Utility of massively parallel sequencing for outbreak tracking

Abstract

The use of massively paralleled sequencing for outbreak tracking was assessed and compared to the currently used amplicon Sanger sequencing based method. Overall both methods showed significant correlation. However, MPS provided greater depth and the ability to identify minority variants among samples within an outbreak that represented consensus changes in one or more samples from the same outbreak. This meant that the MPS data would have been able to link all the samples into a single outbreak or transmission network, where the current Sanger sequencing may not have been able to link them all.

Introduction

In a suspected norovirus outbreak, the detection of norovirus alone or the characterisation of the strain present among those affected is often not enough to determine the source or transmission route. Sequencing of the hypervariable region and stringent conservation cut off values have been used to infer the likelihood of linked transmission events (1-3). Conversely, diversity in the form of presence of different genotypes or variants of the same genotype among those affected in foodborne outbreaks is a strong indicator of either multiple sources of transmission, or typically of infection from shellfish contaminated at source. Current methods rely on the amplification, using genotype-specific oligonucleotide primers, and Sanger sequencing, and therefore requires prior knowledge of the genotype/s involved in the samples under investigation. This method has proven reasonably sensitive for application to acute clinical samples which contained high viral loads, but it use for linking environmental contamination to transmission is severely hampered by the limiting viral loads typically present environmental swabs.

Here we aimed to assess the utility and sensitivity of massively parallel sequencing (MPS) methods in outbreaks investigation, using clinical samples from a small number samples associated with of food borne or healthcare outbreaks, and compare it to the current Sanger based methods.

Clinical sample preparation, enrichment, library preparation and sequencing

Clinical samples associated with two food-borne and three hospital based suspected outbreaks for which genotype and variant identity had already been characterised by PCR and partial Sanger sequencing were obtained from the Enteric Virus Unit, PHE, Colindale (Table 1).

Stool samples were prepared as 10-20% suspensions in sterile PBS (Sigma, Dorset, UK) and used for norovirus capture and enrichment using porcine mucin coated magnetic beads.

MagnaBind™ carboxyl derivatized beads (Fisher Scientific, Leicestershire, UK) were prepared as described in the manufacturer's instructions, pooled and stored at 4°C. PGM III (Sigma, Dorset, UK) or BSA as control (Sigma, Dorset, UK) were dissolved in conjugation buffer to a concentration of 7.5 mg/mL for the coupling reaction.

Prior to enrichment, the stool suspension was centrifuged at 5000 rpm for 15 minutes and the PGM-MB stock was reconstituted by pipetting. A total of 100 µL of homogenous PGM-MBs was added to a 1 mL of capture mixture containing stool suspension and 0.1 M citric acid-trisodium citrate buffer at pH 3.6 in a ratio 1:3. The mixture was kept under constant mixing on an SB2 rotator shaker (Fisher Scientific, Leicestershire, UK) at room temperature for 15 minutes. After mixing the PGM-MBs were isolated by the DynaL magnetic separation (MS) rack (Fisher Scientific, Leicestershire, UK) and washed to homogeneity with 1 mL 0.1 M citric acid-sodium citrate buffer, this was repeated a further two times. The PGM-MBs were then eluted into 60 µL PBS (Sigma, Dorset, UK) and transferred to a sterile microcentrifuge tube.

Nucleic acids were extracted from the PGM-MB eluate using the Guanidinium isothiocyanate (Gn) silica method (4).

Prior to library preparation extracted RNA was DNase I (Sigma, Dorset, UK) treated and column purified using the Qiaquick gel extraction kit (Qiagen, West Sussex, UK), following the manufacturer's instructions.

Table 1: Outbreak and sample characteristics

Outbreak	**Setting**	**Sample**	**Genotype**	**RT-PCR Ct value**
A	Restaurant	1	GII.20	23.16
		2		24.86
		3		27.88
		4		NA*
		5		25.33
		6		24.95
B	Hospital	1	GII.4	18.75
		2		21.99
		3		24.86
		4		24.38

Outbreak	Setting	Sample	Genotype	RT-PCR Ct value
C	Pub	1	GII.Pe/GII.4	21.74
		2		26.88
		3	GII.4	32.41
		4		22.54
D	Hospital	1	GII.4	29
		2		31.27
		3		30.24
		4		28.37
E	Hospital	1	GII.4	37.11
		2		14.09
		3		24.89
F	Hospital	1	GII.Pe/GII.4	20.64
		2		33.95
		3		21.15
		4		26.53

Library preparation for sequencing

Each Library was prepared using the ScriptSeq v2 RNA-Seq Library Preparation Kit (Illumina, Essex, UK) with one slight modification to the manufacturer's instructions. HuNoV RNA was denatured at 95°C, and subsequently placed on wet ice before library preparation. At the amplification stage Illumina compatible barcodes replaced the reverse primer, to allow for sample multiplexing. Prior to sequencing on the HiSeq2500 Illumina platform at the Centre for Genomic Research, University of Liverpool, library size and concentration was analysed with the 2100 Bioanalyzer (Agilent Technologies) and Qubit dsDNA High Sensitivity assay (Fisher Scientific, Leicestershire, UK) respectively. A sample with an insufficient amount of library present after 12 cycles of non-specific PCR amplification was amplified for a further 10 cycles. If additional amplification was performed, amplicons were purified with AMPure beads (Beckman Coulter, Buckinghamshire, UK) at 0.7x the volume of the total PCR reaction, to minimise the presence of primer dimer.

Data analysis

Consensus mapping and calling minority variants was done according to the algorithm in Figure 1 and using the software detailed in Table 2.

Figure 1: Bioinformatics pipeline for generating a consensus sequence and calling minority variants (Colour code= De novo assembly, Find a suitable reference, Quality control, Read alignment, Alignment file sorting and PCR duplicate removal, Variant calling software, reformatting of the variant call file (available online at: https://github.com/riverlee/pileup2base), Data analysis in R)

