Investigation of in-hospital Norovirus transmission using whole genome sequencing

Tse Hua Nicholas Wong

New College

Thesis submitted for the degree of DPhil at the University of Oxford

Trinity Term 2014

Nuffield Department of Medicine

University of Oxford

Abstract

Investigation of in-hospital Norovirus transmission using whole genome sequencing

Tse Hua Nicholas Wong New College, University of Oxford Trinity Term 2014

Norovirus is the commonest cause of viral gastroenteritis, affecting all age groups worldwide. Outbreaks frequently occur in semi-closed communities such as schools, cruise ships, prisons and hospitals. Within the healthcare environment, the economic and logistical burdens and the inconvenience caused by norovirus is significant, since ward closure remains central to infection control. The aim of this study was to investigate norovirus transmission dynamics during hospital outbreaks. The ultimate goal was to provide information that could, in future, lead to the development of novel, less disruptive approaches to curtailing the spread of infection.

The study explored the application of 'next generation' high throughput DNA sequencing technologies to the determination of large numbers of norovirus genomes. Whole genome sequences provide the highest possible level of discrimination among viruses, information which is essential to the identification of linked and independent cases of infection. The approach exploits the high norovirus mutation rate, which is typical of RNA viruses. Consequently, viruses within a single ward which differ by more than a few SNVs can be considered to represent independent introductions, rather than a single outbreak.

Whole genome sequence data (determined for noroviruses collected between 2009 and 2013) were combined with epidemiological data, providing further insights into transmission dynamics. These data identified multiple independent virus introductions during single ward outbreaks. The possible origin of such outbreaks in Oxfordshire hospitals were investigated using viruses originating in the local community, and in other healthcare environments distributed throughout the UK. Whole genome sequences of noroviruses from consecutive years were genetically divergent, confirming the rapid evolution of the virus over time and excluding the possibility of prolonged environmental contamination as a reservoir of infection. Such detailed information on norovirus transmission within the healthcare environment could inform alternative future approaches to optimising infection control within the healthcare setting.

Acknowledgements

I have had a wonderful time and experience completing this DPhil. Firstly, I am grateful to my supervisors Professor Derrick Crook, Dr Kate Dingle and Professor Tim Peto for their guidance and ever present supervision. I particularly wish to thank Dr Kate Dingle for guiding me during my humble beginnings in the laboratory, passing me her great wisdom and excitement for the trials and tribulations I was due to encounter during my time. I also wish to send thanks to Professor Tim Peto for allowing me into his house on multiple occasions to sample his rare Luaka tea and discuss the work and analytical approaches.

I am also grateful for many other people who have been involved in this project; Lily O'Connor and the tireless infection control team at the John Radcliffe Hospital, the microbiology lab at the hospital, Kulvinder Kaur, Amy Trebes, and Paolo Piazza at the Wellcome Trust Centre of Human Genetics who have endured receiving my countless extracted samples for sequencing at odd hours, and both Dr Liz Batty and Dr David Wyllie for tolerating my endless "simple" questions in data analysis and bioinformatics.

My appreciation goes out to all patients and staff on the wards who have provided samples for me over the years. I also wish to thank Ali Vaughan, Jessica Wray, Esther Robinson, Louise Pankhurst, Dai Griffiths and Mary Deadman in the laboratory for keeping me on the straight and narrow. Outside of Oxford, I wish to acknowledge Professor Mark Wilcox at Leeds, Dr John Paul and Kevin Cole in Brighton, and Dr Julian Sutton at the Southampton PHE laboratory.

I wish to thank my good colleagues and friends within the research office: John, Tim, Nicole, Bernadette, Claire and David who have provided me with endless laughter and help when the going got tough.

I finally wish to dedicate this thesis to my wonderful wife Tracy and daughter Naomi, the both of whom have patiently endured my late nights, and the latter will be most glad that I will now be able to read her more bedtime stories.

Publications and conference attendances

Publications

Batty EM, Wong TH, Trebes A, et al., A modified RNA-Seq approach for whole genome sequencing of RNA viruses from faecal and blood samples. PLoS One. 2013 Jun 10;8(6):e66129. doi: 10.1371/journal.pone.0066129

Wong TH, Dearlove BL, Hedge J, Giess AP, Piazza P, Trebes A, et al., Whole genome sequencing and de novo assembly identifies Sydney-like variant noroviruses and recombinants during the winter 2012/2013 outbreak in England. Virol J. 2013 Nov 13;10:335. doi: 10.1186/1743-422X-10-335

Oral Presentations

Wong TH, Dearlove B, Crook DW., et al., (2012) Using whole genome sequence data to determine transmission paths and mutation rates in Norovirus outbreaks. Presented at ECCMID London, 31 March 2012

Wong TH, Batty E, Cule M., et al., (2012) A new method for whole genome sequencing of Noroviruses to aid outbreak investigations. Presented at ID Week San Diego, 18 October 2012

Wong TH, Batty E, Dearlove B, et al., (2013) Long term Norovirus surveillance using Whole Genome Sequencing confirms a new variant as the cause of this winter's UK epidemic., Presented at Applied Bioinformatics and Public Health Microbiology, Hinxton 15 May 2013

Wong TH, Batty E, Wyllie D, Crook DW., et al., (2014) Elucidating hospital norovirus transmission using whole genome sequencing. Presented at British Infection Association London 17th Annual Meeting, 13 June 2014

Posters

Wong TH, Walker AS, Finney J et al., (2011) Use of social network analysis to potentially identify patients most at risk early in a norovirus outbreak. IDSA Boston 21 October 2011

Batty E, Wong TH, Dingle K, et al., (2012) A fast and simple approach to generate whole genome sequences of Noroviruses from faecal samples for infection surveillance during outbreaks, Health Protection 2012 Warwick 11 October 2012

Wong TH, Batty, E, Trebes A., et al., (2013) Long term Norovirus surveillance using Whole Genome Sequencing confirms a new variant as the cause of this winter's UK epidemic. Late breaking poster, ECCMID Berlin 28 April 2013

Declaration and attributions

I, Tse Hua Nicholas Wong, designed and conducted all the analyses presented in this thesis.

A detailed description of work undertaken and assistance obtained with processing of laboratory samples and specific methods is set out below.

Chapter 2 and 3 attributions

All RNA was extracted from faecal samples by myself, under the supervision of Dr Kate Dingle. Roche 454 sequencing was designed with Dr Kulvinder Kaur and Jenny Taylor and undertaken at the Wellcome Trust Centre for Human Genetics. RNA-Seq was designed with Paolo Piazza and Amy Trebes and undertaken at the Wellcome Trust Centre for Human Genetics. Initial pre-processing of raw sequence data was undertaken using a data processing pipeline developed by the Department of Statistics, University of Oxford, in particular by Dr Camilla Ip, Dr Tanya Golubchik, Dr Elizabeth Batty, Dr Madeleine Cule and Dr Daniel Wilson. *de Novo* assembly of the raw sequence data was undertaken by Dr David Wyllie and Adam Giess within the Modernising Medical Microbiology Group. I contributed to the refinement and testing of the sequencing data before applying it to final analysis.

Chapter 4 and 5 attributions

Epidemiological data on symptomatic cases in the OUH was collected by myself with help from the Infection Control department at the John Radcliffe Hospital, namely Graham Pike, Lydia Rylance-Knight, Simon Wells, Lisa Butcher, Dayle Kinch, Ruth Mulroney and Lily O'Connor. Data from non-Oxford Locations were also collected and collated by myself with the help of Professor Mark Wilcox at Leeds, Dr John Paul, Kevin Cole and Angeline Boorer in Brighton. The data was analysed by myself, under the supervision of Professor Tim Peto and Dr Kate Dingle.

Chapter 6 attributions

I designed and conducted the analysis of the whole genome sequencing data under the supervision of Professor Tim Peto and Dr Kate Dingle.

List of Abbreviations and Acronyms

bp	base pair
CDC	Centers for Disease Control and Prevention
DNA	deoxyribonucleic acid
EHEC	Enterohaemorrhagic Escherichia coli
EIA	enzyme immunoassay
exp_cov	expected coverage
Gb	gigabase
HBGA	histo-blood group antigen
HCV	Hepatitis C virus
HIV	Human Immunodeficiency Virus
IID	infectious intestinal disease
kD	kiloDalton
kmer	short string of nucleotide reads
Mb	megabases
ML	maximum likelihood
NGS	next generation sequencing
NHS	National Health Service
nm	nanometer
nt	nucleotide
ORF	open reading frame
OUH	Oxford University Hospitals; comprising of the John Radcliffe, Churchill, Horton Hospitals and Nuffield Orthopaedic Centre
PCR	polymerase chain reaction
qPCR	quantitative PCR

Q score	Phred score
RdRP	RNA dependent RNA polymerase
RNA	ribonucleic acid
RT-PCR	reverse transcriptase polymerase chain reaction
SNV	single nucleotide variant
SRSV	small round structured virus
VP1	major capsid protein
VP2	minor structural protein
WGS	whole genome sequencing

Contents

Abstract 2
Acknowledgements 3
Publications and conference attendances 4
Publications 4
Oral Presentations 4
Posters 4
Declaration and attributions 5
Chapter 2 and 3 attributions 5
Chapter 4 and 5 attributions 5
Chapter 6 attributions 5
List of Abbreviations and Acronyms 6
Chapter 1. Introduction 14
Chapter objective 14
1.1 Epidemiology of acute gastroenteritis 14
1.1.1 Caliciviruses 16
1.2 Norovirus background 16
1.3 Structure and genome 19
1.3.1 Structural proteins 20
1.3.1.1 Major capsid protein VP1 20
1.3.1.2 Minor capsid protein VP2 21
1.3.2 Norovirus nonstructural proteins 22
1.3.2.1 Putative scaffold protein for replication complex assembly (p48) 22
1.3.2.2 Nucleoside triphosphatase (NTP or p41) 22
1.3.2.3 Protein p22 22
1.3.2.4 Protein VPg 22
1.3.2.5 Proteinase (3C or 3CL^{pro}) 23
1.3.2.6 RNA dependant RNA polymerase (RdRp) 23
1.4 Classification 23
1.5 Immunology 26
1.5.1 Virus host interactions 27
1.6 Epidemiology of norovirus 28
1.7 Evolution of norovirus 29

1.7.1 Mutation, quasispecies and genetic bottlenecks 29
1.7.2 Recombination 31
1.8 Clinical features 32
1.9 Pathophysiology 33
1.9.1 Viral entry and primary replication 33
1.9.2 Histological appearaces in the Intestine 33
1.9.3 Physiological and biochemical effects of infection 34
1.10 Transmission 34
1.10.1 Person to person 36
1.10.2 Food and Water 36
1.11 Infection control 37
1.12 Diagnostics and molecular epidemiology 39
1.13 Impact on healthcare 40
1.14 Clinical Management 42
1.15 Advances in DNA Sequencing Technology 42
1.16 First generation DNA sequencing 43
1.17 Next generation sequencing 43
1.18 Using NGS in clinical virology 48
1.18.1 Whole viral genome reconstruction 48
1.18.2 Virus discovery and metagenomics 48
1.18.3 Characterisation of intra host variability 49
1.19 Study Aims 51
Chapter 2. Methods 52
Chapter objective 52
2.1 Sampling frame and collection of data 52
2.1.1 Ethical considerations 52
2.1.2 Clinical samples and epidemiological data 52
2.1.3 Definition of a hospital outbreak 54
2.1.4 Collection of Epidemiological Data 55
2.2 RNA extraction 55
2.3 Detection of norovirus in clinical samples 56
2.4 Whole genome norovirus sequence determination via Sanger sequencing 59
2.5 Template independent determination of norovirus genome sequences 60

2.5.1 Norovirus sequencing via 454 pyrosequencing 60
2.6 Norovirus sequencing via Illumina platform 65
2.7 Bioinformatics and Data analyses 68
2.7.1 Sequence assembly 68
2.7.3 de novo assembly 69
2.7.3.1 Pre-assembly filtering 69
2.7.4 Sequence comparison 72
2.7.5 Phylogenetic tree construction 72
2.7.6 Statistical analyses 73
Chapter 3. Evaluation of Three Approaches to Norovirus WGS Determination 74
Chapter Objective 74
3.1 Experiment 1: WGS Determination using Sanger sequencing 74
3.2 Experiment 2: WGS Determination using the Roche 454 Pyrosequencing Platform 76
3.3 Experiment 3: Genome sequencing determination by Illumina sequencing (mapping to a reference genome) 78
3.3.1 Proof of principle experiment using the Illumina MiSeq bench-top sequencer 78
3.3.1.1 Comparison of WGS determined by Sanger sequencing and the MiSeq platform 79
3.3.2 High throughput WGS determination via the HiSeq platform (reference based assembly of reads) 80
3.4 Illumina sequencing (*de novo* assembly approach) 82
3.4.1 Samples 83
3.4.2 *De Novo* assembly of norovirus sequences 84
3.4.2.1 Assembler performance 84
3.4.2.2 Intra-sample diversity 85
3.4.2.3 Determinants of non-assembly 86
3.4.2.5 Reproducibility of the process 87
3.5 Choosing the best analytical approach 87
Chapter 4. Descriptive epidemiology of Norovirus outbreaks 2009-2013 89
Chapter Objective 89
4.1 Background 89
4.2 Method 90
4.3 Results 91
4.3.1 Epidemiologic information from OUH, 2009-2013 91
4.3.2 Epidemiological data from non OUH sites 93

4.3.3 Samples received and diagnostically tested 93
4.3.4 Seasonal effects of norovirus outbreaks in the OUH 93
4.3.5 Age Ranges 94
4.3.6 Gender 96
4.3.7 Wards affected 98
4.3.8 Description of symptoms 100
4.3.9 Duration of illness 105
4.3.9.2 Brighton and Leeds data 108
4.4 Discussion 110
Chapter 5. Description of outbreaks between wards 114
Chapter objective 114
5.1 Background 114
5.2 Epidemiological data from Oxford 114
5.2.1 2009-2010 season 114
5.2.2 2010/2011 season 117
5.2.3 2011/2012 season 120
5.2.4 2012/2013 season 122
5.3 Outbreaks in Brighton and Leeds 123
5.4 Discussion 123
Chapter 6 Understanding norovirus transmission during hospital outbreaks by combining genomic and epidemiological data 126
Chapter objectives 126
6.1 Background 126
6.2 Virus Genomes Sequenced 128
6.3 Oxfordshire Norovirus Genomes 130
6.4 Analysis of OUH norovirus genomes 131
6.5 Within Season Norovirus Genetic Diversity (OUH 2009-2010) 133
6.5.4 Within Season Norovirus Genetic Diversity (OUH 2010-2011) 140
6.5.5 Genomes from the 2011/2012 season 145
6.5.6 Within Season Norovirus Genetic Diversity (Oxfordshire 2012-2013) 151
6.6 Community noroviruses and their relationship to viruses causing hospital outbreaks 155
6.7 Genomic diversity between geographical sites 162
6.8 Overall relationships among local, regional and national noroviruses 165
6.9 Discussion 167

Chapter 7. Final conclusions 170
7.1 Whole genome sequences of norovirus can be determined using next generation sequencing technology 170
7.2 Epidemiological data question the reliability of the Kaplan criteria in outbreak identification 171
7.3 WGS provides enhanced discrimination among norovirus strains 171
7.4 Norovirus genomes and epidemiological data can be combined to characterise virus transmission events 171
7.5 Limitations of the study 172
7.6 Translational benefits and future directions 174
References 176
Supplementary information 187
Genome sequences from other geographical locations 187
S1 Brighton 187
S2 Leeds 192
S3 East Surrey Hospital and other sites 195
Appendices 200
Appendix 1. Details of primers used for producing the seven overlapping amplicon fragments. 200
Appendix 2. List of all 477 norovirus genomes downloaded from NCBI for bioinformatic processing 201
Appendix 3. Coverage and percentage plots of 61 HiSeq sequences. 217
Appendix 4. Epidemiological histograms for outbreaks in Brighton and Leeds 218
Appendix 5. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples season 2009-2010. 226
Appendix 6. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples season 2010-2011 227
Appendix 7. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples season 2011-2012 228
Appendix 7. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples season 2011-2012 (cont'd) 229
Appendix 8. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples season 2012-2013 230
Appendix 9. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples 231
Appendix 9. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples (cont'd) 232

Appendix 9. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples (cont'd) 233
Appendix 9. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples (cont'd) 234
Appendix 9. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples (cont'd) 235
Appendix 9. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples (cont'd) 236
Appendix 10. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples season 2011-2012 (Brighton) 237

Chapter 1. Introduction

Chapter objective

To describe the basic structure, natural history and pathogenesis of norovirus, including our current understanding of its effects in the healthcare setting. Additionally to briefly review current and past approaches to whole genomic sequencing.

1.1 Epidemiology of acute gastroenteritis

Acute infectious gastroenteritis remains a common illness affecting all age groups, with clinical features such as nausea, vomiting, diarrhoea and abdominal cramping commonly seen, accompanied with characteristic inflammation of the mucosal lining of the stomach and the intestinal tract. Agents of infectious gastroenteritis (or infectious intestinal disease) are estimated to cause 1.5 billion cases world-wide and 1.45 million deaths annually [1-3]. Globally, one in ten child deaths result from infectious intestinal disease in the under fives, resulting in 800,000 fatalities worldwide annually [4]. Causes of infectious intestinal disease (IID) may be bacterial, protozoa or viral. The incidence and aetiology of IID in under fives are shown in Figure 1.1.

Within the UK the incidence of IID is substantial, with around 25% of the population suffering from an episode of IID in a year, equivalent to 17 million cases annually [5]. The most commonly identified microorganisms found in stool samples from those with IID were norovirus, sapovirus, Campylobacter *spp*. and rotavirus. Figure 1.2 illustrates the aetiological causes of infectious intestinal diseases in England between 2008-2009 within the community and general practice presentation cases, irrespective of ages.

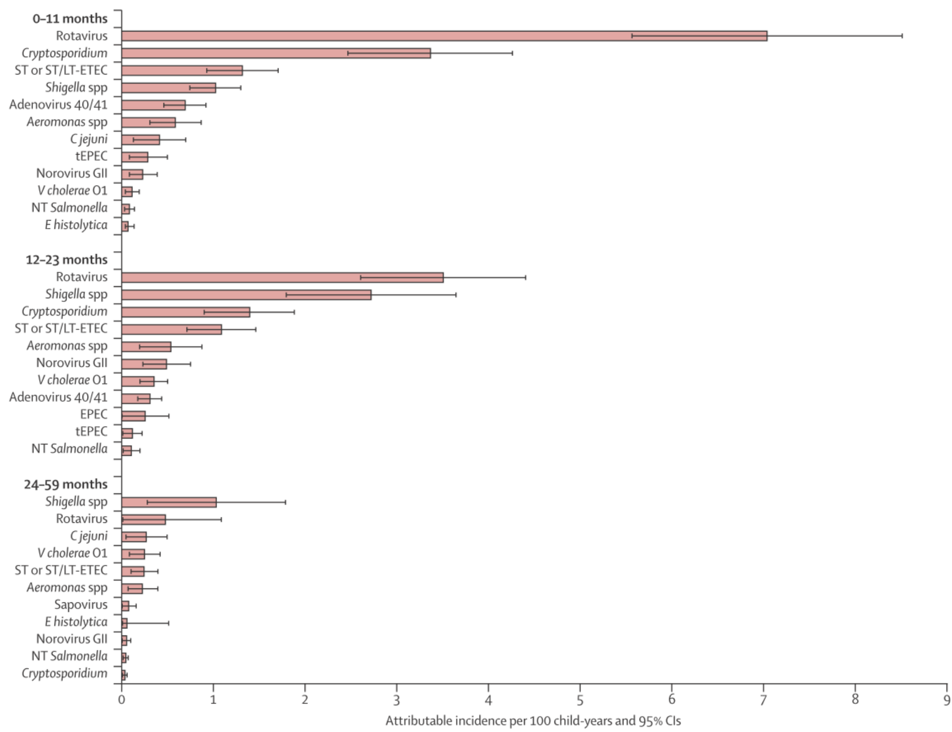


Figure 1.1 Attributable incidence of pathogen specific moderate to severe diarrhoea per 100 child-years by age stratum, all sites combined. Data from Global Enteric Multicenter Study (GEMS) and taken from Kotloff et al., 2013 [4]

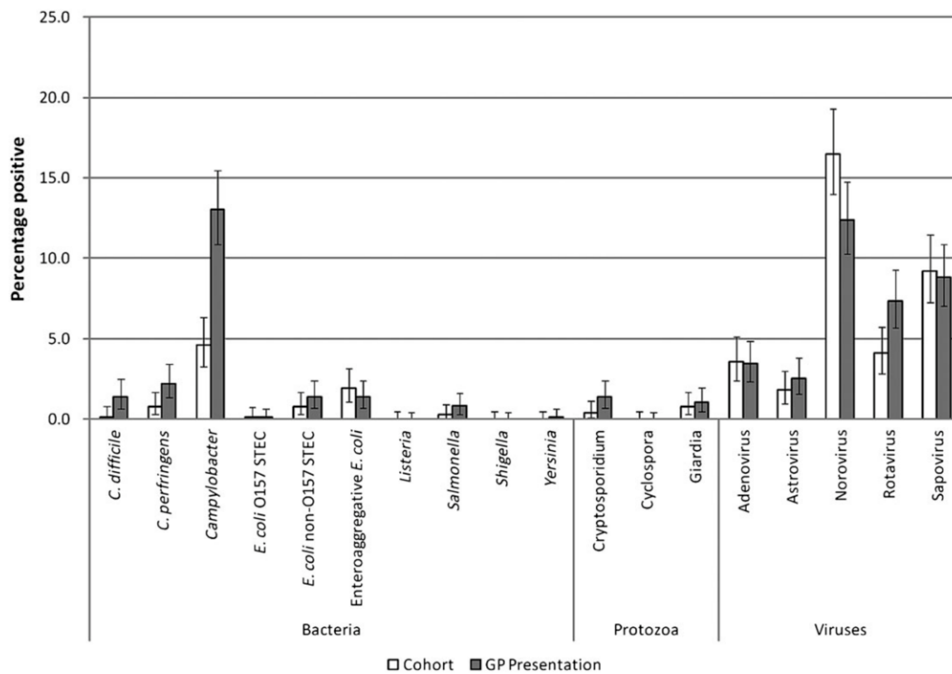


Figure 1.2 Microbiological findings in community and general practice presentation cases in the second study of Infectious Intestinal Disease (2008-2009).This included samples taken from all age groups. Taken from Tam et al., (2012)[5]

1.1.1 Caliciviruses

Noroviruses are members of the *Caliciviridae* family, members of which are small (~27 to 38nm) isosahedral, non enveloped, positive sense, single stranded RNA viruses that can infect a large range of host species [6]. The family also includes *Lagovirus, Vesivirus, Nebovirus, and Sapovirus* [7], but only the latter and norovirus are human pathogens. Norovirus is recognised as the most important cause of epidemic, non bacterial gastroenteritis amongst all age groups worldwide [8], and account for 40-50% of all foodborne gastroenteritis in the United States [9].

