recognition of genotypes, variants, and outbreak events (Table 1). By contrast, in >50% of the outbreak events, unrelated strains will be considered part of the outbreak event when using the commonly applied primers for region C or D as a standardized method. The PCR for region E showed adequate performance in recognizing outbreak events, but low in recognizing variants.

Overall, results of the full alignment are maximal for recognizing outbreak events. Its performance can be most closely approached on the basis of pure clades, irrespective of bootstrap values, in ML trees in the 900–1,400 region, while the 950–1,350 region can be used for the II.4-2006b variant. Lack of specificity is mainly caused by low bootstrap values for pure clades (i.e., <70, criterion 2).
Table 2 Specificity (%) of different genomic regions in clustering genogroups, genotypes, variants, and outbreak events as a group separated from others, as derived from bootstrapped ML trees

| Group               | N  | Full          | Windows (center 1,150)          | PCR regions                  |
|---------------------|----|---------------|--------------------------------|-----------------------------|
|                     |    |               | 1–1,620 (1,620 nt)             | 1–282 (282 nt)              |
|                     |    |               | full Capsid                     | G2SK region C               |
|                     |    |               | 1,000–1,300 (300 nt)           | region C                    |
|                     |    |               | window-300                     | G1/SK region C              |
| Genogroups          | 4  | 100           | 900–1,400 (500 nt)             | region E                    |
| GI genotypes        | 8  | 100           | window-400                     | P2 domain                   |
| GII genotypes       | 20 | 100           | 95–1,284 (280 nt)              | CapC/D1/D3 region D         |
| GIII genotypes      | 2  | 100           | 100–284 (280 nt)               | CapA/B1/B2 region D         |
| Variants            | 12 | 83            | 67–1,439 (143 nt)              |                             |
| Outbreak events     | 11 | 100           | 5–1,459 (143 nt)               |                             |

On the basis of criterion 1: pure branches with support values >70

| Genogroups          | 4  | 100           | 75–1,275 (250 nt)              | 1–282 (282 nt)              |
| GI genotypes        | 8  | 100           | 63–1,025 (250 nt)              | G2SK region C               |
| GII genotypes       | 20 | 100           | 50–1,250 (250 nt)              | region E                    |
| GIII genotypes      | 2  | 100           | 50–1,100 (250 nt)              | P2 domain                   |
| Variants            | 12 | 83            | 50–1,150 (250 nt)              | CapC/D1/D3 region D         |
| Outbreak events     | 11 | 100           | 50–1,200 (250 nt)              | CapA/B1/B2 region D         |

On the basis of criteria 1 and 2: pure branches irrespective of support values

| Genogroups          | 4  | 100           | 50–1,350 (200 nt)              | 1–282 (282 nt)              |
| GI genotypes        | 8  | 100           | 50–1,350 (200 nt)              | G2SK region C               |
| GII genotypes       | 20 | 100           | 50–1,284 (280 nt)              | region E                    |
| GIII genotypes      | 2  | 100           | 50–1,100 (250 nt)              | P2 domain                   |
| Variants            | 12 | 83            | 50–1,150 (250 nt)              | CapC/D1/D3 region D         |
| Outbreak events     | 11 | 100           | 50–1,200 (250 nt)              | CapA/B1/B2 region D         |

Nucleotide position was chosen from the reference strain AB220921, and nt position 1 of the capsid gene corresponds with nt position 5,085 of the GII-4 strain Lordsdale/1,995/UK (GenBank X86557)
However, despite these low bootstrap values, the known clusters appear as pure clades in the ML trees.

Discussion

In a novel approach combining bioinformatics, epidemiologic and virologic data and viral nt sequence data, we identified nt positions 900–1,400 as the informative genomic regions best approaching the full capsid sequence in its ability to correctly assign genotypes, variants, and the outbreaks events used in this analysis simultaneously. The positions 950–1,350 of these norovirus capsid genes can be considered the target and minimum fragment length for laboratory networks aiming to identify outbreak events with specificity >80%. This target lies within the variable P-domain (Fig. 2) coinciding with the exposed part of the capsid, which was expected to contain the most informative region since it shows the largest genetic variation. This is also confirmed with the comparison of currently used regions [14, 28–30] where we found that the P2 domain (795–1,253), the most variable region within the P domain, best approaches the full capsid sequence. The shell domain, including regions C and E, contains the more conserved regions of the capsid. Use of region E (nt 301–577) tends to result in difficulties in distinguishing variants as pure clades from other clades, whereas PCRs for regions C (~nt 1–284) and D (~nt 1,372–1,585) tend to create difficulties in distinguishing at least one-third of the selected outbreak events as pure clades. Although we did not identify a conserved region for consensus primers covering the identified region in the broad range of noroviruses included in our study, target regions for primers are available for the ~600–1,400 nt positions since the fragment is located between regions E and D. Still, further virological research is recommended to identify (degenerate) primers more focused and optimal for amplifying the identified region.