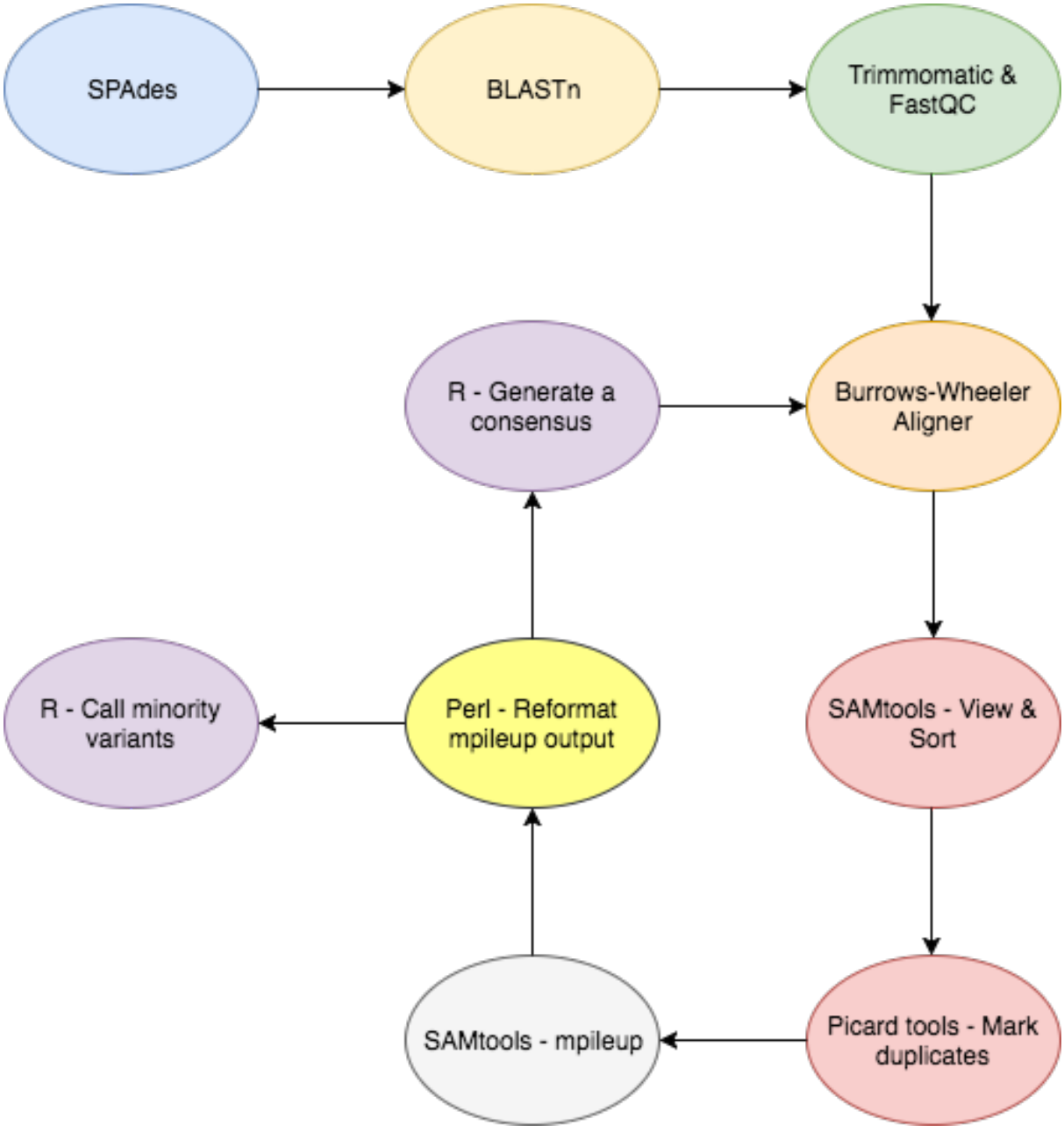


Table 2: References to software used in the consensus generation and minority variant calling pipeline

Software name	Version	Function	Reference
SPAdes	v.3.5.0ls	De novo assembly	(5)
Blastn	2.2.27+	Reference search	(6)
Trimmomatic	0.36	Quality control	(7)
FastQC	V0.10.1	Diagnostics	(8)
Burrows-Wheeler Aligner	0.5.9-r16	Aligner	(9)
Samtools	0.1.18-r580	Sequence alignment map processing and variant calling	(10)
Picard tools	2.1.1	Remove duplicate reads	Available online at: http://broadinstitute.github.io/picard

Multiple alignment analysis was performed in T-Coffee with the Clustalw algorithm under default settings (11, 12), and reformatted with BoxShade (3.2) (http://www.ch.embnet.org/software/BOX_form.html).

Phylogenetic analysis was performed with Molecular Evolutionary Genetics Analysis (MEGA) software (version 7). Nucleic acid sequences were aligned with Clustalw, under the default settings, and phylogenetic relationships were inferred with the Maximum-Likelihood algorithm. Phylogenetic trees were subsequently reformatted using the online Interactive Tree of Life (ITOL) software (13).

Minority variants were called according to the following criteria: detected at ≥ 5 % of the total base calls and ≥ 200-fold coverage or at ≥ 20 % of the total base calls and ≥ 50-fold (Kelly, D *et al,* unpublished).

Results

HiSeq (Illumina) reads from outbreaks A, B, C, D, E and F were assembled into contiguous sequences *de novo*, and a local nucleotide alignment showed agreement with the partial sequence data obtained at PHE.

Complete or near complete HuNoV genomes were recovered from most stool samples (22/25) using the mucin enrichment method described (Figure 2) (Kelly D et al, unpublished), and no library preparation failures were seen. For each sample the coverage standard deviation across the whole virus genome varied from 44.16 to 82% around the mean (Table 3).

Figure 2: Summary norovirus genome coverage plots by outbreak and sample

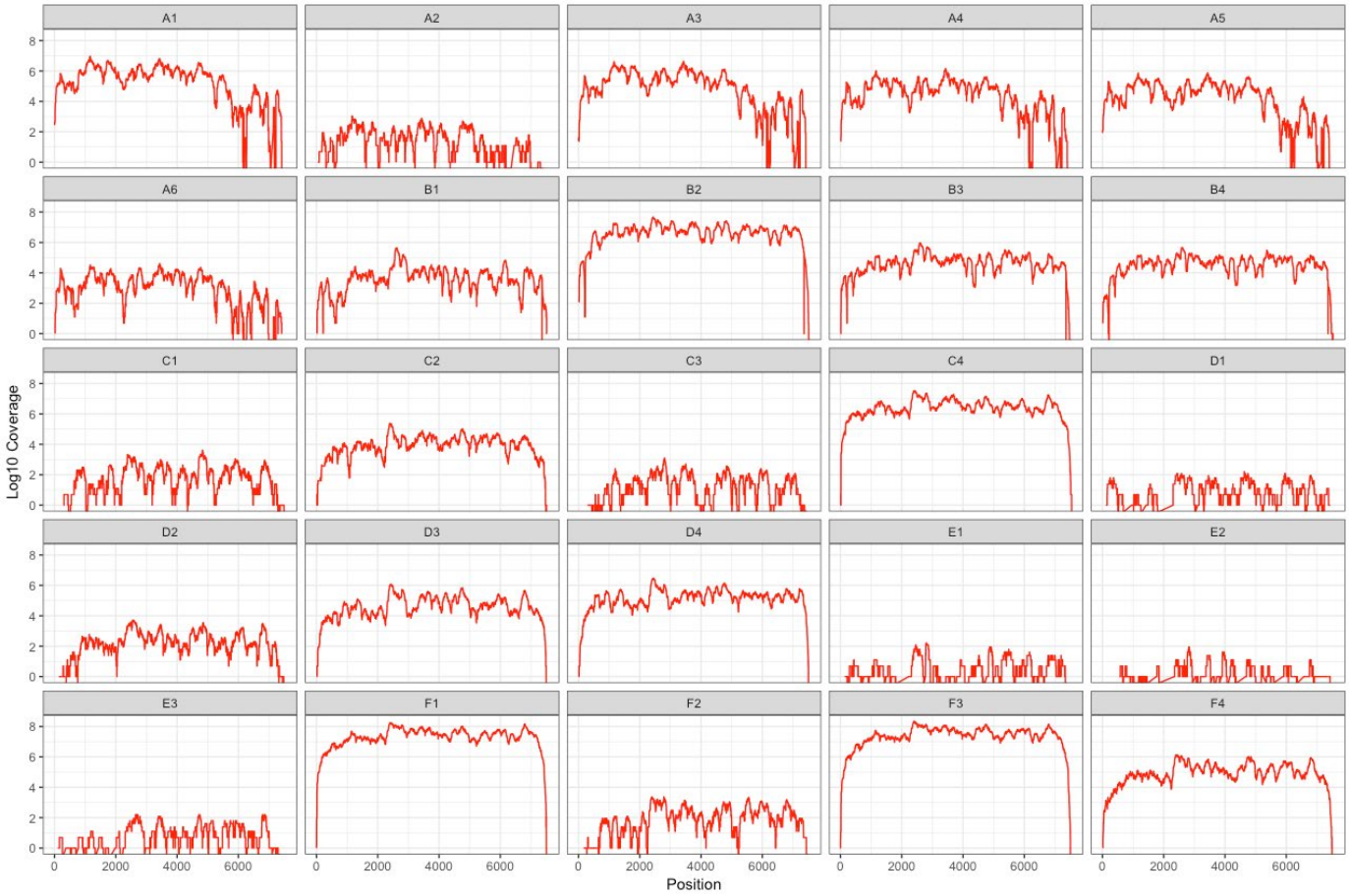


Figure 3. Linear regression analysis of mean sample coverage against the GII HuNoV Ct value of the RNA extract