1.2 Norovirus background

Norovirus is now well established as the most common cause of acute non-bacterial gastroenteritis outbreaks [10, 11]. An infectious cause for many such outbreaks was established with the challenge of volunteers in 1972 using faecal filtrates collected during a 1968 elementary school outbreak in Norwalk, Ohio [12]. Subsequently, the "Norwalk virus" was visualised via electron microscopy of these volunteer stools. The history of norovirus, as it was subsequently designated by the International Committee on the Taxonomy of Viruses [13] has been summarised in Table 1.1.

Table 1.1. Important events in the chronology of norovirus gastroenteritis. Adapted from Goodgame et al., (2006) with modifications [10].

Year	Event	Reference
1968	Outbreak of acute gastroenteritis among teachers and students at a school in Norwalk, OH; no organism could be cultured.	[12]
1971	Volunteers infected with bacteria-free filtrates of stool from Norwalk outbreak.	[14]
1972	Virus identified (by immune electron microscopy of stool) as a cause of acute gastroenteritis by volunteer challenge.	[12]

1977	Other viruses morphologically similar to Norwalk virus (small round-structured viruses (SRSV)) seen by immune electron microscopy in patients with symptoms suggestive of viral gastroenteritis.	[15]
1978	Immune adherence haemagglutination assay demonstrates antibody to Norwalk virus in 50% of adults.	[16]
1981	Detection of a single SRSV structural protein with an estimated mass of 59 kD suggests that SRSV could be a calicivirus.	[17]
1982	Radioimmunoassay detection of SRSV antigen in stool begins to establish SRSV as a common cause (40%–60%) of outbreaks of acute nonbacterial gastroenteritis.	[18]
1985	Enzyme immunoassay for SRSV antigen in stool and antibody in blood is efficient and quick and eliminates need for radioisotopes.	[19]
1989	Enzyme immunoassay used to investigate an outbreak of SRSV aboard a cruise ship.	[20]
1990	"Norwalk"-like virus identified as a calicivirus based on its genome consisting of positive-sense, single-stranded, polyadenylated RNA.	[21]
1992	RT-PCR detection of Norwalk-like virus nucleic acid in stool.	[22]
1992	SRSV capsid protein (produced in large quantities in a baculovirus expression system) spontaneously forms virus-like particles that are morphologically and antigenically similar to naturally occurring virus.	[23]
1993	RNA genomes of Norwalk-like virus and other SRSVs designated Southampton virus and Lordsdale virus are sequenced.	[24-26]
1995	Through careful selection of primer pairs that each target a conserved area of the viral genome, broadly reactive RT-PCR assays can overcome the genetic diversity of the noroviruses, facilitating development of a stool test sensitive enough for clinical specimens.	[27]
1996	Use of RT-PCR detection assays in The Netherlands shows that 91% of outbreaks of acute nonbacterial gastroenteritis are due to Norwalk-like viruses.	[28]
1998	Broadly reactive RT-PCR analysis of 90 outbreaks involving 5000 people demonstrates that SRSV are the cause of 96% of foodborne and epidemic nonbacterial gastroenteritis in the United States.	[29]
1998	Use of broadly reactive RT-PCR to detect SRSV on environmental surfaces in investigation of hospital epidemic of acute gastroenteritis.	[30]

2000	First time RT-PCR used to detect "Norwalk-like viruses" in food samples (ham from delicatessen bar) contaminated by a food handler, suspected to be the source of an epidemic of acute gastroenteritis.	[31]
2000	Norwalk-like viruses named as one of two human genera in the calicivirus family.	[7]
2000	Molecular epidemiology shows frequency, severity, and economic importance of Norwalk-like viruses in outbreaks of diarrhoea in hospitals and nursing homes.	[32]
2000	New diagnostic tests provide insights into reasons underlying high Norwalk-like virus prevalence: large human reservoir, many asymptomatic cases, viral stability in the environment.	[33]
2002	Nine outbreaks of norovirus gastroenteritis occur on cruise ships; some persist after cleaning and require the ship to be put out of service for outbreaks to cease.	[34]
2002	100% to 200% increase frequency of norovirus outbreaks in 10 European countries possibly related to new variant strain that rapidly becomes the predominant norovirus throughout Europe.	[35]
2003	European Union–funded Foodborne Viruses in Europe Network and US CDC Calicinet established as surveillance structures, early warning tools, and research centers for norovirus outbreaks.	[36]
2003	Broadly reactive and highly sensitive assay for Norwalk-like virus based on real time quantitative PCR.	[37]
2004	Attempts to develop a method for the cultivation of Norovirus are unsuccessful.	[38]
2004	Asymptomatic and symptomatic excretion of norovirus found during hospital outbreak.	[39]
2005	Norwalk-like virus replaced with Norovirus at International Committee on Taxonomy of Viruses	[40]
2006	GII.4 norovirus is the predominant circulating genotype identified in norovirus outbreaks worldwide.	[41]
2008	Norovirus found to cause 900,000 cases of paediatric gastroenteritis and at least 1.1 million episodes and 218,000 deaths in developing nations.	[8]
2008	Infectious dose estimated to be between 18-1,000 virus particles.	[42]

2012	New GII.4 strain Sydney found to be new emergent variant, displacing previous New Orleans variant.	[43]
2014	As of February 2014, Norovirus still cannot be cultured in vitro and a straightforward animal model is lacking	[44]

1.3 Structure and genome

Norovirus is a non-enveloped positive sense single stranded RNA virus. Morphologically it measures 27-38nm (Figure 1.3).

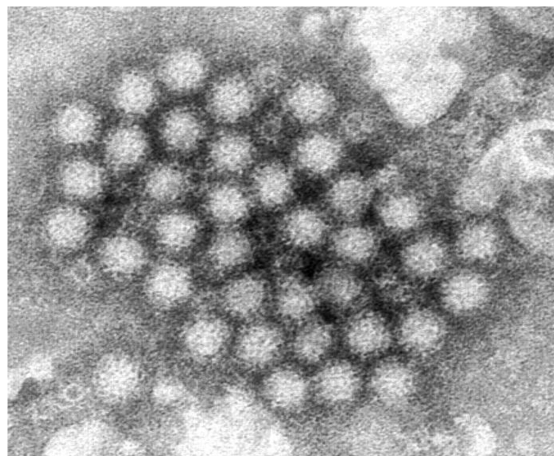


Figure 1.3 Electron microscope image of norovirus.
Content provider CDC/Charles D. Humphrey with permission

The norovirus genome is between 7.4 to 7.7 kilobases in length and organised into three overlapping open reading frames (ORF1-ORF3). The 5' proximal end of the genome is protein linked, whilst the 3' end is polyadenylated (Figure 1.4) [45]. ORF1 encodes the non structural polyprotein which is processed co-and post-translationally by the viral 3C-like protease ($3CL^{pro}$). The absolute number of functional precursors and non-structural proteins is not clear but the functions of some are known [46]. The ORF1 coding order (N to C-terminal) are: putative scaffold protein (p48); nuceloside triphosphatase (NTP); protein 22 (p22); viral protein (VPg); protease 3CL pro (3C); and RNA dependent RNA

polymerase (RdRP). ORF2 encodes the major capsid protein (VP1) and ORF3 encodes a minor structural protein (VP2) [47, 48].

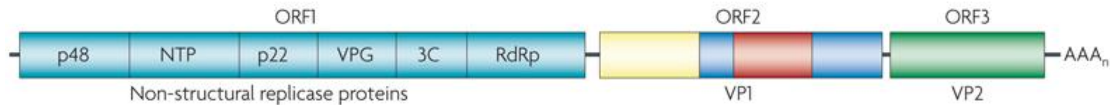


Figure 1.4. Genome organisation of Norovirus. Consists of 3 open reading frames (ORF) encoding non structural polyprotiens (ORF1), major capsid protein (ORF2) and minor capsid protein (VP2). Taken from Donaldson et al., (2010)[49].

1.3.1 Structural proteins

The norovirus capsid contains 90 dimers of the major capsid protein VP1 and relatively few (one or two) copies of the minor structural protein VP2 [50].

1.3.1.1 Major capsid protein VP1

The ORF2 encoded VP1 ranges in size in different noroviruses from ~530-555 amino acids with a calculated molecular weight of 58-60 kiloDaltons. A central variable domain carrying antigenic determinants defining strain specificity is flanked by conserved domains at both ends. VP1 can assemble into virus like particles (VLPs) when expressed in baculovirus, and this recombinant material has allowed the structural and functional domains of the capsid to be defined [51]. The VP1 protein consists of an internal N-terminal (N) domain, shell (S) domain and a protruding (P) domains [52] (Figure 1.5). The N terminal 225 amino acids constitute the S domain and contain elements essential for formation of the icosahedron. Baculovirus studies have revealed that expression of the S domain alone can still allow VP1 dimer formation to occur, but with the production of smaller, smooth VLPs [53]. The P domain is divided into P1 and P2 sub-domains. The P2 domain is especially important as it forms the most distal surface of the folded monomer,

whereas the P1 sub-domain forms the sides of the arches. Consequently, the hyper variable region within the P2 sub-domain is believed to play an important role in receptor binding and immune reactivity, and may be responsible for Histo Blood Group Antigen (HBGA) interactions associated with norovirus infections [54, 55].

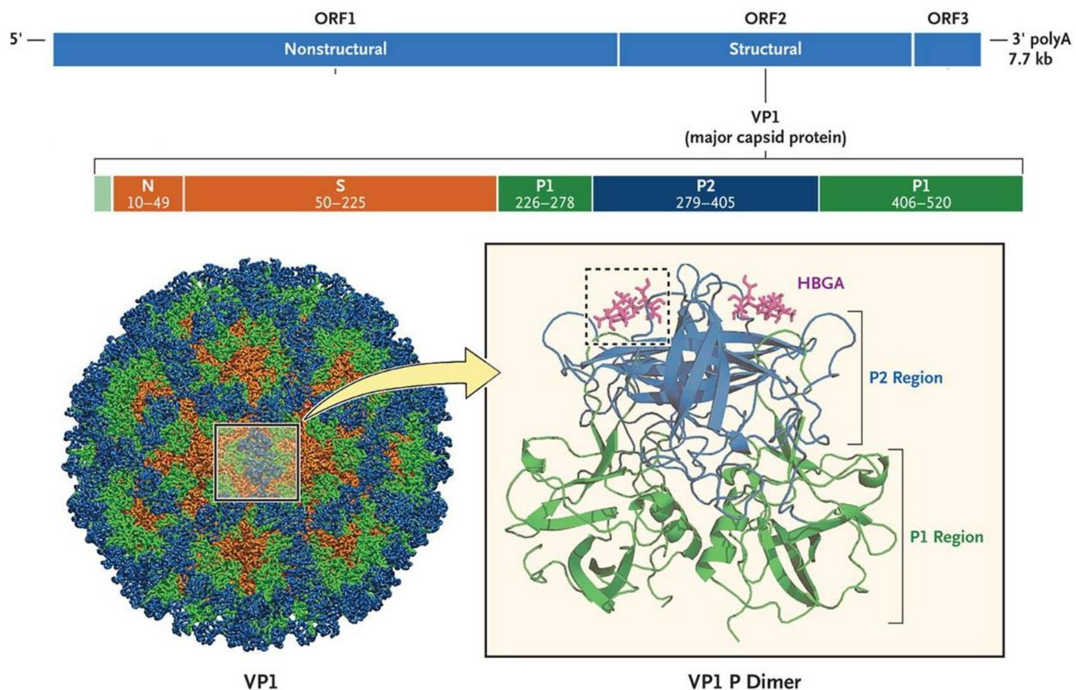


Figure 1.5. Norovirus capsid. VP1, the major capsid protein is organised structurally into the N-terminal (N), shell (S), protruding (P) domains. The P2 region of the P domain (blue) form the outermost surface of the capsid protein and is the site where HBGA (magenta) interacts with the virion (dashed line in box).Taken from Bok (2012)[56]

1.3.1.2 Minor capsid protein VP2

ORF3 encodes the minor structural protein designated VP2 which ranges from 208-268 amino acids in size, with a calculated molecular weight of ~22-29kDa, and exhibits extensive genetic diversity among strains. The role of VP2 beyond that of a minor structural protein has not been well defined. VP2 has been shown to possibly play a role in particle stability, although VLPs can assemble in the absence of VP2 [53].

1.3.2 Norovirus nonstructural proteins

Norovirus ORF1 encodes a large polyprotein which is cleaved into non-structural proteins. Each protein and its function is explained briefly below.

1.3.2.1 Putative scaffold protein for replication complex assembly (p48)

Unlike other norovirus non-structural proteins, the amino acid sequence of p48 shows no similarity to known viral or cellular proteins in the public databases. It is postulated that p48 may function as a scaffold protein for replication complex assembly [57].

1.3.2.2 Nucleoside triphosphatase (NTP or p41)

Purification of NTP (or p41) revealed binding as well as hydrolysis of ATP in vitro [58].

1.3.2.3 Protein p22

There are as yet no data on the possible functions of p22, but it does occupy a position within the norovirus genome which corresponds to the 3A protein in *picornaviruses*. The 3A protein in *picornavirus* is important for membrane localization of replication complexes, and targeted mutations in 3A yield viruses defective in RNA synthesis with decreased ability to inhibit cell secretory pathways [59].

1.3.2.4 Protein VPg

VPg is covalently linked to norovirus genomic and subgenomic mRNAs. In other viruses, such as *Picornaviridae, Potyvirdae, and Comoviridae*, VPg has a variety of functions in their replication cycles. It is known that Calicivirus genomic RNA lacking VPg is not infectious [60]. In feline calicivirus, removal of the VPg resulted in much reduced viral

protein synthesis in vitro [61]. Hence it is believed that norovirus VPg functions in recruiting the translation machinery to the viral genomic RNA [62].

1.3.2.5 Proteinase (3C or 3CL^{pro})

Norovirus encodes a single protease called 3C-like (3CL^{pro}) due to its similarity to the picornavirus 3Cm protein. Studies of various norovirus strains suggest that the 3CL^{pro} modulates both viral and cellular gene expression [63].

1.3.2.6 RNA dependant RNA polymerase (RdRp)

The norovirus RNA dependent RNA polymerase (RdRp) extends from amino acid 1,281 to the C-terminus of ORF1. The RdRp is predicted to play an integral role in the synthesis of negative sense genomic RNA as well as positive sense subgenomic and genomic RNA [64].

1.4 Classification

Determination of the complete norovirus genome sequence and hence its genetic organisation, was originally delayed by the lack of a norovirus cell culture system. Once molecular techniques became available, norovirus sequences were cloned using baculovirus recombinants and later amplified using total stool cDNA [22, 23]. In the 1990s the complete genome sequences were determined for both major genogroups [24, 25]. This facilitated a large number of studies of norovirus molecular epidemiology which revealed the diversity of the genus. The norovirus genus is currently sub-divided into six main genogroups (designated GI-GVI) each of which is further subdivided into a total of approximately 30 genotypes (8 GI, 19 GII, 2 GIII, 1 GIV, and 1 GV). Assignment to a genotype is based on sequence similarity within the VP1 protein region of the genome (Figure 1.6). Members of GI and GII are by far the most frequent cause of human

infections, with GII.4 (Genogroup II, Genotype 4) being the commonest individual strain responsible for 70-80% of human norovirus outbreaks worldwide since the mid-1990s [65]. Significant genetic variation of ORF1 and viral capsid amino acid sequence exists within a genogroup (14-44%) and between genogroups (45-61%)[45]. Recently a standardised nomenclature has been published which includes the organism, host, country code, year of sampling, genogroup and genotype and variant name [66], e.g norovirus GII/Hu/GB/2010/GII.P4_GII.4_NewOrleans2009/London48. New variants of the virus, especially within GII.4 has been named according to the year and location of the first full length capsid sequence in the public domain, e.g. GII.4 New Orleans 2009 or the latest variant, GII.4 Sydney 2012 [67].

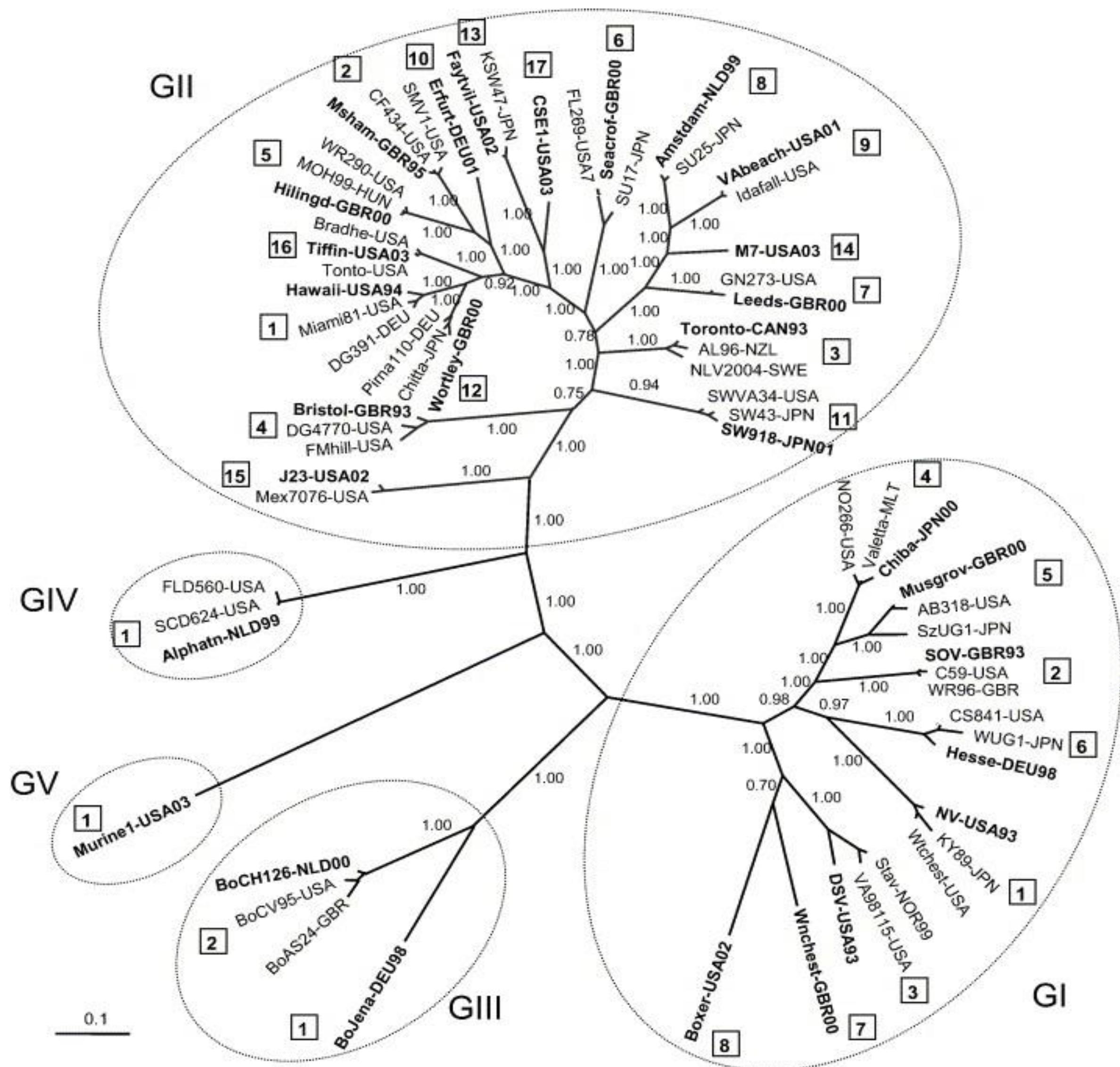


Figure 1.6. Phylogenetic analysis of complete capsid amino acid sequences of 141 norovirus strains. Taken from Zheng et al., 2006 [45]. The tree was constructed with the structural alignment of 141 norovirus sequences (column gaps removed) using the MrBayes program. Tree topology was evaluated on the basis of 1,000,000 generations (posterior possibility, 1.00 equals 100% out of 1,000,000). Numbers in square boxes are the cluster number within the genogroup (G). Country codes: AUS, Australia; CAN, Canada; DEU, Germany; FRA, France; HUN, Hungary; JPN, Japan; MLT, Malta; MEX, Mexico; NLD, Netherlands; NZL, New Zealand; NOR, Norway; SWE, Sweden; GBR, the United Kingdom; USA, the United States.