Non-parametric bootstrapping, i.e., randomized selection of columns in the alignment with replacement, is commonly accepted as a method for assessing confidence of phylogenetic analysis [32]. It was proposed as a method for obtaining confidence limits on phylogenies that would be estimated from repeated sampling of many characters from the underlying set of all characters, and not the true phylogeny [33, 34]. Our analysis illustrates that pure clades of closely related strains based on fragments may, irrespective of their bootstrap value, reflect the phylogeny in the full capsid sequences from which these fragments were derived. Pure clades were found in ML trees on the basis of subalignments, and could be confirmed in the full capsid sequence alignment ML tree, where bootstrap values were well over 70. The low bootstrap values can be understood when considering the low number of informative sites at these levels of resolution, where few mutations may be informative of the common ancestor. Randomization may thus exclude the informative sites, which will have a stronger effect on closely related strains in outbreak events or variants, when compared to genotypes or genogroups.

The approach chosen provides a method for comparing quality of results and weight of the conclusions drawn on the basis of the use of different genotyping targets. However, we caution against over-interpretation, since the noroviruses that are selected for full capsid sequencing may not represent the general norovirus population. Most of our known clustered sequences (to represent outbreaks) were from patients chronically shedding the virus. This may explain the fact that the region subject to selective pressure was again identified as the most informative region for grouping outbreak strains together [19, 31], although Xerry et al. [31] considered strains with one or more mutations in this hyper variable region as representing unrelated transmission events. Still, random mutations will remain informative in linking outbreak strains. It will be interesting to see if the P2 domain and nt 900–1,400 remain the informative regions during prospective analysis of outbreaks with systematic full capsid sequencing of all patient samples, and compared to a set of strains randomly selected for full capsid sequencing.

The current analysis was only possible through the availability of the systematically collected FBVE data. Nevertheless, our findings may have consequences for such networks’ conclusions with respect to the identification of international outbreak events. For example, in the FBVE network, 98% GI and 7% GII partial capsid sequences covered region E [35]. In CaliciNet the conjunction between ORF1 and ORF2 (i.e., including region C of the capsid gene) and region D are considered best practice for norovirus detection as this region is highly conserved [14, 36]. Thus, if no secondary typing protocol was used, these networks are likely to have not recognized outbreak events within the background sequence noise. In NoroNet, a network aiming to detect emergence of new variants, regions C and D of the capsid gene were used for collection of representative sequences. However, variant assignment needed to be based on the full capsid sequence [6]. Future confirmation of our findings is likely to serve laboratory efforts in identifying outbreak events by cutting down the number of clustering sequences to the most likely related ones.

In our aim to develop a generic method that should be applicable to multiple sequence-based data sets of other pathogens as well, we included a large variety of norovirus sequences in the alignment. Alignments within genotypes, however, will logically show better performance of the ML trees. Therefore, aside from this method, in real-time
analysis for confirmation of an outbreak event, additional phylogenetic analysis is needed. Nevertheless, in real-time the information of genogroup, genotype or—new—variant is not known either. Thus, a consensus typing region should be able to distinguish outbreak events irrespective of the genogroup, genotype or variant involved. Therefore, we considered the inclusion of a broad range of sequences justified.

Outbreaks caused by sewage contaminated foods may involve multiple strains of different genotypes [37] potentially resulting in recombination events. Recombination is a frequent event in noroviruses [38] that may be missed if linking is only focused on the identification of closely related strains. However, each of these multiple or recombined strains is likely to evolve within outbreaks as well, and phylogenetics may still add to the identification of such outbreaks. Nevertheless, the epidemiological focus should not be omitted in order not to miss outbreak events involving multiple genotypes.

Our results, like other studies, point toward considering an agreement among all laboratories to sequence the P2 region of the NoV capsid of outbreak specimens. Subsequently, specimen aliquots can be sent to specialist laboratories where full length capsid gene sequencing can be carried out if the linkage of international outbreaks is suspected. This procedure may also be the way forward to obtaining a larger number of full capsid sequences for periodic reassessment of such a procedure and further confirmation of the P-domain as containing the most informative genomic region. We will make the method available in the public domain. Despite the calculation time needed, and the development toward whole genome sequencing, we are of the opinion that our method will remain of use for public health purposes. The analysis can be performed as an evaluation point to guide laboratory efforts in recognizing international outbreaks once a large enough data set of reference sequences of substantial length is available. Whether this method is applicable to other pathogens for which full length sequences together with epidemiological information are available should be subject of future research. Although the costs of whole genome sequencing are decreasing, obtaining a full genome sequence of pathogens that cannot be cultured will remain a challenge for regional diagnostic laboratories. Therefore, using shorter fragments sufficiently specific in recognizing outbreak events as a consensus is likely to improve the identification of international outbreak events.