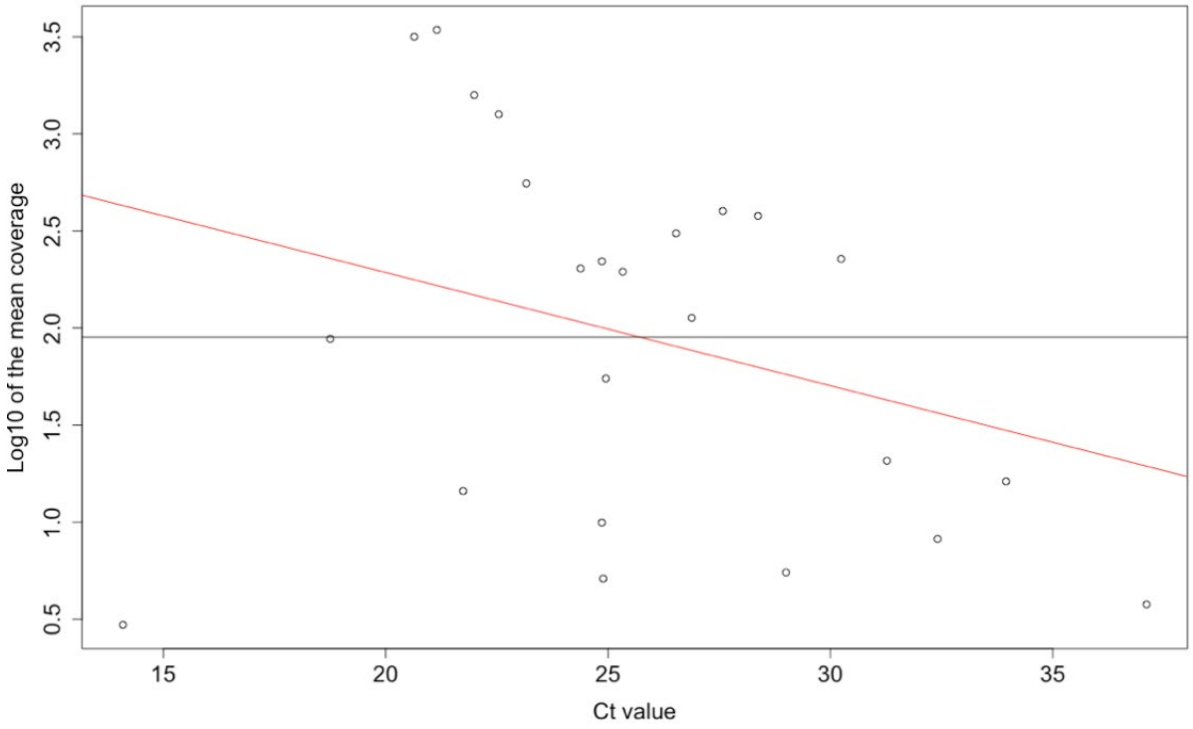


Table 3. Summary of sequence reads recovered from stool samples

Outbreak	Sample	Coverage		Percentage of the genome covered
		Mean	Standard deviation	
A	1	556.02	396.45	99.67
	2	9.95	7.43	95.33
	3	400.68	310.53	99.57
	4	230.12	161.36	99.83
	5	194.94	141.24	99.73
	6	54.96	39.08	98.54
B	1	87.98	72.27	99.44
	2	1584.81	676.05	99.88
	3	220.56	121.35	99.74
	4	202.55	90.14	99.56
C	1	14.47	11.44	95.75
	2	112.81	61.93	99.54
	3	8.2	6.39	94.12
	4	1261.63	620.9	100
D	1	5.51	3.42	85.06
	2	20.72	14.99	98.19
	3	226.99	146.36	99.91
	4	378.11	188.86	100
E	1	3.77	2.88	83.85
	2	2.96	2.02	54.25
	3	5.12	3.66	90.24
F	1	3163.83	1397.26	99.99
	2	16.22	11.73	95.7
	3	3434.15	1538.43	100
	4	307.26	181.65	99.77

Linear regression identified a negative trend was observed between the mean coverage and Ct value, but was not significant (Figure 3).

Over half (58%) of the amplicons used in P2 domain Sanger sequencing were identical to their respective outbreak consensus sequence. If MPS was applied to the complete genome instead, the number of complete P2 domain matches decreased to 48% (See Table) Overall, the P2 domain of samples sequenced by Sanger method differed from the outbreak consensus sequence by 0-0.11%, in contrast, when P2 domain sequence was analysed from the MPS data it ranged from 0-3.94%, and when samples with a mean coverage below 21 were excluded the range

was reduced to 0-0.87. Diversity across complete genomes was greater across samples within outbreaks.

Table 4: Consensus identity measurements across the P2 domain (from Sanger and MSP methods) or the whole genome (MPS only). Cells shaded in dark grey indicate data not obtained. Colour shading identified 100% identity among cases/samples within an outbreak

Outbreak	Sample	Sanger sequence P2 identity (%)	MPS P2 identity (%)	MPS Overall identity (%)
A	1	100	100	99.34
	2	99.69	99.46	95.75
	3	100	100	99.16
	4	100	100	99.07
	5	100	100	99.02
	6	100	100	98.43
B	1	99.75	100	99.54
	2	100	100	99.9
	3	100	100	99.81
	4	100	100	99.64
C	1		99.08	96.04
	2		99.42	99.42
	3	100	97.93	93.97
	4	100	99.42	99.77
D	1	99.88	99.79	88.01
	2	99.58	98.09	96.57
	3	98.92	99.79	99.7
	4	99.78	99.79	99.77
E	1	100	99.9	99.64
	2	99.89	95.95	65.14
	3	99.9	98.44	91.03
F	1		100	99.94
	2		99.68	96.01
	3		100	99.79
	4		100	99.94

By pairwise comparison of all outbreaks, there was > 98% agreement between the P2 domain obtained by MPS and the partial Sanger sequencing with four exception: outbreak C sample 3 (97.22%), outbreak D sample 2 (97.88%), outbreak E sample 2 (88.91%) and outbreak E sample 3 (97.92%) for which the mean coverage ranged from 2.96-20.72. Although this may be attributable to low coverage, among other samples with similar levels of coverage heterogeneity between the two techniques at the P2 domain was not observed. In outbreak A and B conserved mismatches were observed between MPS and Sanger method that were near the 5' and 3' end of the amplicon (data not shown), and these may potentially be due to primer induced error with the Sanger method.

Phylogenetic analysis also confirmed clustering regardless of method and genome length used for the analysis when coverage was high for whole genomes (Figures 4 to 8). However, in agreement with the mismatch analysis between Sanger method and MPS, those samples with low coverage and < 98 % similarity did not remain in the clusters assigned by partial amplicon sequencing of the P2 domain, highlighting that coverage is likely to be the critical limiting factor in the application of MPS for outbreak investigation.