1.5 Immunology

Our understanding of human immunity to norovirus infection is complicated by the heterogeneous responses of the population and the transient nature of immunity in some individuals. Unravelling the nature of human protective immunity against norovirus is challenging for several reasons. Firstly, noroviruses are highly prevalent globally meaning that nearly everyone has been previously exposed to one or more norovirus strains in their lifetime [49]. Secondly, human noroviruses do not appear to infect small animals [68, 69] and cannot be cultured in vitro. Hence much of our current understanding of human immunity to norovirus infection has been derived from human challenge studies, of which there have been a few since the 1970s [70-72]. Five strains from two different genogroups have been used in such studies to date: GI.1 (Norwalk), GI.5 (Montgomery), GII.2 (Snow Mountain), GII.1 (Hawaii) and recently GII.4 (Farmington Hills)[73]. In the initial challenge studies, involving GI.1 Norwalk virus, not all individuals were susceptible or symptomatic. It was postulated that continuous susceptibility may be due to the failure to generate a virus specific immune response. However, a virus specific antibody response does develop, the presence of which does not correlate with protection against subsequent infection [71]. Virus specific serum IgA and IgM responses are short lived, whilst serum IgG persists for months following infection [72]. It has been difficult to determine whether norovirus specific antibodies are neutralising; surrogate cytokine studies have been performed to investigate specific T cell induction in infected individuals. Peripheral blood mononuclear cells (PBMC) were collected from norovirus challenged volunteers at pre challenge and post challenge and secretion of cytokines measured. However, that particular study was deemed inconclusive due to the difficulty

of finding true unexposed uninfected individuals as controls, as both the pre-challenge and post-challenge controls exhibited similar increases in cytokine levels [74].

It has recently been suggested that norovirus strains can elicit herd immunity. This finding is based on the theory of "epochal evolution" of the GII.4 strain pandemics, with the emergence of a pandemic virus strain for a 1-2 year period followed by a period of relative GII.4 low activity prior to the emergence of a new pandemic GII.4 strain [75]. It is postulated that the emergence of a new pandemic strain results from (i) antigenic drift mediated altered carbohydrate receptor usage and (ii) altered antigenicity facilitating virus escape of herd immunity [65, 76].

Considerable hurdles such as strain heterogeneity have yet to be properly addressed. Vaccine research is currently still at an early phase, with some multivalent vaccine trials underway which may show some promise [77, 78].

1.5.1 Virus host interactions

Recently, two genetically determined factors, a person's ABO blood type and their secretor status (ABO and FUT2 (alpha-1,2 fucosyltransferase) gene products) has been identified as conferring susceptibility or resistance to symptomatic Norwalk virus infection [70, 72]. Norwalk virus (GI.1) appears to infect individuals who have the gene that encodes the functional FUT2 enzyme allowing expression of histo-blood group antigens (HBGAs) on mucosal surfaces and a secretor-positive phenotype. The glycosyltransferases that control their synthesis are encoded by the highly polymorphic ABO, Lewis, and secretor gene families. Individuals bearing a mutation in the FUT2 enzyme (~20% of the population) will not express the HBGAs necessary for docking and

entry, and are therefore resistant to infection [72]. This association between the HBGA and norovirus has been shown to be an important modality for norovirus infectivity with Norwalk as well as some GI and GII strains [70, 79]. However, secretor negative individuals challenged with GII.2 strain (Snow Mountain virus) do develop clinical illness. Therefore, other important modalities, such as attachment to Lewis carbohydrates, may be important as co-modulators and need further investigation [74]. More recently, a challenge study where individuals were inoculated with donor challenged units of GII.4 norovirus, showed that up to 70% of secretor positive individuals were infected with norovirus [73]. In this study, only one non-secretor out of seventeen (5.9%) became ill when challenged, whilst another shed virus for a single day.

More recently, a study by Jones et al., (2014) have demonstrated in mouse models, that B cells may act as a cellular target of noroviruses, requiring HBGA expressing enteric bacteria as a stimulatory factor for norovirus infection. This may lead to the development of an *in vitro* infection model for human noroviruses in the future [80].

1.6 Epidemiology of norovirus

Norovirus outbreaks occur mostly in semi closed environments such as nursing homes, schools, hospitals, cruise ships, and military settings [81]. There appears to be a peak during winter (hence norovirus being referred to as "winter vomiting virus") although outbreaks do occur throughout the year [82]. In 2002, a large rise in norovirus outbreaks was reported worldwide which correlated with the emergence of a new variant of the GII.4 strain [35]. As discussed in Chapter 1.5, successful variants of this strain which is thought to evolve continually through epochal evolution, have caused global pandemics in 2002 (GII.4 Farmington Hills 2002 strain), 2004 (GII.4 Hunter 2004 strain), 2006 (GII.4

Yerseke 2006a and GII.4 Den Haag 2006b), 2008 (GII.4 Apeldoorn 2007), 2010 (GII.4 New Orleans 2009) and 2013 (GII.4 Sydney 2012). Each of these GII.4 variants rapidly spread globally and overall they are believed to account for a minimum of 70-80% of all norovirus outbreaks since 2002 [83].

1.7 Evolution of norovirus

The availability of multiple nucleotide sequences representing complete and partial norovirus genomes is gradually improving understanding of the emergence and evolution of new human norovirus strains. The GII.4 strain has predominated for at least 35 years (and possibly longer as prior to this viruses were not characterised) and it appears that a new GII.4 variant appears on average every 2-3 years, displacing the previous strain near to extinction [35, 84]. The genetic mechanisms underlying the emergence of new GII.4 strains are not fully understood, but epochal evolution and herd immunity may play a large part. Norovirus utilise two main mechanisms of variation: mutation and homologous recombination [85]. Both have been proposed to explain why the pandemic GII.4 lineage has been so influential during outbreaks.

1.7.1 Mutation, quasispecies and genetic bottlenecks

RNA dependent RNA polymerases lack a proof-reading mechanism, causing RNA viruses to have the highest mutation rate of all organisms [86]. A review of the literature by Drake et al., (1999) revealed a mutation rate of between 10^{-3} to 10^{-5} per generation per nucleotide. In comparison, DNA dependent DNA polymerases have an error rate of around 10^{-9} per generation per nucleotide [87]. The mutation rate of norovirus has been estimated to be in the range of 1.9-9.0 x 10^{-3} substitutions per nucleotide per year [84, 88, 89], favouring the generation of variants on which selection pressures, including host

or herd immunity, can act [90]. Interestingly, it has been shown that the fidelity of the norovirus RdRp exhibits an inverse relationship with the prevalence of its genotype [89]. Consequently, genotype GII.4 and GII.b/GII.3, the two predominant strains, appear to have a lower RNA polymerase fidelity than the other strains, suggesting, like in the example of poliovirus, that low fidelity could provide a fitness benefit, enabling viruses to avoid immune recognition by rapidly altering their antigenic properties [91], thus becoming more prevalent.

The point mutations that noroviruses accumulate during infection of an individual host may give rise to what is termed a "quasispecies population" consisting of a quantitatively dominant genome, surrounded by a cloud of many sequences which differ from the majority sequence to various extents [90]. The presence of heterozygous variants within samples has been reported, particularly with chronically infected immunocompromised patients [92].

Transmission bottlenecks occur when only a few individual pathogens are transmitted from one infected host to another to initiate new infection [93]. This has dramatic effects on the evolution of virulence in certain pathogens such as RNA viruses. It has been reported in both HIV and HCV that this "bottleneck" occurs following a transmission event, where on average 1 to 3 viruses are transmitted to the new host [94, 95]. Strong functional constraints on the transmitted variants are believed to drive this event [96]. The extent to which this effect occurs in norovirus is being questioned. Both HCV and HIV are chronic infections, and the potential for the development of quasispecies is high. Norovirus infections are acute in almost all patients, hence the opportunity for the development of quasispecies and subsequent bottlenecking effects may be less relevant.

It may be important to understand the role played by chronically infected immunocompromised patients as reservoirs of norovirus infection, and generators of potentially novel variants, which may go on to infect the immunocompetent population [97].

1.7.2 Recombination

Evidence of inter-genotypic recombination has been found in norovirus [98], as in other RNA viruses [99], providing a further mechanism whereby new variants may be generated, in addition to point mutation. The recombination break point most frequently identified to date lies within the overlapping 17nt of ORFs 1 and 2. Consequently, recombination often results in new combinations of polyprotein and capsid genes [98]. Inter-genotypic norovirus recombination events have been reported and this could play an important role in the emergence of some of the epidemic GII.4 variants [100, 101]. As reviewed by Eden et al., (2013) [67], studies have reported recombination within ORF1 [102, 103] and ORF2 [65, 104]. Compared to certain bacteria and other RNA viruses, the absence of very large numbers of genome sequences for norovirus (of a single genotype) has hampered the detection of potential intra-genotype recombination, particularly within the important GII.4 lineage. This limitation of the available data is illustrated by the fact that most studies have sequenced only a small proportion of the genome flanked by conserved sequences used for PCR oligonucleotide primer design; the capsid or the region of ORF1 encoding the RNA dependent RNA polymerase [67]. Furthermore, it has been noted [67] that the lack of multiple full genome sequences of individual norovirus genotypes also makes it difficult to distinguish true recombination from other processes

shaping the evolution of the GII.4 lineage, such as a high level of immune selection acting on non-synonymous point mutations within the capsid gene [67].

1.8 Clinical features

Norovirus affects both genders and all ages [81]. Symptoms are characterized by an acute onset of nausea, vomiting, abdominal cramps, myalgias, and non-bloody diarrhoea. The incubation period is believed to be between 15-48hours, and symptoms can continue for up to 10 days [105, 106], although they usually resolve within 48 hours [107]. However, recent studies have shown that the median duration of symptomatic illness may be longer (4–6 days) in patients affected during hospital outbreaks and in children <11 years of age [108]. A long period of norovirus excretion post infection has been demonstrated, continuing for a median of 28 days after inoculation (range 13-56 days) [109]. Vomiting is more prevalent in persons >1 year of age, whereas children <1 year more often develop diarrhoea alone. Fever, which is reported in 37–45% of the patients, typically resolves within 24 hours. Reports have suggested that vomiting is the prominent symptom in more than 50% of affected individuals; whilst diarrhoea tends to be short lived and less severe than that due to other causes of gastroenteritis [18]. Deaths have been reported during outbreaks in nursing homes and norovirus has been associated with necrotizing enterocolitis [110]. In community-based studies, clinical disease caused by norovirus appears to be milder than rotavirus disease; however, children hospitalized with norovirus infection tend to have similar severity scores to those with rotavirus infections [111]. In the hospital ward setting, nosocomial outbreaks of norovirus gastroenteritis are very disruptive and costly to effective daily management [112]. Attack rates have been estimated and may be very high, on occasion affecting more than 50% of ward patients

and staff [113]. The disease is usually self limiting, but symptoms may persist in vulnerable populations such as the elderly and the immunocompromised, with symptoms lasting over two years in the latter category [114, 115].

1.9 Pathophysiology

Data concerning the pathophysiology of norovirus infection have been derived from physical, histological and biochemical studies of infected human volunteers [116, 117].

1.9.1 Viral entry and primary replication

Norovirus is spread mainly via the faecal oral route and infectious virions entering via the mouth and are then transported to the digestive tract. The exact location for primary replication is unknown, but it is believed that it may occur in the upper gastrointestinal tract, such as the duodenum or the jejunum of the small intestine. There have also been reports of extra-intestinal spread, although the mechanism is unclear [118].

1.9.2 Histological appearaces in the Intestine

Histologically, the proximal intestinal mucosa of human volunteers who became ill after administration of both Norwalk (GI.1) and Hawaii (GII.1) viruses showed broadening and blunting of the villi, shortening of the microvilli, enlarged and pale mitochondria, increased cytoplasmic vacuolisation, and intercellular oedema [116, 117, 119]. Crypt cell hyperplasia has also been reported following norovirus infection. Since only proximal intestinal biopsies were taken from the volunteers it is unknown whether the distal intestine is also affected by norovirus. Some studies have suggested that norovirus causes apoptosis of enterocytes in humans, pigs and mice [116]. The precise cell tropism of human norovirus is unknown. Interestingly, when intestinal biopsies from human

volunteers were analysed by electron microscopy, virus particles were not observed, leading many to postulate that human norovirus infect very few cells in vivo [117].

1.9.3 Physiological and biochemical effects of infection

It has been reported from human volunteer studies that norovirus infection causes a marked delay in gastric emptying, possibly related to the high frequency of vomiting episodes [120]. Malabsorption of elements such as D-xylose, fat and lactose also occurs, correlated with shortened microvillli and decreased activity of specific brush border enzymes on enterocytes, including sucrase, trehalase, and lactase [121]. A secretory component to norovirus induced diarrhoea is as yet unknown.

1.10 Transmission

Several properties of Norovirus facilitate its spread: its low infectious dose; its prolonged period of post infection shedding; its high environmental stability and the inadequate long term immunity of many patients (table 1.2). Faecal oral transmission may involve prior contact of the patient with a contaminated environment, or the consumption of contaminated food or water. In addition, direct person to person spread via aerosols has been described (figure 1.7)[122].

Table 1.2. Characteristics of noroviruses that facilitate their spread during epidemics.

Feature	Observation	Consequences
Low infectious dose	$<10^2$ viral particles	Permits person-to-person spread via aerosols or contaminated surfaces, or spread by foodhandlers or contaminated water.
Prolonged asymptomatic shedding	≤2 weeks	Increased risk for secondary cases; a particular issue regarding post-symptomatic foodhandlers.
Environmental stability	Survives up to 10 ppm chlorine, freezing, and heating to 60 C	Difficult to eliminate for example from contaminated water; virus maintained in ice and steamed oysters
Substantial strain diversity	Multiple genetic and antigenic types	Hampers diagnosis of community acquired infections due to the prevalence of a wide range of genetically diverse strains, a property linked to multiple infections over the lifetime of an individual.
Lack of lasting immunity	Disease can occur on re-infection	Childhood infection does not protect from disease in adulthood even by the same strain; difficult to develop vaccine affording lifelong protection.

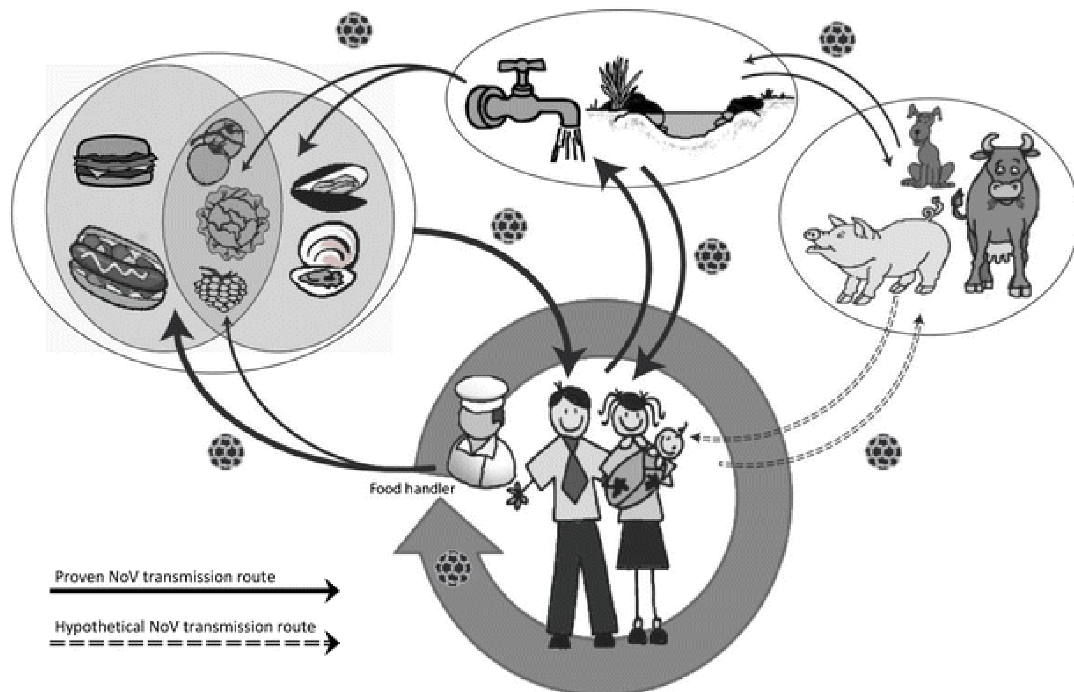


Figure 1.7 Schematic overview of the transmission routes of human and animal norovirus. Solid and dashed arrows indicate proven and hypothetical transmission routes, respectively. The thickness of the arrows is related to the probability of transmission via a given route. Taken from Mathijs et al., (2012) [122]

1.10.1 Person to person

Person to person transmission was cited as the cause of 83.7% of nearly 3,000 norovirus outbreaks reported in the USA between 2009-2013 [123]. Similarly, in Europe, 74% of almost 900 confirmed norovirus outbreaks between 2002-2006 were identified by the Food borne viruses in Europe network as involving predominantly person to person spread. Lopman et al., (2003) [124] and Siebenga et al., (2007) [125] reported that person to person spread was the key transmission route in 85% and 82% of norovirus outbreaks in the UK and The Netherlands, respectively.

Numerous reports have described surface contamination in hospital wards including switches, televisions, mobile telephones, public phones, water taps, toilet switches, microwave ovens, keyboards, bed frames and chairs [126, 127]. In schools, computer keyboards and mice have been implicated as major vectors of transmission [128]. It is likely that both aerosols and contaminated environmental surfaces facilitate transmission in semi closed communities.

1.10.2 Food and Water

Norovirus outbreaks, which can be large scale, may also result from contamination of drinking or recreational water sources infected with human faecal material [129, 130]. Contaminated surface waters, including lakes and bays which are used recreationally, represent a known route for water-borne norovirus transmission [131]. A review of water borne gastroenteritis outbreaks (2001 to 2004) demonstrated norovirus as the cause in 39.2% where an aetiology was confirmed [132, 133].

The majority of cases due to contaminated food occur following its preparation by infected handlers. However, contamination can also occur during food production, for instance when contaminated irrigation water is applied agriculturally [134] or when shellfish are harvested from water contaminated with human effluent [135] and consumed following insufficient cooking. These circumstances provide the potential for outbreaks on a worldwide scale due to the transportation of food products internationally [136-138]. Shellfish are particularly susceptible to contamination, as they are filter feeders which concentrate the virus from marine and estuarine waters.

Other foods, ranging from fresh lettuce to frozen strawberries have been reported as potential sources of transmission [139, 140].

1.11 Infection control

Effective infection control requires precise epidemiological information together with accurate pathogen detection ideally close to real time. [141-144].

In a hospital setting, the main aim is to minimise disruption of essential daily services and inconvenience to patients, whilst maximising the organisation's resources. In the past, national guidelines for managing hospital outbreaks recommended ward closure as a key control measure, with re-opening delayed until 72 hours after the last symptomatic case [145]. Recent evidence has suggested that entire ward closure is not always necessary and that efficient control may be achieved by the closure of bays [146]. Current UK PHE guidelines recommend the adoption of a pragmatic approach to norovirus infection control, isolating areas where patients have been symptomatic and carefully cohorting

those who have been exposed (http://www.hpa.org.uk/webc/HPAwebFile/HPAweb_C/1317131639453).

The education of staff and patient's relatives has also been a recent focus of attention (http://www.hpa.org.uk/webc/HPAwebFile/HPAweb_C/1317131639453).

However, despite a few comparative trials investigating the duration of outbreaks with different infection control interventions, no meta-analysis has successfully proven that the newer measures help or hinder outbreak control [147].

There are still many uncertainties about norovirus transmission; namely, the relative importance of contaminated hospital environments [148], asymptomatic carriers [149], and asymptomatic post-infection shedders [150]. The extent to which hospital outbreaks may be caused by repeated introductions from a currently unidentified point source within or outside the hospital, or whether they derive from chains of secondary transmissions is unknown. This issue cannot be investigated using epidemiological data and laboratory diagnosis alone. For example, a recent study employing both real time laboratory detection and rapid genotyping found that a hospital outbreak was in fact due to independent introductions of four different norovirus genotypes as opposed to ongoing transmission of a single virus strain between patients [151]. Such data are essential to evaluate the importance of high impact infection control interventions such as ward closure [151].

Therefore, recording norovirus epidemiology and transmission dynamics, and characterising outbreak-associated strain to the level of the full genome should further

improve our understanding of norovirus transmission by addressing the outstanding transmission questions outlined above with a view to optimising infection control.

1.12 Diagnostics and molecular epidemiology

During the majority of norovirus outbreaks, highly sensitive and specific molecular detection methods are unlikely to be available in real time. Hence, the 1982 Kaplan criteria are frequently used by healthcare professionals worldwide [152] to identify likely outbreaks. The following conditions should be met if an outbreak is caused by norovirus:

1. A mean (or median) illness duration of 12 to 60 hours,
2. A mean (or median) incubation period of 24 to 48 hours,
3. More than 50% of people with vomiting, and
4. No bacterial agent found.

These criteria have been associated with a sensitivity of ~70% and a specificity of up to 99% [153]. The low sensitivity means norovirus cannot be excluded as the aetiological agent even if an outbreak fails to meet the criteria and suspicion of a viral cause therefore remains.