Acknowledgments This study was supported by the Dutch Food and Consumer Product Safety Authority, the European Commission DG Research Quality of Life Program, 6th Framework (EVENT, SP22-CT-2004-502571) and SG SANCO (DIVINE-net, 2003213). We thank Adam Meijer, Ingeborg Boxman, Erwin Duizer, Harry Vennema and Marijn van Ballegooijen for many constructive discussions that helped us to improve the manuscript.

Conflict of interest None of the authors have reported a conflict of interest or declared commercial affiliations.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

Appendix

Members of the FBVE network who contributed data were: D. Brown, B. Adak, J. Gray, J. Harris, M. Iturriza (United Kingdom); B. Böttiger, K. Mølbak, C. Johnsen (Denmark); K.-O. Hedlund, Y. Andersson, M. Thorhagen, M. Lysén, M. Hjertqvist (Sweden); P. Pothier, E. Kohli, K. Balay, J. Kaplon, G. Belliot, and S. Le Guyader (France); F. Ruggeri, and I. Di Bartolo (Italy); E. Schreier, K. Stark, J. Koch, M. Höhne (Germany); K. Vainio, K. Nygard and G. Kapperud (Norway).

References

1. M. Formiga Cruz, G. Tofino Quesada, S. Bofill Mas, D.N. Lees, K. Henshilwood, A.K. Allard, A.C. Conden Hansson, B.E. Hernroth, A. Vantarakis, A. Tsibouxi, M. Papapetropoulou, M.D. Furones, R. Girones, Appl. Environ. Microbiol. 68, 5990–5998 (2002)
2. M. Steele, J. Odumeru, J. Food Prot. 67, 2839–2849 (2004)
3. F.K. Kaferstein, Y. Motarjemi, D.W. Betcher, Emerg. Infect. Dis. 3, 503–510 (1997)
4. M. Petignani, M. Harms, L. Verhoef, R. van Hunen, C. Swaan, J. van Steenbergen, I. Boxman, I.S.R. Peran, H. Ober, H. Vennema, M. Koopmans, W. van Pelt, Euro. Surveill. 15, 19572 (2010)
5. M. Petignani, L. Verhoef, R. van Hunen, C. Swaan, J. van Steenbergen, I. Boxman, H.J. Ober, H. Vennema, M. Koopmans, Euro. Surveill. 15, 19512 (2010)
6. J.J. Siebenga, H. Vennema, D.P. Zheng, J. Vinje, B.E. Lee, X.L. Pang, E.C. Ho, W. Lim, A. Choudekar, S. Broor, T. Halperin, N.B. 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, J. Infect. Dis. 200, 802–812 (2009)
7. M. Koopmans, H. Vennema, H. Heersma, E. van Strien, Y. van Duynhoven, D. Brown, M. Reacher, B. Lopman, Emerg. Infect. Dis. 9, 1136–1142 (2003)
8. M.A. de Wit, M.P. Koopmans, L.M. Kortbeek, W.J. Wannet, J. Vinje, F. van Leusden, A.I. Bartelds, Y.T. van Duynhoven, Am. J. Epidemiol. 154, 666–674 (2001)
9. J.G. Wheeler, D. Sethi, J.M. Cowden, P.G. Wall, L.C. Rodrigues, D.S. Tompkins, M.J. Hudson, P.J. Roderick, Br. Med. J. 318, 1046–1050 (1999)
10. A. Bertolotti-Ciarlet, S.E. Crawford, A.M. Hutson, M.K. Estes, J. Virol. 77, 11603–11615 (2003)
11. K.Y. Green, Caliciviridae: The Noroviruses, in Fields Virology, 5th edn., ed. by D.M. Knipe, P.M. Howley (Lippincott, Philadelphia, 2007), pp. 949–979

12. D.P. Zheng, T. Ando, R.L. Fankhauser, R.S. Beard, R.I. Glass, S.S. Monroe, Virology 346, 312–323 (2006)

13. M. Koopmans, E. Strien van, H. Vennema, Molecular epidemiology of human caliciviruses, in Viral gastroenteritis, ed. by U. Desselberger, J. Gray (Elsevier, Amsterdam, 2003), pp. 523–554