Figure 4: Outbreak A Maximum-Likelihood phylogenetic trees (A=Sanger method of the P2 domain, B=MPS of the P2 domain, C=MPS of the complete genome, boot strap values of nodes over 0.8 of 1000 bootstrap replicates are shown as red squares)

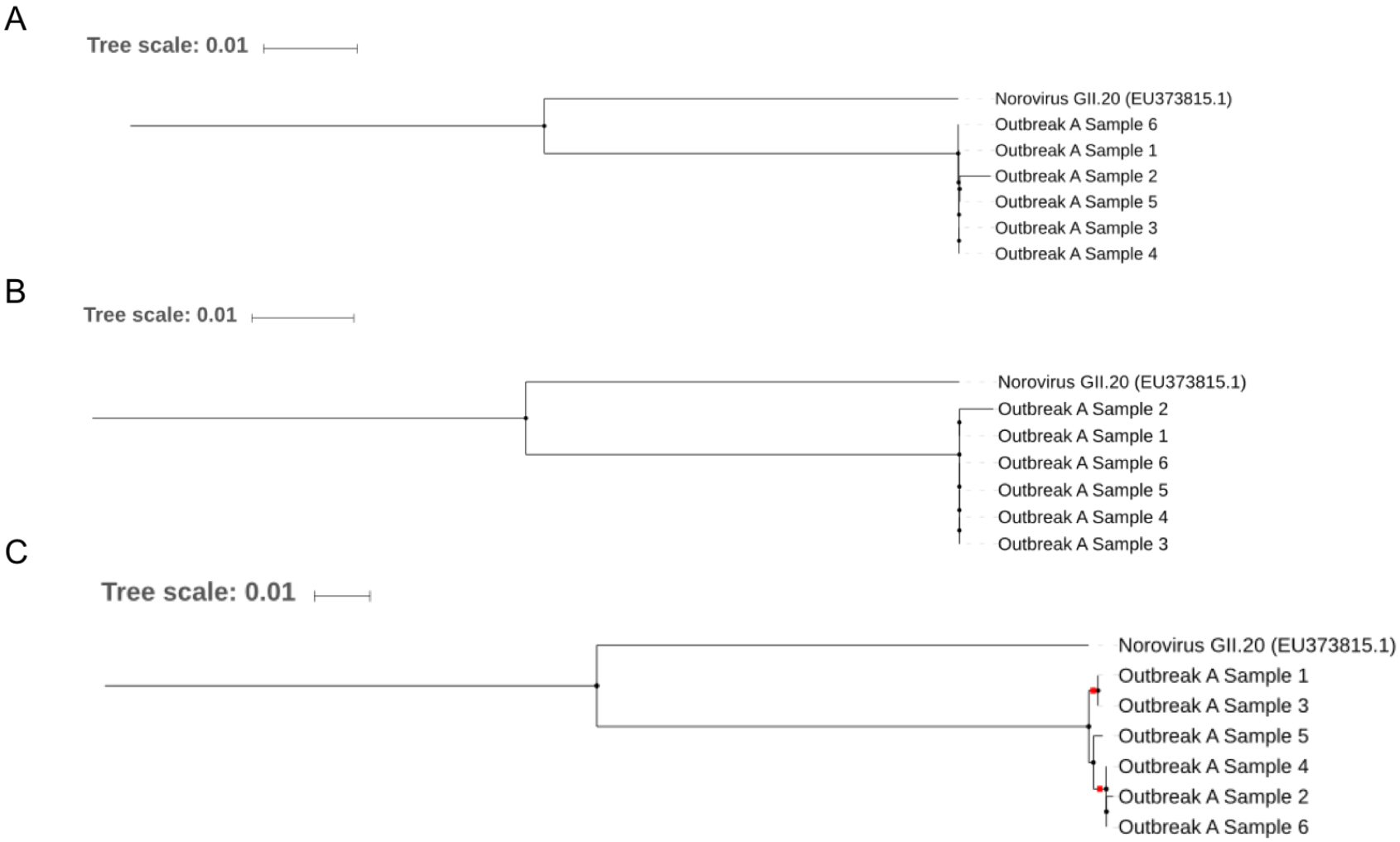


Figure 5: Outbreak B, C and D Maximum-Likelihood phylogenetic tree of P2 domain Sanger method (boot strap values of nodes over 0.8 of 1000 bootstrap replicates are shown as red squares)

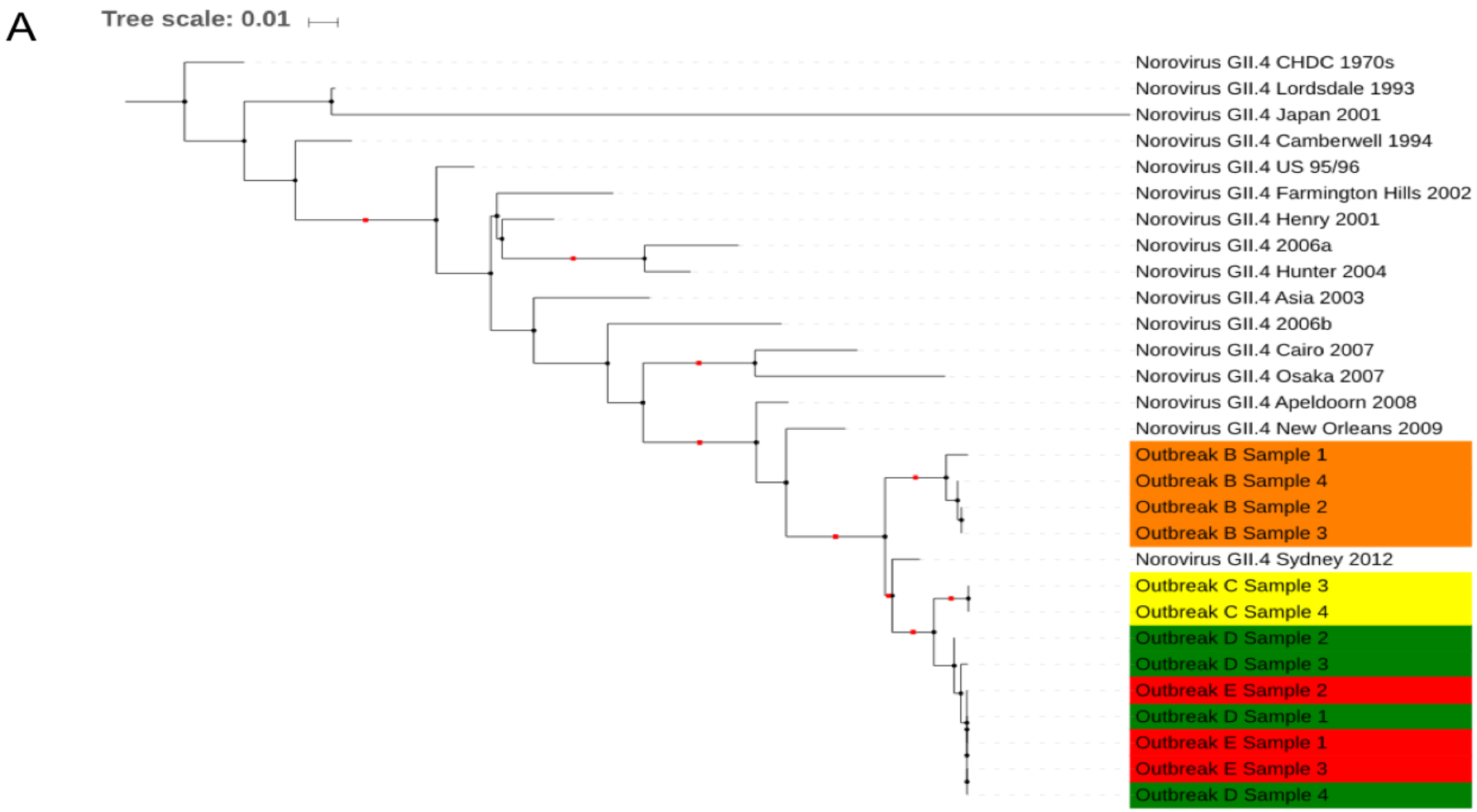


Figure 6: Outbreak B, C and D Maximum-Likelihood phylogenetic tree of P2 domain MPS (boot strap values of nodes over 0.8 of 1000 bootstrap replicates are shown as red squares)