The mainstay of norovirus in vitro diagnostics for more specialist and reference laboratories in recent years has been reverse transcriptase polymerase chain reaction (RT-PCR) assays which target the RNA polymerase region of the genome (RdRP) due to the relatively high level of nucleotide sequence conservation in this region [28]. An alternative region targeted in the same way has been the P2 capsid region of VP1 within ORF2; a highly genetically diverse sequence flanked by conserved sequences used in oligonucleotide primer design [154]. Routine clinical laboratories more frequently send

suspected samples to regional reference laboratories or use commercial enzyme immunoassays (EIA). EIA and transmission electron microscopy demonstrate high specificity but lower sensitivities against RT-PCR gold standard testing [155].

1.13 Impact on healthcare

Norovirus infections are a significant economic and logistical burden to healthcare systems, as well as extremely unpleasant for the patients concerned. In the UK, norovirus was estimated to cost the NHS up to £110 million per annum in years of high incidence (2002-2003) [112]. Furthermore, the average cost of a single laboratory confirmed nosocomial infection in the USA is estimated to be over USD $15,000. A 2007 norovirus outbreak at Johns Hopkins Hospital, (946 beds), cost an estimated USD $650,000. According to Public Health England, approximately 3,000 people a year are admitted to hospital with norovirus in England. The seasonality of laboratory confirmed cases of norovirus including healthcare and community infections is shown (figure 1.8 and 1.9). Between July 2012 to July 2013, 1,459 outbreaks were reported within England leading to the closure of 1,311 (90%) bays or wards (from Public Health England Hospital Norovirus Outbreak reporting System (HNORS) page 2: http://www.hpa.org.uk/webc/HPAwebFile/HPAweb_C/1287143931777).

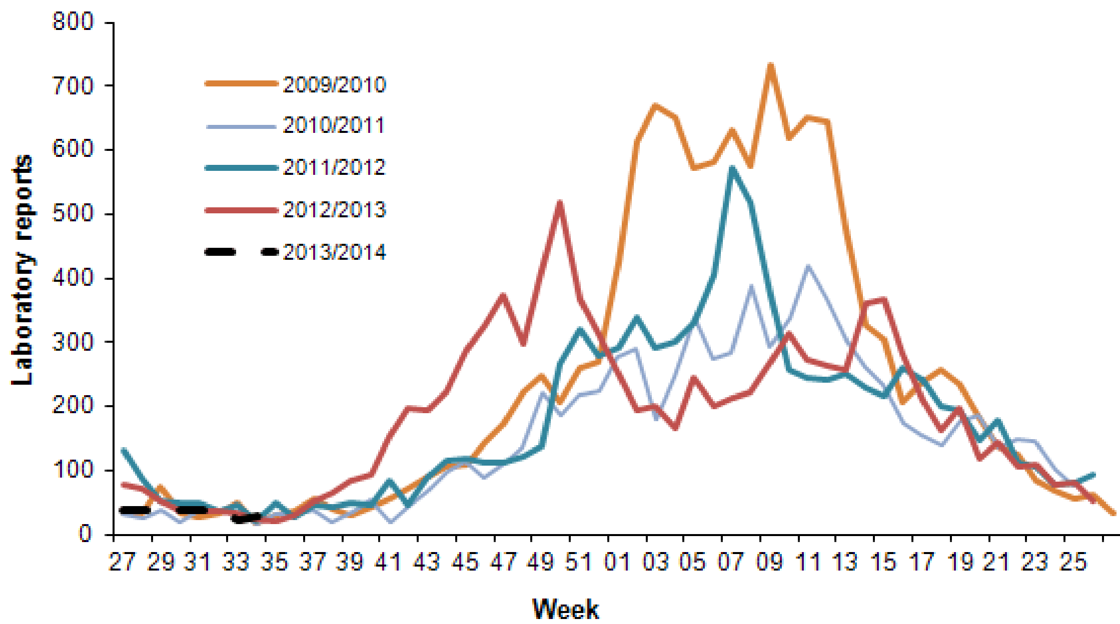
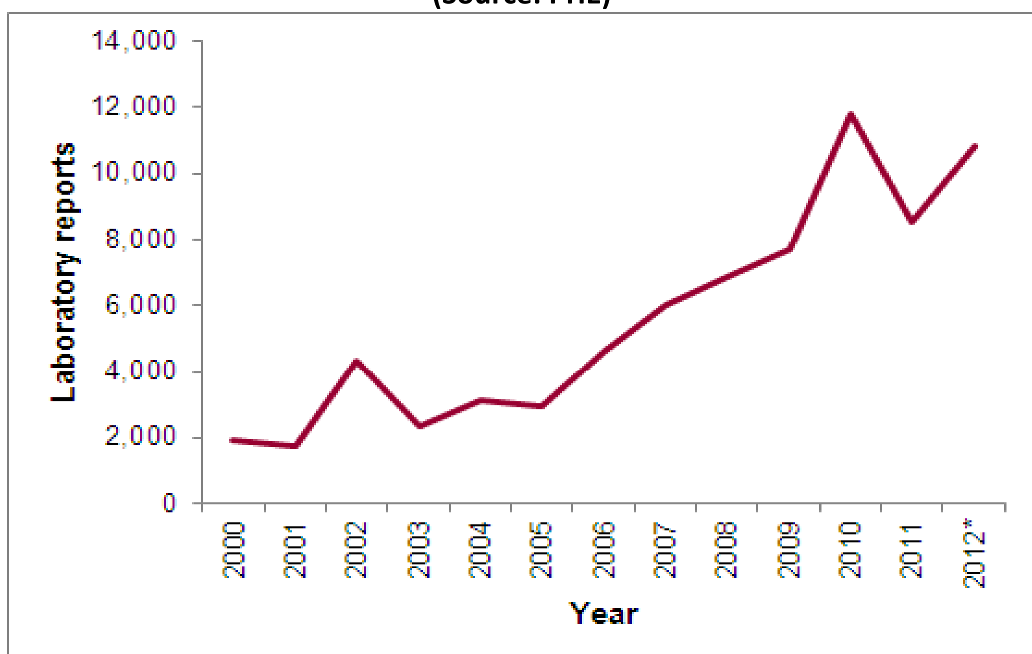


Figure 1.8. Seasonal comparison of laboratory reports of norovirus (England and Wales). Source: PHE

Using a stochastic susceptible exposed infectious recovered (SEIR) compartment model to stimulate the economics affected of various norovirus outbreaks from the hospital perspective, Lee et al., (2011) predicted that detecting and containing a norovirus outbreak in a 10-bed ward on day 3 instead of day 5 could save USD $11,623 [156].

Figure 1.9. Laboratory reports of norovirus infection in England and Wales 2000-2012 (Source: PHE)



1.14 Clinical Management

Treatment is generally conservative during norovirus infection, with dehydration being the classical complication that needs attention. Oral rehydration solutions providing essential electrolyte replacement may be first line. Parental and intravenous electrolyte replacement may be required in those with severe dehydration. Although few people die from norovirus in the developed world, economic costs of outbreaks in the hospital setting are high [156], and the development of a norovirus vaccine would reduce economic costs.

1.15 Advances in DNA Sequencing Technology

Although relatively recent technical advances in the molecular detection of human pathogens, such as nucleic acid amplification and mass spectrometry have been implemented in many routine clinical microbiology laboratories, many of the techniques which remain in general use would be familiar to a 19th century microbiologist. The advent of next generation DNA sequencing has facilitated the rapid sequencing of whole pathogen genomes [144, 157], particularly in the field of bacteriology. For technical reasons, including the inability of some viruses to grow in culture, the high throughput sequencing of viral genomes has lagged behind, despite generally being smaller in genome size. Next generation DNA sequencing has been used largely in the research laboratory thus far; its application in the clinical laboratory is highly likely to occur in the next few years [158]. This will facilitate the simultaneous detection and characterisation of pathogens to the genome level, together with molecular epidemiological surveys in real time. The potential benefits of this technology for patients and infection control in both hospitals and communities are therefore apparent [159, 160].

1.16 First generation DNA sequencing

Sanger pioneered the original concept for DNA sequencing in 1975 and later that year published a rapid method for determining sequences in DNA by primed synthesis with DNA polymerase [161, 162]. Two years later, Sanger again published a landmark paper on DNA sequencing based on the chain terminating dideoxynucleotide analogues [163]. Accompanied with Maxam and Gilbert's chemical degradation DNA sequencing technique in which terminally labelled DNA fragments were chemically cleaved at specific bases and separated by gel electrophoresis [164], these two laboratories initiated the birth of the first automated DNA sequencer, made by Caltech [165], and later commercialised by Applied Biosystems (ABI), the European Molecular Biology Laboratory (EMBL) and Phamacia-Amersham [166]. This later commercialisation has led to the worldwide generalisation of Sanger sequencing, or capillary electrophoresis sequencing.

1.17 Next generation sequencing

In 2000, the company 454 Life Sciences was created which went on to develop the first commercially available next generation sequencing (NGS) platform, known as the GS20 in 2005 [167]. This platform employs a technology known as pyrosequencing which is based on the "sequencing by synthesis" approach [167]. This differs from traditional Sanger sequencing, which is based on the eletrophoretic separation and detection of DNA fragments terminated by the incorporation of dideoxynucleotide analogues [163]. In contrast, pyrosequencing relies on the detection of pyrophosphate released on nucleotide incorporation (Figure 1.8a-c). The pyrosequencing DNA synthesis reaction mixture includes the enzymes *ATP sulfurylase* and *luciferase* together with the substrates adenosine 5' phosphosulfate and luciferin optimised such that the release of a

pyrophosphate group upon addition of a nucleotide, results in the production of detectable light.

During 2008, an alternative NGS technology was released by Illumina [168] (Figure 1.8d-h). This relies on a process termed "bridge amplification". Successive rounds of PCR result in the generation of tiny islands or clusters of amplified molecules (Figure 1.8h), which serve as clones for subsequent sequencing using chain termination, similar to traditional Sanger sequencing. Unlike Sanger sequencing, however Illumina sequencing uses fluorescently labelled *reversible* terminators, which stop the reaction only temporarily upon single base incorporation. The reaction occurs again after a high resolution digital image is taken. Two platforms comprising a large, high throughput platform (HiSeq) and a lower throughput benchtop platform (MiSeq) are commercially available (http://www.illumina.com/systems/hiseq_2500_1500.ilmn and http://www.illumina.com/systems/miseq.ilmn).

Differences between the two include: (1) read length (HiSeq up to 150bp versus MiSeq up to 300bp); (2) run time (HiSeq up to 11 days versus MiSeq up to 48 hours) and (3) amount of output data (HiSeq up to 600 gigabases per run, and MiSeq up to 15 Gb) (http://www.illumina.com/systems/sequencing.ilmn).

Other platforms that have been used in next generation sequencing, such as Ion Torrent (Life Technologies, CA, USA), SOLiD (Applied Biosystems, CA, USA), and PacBio (Pacific Biosciences, CA, USA) will not be described in detail here but their general principles have been illustrated in Figure 1.10 and Table 1.3.

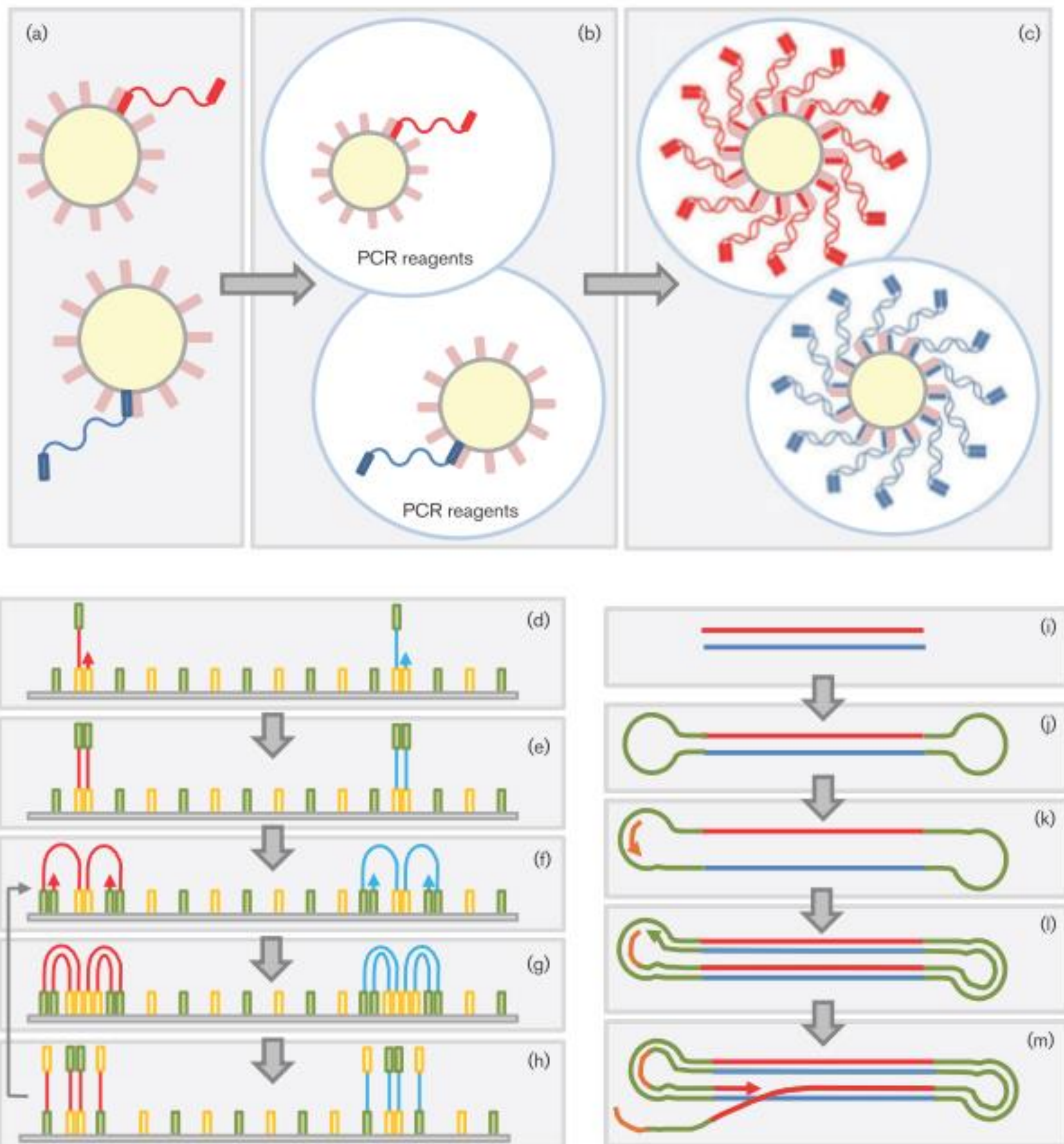


Figure 1.10. General principles of template amplification. (a–c) Emulsion PCR (Roche 454, SOLiD and Ion Torrent). (a) Adapters are used to capture single molecules of template onto microbeads by primer hybridization. (b) Beads are incorporated into a carefully controlled emulsion, in which each bubble constitutes a microreactor containing DNA template, primer and reagents for PCR. (c) Following amplification, each bead is coated with clonally amplified molecules. (d–h) Bridge amplification (Illumina). (d) Single-stranded template annealed to a glass plate by hybridization to a complementary primer. (e) The primer forms the basis for extension. (f) The free end of each single-stranded molecule can anneal to a second anchored primer in close spatial proximity, forming a 'bridge' that acts as a template for (g) a second round of amplification. This results in (h) four linear molecules. Stages (f)–(h) are essentially repeated to generate clonally amplified islands or clusters for subsequent sequencing. (i–m) Linear amplification (PacBio). (i) Template dsDNA. (j) Bound hairpin adapters create a single-stranded circular template. (k) Binding of a primer complementary to hairpin sequence. (l–m) Linear amplification and strand displacement create a single strand of DNA containing multiple copies of plus- and minus-strand sequences that serves as template for sequencing.

The potential output, in terms of the amount of nucleotide sequence data generated, range from approximately five hundred million bases per run (Roche 454) to billions of bases per run (Illumina, SOLiD). The costs of these technologies has declined rapidly, hence they have become affordable for even smaller research and clinical laboratories (Figure 1.11) (NHRGI). Other next generation sequencers are available, but less widely used; their specifications are described briefly (Table 1.3).

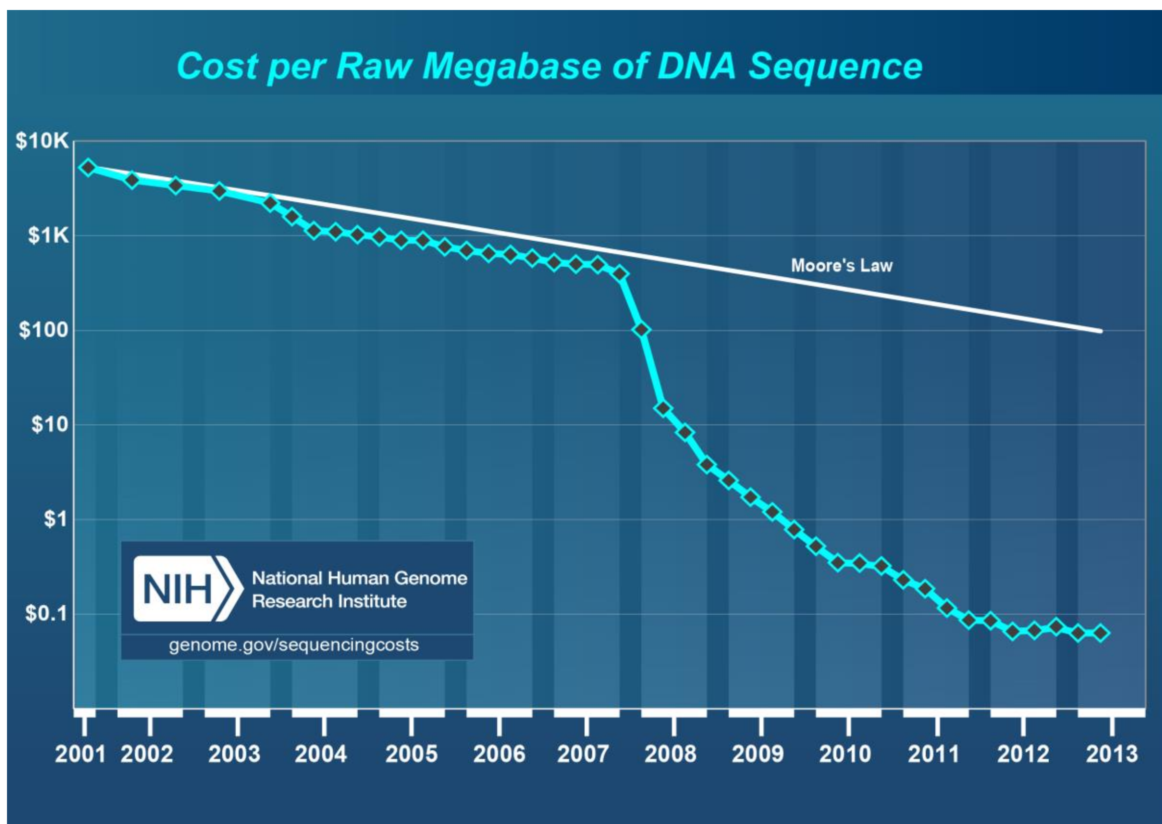


Figure 1.11. Costs associated with DNA sequencing . Moore's Law describes a long term trend in the computer hardware industry that involves the doubling of "compute power" every two years. Technology improvements that keep up with Moore's Law are widely regarded to be doing exceedingly well, making it useful for comparison. ***Figure used with permission from NHGRI***

Table 1.3. Summary of next-generation DNA sequencing technologies: methods used in sample preparation, molecule separation and sequencing, and advertised outputs

First step for all methods is template fragmentation.

Method	Adapter type	Amplification?	Separation	Sequencing chemistry	Approximate read length (bases)#	Approximate maximum amount of data per run
Roche 454*	Adapters	Emulsion PCR	Microbeads and 'picotitre' plate	Pyrosequencing	400–700	700 Mb
SOLiD	Adapters	Emulsion PCR	Beads on glass slide	Ligation	50–75	20 Gb
Illumina*	Adapters	Bridge amplification in situ	Glass slide hybridization	Reversible terminators	25–500	600 Gb
PacBio	Hairpin adapters	Linear amplification	Captured by DNA polymerase in microcell	Fluorescently labelled dNTPs	1,000	Not available
Ion Torrent*	Adapters	Emulsion PCR	Ion Spheres and high-density array	Detection of released H^+	35–400	1 Gb

*These technologies are available on platforms with different scales of throughput. Adapted from Loman et al., 2012 [169]

#Approximate values based on data published on the companies' websites on 9 March 2012. These data are for guidance only and are subject to change; readers interested in the details should consult either the manufacturers or those that are offering the sequencing service.

1.18 Using NGS in clinical virology

Viruses exist in their host as highly dynamic entities characterised by a rapid turnover, high mutations rate and large population size [92]. As a consequence, novel variants constantly evolve, creating a pool of genetic diversity on which natural selection can act. NGS provides a number of new opportunities with respect to the study of viral genomes, which are summarised below.

1.18.1 Whole viral genome reconstruction

Compared to bacteria, virus genomes are compact. However, high throughput sequencing of whole virus genomes using established methods such as PCR amplification and Sanger sequencing has been marred by cost, labour and high failure rates [159]. Large numbers of whole genome sequences are imperative for basic research projects aiming to understand virus evolution and molecular epidemiology. An example is influenza virus, where the requirement for genome sequences has been driven by the emergence of H5N1 and an H1N1 pandemic [170]. In addition influenza viruses can re-assort or recombine, meaning that phylogenetic analyses based on partial sequence may never pick up the full complexity of historical recombination events that often lead to pandemic emergence.