14. K. Mattison, E. Grudeski, B. Auk, H. Charest, S.J. Drews, A. Fritzinger, N. Gregoricus, S. Hayward, A. Houde, B.E. Lee, X.L. Pang, J. Wong, T.F. Booth, J. Clin. Microbiol. 47, 3927–3932 (2009)

15. E.C. Holmes, B.T. Grenfell, PLoS Comput. Biol. 5, e1000505 (2009)

16. A. Kroneman, H. Vennema, K. Deforce, H. v d Avoort, S. Penaranda, M.S. Oberste, J. Vinje, M. Koopmans, J. Clin. Virol. 51, 121–125 (2011)

17. J.J. Siebenga, H. Vennema, B. Renckens, E. de Bruin, B. van der Veer, R.J. Siezen, M. Koopmans, J. Virol. 81, 9932–9941 (2007)

18. M. Rondy, M. Koopmans, C. Rotsaert, T. Van Loon, B. Beljaars, G. Van Dijk, J. Siebenga, S. Srivaka, J.W. Rossen, P. Teunis, W. Van Pelt, L. Verhoef, Epidemiol. Infect. 139, 453–463 (2011)

19. J.J. Siebenga, M.F. Beersma, H. Vennema, P. van Biezen, N.J. Hartwig, M. Koopmans, J. Infect. Dis. 198, 994–1001 (2008)

20. L. Verhoef, E. Duizer, H. Vennema, J. Siebenga, C. Swaan, L. Isken, M. Koopmans, K. Balay, P. Poither, P. McKeown, G. van Dijk, P. Capdepon, G. Delmas, Euro. Surveill. 13, 19025 (2008)

21. M. Nilsson, K.O. Hedlund, M. Thorhagen, G. Larson, K. Johansen, A. Ekspung, L. Svensson, J. Virol. 77, 13117–13124 (2003)

22. A. Gallay, H. De Valk, M. Cournot, B. Ladeuil, C. Hemery, C. Castor, F. Bon, F. Megraud, P. Le Cann, J.C. Desenclos, Clin. Microbiol. Infect. 12, 561–570 (2006)

23. R.C. Edgar, BMC Bioinform. 5, 113 (2004)

24. T. Farkas, K. Sestak, C. Wei, X. Jiang, J. Virol. 82, 5408–5416 (2008)

25. A. Stamatakis, Bioinformatics 22, 2688–2690 (2006)

26. Z. Yang, Trends Ecol. Evol. 11, 367–372 (1996)

27. Z. Yang, J. Mol. Evol. 39, 306–314 (1994)

28. S. Kojima, T. Kageyama, S. Fukushi, F.B. Hoshino, M. Shinohara, K. Uchida, K. Natori, N. Takeda, K. Katayama, J. Virol. Methods 100, 107–114 (2002)

29. J.S. Noel, T. Ando, J.P. Leite, K.Y. Green, K.E. Dingle, M.K. Estes, Y. Seto, S.S. Monroe, R.I. Glass, J. Med. Virol. 53, 372–383 (1997)

30. J. Vinje, R.A. Hamidjaja, M.D. Sobsey, J. Virol. Methods 116, 109–117 (2004)

31. J. Xerry, C.I. Gallimore, M. Iturriza Gomara, D.J. Allen, J.J. Gray, J. Clin. Microbiol. 46, 947–953 (2008)

32. D.M. Hillis, J.J. Bull, Syst. Biol. 42, 11 (1993)

33. P.S. Soltis, D.E. Soltis, Stat. Sci. 18, 12 (2003)

34. J. Felsenstein, Evolution 39, 9 (1985)

35. L. Verhoef, R.D. Kouyos, H. Vennema, A. Kroneman, J. Siebenga, W. van Pelt, M. Koopmans, Emerg. Infect. Dis. 17, 412–418 (2011)

36. T. Kageyama, S. Kojima, M. Shinohara, K. Uchida, S. Fukushi, F.B. Hoshino, N. Takeda, K. Katayama, J. Clin. Microbiol. 41, 1548–1557 (2003)

37. F.S. Le Guyader, F. Bon, D. DeMedici, S. Parmaudeau, A. Bertone, S. Crudeli, A. Doyle, M. Zadane, E. Suffredini, E. Kohli, F. Maddalo, M. Monini, A. Gallay, M. Pommepuy, P. Poither, F.M. Ruggeri, J. Clin. Microbiol. 44, 3878–3882 (2006)

38. R.A. Bull, M.M. Tanaka, P.A. White, J. Gen. Virol. 88, 3347–3359 (2007)