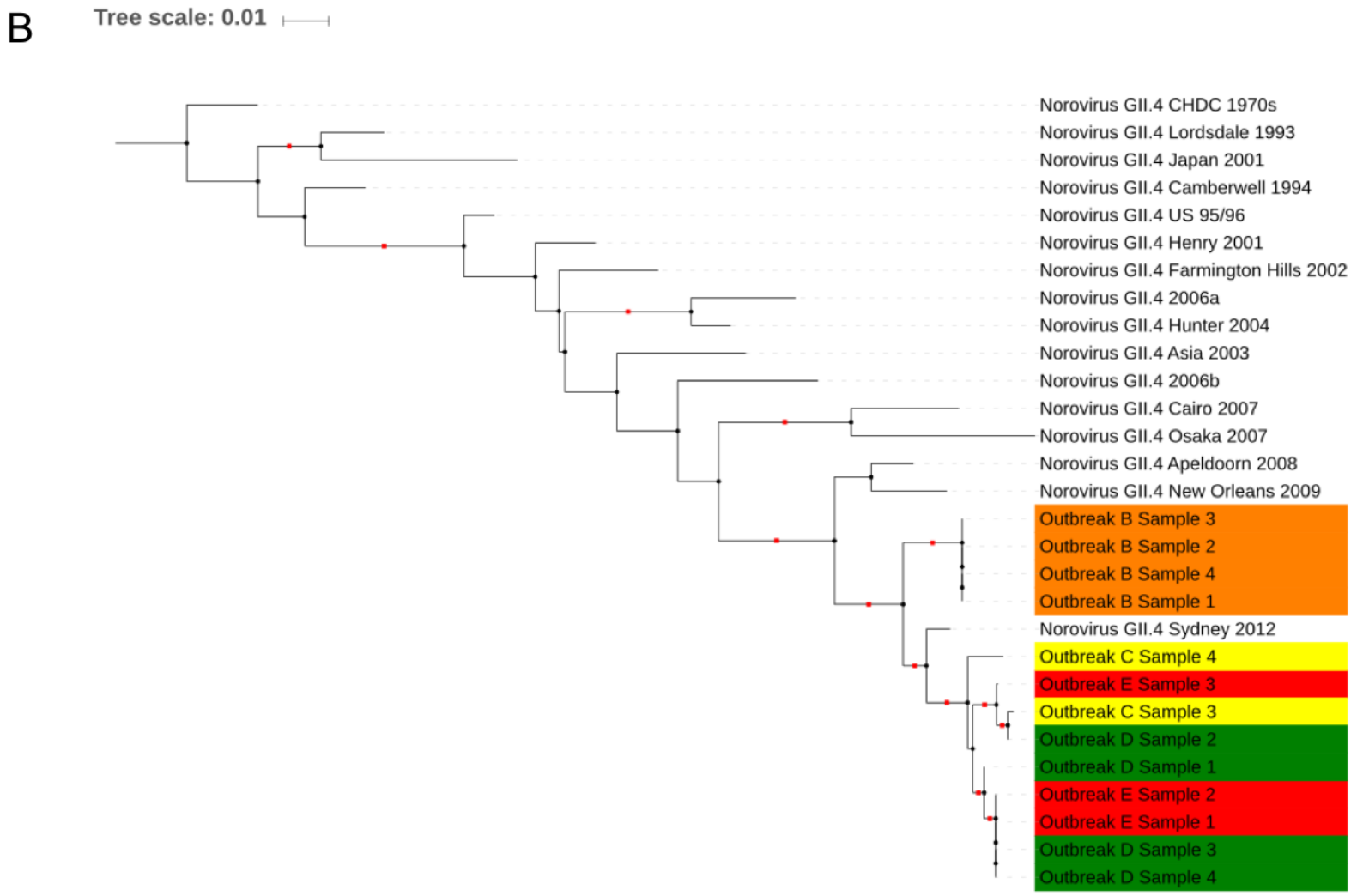


Figure 7: Outbreak B, C and D Maximum-Likelihood phylogenetic tree of complete genome MPS (boot strap values of nodes over 0.8 of 1000 bootstrap replicates are shown as red squares)

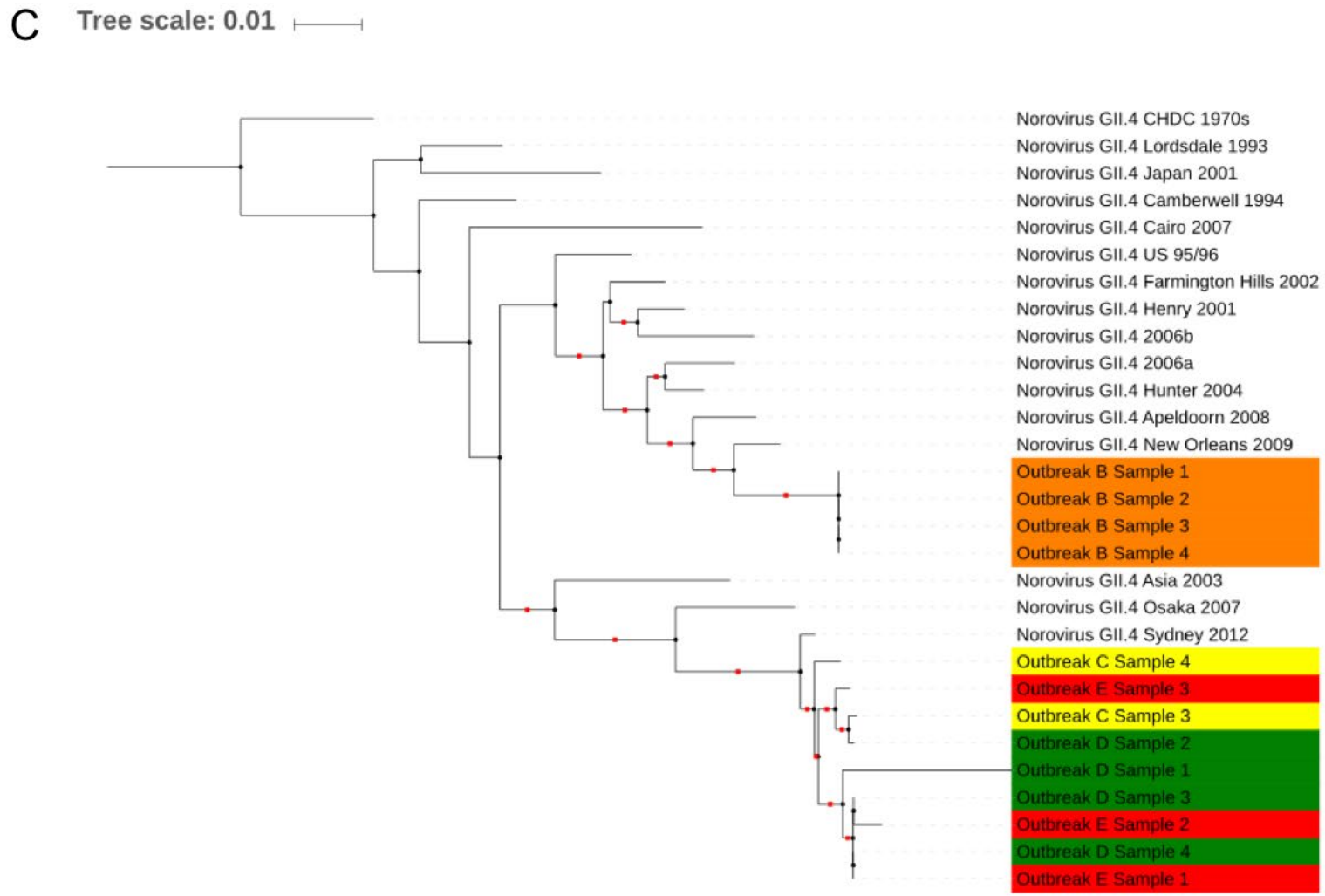
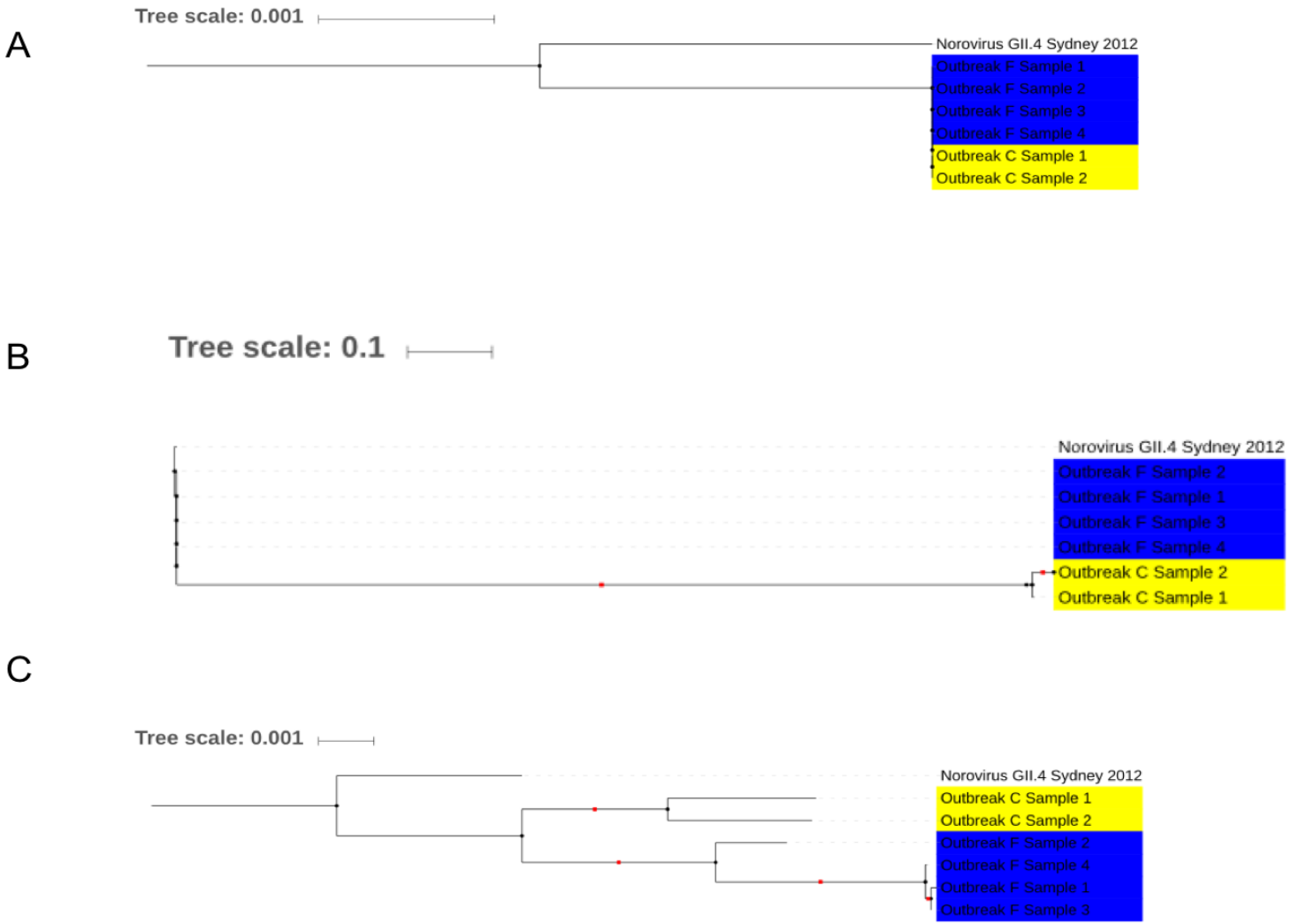