1.18.2 Virus discovery and metagenomics

The rate of discovery of new viruses capable of infecting humans and animals has been constant over the past few decades. However, many viruses are likely to remain undiscovered [171]. There is a need to rapidly characterise newly discovered viral pathogens to understand their pathogenesis, evolution and epidemiology, as well as to identify effective measures of control. An example is the SARS coronavirus, where

collaborative efforts after the syndrome was originally described, allowed the novel virus to be described in only one month [172, 173]. It is believed that NGS will in the future allow this type of research to be performed in a fraction of the time. We have also seen this in the bacterial world, when the 0104 EHEC *E.coli* was sequenced using next generation sequencing technology within a very short space of time [174]. Metagenomics is defined as the characterisation of total genetic information directly from a sample. Unlike bacterial genomics which require previous culture, the ability to sequence all pathogen (or even all microbial) nucleic acid from the clinical sample in question is attractive because of its speed and totality, although the large amount of data generated will need to undergo careful filtering [175].

1.18.3 Characterisation of intra host variability

One of the advantages of NGS over Sanger sequencing is the potential for increased depth of coverage. As described above, viruses have high evolutionary rates and the diversity of closely related variants within a single samples cannot be determined using Sanger sequencing which determines the majority, or consensus sequence for these populations, while the minority variants not determined [176]. In evolutionary terms, minority variants provide the raw material for selection of variants best adapted to the prevailing environmental conditions. NGS has already been used to monitor population diversity in HIV [177], Hepatitis B [178] and other viruses [179]. Varghese et al., [180] have used NGS in HIV to potentially dissect the viral bottleneck found within host and attempt to find the differences between resistance acquired by transmission and that which may have evolved within the patient. Some studies have already moved away

from sequencing specific genomic regions to estimating the diversity across the whole genome [181-183].

The challenges associated with purifying viral RNA from heterogeneous samples such as faeces has severely limited the application of high-throughput sequencing [8]. Unlike whole genome sequencing of bacteria [157, 184, 185], which is now well developed, sequencing of viruses has not yet benefitted from dramatic advances in sequencing capacity afforded by NGS.

Successful approaches for sequencing RNA viruses have previously been dependent on template-specific primer-based amplification overlapping genome fragments [186]. These amplicons were then sequenced either by the Sanger method, or by NGS. Recent examples of this approach include studies of Norovirus evolution and HCV diversity [187]. Next-generation sequencing of amplicons and partial genome fragments from RNA viruses have been used to investigate variants within populations of Human Immunodeficiency Virus and HCV infecting individual patients [188, 189]. An enhancement of the amplification approach employs sample enrichment for target viral sequences incorporating primers as "bait" to capture larger genomic fragments [190]. All these approaches are expensive, labour-intensive, slow and inflexible, and generally require *a priori* knowledge of partial or closely related virus sequences, and multiple and complex sets of oligonucleotide primers for different virus strains [191]. In turn, prior assumptions about the sequences present can bias resulting data, altering the representation of the virus population present in a sample [182].

1.19 Study Aims

The aims of this study were firstly to adapt a range of NGS approaches to norovirus in order to identify the most efficient and cost-effective method of generating a large number of whole virus genomes representing the predominant GII.4 strain (chapters 2 and 3). Then to use both conventional epidemiological data and the whole GII.4 genome data to investigate norovirus transmission dynamics and intra-genotype evolution on a number of different levels ranging from individual patients and wards, to hospital, community and national scales (chapters 4 to 6). These aims should provide a rational basis for the guidance issued to hospitals for the control of norovirus outbreaks.

Chapter 2. Methods

Chapter objective

To describe and define the sampling frame used for collection of samples. To describe the different methods attempted to produce whole genomic sequences of norovirus.

2.1 Sampling frame and collection of data

2.1.1 Ethical considerations

These studies were conducted in compliance with the protocol described below, the Data Protection Act (DPA number: Z5886415), and NHS research governance. The Modernising Medical Microbiology study protocol version 1.0 was approved by the Berkshire Research Ethics Committee on the 1st October 2010.

2.1.2 Clinical samples and epidemiological data

All epidemiological data and clinical samples were collected between September 2009 and August 2013. The study was prospective and observational, and it took place at three NHS Trusts with additional clinical samples source from a regional PHE reference laboratory:

(1) Oxford University Hospitals NHS Trust serving a population of 650,000. The trust includes four different hospitals within Oxfordshire County and provides a total of 1,300 beds;

(2) Leeds Teaching Hospitals NHS Trust serving a population of 1.8 million, and which includes 3 hospitals within the region and provides a total of 1,900 beds;

(3) Brighton and Sussex University Hospitals NHS Trust serving a population of 200,000 and providing 800 beds. Their laboratory also act as a regional laboratory for the Surrey and Sussex Healthcare NHS Trust.

(4) Southampton Public Health Laboratories Virology Reference Laboratory, serves as a reference microbiology and virology laboratory for the South East of England providing them with a regional norovirus diagnostic testing service.

Epidemiological data, varying in detail according to the originating laboratory, were recorded and clinical samples were collected by clinical or primary health care teams managing patients with gastroenteritis illness. In a non-outbreak situation, information and stool samples were collected routinely, for example to facilitate contact tracing or to fulfil mandatory obligations for individual case investigations. During outbreak situations, in addition to the above, further data and samples were collected for clinical surveillance and to inform infection control staff. Faecal samples were submitted to the NHS Trust or PHE microbiology laboratories indicated above for norovirus diagnostic testing.

Only norovirus positive clinical samples and a limited set of epidemiological data were obtained from non-Oxford NHS trusts (from 2011 to 2013, diagnostically RT-PCR tested at own microbiology laboratory). The PHE Southampton reference laboratory provided pseudo-anonymised clinical samples without associated epidemiological data. The only information recorded was the geographical location (hospital name) in which the sample was collected and the date. In some cases, multiple longitudinal samples were received from a single patient. Faecal samples were stored at 4°C and disposed of by autoclaving at the conclusion of the study or kept long term.

2.1.3 Definition of a hospital outbreak

The following Public Health England definitions were used:

Criteria for a suspected case of norovirus:

a) Vomiting: two or more episodes of vomiting of suspected infectious cause* occurring in a 24 hour period;

b) Diarrhoea: two or more loose stools in a 24 hour period*;

c) Diarrhoea and vomiting: one or more episodes of both symptoms occurring within a 24 hour period *;

**not associated with prescribed drugs or treatments and not associated with reaction to anaesthetic or an underlying medical condition or existing illness*.

A confirmed case of norovirus:

a, b or c above with microbiological confirmation (via Enzyme Immunoassay (EIA) or Reverse Transcriptase Polymerase Chain Reaction (RT-PCR)

Norovirus outbreaks:

Suspected outbreak: two or more cases, as defined above, occurring in a functional care unit within the hospital without laboratory confirmation.

Confirmed outbreak: as above with laboratory confirmation

Outbreaks are considered to be over if no new cases arise within 7 days of the last case being reported symptom free.

2.1.4 Collection of Epidemiological Data

A template was used for the recording of detailed epidemiological data for Oxford inpatients. Data included:

- Patient;
- Age of patient;
- Gender;
- Location of patient including bed number (including details of any movements);
- Date of ward admission;
- Brief medical history;
- Medications (including antibiotics and laxatives);
- Date of onset of symptoms and their duration;
- Number of episodes of diarrhoea/vomiting per day during onset

Clinical samples collected outside of Oxford were assigned a unique study number and anonymised, patient names and addresses were not recorded.

2.2 RNA extraction

Total RNA from faecal samples was isolated using the Fujifilm Quickgene DNA tissue kit SII under the manufacturer's RNA extraction from stool protocol for the Fujifilm Quickgene Mini-80 nucleic acid isolation system (Fujifilm Corp., Tokyo, Japan). Three hundred microlitres of supernatant from a 10% clarified emulsion of faeces centrifuged at 13,200 rpm for 10 minutes was used to prepare the lysate. Fifty microlitres of RNA was eluted from the Mini-80 device. The resulting RNA samples were quantified by Nanodrop 1000 Spectrophotometer (Thermo Fisher Scientific, MA, USA) to estimate concentrations and to check its purity using the ratio of absorbance at 260 and 280nm.

2.3 Detection of norovirus in clinical samples

2.3.1 Non-quantitative RT-PCR

A non-quantitative RT-PCR assay [192] was adopted for the detection of norovirus in clinical samples. This assay was used to confirm the presence of norovirus in stool samples, which had been previously designated positive by RidaQuick EIA or stored and untested in the clinical microbiology laboratories at the John Radcliffe Hospital. This RT-PCR assay was chosen because it performed well when a different panel of assays were compared [193] since it employs oligonucleotide primers (JV12Y, 5'-ATACCACTATGATGCAGAYTA-3', and JV13I, 5'-TCATCATCACCATAGAAIGAG-3'), containing a level of degeneracy, designed to facilitate binding to relatively conserved sequences of the RdRP and to allow the amplification of a 327 bp sequence.

Extracted total stool RNA was reverse transcribed in a 20μl reaction mixture containing 11.5μl of RNA and 0.5μl (250ng) of random hexamers (Promega, UK). The RNA and primers were denatured at 70°C for 5 minutes, followed by immediate transfer to ice, on which the remainder of the reaction was assembled: 4μl of first strand cDNA synthesis buffer (Invitrogen, UK), 2μl of 0.1M dithiothreitol (Invitrogen, UK), 1μl of deoxynuceloside triphosphates (10mM) (Invitrogen), and 1μl (200U) of Moloney murine leukameia virus reverse transcriptase (Invitrogen, UK). Incubation was performed at 37°C for 60 minutes.

Each 50-μl PCR was assembled on ice and comprised 5 μl of cDNA, 36.75μl of water, 1μl (0.2 μM) of each PCR primer, 5μl of 10x HotStarTaq PCR buffer (Qiagen), 1μl (0.2μM) deoxynucleoside triphosphates, and 0.25μl (2.5U) of HotStarTaq DNA polymerase

(Qiagen). The reaction conditions were denaturation and DNA polymerase activation at 95°C for 15 min, followed by 35 cycles of 94°C for 30s, 37°C for 40s, 72°C for 40 sec, and a final extension cycle at 72°C for 5 minutes. The amplification products were then visualised using a 2% agarose gel (Life Technologies, USA) stained with ethidium bromide (Life technologies, USA) and run at 110 volts for 40 minutes.

2.3.2 Genotyping by Sanger sequencing

The identity of the 327 bp amplification products was confirmed by Sanger nucleotide sequencing. A BLAST search of each sequence against GenBank (http://blast.ncbi.nlm.nih.gov/Blast.cgi) confirmed the presence of norovirus and identified the genotype.

PCR amplification products were purified by precipitation with 60µl of 20% polyethylene glycol (molecular weight 8,000) and 2.5 M sodium acetate, and then incubated at room temperature for 1 hr. DNA was then pelleted by centrifugation at 2,750rpm for 1 hr. The resultant pellet was washed twice with 150µl of 70% ethanol and air dried for 30 minutes. The purified DNA was then resuspended in purified water dependent on its intensity on the original gel electrophoresis. 25µl of water was added for strong bands, 15µl for medium bands, and 10µl for weak bands. Their nucleotide sequences were determined at least two times on each DNA strand using the amplification primers (JV12Y and JV13I) [188] and BigDye v3.1 Ready Reaction Mix (Applied Biosystems, Warrington, United Kingdom). Each reaction mixture consisted of: 0.25µl BigDye, 1.875µl of 5 x BigDye Sequencing Buffer, 1.875µl water, 4µl of each primer (0.6µM), and 2µl of DNA. The reaction conditions were 30 cycles of denaturation at 96°C for 10s, annealing at 50°C for 5s, and extension at 60°C for 2 min.

Unincorporated dye terminators were removed by precipitation of the termination products with 2 volumes of 100% ethanol and 1/10 volume of 3 M sodium acetate, pH 5.2, incubation at room temperature for 1 hour and centrifugation at 2,750rpm for 1 hour. The resultant DNA pellet was washed with 70% ethanol and dried. The reaction products were re-suspended in formamide buffer and analysed using the Prism 3730 automated DNA sequencer (Applied Biosystems, UK). Sequences were assembled from the resultant chromatograms using the STADEN suite of computer programs [194]. This method of genotyping was ultimately superseded by NGS.

2.3.3 Quantitative RT-PCR

A further RT-PCR assay was adopted to allow quantitation of the norovirus load present in clinical samples. A published assay was used which employed Taqman probes and primers detecting the ORF1/ORF2 junction [37] (Genogroup 1 (5279-5371nt): COG1F, 5'-CGYTGGATGCGNTTYCATGA-3' and COG1R, 5'-CTTAGACGCCATCATCATTYAC-3', RING1A 5'-[FAM]-AGATYGCGATCYCCTGTCCA-[BHQ1]-3', RING1B 5'-[FAM]-AGATCGCGGTCTCCTGTCCA-[BHQ1]-3'; Genogroup 2 (5358-5381nt): COG2F, 5'-CARGARBCNATGTTYAGRTGGATGAG-3', COG2R, 5'-TCGACGCCATCTTCATTCACA-3', RING2, 5'-[FAM]-TGGGAGGGCGATCGCAATCT-[BHQ1]-3'). Positive controls and standards for real time PCR utilising these primers were DNA plasmids containing cloned GI and GII capsid gene fragments (pJPGI and pJPGII) gifted from Dr Kazuhiko Katayama [37] Department of Virus Diseases and Vaccine Control, National Institute of Infectious Diseases, Murashi-murayama, Tokyo, Japan).

2.4 Whole genome norovirus sequence determination via Sanger sequencing

The determination of whole virus genome sequences by Sanger sequencing is dependent on prior knowledge of a closely related genome because it relies upon oligonucleotide primer dependent RT-PCR. Multiple overlapping fragments of the norovirus genome must be amplified by PCR because a single Sanger sequencing reaction can determine only 700-900nt. The viral genotype must therefore be determined prior to Sanger sequencing to enable appropriate oligonucleotide primers to be used. In the present study, these were designed using a reference sequence Norovirus Hu/GII.4/Orange/NSW001P/2008/AU (Genbank accession number GQ845367), chosen because it represents a recent variant of the predominant GII.4 genotype. The oligonucleotide primers used are listed in Appendix 1. Sanger sequencing was therefore restricted to noroviruses of the GII.4 genotype.

2.4.1 Sanger sequencing of seven overlapping DNA fragments

Reverse transcription and first strand cDNA synthesis were performed using the Accuscript High Fidelity 1st strand cDNA synthesis kit (Agilent, Santa Clara, California, USA) following the manufacturer’s published protocol.

PCR amplification was then performed using the following reaction composition: each 50µl PCR was assembled on ice and comprised 2 µl of cDNA, 31.5µl of water, 2.5µl (0.2 µM) of each PCR primer (see appendix 1), 0.5µl of Phusion High Fidelity DNA polymerase (New England Biolabs, USA), 1µl (10mM) deoxynucleoside triphosphates, and 10µl of 5x Phusion HF buffer (New England Biolabs, USA). The reaction conditions were denaturation and DNA polymerase activation at 98°C for 1 min, followed by 35

cycles of 98°C for 10s, 64°C for 30s, 72ºC for 30 sec, and a final extension cycle at 72°C for 5 minutes. Additional oligonucleotide primers used for sequencing only (rather than amplification and sequencing) were designed to bind within PCR amplicons, the aim being to increase the depth, and therefore the reliability of the sequence data obtained. The size and presence of a single DNA species within the amplification reaction products was confirmed by 1% agarose gel electrophoresis. PCR amplicons were purified by precipitation and resuspended in aqueous solution. The re-suspension volume was dependent on the yield of the PCR as estimated from the agarose gel. The sequencing reaction products were purified, prepared with a similar protocol to that described in section 2.3.2, except for an adjusted annealing temperature of 55ºC and analysed with a Prism 3730 automated DNA sequencer (Applied Biosystems, Foster City, CA, USA). The quality of the resultant chromatograms was inspected and consensus sequences were assembled using the STADEN suite of computer programs [194]. The sequences at the 5' and 3' termini of the overlapping PCR amplicons were oligonucleotide primer dependent and were therefore excluded from the final consensus.

2.5 Template independent determination of norovirus genome sequences

Both the Roche 454 and the Illumina MiSeq and HiSeq sequencers were evaluated with the aim of ultimately generating norovirus WGS directly from total stool cDNA.

2.5.1 Norovirus sequencing via 454 pyrosequencing

454 pyrosequencing contrasts with Sanger sequencing in that it requires no prior knowledge of the virus genome sequence. However, since norovirus cannot be cultured, RT-PCR was still used to generate the required quantity of sufficiently pure norovirus

double stranded cDNA. This was obtained using only two overlapping PCR amplicons (4.3kb and 3.4kb respectively) spanning the entire genome (Figure 2.1).

To obtain the two overlapping PCR amplicons, total RNA of high quality in terms of both length and purity was extracted from stool using both Vertrel XF (DuPont, USA); a freon subsitute used in viral purification methods to separate virus from infected cell debris and Trizol LS reagent (Life Technologies, USA), a monophasic solution of phenol, guanidine isothiocyanate designed to isolate high quality total RNA. The use of these compounds have been used previously for the extraction of viruses [195]. This differed to the protocol described in chapter 2.2. In addition, the Purelink RNA Mini Kit (Life technologies, USA) was used to isolate total RNA following the manufacturer's published protocol. A large volume of faeces was required (3ml) for the downstream amplification to be successful. Briefly, 3ml of faeces was vortexed with 7ml of phosphate buffered saline. Ten millilitres of Vertrel XF was then added and vortexed for 3 minutes before this mixture was centrifuged for 15 minutes at 3,220rpm. The upper aqueous layer was removed carefully and filtered through Amicon Ultra-4 Centrifugal Filter units at 3,220rpm for 15 minutes. 0.25ml of the filtrate was then added to 0.75ml of Trizol LS and vortexed for 1 minute to completely dissociate the nucleoprotein complexes. Two hundred microlitres of chloroform was then added and shaken for 15 seconds. The mixture was then centrifuged at for 15 minutes at 12,000rpm at 4°C. Four hundred microlitres of the aqueous layer was removed and cleaned with 400ml of 70% ethanol. Seven hundred microlitres of the aqueous layer was then added to the Purelink cartridges and centrifuged under the manufacturer's protocol (12,000rpm for 15 seconds, 700ml wash buffer 1 added and spun 12,000rpm for 15 seconds, followed by

500µl of wash buffer 2 added and re-spun at 12,000rpm for 15 seconds, before a final spin at 1,000rpm for 1 minute). The RNA was reverse transcribed using the kit described above (Chapter 2.4.1). The primers used for this approach were:

Name	Primer sequence	Polarity	Binding site (nt)
GII5'2	5’-GTGAATGAAGATGGCGTC-3’	+	1
JV13I	5’-TCATCATCACCATAGAAIGAG-3’	-	4,586
GII4pol_1	5'-GTGCCCTGCACCTCCCAATG-3'	+	4130
NVT30	5’-TTTTTTTTTTTTTTTTTTTTTTTTTTTTTT-3’	-	7550

The cDNA was used as the template for the two PCRs. The amplification products from up to 10 PCRs were pooled to obtain the 200ng of pure norovirus ds cDNA required for pyrosequencing. The challenge was that not all faecal material received had this large amount of faecal material provided. Consequently, a large number of samples were deemed insufficient in quantity for full genome sequencing.