Figure 8: Outbreak C and F Maximum-Likelihood phylogenetic trees (A=Sanger method of region C, B=MPS of region C, C=MPS of the complete genome, boot strap values of nodes over 0.8 of 1000 bootstrap replicates are shown as red squares)



In total 45 mutations were present in minority variants within the HNoV sequences of outbreaks A-F. It was not possible to detect minority variants in outbreak E, due to the low coverage. In the samples within outbreak B despite high coverage, fewer minority variants were identified when compared with the rest of samples and outbreaks (Table 5 and figures 9 to 12).

Table 5: Frequency of minority variants identified in each ORF from outbreaks A-F, excluding samples in outbreak E due to insufficient coverage

ORF	**Outbreak**				
	A	**B**	**C**	**D**	**F**
1	10	0	7	8	2
2	6	0	2	7	2
3	0	1	0	0	0
Total	**16**	**1**	**9**	**15**	**4**

Among the 45 minority variants identified, only 3 positions and substitutions were conserved among cases within an outbreak.

Figure 9: The frequency of minority variants identified in stool samples from outbreak A compared to each consensus sequence (Blue = Adenine, Red = Cytosine, Green = Guanine, Yellow = Thymine)

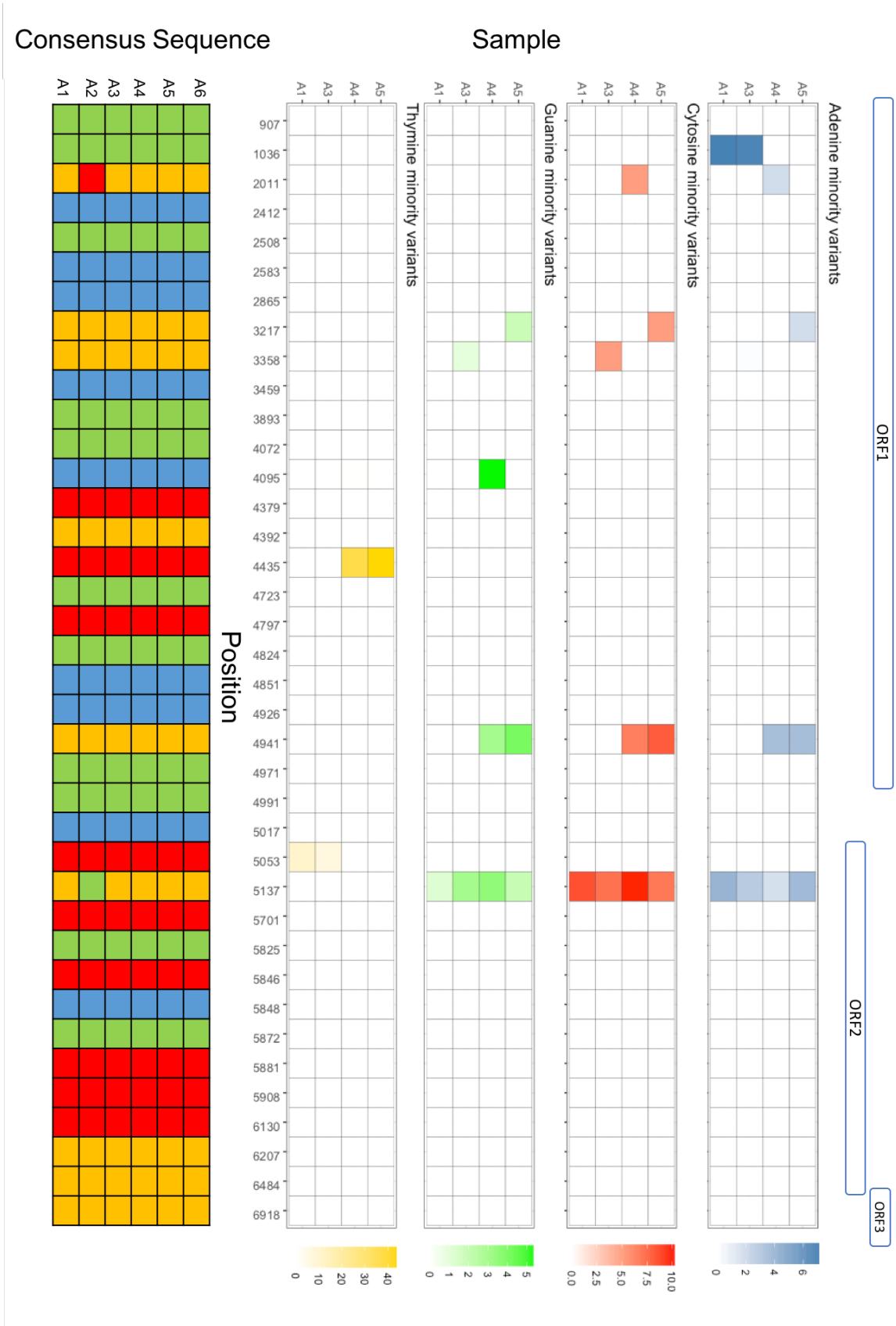


Figure 10: The frequency of minority variants identified in stool samples from outbreak C compared to each consensus sequence (Blue = Adenine, Red = Cytosine, Green = Guanine, Yellow = Thymine)