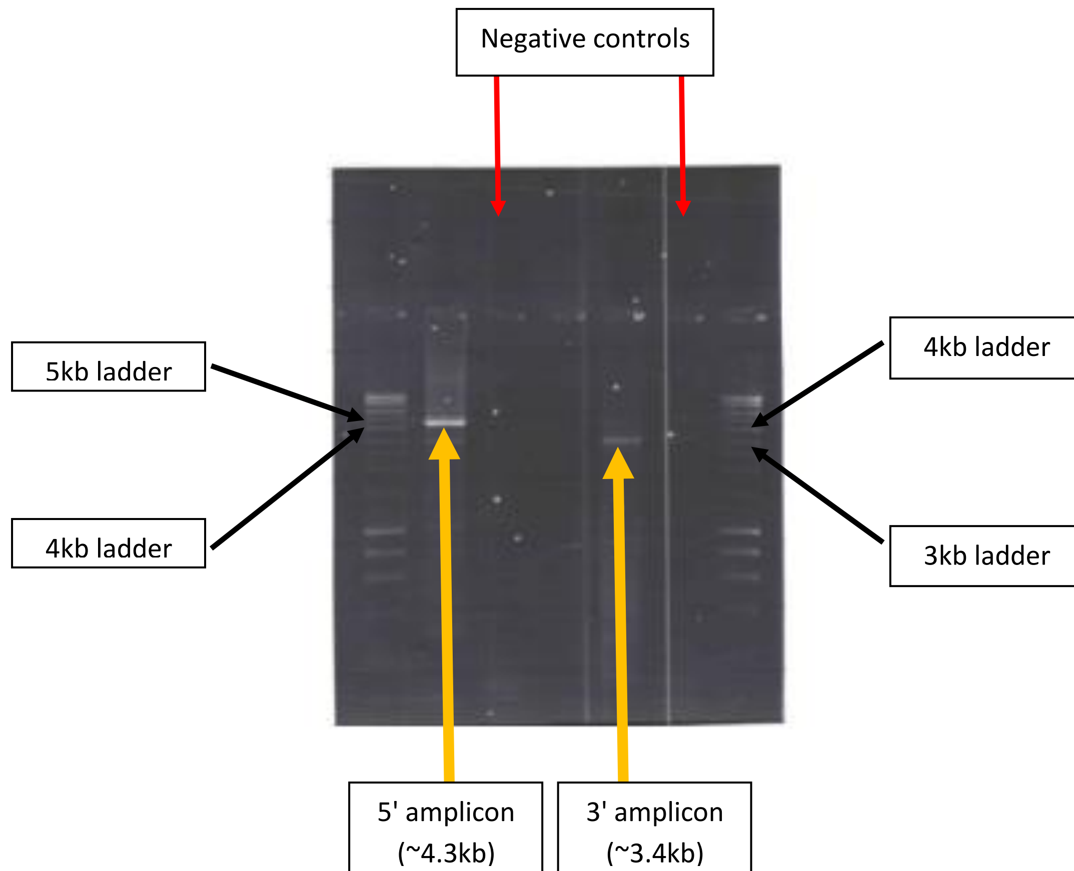


Figure 2.1. Example of an agarose gel electophoresis (1% agarose) revealing both the larger 4kb and smaller 3 kb PCR amplicons

Pooled PCR products were prepared for sequencing on the Roche 454 Genome Sequencer FLX TItanium (Roche) using the Roche GS FLX Titanium 454 Rapid library preparation kit and following standard protocols without modifications. The workflow for this kit are summarised below in Figure 2.2. Further details of the procedural protocol can be downloaded from the Roche 454 Rapid Library Preparation Method Manual available here:

www.454.com/downloads/my454/documentation/gs-flx-plus/Rapid-Library-Preparation-Method-Manual_XLPlus_May2011.pdf

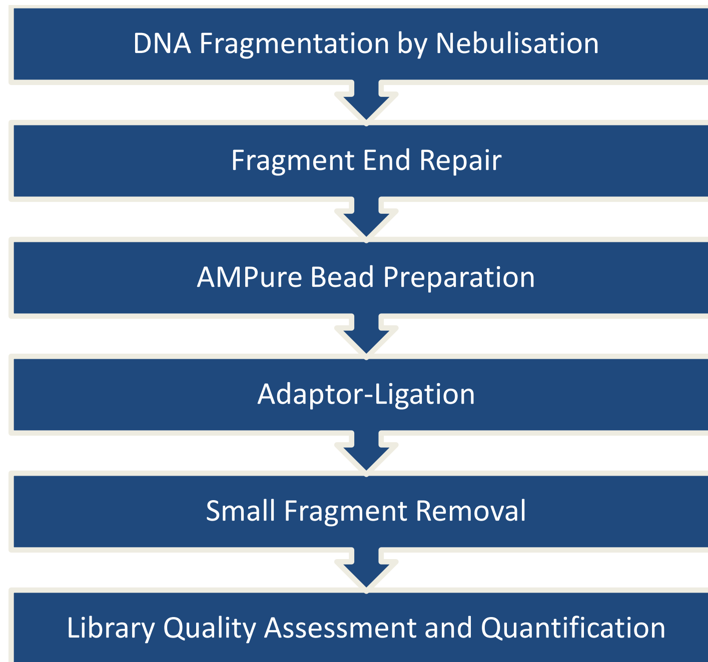


Figure 2.2. Workflow of the Rapid Library Preparation. Taken from Roche Rapid Library Preparation Manual (http://454.com/downloads/my454/documentation/gs-flx-plus/Rapid-Library-Preparation-Method-Manual_XLPlus_May2011.pdf)

The samples were barcoded using Rapid library barcoded adaptors (multiplex identifier or MID) during the adaptor ligation step [196]. Libraries ranging in concentration from to $3.64x10^8$ molecules/μl to $1.79x10^9$ molecules/μl were generated using this kit.

Prior to adding the samples to the sequencer, the DNA library samples underwent emulsion based clonal amplification (emPCR amplification). This was performed according to manufacturer's protocols and involves seven key steps as outlined below in Figure 2.3. 3μl of a $1x10^7$ molecule/μl dilution was used for emPCR amplification using a MV emPCR kit. The emPCR products were sequenced on one lane of a four lane gasket using a GS Titanium Sequencing Kit XLR70.

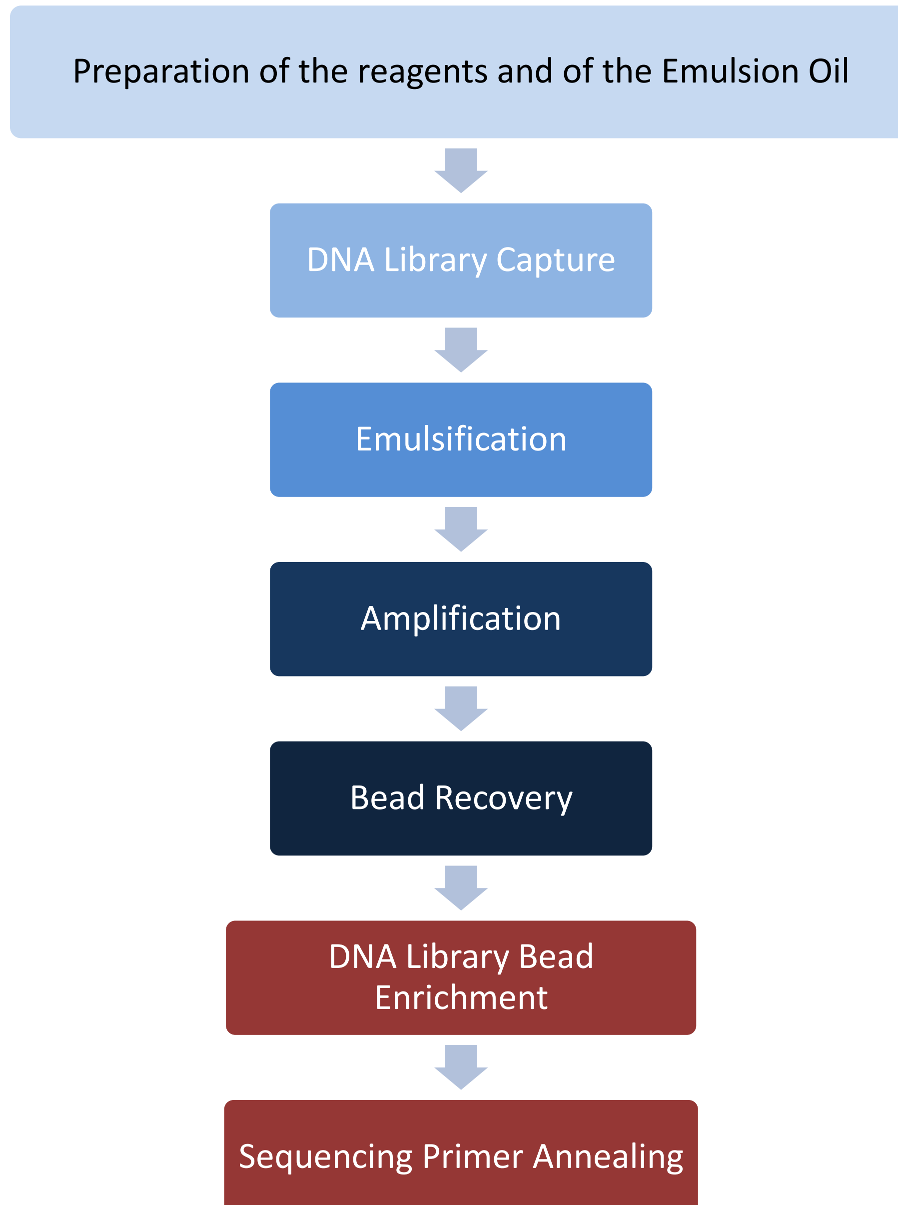


Figure 2.3. Workflow of the emulsion based clonal amplification (emPCR) method. Taken from Roche emPCR Amplification Method Manual (http://454.com/downloads/my454/documentation/gs-junior/method-manuals/GSJunior_emPCR_Lib-A_RevApril2011.pdf)

2.6 Norovirus sequencing via Illumina platform

2.6.1 Illumina RNA-Seq Library Preparation

Illumina sequencing also requires no prior knowledge of the genomes to be determined. In this case, PCR amplification was also not required so the method was truly primer independent.

Total RNA was extracted as described in Chapter 2.2. Its quantity and integrity were assessed using Quant-IT RiboGreen RNA Assay Kit (Invitrogen, Carlsbad, CA, USA) and Agilent Tapestation 2200 R6K. Libraries for Illumina sequencing were constructed from 100ng of total RNA using the NEBNext mRNA Sample Prep Kit 1 (New England Biolabs, Ipswich, MA, USA), following the manufacturers' guidelines with minor modifications: end repair in 50µl reaction volume (40µl DNA, 5µl buffer and 5µl enzyme), post fragmentation clean-up with 1:2.8x volume Agencourt Ampure RNAClean XP (Beckman Coulter, Pasadena, CA, USA); post cDNA synthesis clean-up with 1:1.25x volume Ampure XP Beads; post end repair, A-Tailing and adapter ligation clean-ups with 1:1.8x volume Ampure XP Beads and post-PCR library clean-ups with 1:1x volume Ampure XP Beads. Additionally, upon ligation of Illumina Adapters (Multiplexing Sample Preparation Oligonucleotide Kit) each library underwent DNA size selection with two Ampure Bead steps (firstly, 1:0.7x volume and secondly, the supernatant from the first bind was taken for a 1:1.7x volume clean-up), selecting 200-600bp fragments in 30µl 10 mM Tris-Cl, pH 8.5. Pre-PCR workflow was partially performed using a Beckman Biomek FX and post-PCR steps were performed using a Beckman Biomek NX^{P} and Biomek 3000.

The following custom primers (25 µM each) were used for the PCR enrichment step:

Multiplex PCR primer 1.0

5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'

Index primer

5'CAAGCAGAAGACGGCATACGAGAT[INDEX]CAGTGACTGGAGTTCAGACGTGTGCTCTTCCG

ATCT-3'

Amplified libraries were analysed for size distribution using the Agilent Tapestation 2200 D1K. Libraries were quantified by quantitative RT-PCR using Agilent qPCR Library Quantification Kit and a Mx3005P instrument (Agilent) and relative volumes were pooled accordingly. Finally, a second quantitative RT-PCR was performed to measure the relative concentration of the pool compared to a previously sequenced mRNA library in order to determine the volume to use for sequencing.

2.6.2 Illumina Amplicon Library Preparation

Amplicons were quantified and quality assessed using Quant-IT Qubit dsDNA High Sensitivity Assay (Invitrogen) and 1% E-gel (Invitrogen) respectively. Fifty nanograms were sheared to 400bp using Covaris^{TM} DNA Shearing (Woburn, MA, USA) and concentrated using 1:1x volume Ampure XP Beads. Eluted DNA fragments were then processed through standard library preparation procedures (end repair through to Ampure Bead size selection using the NEBNext DNA Library Prep Master Mix Set for Illumina). Post-reaction clean-ups were 1:1.8x volume Ampure XP Beads and post-PCR libraries were cleaned with 1:1x volume Ampure XP. Ten cycles of PCR amplification were performed using custom primers as above.

2.6.3 Sequencing

Multiplex libraries were prepared using barcoded primers and a median insert size of 150bp (increased to 200bp with the modified method). The pooled libraries from three initial Norovirus were sequenced on an Illumina MiSeq with 150bp paired end reads following standard Illumina protocols. An average of 0.5Gb of sdata was produced per sample. A larger pool of Norovirus samples were sequenced on an Illumina HiSeq 2000 with 100bp paired end reads following standard Illumina protocols.

2.7 Bioinformatics and Data analyses

2.7.1 Sequence assembly

The following programs were used to generate consensus sequences from nucleotide sequences generated by Sanger and next generation sequencing.

2.7.1.1 Staden

Staden is an open source suite of programs for DNA sequence assembly, editing and analysis [194]. This package was used solely for the viewing and assembly of contigs generated by Sanger sequencing, to derive a consensus. Prior to assembly, sequence quality was assessed by viewing chromoatograms in Trev. Poor quality sequence traces were discarded. Then the sequences were uploaded to pregap using Contig Selector on gap4 and a single contiguous sequence was assembled, all using the package's default settings.

2.7.2 Mapping assembly of Illumina data

The norovirus Hu/GII.4/Orange/NSW001P/2008/AU was used as reference (accession number GQ845367). This was selected as it was the closest match using NCBI BLAST to the first 20 Sanger sequenced 327bp RT-PCR amplicon used in the diagnostic RT-PCR stage.

Sequences were mapped to the norovirus reference using Stampy v1.0.14 with no BWA pre-mapping. Bases and single nucleotide variants (SNVs) were called using the SAMtools “mpileup” command with options '-E -M0 -Q30 -q30 -o40 -e20 -h100 -m2 -D -S' and BCFtools. Sites were filtered to avoid unreliable calls using the following criteria (settings available in Stampy):

a) A minimum depth of five reads at each position;

b) A minimum average Phred base quality of 10;

c) A minimum SNV quality of 25;

d) At least 75% of reads at the position support the call and the position was called as homozygous.

Positions were called as heterozygous based on the SAMtools genotype calls.

2.7.3 de novo assembly

2.7.3.1 Pre-assembly filtering

Prior to the assembly, the Illumina generated reads were "filtered" so that reads lacking similarity to known norovirus genomes were removed. A total of 477 full length norovirus genomes were downloaded from NCBI on 11 January 2013 (Appendix 2). 91 sequence clusters were identified with <=2% divergence over an alignment of at least 90% of the norovirus genome using BLASTClust, which is part of the NCBI Blast tools package (http://www.ncbi.nlm.nih.gov/BLAST/). One representative genome was chosen from each cluster.

2.7.3.2 Identification of norovirus-like sequences among Illumina reads generated from total cDNA

Illumina paired-end reads in which either or both sequences matched any cluster representative with a blastn e-value of $<1x10^{-8}$ were retained. These sequences were identified as Norovirus-like sequences (NLS). For each sample, the most similar reference sequence (MSRS) was identified on the basis of the number of matching read pairs.

2.7.3.3 *Read processing*

Reads were quality trimmed using the FastX Toolkit (version 0.0.13, Cold Spring Harbour Laboratory, New York, USA (http://hannonlab.cshl.edu/fastx_toolkit/)) with a Q-score cut-off of 15, adapters were removed using CutAdapt [197], and duplicate read-pairs were removed using custom Python scripts. We screened reads for library artefact sequences of the form 5' ABA, where A and B are sequences from the same strand of a norovirus reference genome, A is a sequence of at least 18 nucleotides, and ABA is a short read of either 100 (for HiSeq runs) or 150nt (for MiSeq runs). The ABA pattern is not present in any of the canonical norovirus sequences and we found a low (<1/1000) frequency of such reads. Singleton reads without matching partners following processing, were also removed.

2.7.3.4 *Choice and evaluation of the De Novo assembly algortihm*

Four de-novo assembly algorithms were compared, Velvet [198], Celera [199], Edena [200], and Vicuna [201], adapting their default settings as follows: in Velvet, k-mer and exp_cov were explored in a pairwise manner; for k-mer, values of 23 to 79 in steps of 4 were evaluated for the following values of exp_cov: auto, 1, 3, 10, 20, 30, 40, 50, 75, 100, 125, 150, 175, 200, 225, 250, 300, 400, 500, 600, 700, 800, 1000, 1200, 1500, 1800, 2000. Ins_length was specified at 250, based on known insert size distribution. For Edena, all read trimming lengths between 35 and 150 in steps of 5 were evaluated. In Vicuna, the following settings were used: minMSize=9, maxOverhangSize=2, Divergence=15, Max read overhand=4, Max contig overhang=6, Seed kmer length=9, Min contig overlap=25, Min contig links =2, min identity= 90. In cases where multiple assembly parameters were compared, the setting(s) producing maximal N50 were

selected. For all assemblers, contigs less than 300nt were removed from further analysis.

To assess assembly quality and performance, each contig was aligned to its MSRS with blastn and a minimum e-value of 1×10^{-8}.

Additionally, NLS reads were mapped to either the Vicuna-derived contigs, or full genomes produced by capillary sequencing (using the Staden package [194]) with Geneious Pro software (version 6.1.6; Biomatters, New Zealand), using the built-in mapping algorithm using the 'highest sensitivity' setting. The results of the mapping, and the alignment of the contigs to the reference, were inspected manually.

2.7.3.5 Within-patient variant discovery

The NLS reads were mapped to the *de-novo* contig if a single contig was derived from the assembly. Mapping was performed with BOWTIE2 (version 2.1.0) [202] using the --very-sensitive flag. Samtools (version 0.1.19) [203] was used to covert alignments to pileup format. Pileup files were parsed to produce variant calls, excluding variants with base quality scores of less than 30, or mapping quality scores of less than 40. Additionally, an additional step was designed to assess credibility of variants supported by low read numbers: variants were ignored for which, using an exact Binomial test (R 2.1.5, binom.test function), the proportion of reads supporting a variant relative to the reference was not significantly different ($p>0.01$) than an expectation of 1×10^{-4}. Finally, only variants present at 1% or more of the reads were considered for further analyses.

2.7.4 Sequence comparison

A range of different applications were used to understand the evolutionary relationships of the norovirus genomes. Contigs produced by Sanger sequencing were compared using BLAST as well as CLUSTALW or MUSCLE (http://www.genome.jp/tools/clustalw/).

Sequences produced by both 454 and Illumina were compared using the suite of programs described above within Geneious Pro R6 (Biomatters Ltd, New Zealand). In addition, MAFFT version 7.130 [204] was available as a plugin within the sequence comparison suite and this was used as an alternative comparator.

2.7.5 Phylogenetic tree construction

Maximum likelihood trees were used to obtain the optimal relationship between the sequences and possible evolutionary history of the virus. The mapping assembly and filtering produced ambiguous nucleotides, designated Ns, which may be a result of heterozygosity or poor coverage around this region. Unlike multiple imputational algorithms, maximum likelihood gives unbiased parameter estimates and standard errors over the missing data, and therefore offers an optimal probabilistic approach to overcome this.

An evolutionary tree was created using BEAST (Bayesian evolutionary analysis sampling trees) depicting all the full genomic sequences with relatedness. Analysis was performed using BEAST v.1.7.5 [205] combining two random number seed chains (10 million iterations each, saving 1 in 1,000 iterations, with a 1 million iteration burn-in) using: HKY substitution; estimated frequency; strict clock; and constant population size coalescent tree prior. This maximum clade credibility tree was computed using

TreeAnnotator v.1.7.5 (plugin within BEAST) and plotted with Figtree v.1.4.0 (available at http://tree.bio.ed.ac.uk/software/figtree/).

2.7.6 Statistical analyses

Stata 13 (StataCorp LP, USA) was used for the statistical analyses in this thesis. Information from the epidemiological data was collated from a Microsoft Excel spreadsheet and imported as a dataset on Stata 13. Syntaxes and commands (such as *mean, tabulate, summarize*) were used in Stata for statistical analyses as well as producing histograms and bar charts depicting the information.

Chapter 3. Evaluation of Three Approaches to Norovirus WGS Determination

Chapter Objective

As described in Chapter 2, three different experimental approaches to norovirus whole genome sequence determination were described and the results of this evaluation are presented here. These proof of principle experiments facilitated the choice of optimal cDNA sequencing methodology to address the aims of this thesis. Figure 3.1 is a flow diagram which summarises the total number of samples used in each evaluation experiment, together with their relative success rates defined by their reliability in yielding norovirus whole genome sequences.

3.1 Experiment 1: WGS Determination using Sanger sequencing

Clinical Samples: 84 RT-PCR positives from OUH inpatients (2009-2011)

Norovirus genome sequences were determined initially by the Sanger sequencing method (Chapter 2.4), to obtain baseline data for comparison with genome sequences generated by NGS. A total of 84 ward based norovirus RT-PCR positive stool samples collected from outbreaks occurring during two consecutive seasons (2009/2010 and 2010/2011) were prepared for sequencing using seven overlapping PCR amplicons. Of the 84, only samples yielding a single PCR amplicon for each of the seven reactions were sequenced. A total of 25 samples (29%) yielded high quality nucleotide sequence data spanning full norovirus genomes. Fifteen samples yielded partial genomes with >70% coverage, whilst the remainder (n=44) could not be sequenced due to a lack of one or more amplicons. This may have been a consequence of oligonucleotide PCR primer mismatches with the target, competing homologous sequences, or the presence of PCR inhibition within the samples.

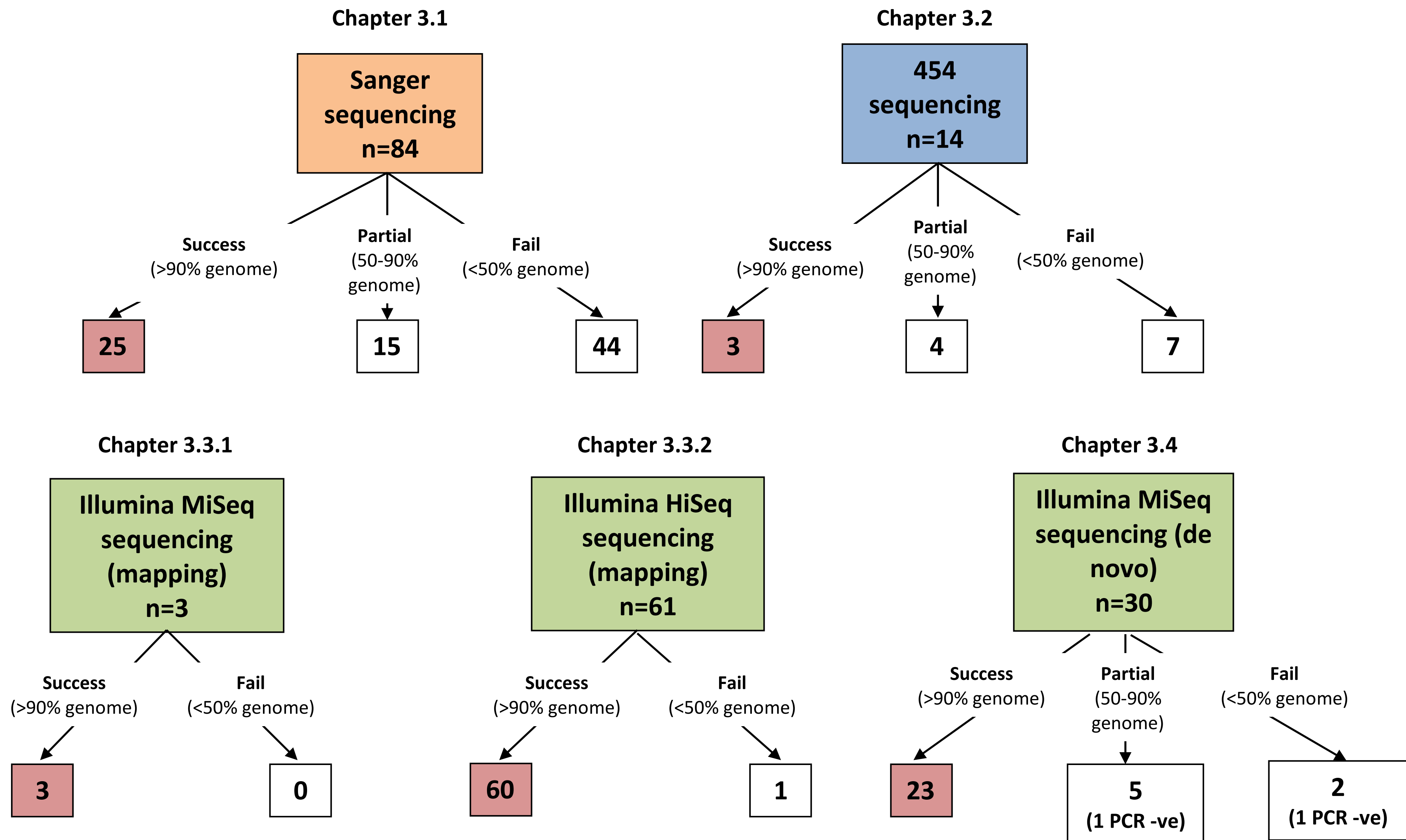


Figure 3.1 Flow diagram to show the total number of samples used in each of the experiments described in this chapter. An experiment was regarded as successful if a genomic sequence spanning >90% of the genome was obtained. Experiments were designated partially successful if coverage was in the range 50-90%. Experiments were scored as 'failures' if coverage spanned <50% of the whole genome.

3.2 Experiment 2: WGS Determination using the Roche 454 Pyrosequencing Platform

Samples used: 14 RT-PCR positive samples from Oxford inpatients (2009-2011)

The 454 pyrosequencing platform was evaluated using a set of 14 high titre norovirus positive stool samples dating from 2009-2011, which were chosen solely on the basis of their strong gel-based RT-PCR positivity (independent of the samples described above). This was important because the two large (4.3kb and 3.4kb) overlapping norovirus genome-spanning PCR amplicons were required at a high concentration post-purification for 454 pyrosequencing (Chapter 2.5.1).

Seven of the fourteen clinical samples (50%) had yielded norovirus-related sequences which could be mapped to the norovirus reference genome (GQ845367). However, only a small proportion of the total reads (0.42-8%) mapped to this reference and the remainder were discarded because BLAST searches (http://blast.ncbi.nlm.nih.gov/Blast.cgi) confirmed they represented contaminating, non-norovirus sequences. The percentage of reference genome coverage by the seven samples ranged from 35% - 100%. The genomic distribution of reads varied among samples, without any specific coverage "hotspots" (Figure 3.2). Only one of the seven samples yielded a consensus sequence spanning the full genome. Due to the high financial cost and labour intensive nature of the 454-based approach, combined with the high failure rate (50%), an alternative sequencing platform underwent validation, with the aim of improving genome coverage and reducing costs.

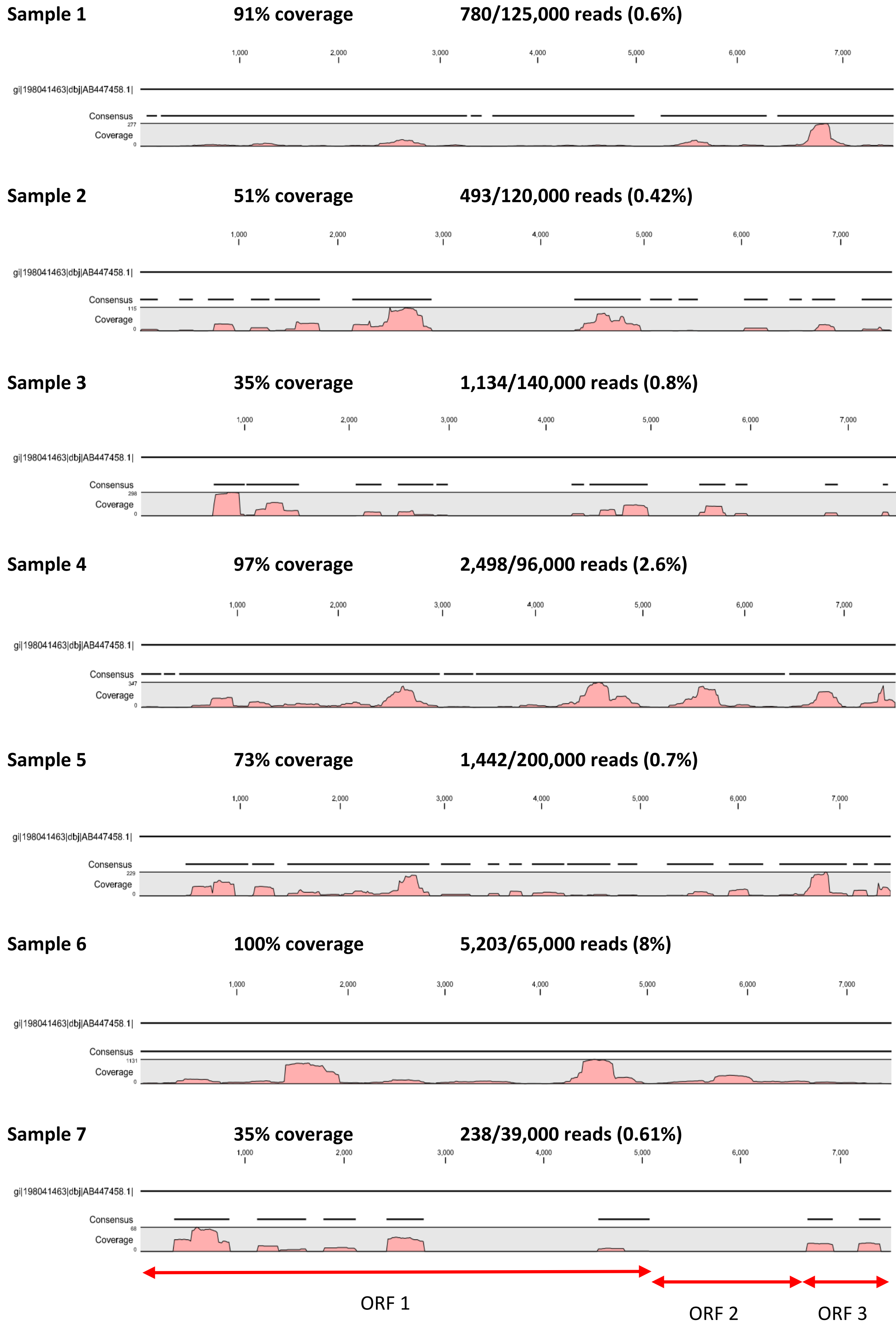


Figure 3.2 Distribution of reference genome coverage by the 454-generated norovirus reads. The consensus line reveals areas on the genome where coverage occurred. Quality statistics regarding the percentage of coverage and number of reads which mapped are shown above each sample.

3.3 Experiment 3: Genome sequencing determination by Illumina sequencing (mapping to a reference genome)

3.3.1 Proof of principle experiment using the Illumina MiSeq bench-top sequencer

Samples used: Three RT-PCR positive samples from Oxford inpatients (2010-2011)

Three norovirus RT-PCR positive faecal samples were used in an initial proof of principle experiment. These samples had been previously identified as GII.4 norovirus positive by quantitative RT-PCR followed by attempted whole genome sequencing using Sanger sequencing as described in Chapter 2.4.1). These three samples were estimated to contain 20,000 (sample 1), 3,000 (sample 2) and 3 million (sample 3) copies of the genome/μl. The MiSeq platform yielded a mean of 3.8 million reads per sample (range 2.6-5.9 million). Of these, 0.85-1.90% (25,141/3,771,150 reads – 50,138/2,671,058 reads) of the total reads mapped to the reference genome (Appendix 3), indicating that a low percentage of the total RNA was from norovirus. However, although the percentage of total reads which mapped was low, reference genome coverage was sufficient to achieve a near complete genome sequence in all three samples (range 97.4-99.1%). Coverage varied across the genome, being lowest at the 5' and 3' termini of (Figures 3.3). Single nucleotide variants (SNVs) were called using SAMtools and BCFtools [203].

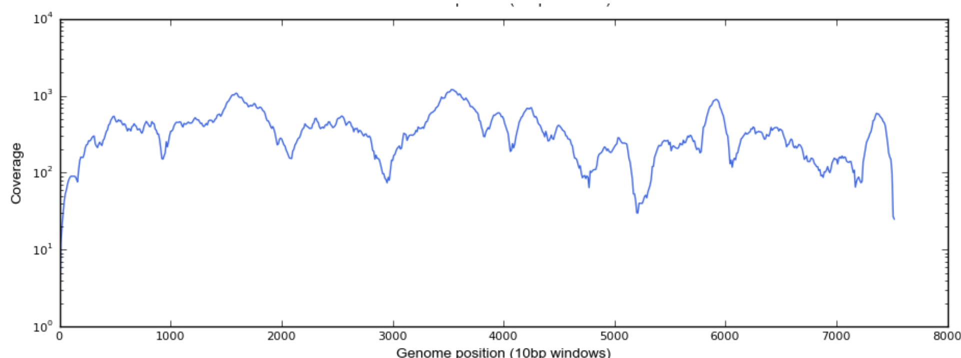


Figure 3.3. Coverage profile of one Norovirus sample from RNA sequencing.

3.3.1.1 Comparison of WGS determined by Sanger sequencing and the MiSeq platform

The reliability of the novel technology represented by the Illumina MiSeq platform was validated by comparing the three MiSeq-determined genomes with available sequences for the same three samples generated previously by conventional Sanger sequencing (section 3.1). For sample 1, 99.1% of the genome was available for comparison, having been determined by both methods, and agreement was 100%. For sample 3, 86.4% of the genome was available for comparison, and once more agreement was 100%. Although sample 2 was diagnostically quantitative RT-PCR positive, it failed to yield the PCR amplicons required for Sanger sequencing. This was likely related to the low norovirus titre 3,000 copies/μl of this sample (estimated by qPCR). However, its inclusion in the experiment was still valid because it demonstrated the enhanced sensitivity of the MiSeq-based approach, relative to Sanger sequencing. The sample 2 norovirus was successfully sequenced using the Illumina platform (98.8% of genome sequenced), indicating that this approach is technically feasible when a low viral load precludes traditional amplification based methods.

In summary, no differences in nucleotide sequence were found in the 14,019 nt of norovirus sequences generated by Sanger and Illumina custom RNA-Seq sequencing methods and the RNA-Seq approach showed improved sensitivity. The Illumina chemistry showed the greatest potential for high throughput, *de novo* norovirus genome sequence determination, by eliminating firstly the requirement for at least some prior knowledge of the sequence to be determined and secondly the need for a relatively high virus titre; two issues which beset the Sanger and 454-based approaches.

3.3.2 High throughput WGS determination via the HiSeq platform (reference based assembly of reads)

Samples used: 61 quantitative RT-PCR positive samples from Oxford inpatients (2009-2011)

The feasibility of high throughput norovirus sequence determination using the Illumina HiSeq platform was assessed using 61 Norovirus positive stool samples. These samples were obtained during outbreaks dating from 2009-2011 and some had previously been used for Sanger sequencing (chapter 3.1). To assess the reproducibility of the HiSeq platform, 15 out of the 61 samples were sequenced twice by this method. Such paired libraries were prepared either from the same total stool RNA extract (9 pairs), or from two independent extractions of the same faecal sample (6 pairs).

Of the 61 samples, 16 had been included in the 25 which previously yielded complete genomes by Sanger sequencing (section 3.1), thus enabling comparison and validation.

Using the HiSeq platform, a mean of 4.6 million reads were produced per sample (Range=2.6-42 million). The proportion of reads that mapped to the reference genome varied across the samples, ranging from 0.01% to 97.98%; a wider range than seen in the three pilot MiSeq samples. Although greater variability would be expected in a larger sample collection, this may also reflect variation in the ratio between norovirus and non-norovirus RNA in the original sample. We observed a significant correlation between the percentage of reads which mapped to the reference genome and the viral titre estimated by qPCR ($rho=0.4$, $p<0.0001$, Figure 3.4).

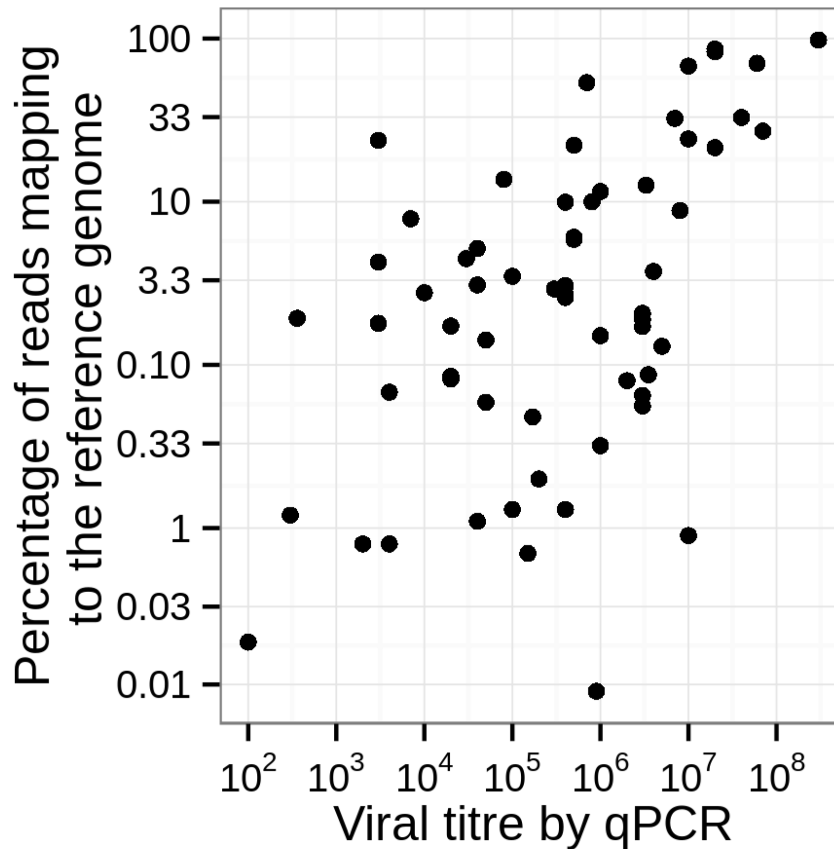


Figure 3.4 A comparison of the viral titre of each sample assessed by qPCR, and the percentage of reads obtained from each sample which mapped to the Norovirus reference genome (GQ845367), showing the correlation between the two measures (rho=0.4, p<0.0001).

Sixty out of sixty-one samples yielded reads sufficient to allow 90-98.9% of the genome to be assembled (mean 97.0%) versus the reference GII.4 (Genbank GQ845367). The one remaining sample yielded a low number of reads (744) so only 4,449 (58.9%) of the genome could be successfully called (sample 6). As before (chapter 3.3.1), bases which could not be determined clustered at either end of the genome with very specific regions consistently yielding low coverage (Appendix 3). No difference in the consensus sequence which was determined for the 15 duplicate pairs was observed. Among the 16

samples where sequence data determined by the Sanger method was also available for comparison, no discrepancies were found (total 118,757nt).

3.4 Illumina sequencing (*de novo* assembly approach)

The development of a *de novo* (i.e. non-reference based) assembly method for NGS norovirus reads is important because the virus is highly genetically variable and evolves rapidly, even within a genotype. Consequently, successful routine surveillance based on these technologies would require the choice of reference genome to be constantly monitored and updated. The predominant GII.4 variant is particularly prone to periodic recombination events that are associated with a sudden, marked increase in incidence and displacement of the previously circulating GII.4 strain [67]. Such reports involving an emergent recombinant GII.4 norovirus occurred in Australia during 2013 [206]. The response was to develop a *de novo* norovirus WGS assembly method with the dual aims of testing the feasibility of this approach for norovirus, and determining whether the novel Australian GII.4 virus had reached the UK. If successful, this approach could have wide applicability, providing an assembly option for virus discovery studies and for known viruses where a reference genome is either unavailable, or is incomplete [200, 201].

The aim was to develop a novel bioinformatical approach to the *de novo* assembly of noroviruses contained in faecal samples collected from 11 separate UK locations which occurred between February 2012 and January 2013.

3.4.1 Samples

Samples used: 30 (28 RT-PCR +ve/ 2 RT-PCR -ve) samples from 2011/2012 and 2012/2013

Faecal samples were collected from hospital and community outbreaks in eleven geographically widespread locations between February 2012 and January 2013. An unusually early increase in the norovirus infection rate occurred during this period (Figure 3.5). A total of 28 norovirus positive samples from symptomatic patients were sequenced using the high throughput approach which was described and validated above (Chapter 3.3.2). The 28 samples had qPCR-estimated norovirus titres of between 10^3 to 10^8 copies/μl. These figures were statistically similar to prior samples obtained during winter 2010/2011 outbreaks (mean $\log_{10}$ copy number for 2013 5.71 vs. 6.13, $p=0.25$ by t-test). To further confirm reproducibility, one sample was sequenced a total of three times from its total RNA extract. Two remaining clinical samples included in this analysis were qPCR negative, although they originated from one of the suspected norovirus outbreaks included in the study. The Illumina MiSeq yielded an average of 2.0 x 10^6 (range 0.8 x 10^6 to 3.5 x 10^6) reads per sample.

Figure 3.5 Location of samples used in proof of principle *de novo* assembly

3.4.2 *De Novo* assembly of norovirus sequences

3.4.2.1 Assembler performance

De novo assembly of norovirus WGS was performed using four different published assembly packages: Velvet [198], which is based on de Bruijn graph construction, and three assemblers (Edena [199], Celera [200] and Vicuna [201]) which use various implementations of an overlap-layout consensus based approach, and a range of assembly parameters. The relative performances of the assemblers are shown (Figures 3.6a and 3.6b). The Vicuna assembler generated the most single-contig assemblies covering >97% (7,321-7,552nt), in 23/28 samples. The best Velvet assemblies were comparable to Vicuna, but the Velvet assembly process was found to be highly sensitive to both short strings of nucleotide reads (known as k-mer) and particularly to the

expected coverage (exp_cov) parameter, necessitating a large parameter space search to find optimal assemblies. Figure 3.6a and 3.6b details a comparison of the four assemblers.

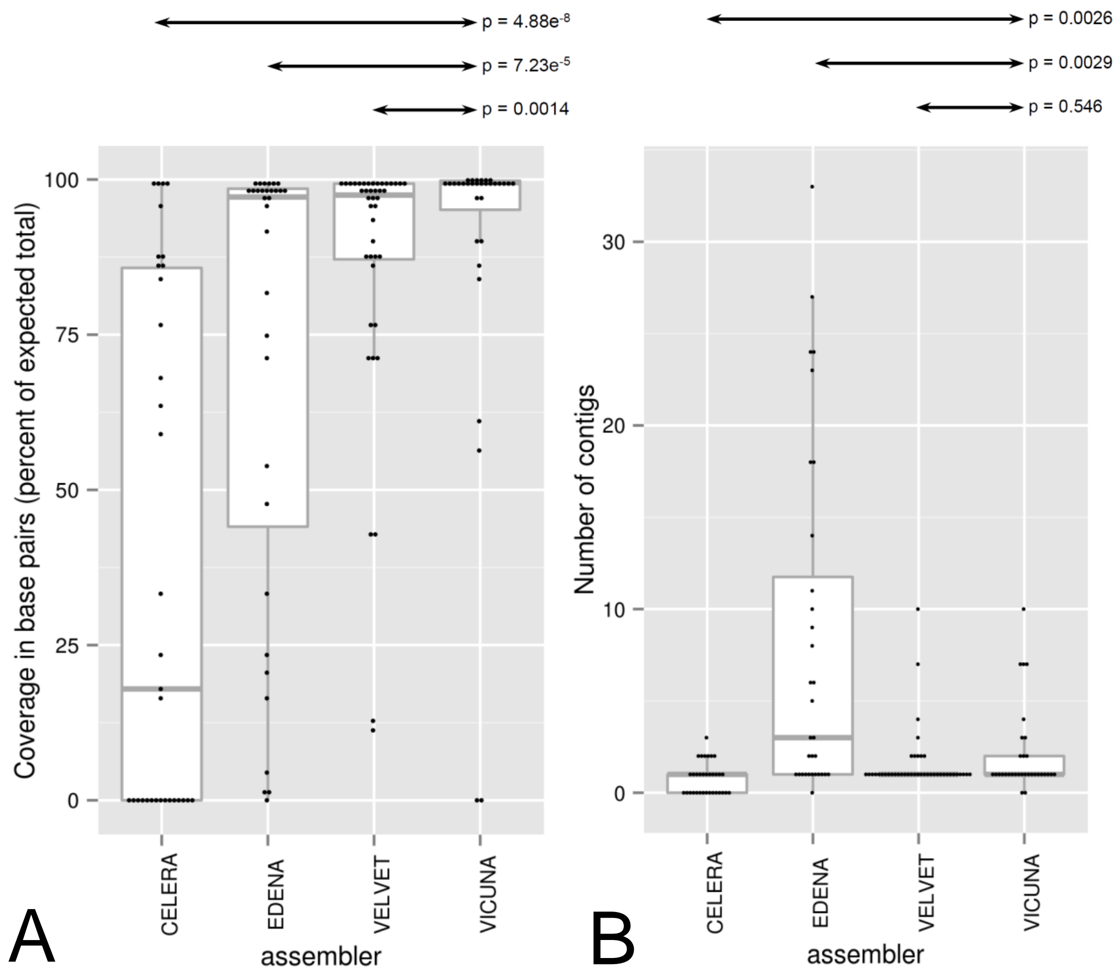


Figure 3.6 (A) The estimated coverage of the contigs produced and (B) the number of contigs produced by four assemblers, on a 32 sample test dataset. One dot signifies the result for one sample. Box and whiskers plots indicating median (thick grey line), 25th and 75th centiles (edges of box) and 5th and 95th centile (ends of whiskers) are shown. Mann-Whitney U test are shown comparing each assembler with Vicuna (p values given at the top of each figure).