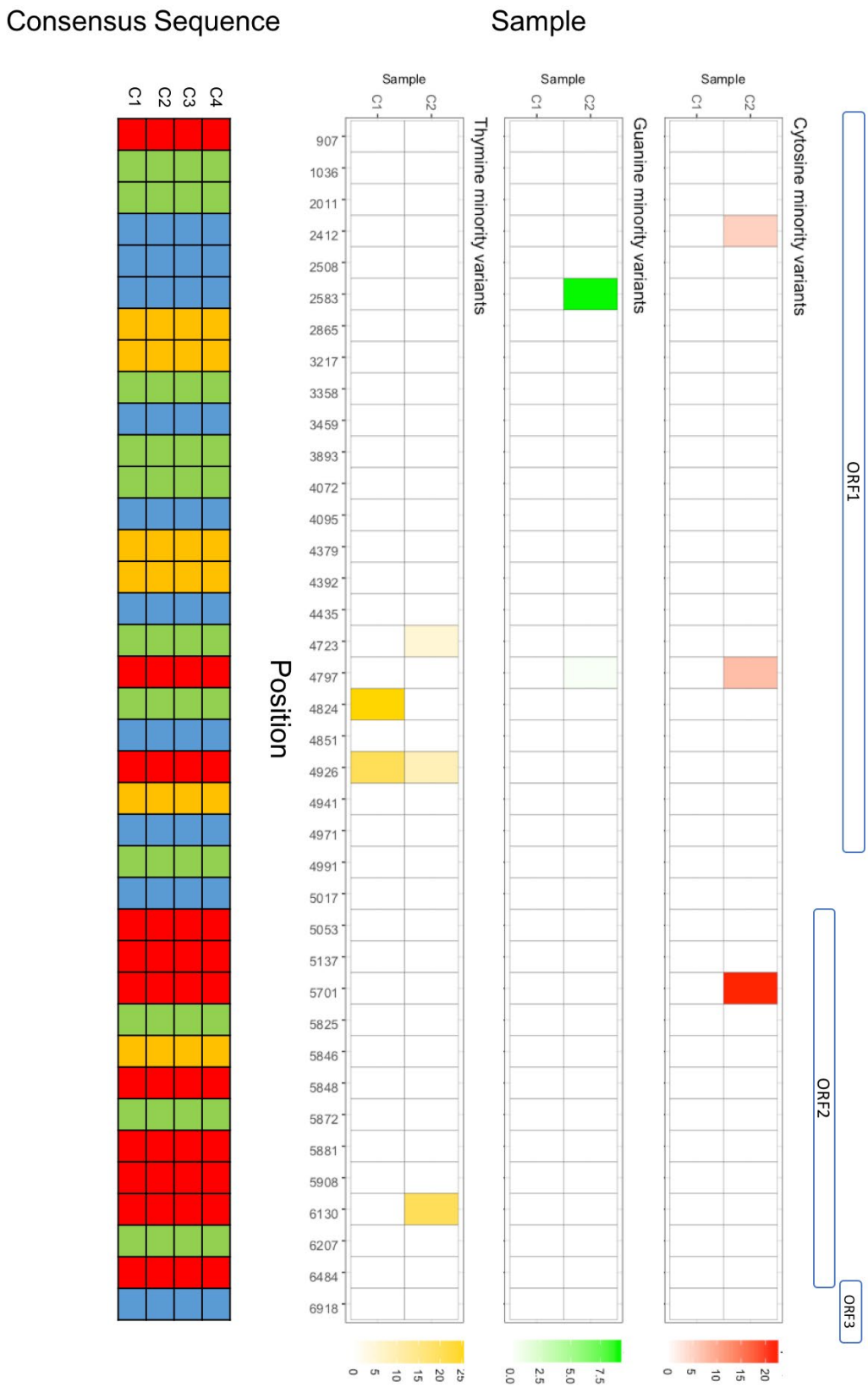


Figure 11: The frequency of minority variants identified in stool samples from outbreak D compared to each consensus sequence (Blue = Adenine, Red = Cytosine, Green = Guanine, Yellow = Thymine)

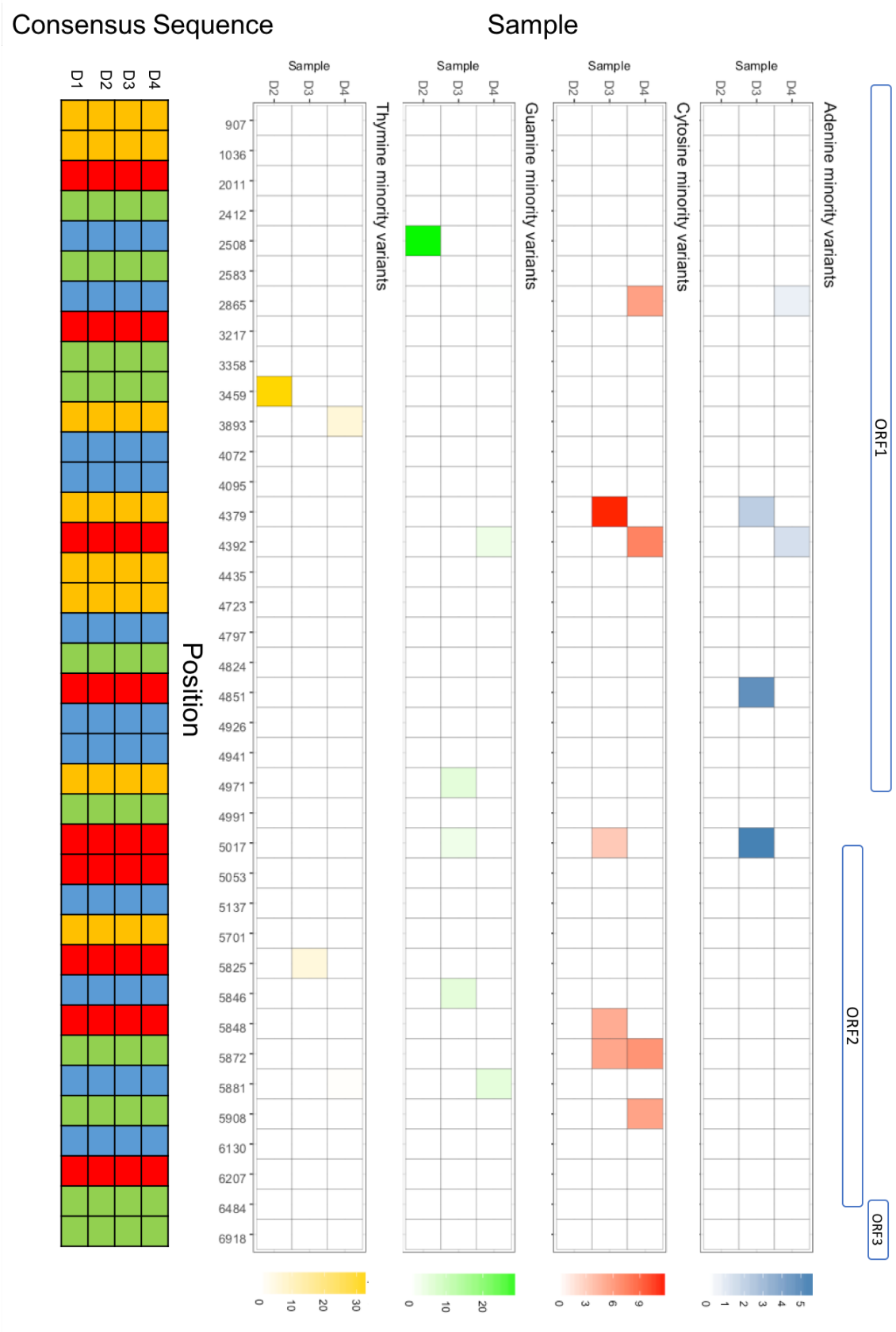
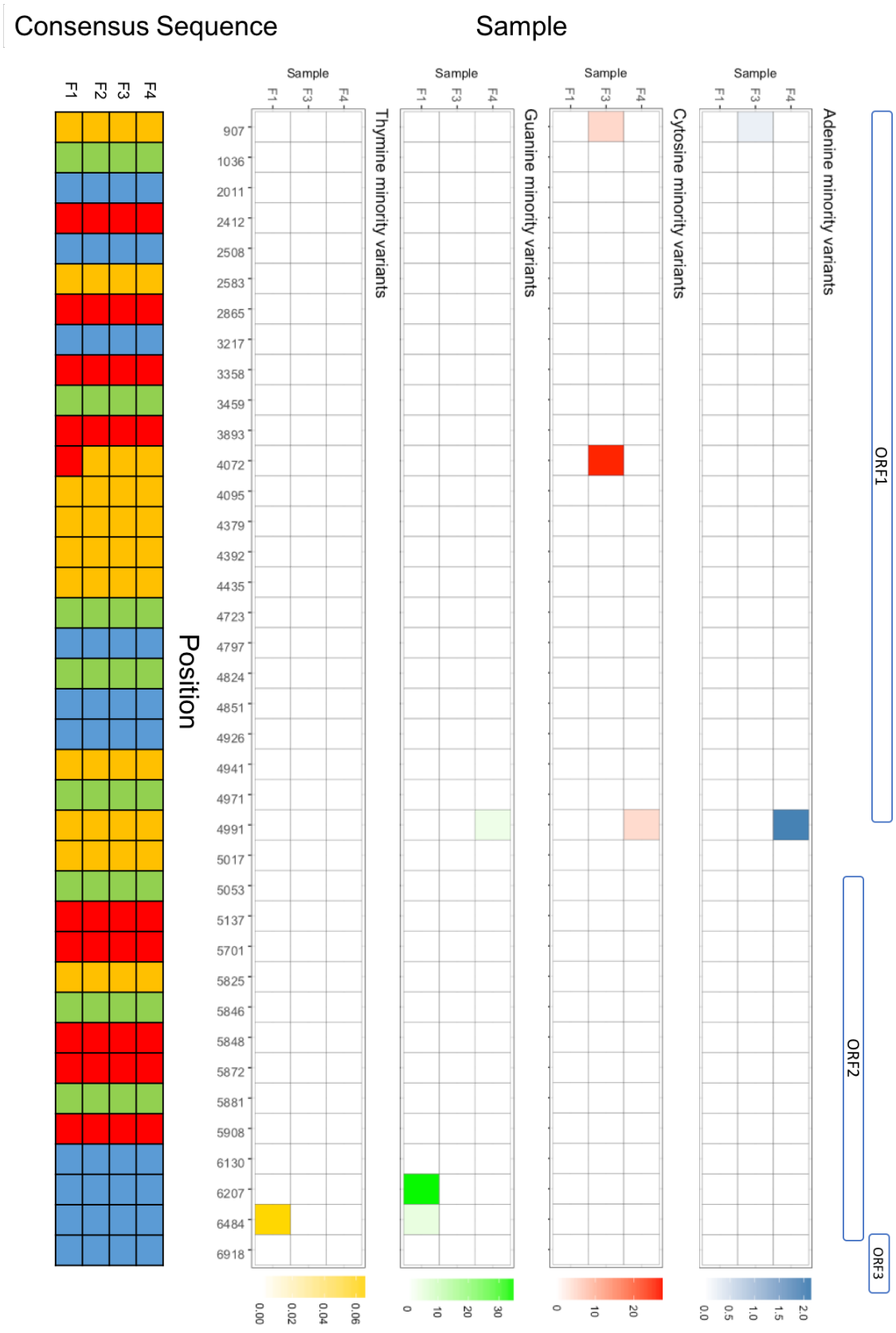


Figure 12: The frequency of minority variants identified in stool samples from outbreak F compared to each consensus sequence (Blue = Adenine, Red = Cytosine, Green = Guanine, Yellow = Thymine)



Discussion

This first attempt to applying MPS for outbreak investigation has demonstrated that stool samples higher viral loads yield sufficient genome coverage and depth to allow discrimination of minority variants. The method of norovirus sequence enrichment based on capture with porcine mucin offered increase sensitivity, and enable sequence to be obtained for all samples tested, in contrast to the amplicon based sanger methods that failed to yield sequence for some of the samples.

Pairwise comparison between of the P2 domain using Sanger or MPS yielded comparable results and if the 100% identity criteria was used to link samples within an outbreak (1-3), interpretation of linkages among samples within an outbreak did not differ significantly when only analyzing P2 domain, however greater diversity was observed across the whole length of the virus genome. Some of these diversity may be introduced by the low coverage obtained in some of the samples.

This study has however highlighted some potential advantages of deep sequencing, over Sanger sequencing, by providing complete resolution of the HNoV genome in each stool sample and detecting minority variants which can further characterize differences in virus populations of individual hosts. The advantage of MPS in HNoV outbreak control has been described previously, where the 454 instrument was used to show nosocomial transmission of minority alleles from one immunosuppressed patient to another, which then became predominant in the latter (14). In a similar manner, outbreak F occurred in a hospital, and a minor allele observed in an individual was observed to be the dominant allele in a separate individual. However, without further epidemiological data it can only be speculated whether this was due to a bottleneck transmission event or exposure to a common source of infection and differential selection in each host. A retrospective MPS investigation of a GII.Pg/GII.3 recombinant point source outbreak in 1972 was recently performed with the Ion torrent instrument (15). Johnson and colleagues described the presence of significant subpopulation heterogeneity in the GIIPg/GII.3 recombinant population among several individuals at amino acid positions 315 and 1293, and proposed to be a consequence of host selection of variants following a common seeding event. The data presented here also suggest a similar scenario in outbreak A, this was a suspected foodborne outbreaks linked to a restaurant; A heterogeneous subpopulation was present in 4/6 individuals affected at position 5137. Moreover, a mutation in the consensus sequence detected in one of the cases (A2), that based on the P2 domain 100% criteria would have excluded it from being linked to the remaining cases, was detected as a minority variant in several of the other cases within this outbreak, hence demonstrating the potential advantages of MPS for fine discrimination and linkage of transmission events.

The advantages highlighted by this work need to be put in the context of the high economic and required time for sequencing and data analysis associated with MPS sequencing. The methods described here do not represent a practical approach in

real time outbreak investigation, but they may have a place in large scale food borne outbreaks. However, before this method can be rolled out, further work would need to be conducted in particular in relation calibrating cutoff values for minority variant call and the limitations associated with different levels of coverage.