3.4.2.2 Intra-sample diversity

The extent of nucleotide sequence diversity between and within the samples studied might be underestimated by the Vicuna-derived consensus: within sample variation is

not revealed. The extent of such intra-sample variation detected is shown in Figure 3.7. This revealed a limited number of within-host variants in samples that assembled of around one site per genome.

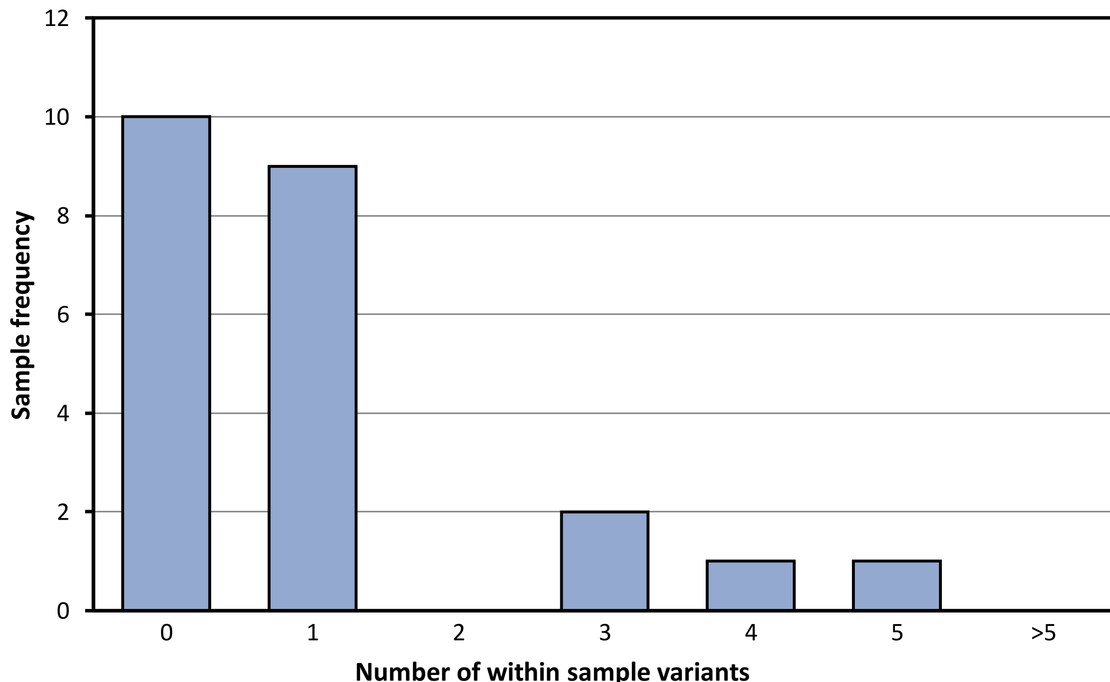


Figure 3.7 Intra-sample variation. Bar chart displaying the frequency of samples and the number of intra-sample variants. Ten out of the 28 samples had no evidence of intra-sample variation.

3.4.2.3 Determinants of non-assembly

Five samples did not assemble to near full genome (defined as >7,500nt or 99% of genome). Read numbers did not differ between samples which could not be fully assembled and samples that could be assembled (mean 1.8 $x10^6$ vs. 2.1 x 10^6 reads (p=0.22, t-test)). In contrast, the total number of reads matching norovirus sequences differed markedly (0.17 x 10^6 vs. 1.04 $x10^6$ reads, p=0.009). Read depth appears, therefore, to be a determinant of the ability to generate *de novo* assemblies using the technique developed here. The two qPCR-negative samples did not yield full genomes,

although partial assemblies were produced (<59% of the genome in one sample, 0% in the other), confirming that one of the affected patients may have been part of the norovirus outbreaks.

3.4.2.4 Mis-assembly at the 5' and 3' termini of contigs

In three samples, although >99.5% coverage was obtained, two contigs were produced that differed at their termini. In each case, one contig had high (>500 reads) read support and represented part of a canonical norovirus genome, while the other was a rearrangement of a norovirus genome, with low (2–4 reads) support. These rearrangements disrupted open reading frames and were therefore unlikely to originate from a replication-competent virus (data not shown). In view of this, and the low read support for them, the re-arranged low-read terminus were manually removed from these three samples and a single contig was formed by manual editing.

3.4.2.5 Reproducibility of the process

To assess the reproducibility of the Illumina sequencing and Vicuna *de novo* assembly process, one sample was processed and sequenced independently three times using the Illumina MiSeq, starting from the same total RNA extract. *De novo* assembly using Vicuna revealed no nucleotide differences between these replicates.

3.5 Choosing the best analytical approach

The data described in this chapter validate the Illumina sequencing platform, combined with the RNA-Seq approach and Vicuna mediated *de novo* assembly as the method of choice for high throughput norovirus WGS determination [160]. The only benefit of the Illumina HiSeq over the MiSeq platforms was the rate of throughput (96 HiSeq samples vs. 16 MiSeq samples per run). Vicuna *de novo* assembly facilitated the construction of

genomes which were more than 7,500nt in length (99% of total genome) compared to the reference-based assembly approach which necessitated the acceptance of >90% coverage as ‘successful’.

Chapter 4. Descriptive epidemiology of Norovirus outbreaks 2009-2013

Chapter Objective

To describe and compare the patient characteristics of norovirus ward outbreaks in Oxfordshire between September 2009 to August 2013. Additional epidemiological data collected from national sites (such as Brighton and Leeds) will also been described.

4.1 Background

Current hospital infection control practice during norovirus outbreaks is based upon the Kaplan criteria, developed in 1982 [18]. Although literature exists which describes norovirus cases and outbreaks affecting semi-closed environments such as schools, restaurants and nursing homes, little has been recorded about the characteristics of norovirus associated gastroenteritis outbreaks impacting on individual wards or hospitals [8, 207].

Using the epidemiological information collected during outbreaks in the Oxford University Hospitals (OUH) between 2009-2013, and two other UK NHS Trusts between 2011-2013, the general characteristics of outbreaks occurring in distinct norovirus seasons will be described. These will include factors such as age, gender, symptom description, duration of symptoms and location of symptomatic patients relative to speciality. For this thesis, each norovirus season is defined as a full year, from September 1st until August 31st the following year.

4.2 Method

The Oxford University Hospitals (OUH) NHS Trust is an acute trust based in Oxfordshire providing both acute and specialist services. The trust comprises four hospitals; John Radcliffe Hospital (800 beds), Churchill Hospital (330 beds), Horton Hospital (250 beds) and the Nuffield Orthopaedic Centre (300 beds) (Figure 4.1). Clinical samples and data were collected from all but the Nuffield Orthopaedic Centre.

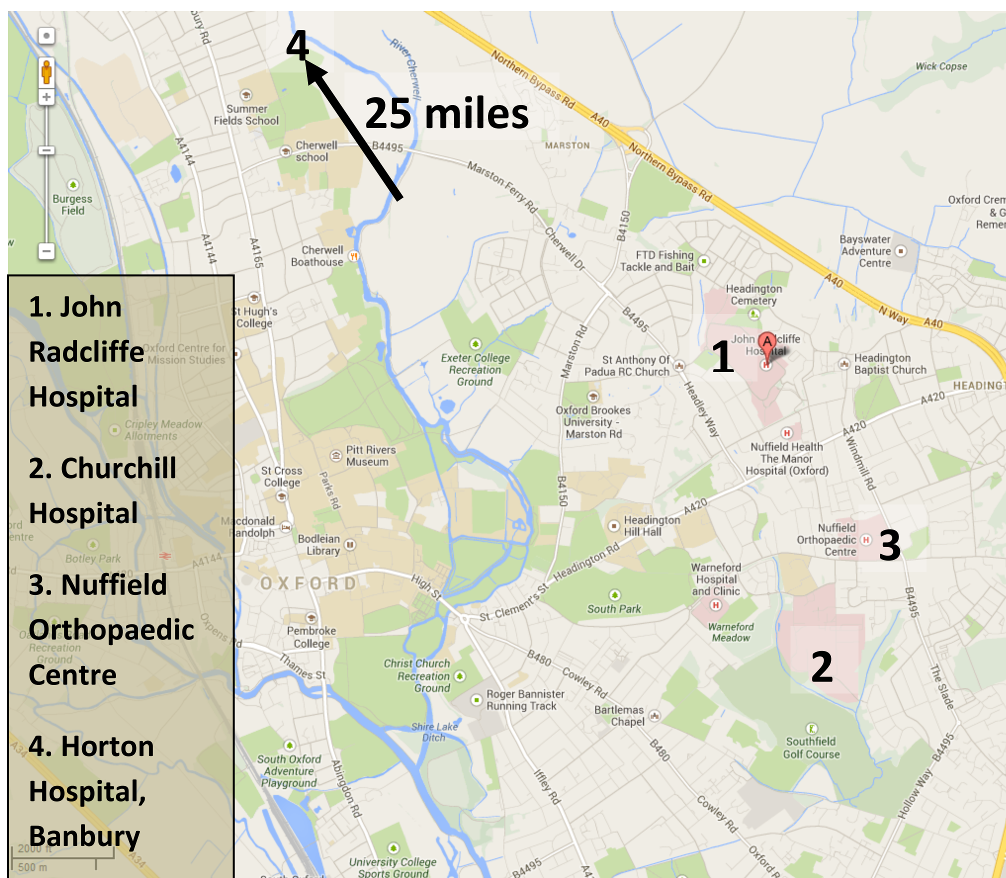


Figure 4.1. Location of the four hospitals comprising the OUH Trust

Norovirus cases were identified during an active prospective study of gastroenteritis outbreaks within hospital wards by myself or the OUH Trust hospital infection control team. Clinical samples from symptomatic patients in the community were also received (at the Clinical Microbiology Laboratory, John Radcliffe Hospital) and available to this study; these included, for example, nursing home or community hospital outbreak

samples or samples sent from sporadic symptomatic cases via their general practitioner. Despite these samples having limited epidemiological information about symptom duration and description, they have been used in combination with genomic information described in both Chapters 5 and 6.

Additional limited epidemiological information was available from other participating sites within England. Samples and accompanying epidemiological information was available from Brighton and Sussex University Hospitals NHS Trust and Leeds Teaching Hospitals NHS Trust. Only norovirus RT-PCR positive samples (tested prior to receipt in Oxford) and accompanying epidemiological information from two seasons (2011-2013) were available to this study from these two sites.

4.3 Results

4.3.1 Epidemiologic information from OUH, 2009-2013

Between 1st September 2009 and 31st August 2013 (four norovirus seasons), a total of 512 ward inpatients were identified with symptoms of gastroenteritis within the trust. The clinical criteria for inclusion have been described in Chapter 2.1.3. The majority of these patients (41% or 210/512) were symptomatic during the 2009/2010 season. The age, gender, ward and date of onset of symptoms were recorded.

The wards containing symptomatic patients were visited each day and details of diarrhoea and vomiting were recorded. Complete information was collected from 469 (92%) of these patients. In 43 cases, information was incomplete mainly due of inadequate nursing records and poor recall from the patient. These patients were excluded in the analysis of symptoms. Figure 4.2 gives a detailed breakdown of the numbers described.

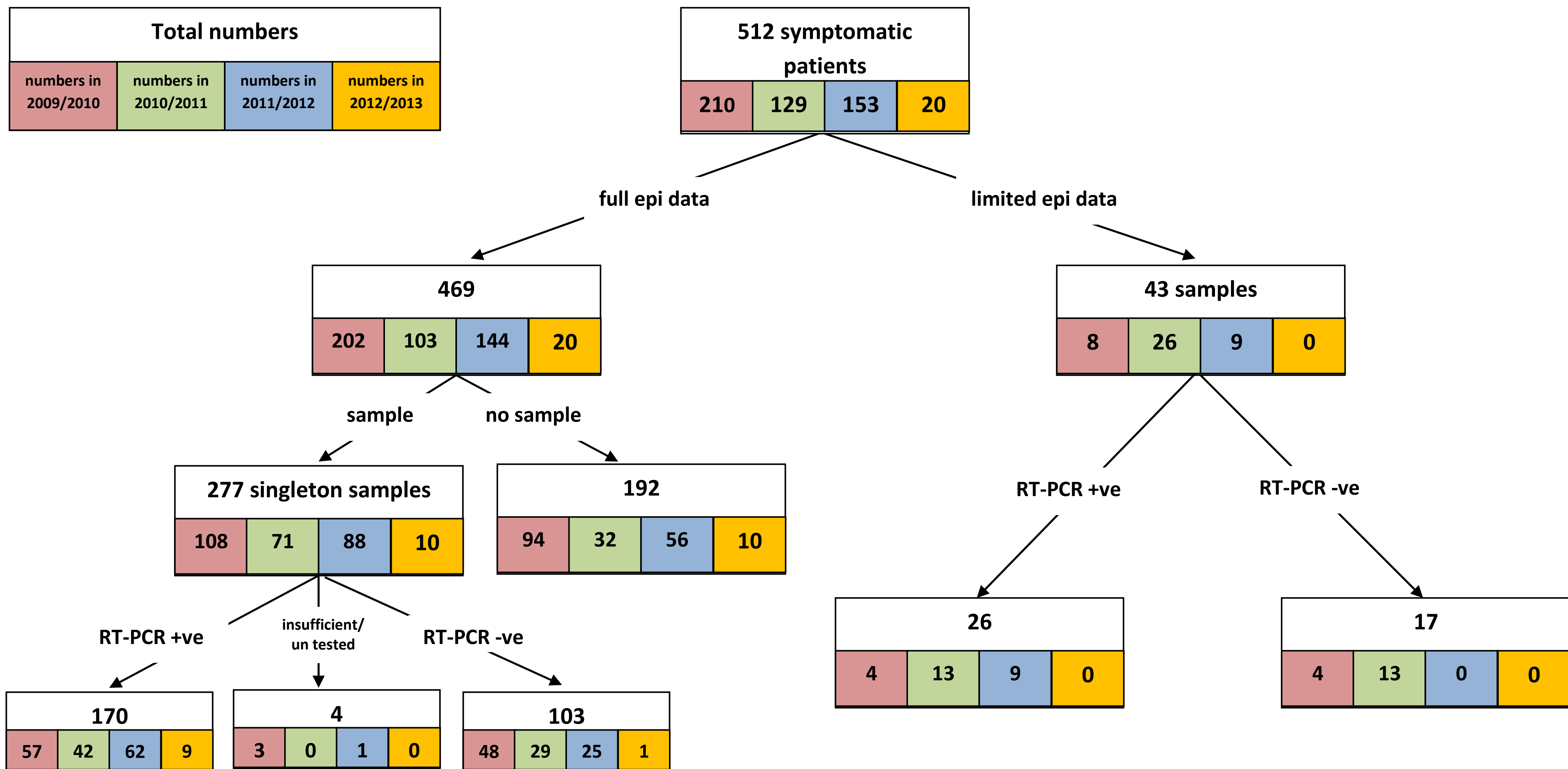


Figure 4.2. Flow chart depicting number of symptomatic patients, number of samples received, tested and sequenced in OUH over 4 seasons. *Samples collected longitudinally have been excluded*

4.3.2 Epidemiological data from non OUH sites

Norovirus RT-PCR positive samples from two seasons (2011/2012 and 2012/2013) were available from Brighton and Sussex University Hospitals NHS Trust and Leeds Teaching Hospitals NHS Trust. Epidemiological and symptom data were also available for the majority of these samples. This information will be discussed throughout the chapter independently, as well as in combination with the OUH data.

4.3.3 Samples received and diagnostically tested

A total of 320 of 512 (63%) of symptomatic patients provided at least one faecal sample (277 with epidemiological data; 43 without epidemiological data). As described above, a small cohort of patients could not produce a faecal sample, whilst others only exhibited vomiting and did not have diarrhoea, precluding them from having a sample taken for diagnostic testing. Of the 320 patients providing samples, 196 were diagnostically RT-PCR positive of which 170 had symptom data. In addition 61 patients provided more than one sample during the course of their illness. These longitudinal samples will be discussed further in the genomic analyses in chapters 6.

4.3.4 Seasonal effects of norovirus outbreaks in the OUH

Norovirus has been referred to as "winter vomiting disease", mainly due to its predominance in winter months [81, 82]. However, outbreaks may also occur in warm summer months [208]. Epidemiological data from Oxford confirms the winter seasonality of norovirus throughout the four annual seasons studied. In 2011 and 2012, confirmed outbreaks occurred within wards during spring and summer months, namely April, June and July in 2011 and April and May in 2012. Figure 4.3 details the distribution of symptomatic patients according to month and season.

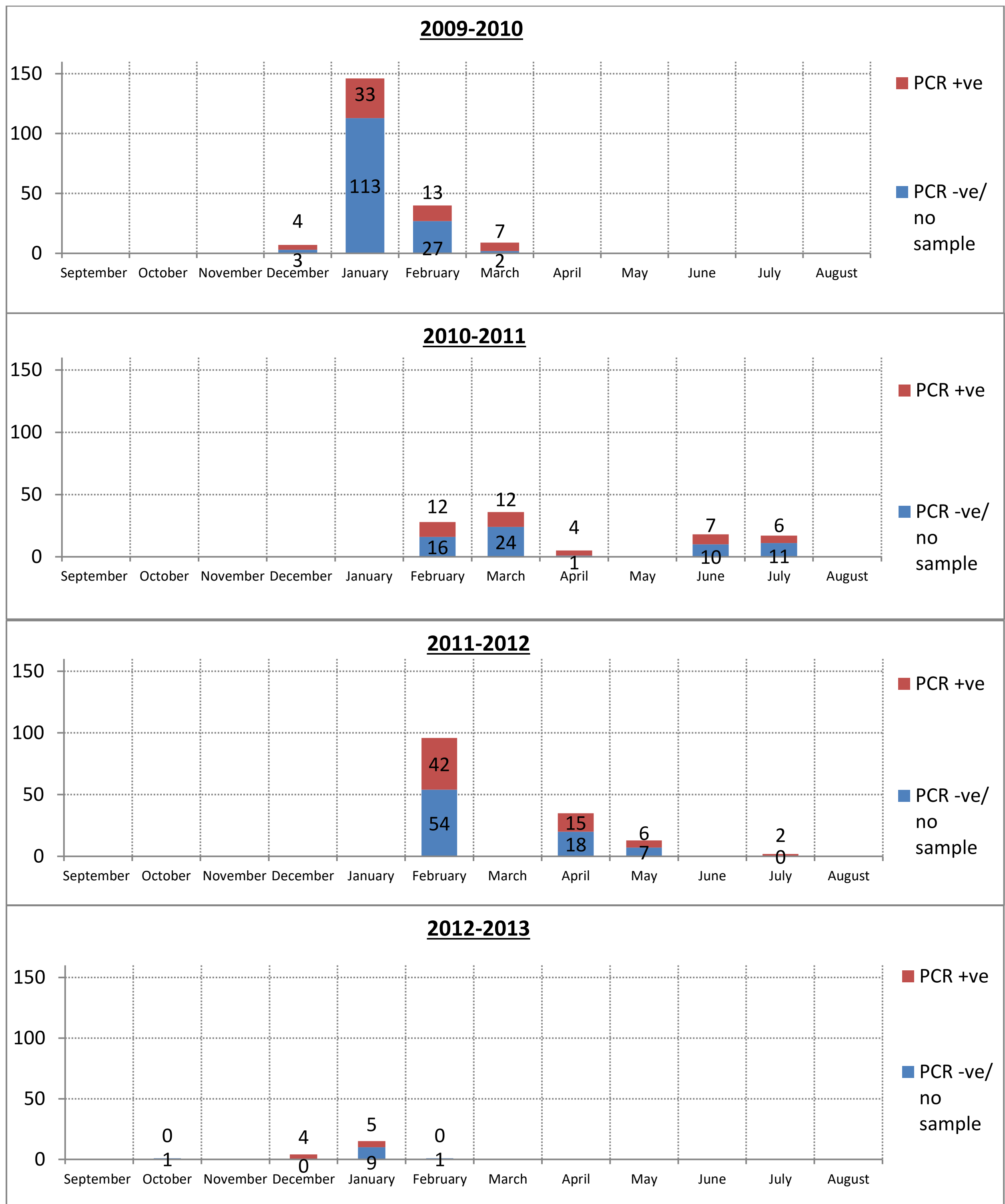


Figure 4.3 Distribution of symptomatic patients in the OUH between 2009-2013. Each Season runs from 1st September of one year to 31st August the following year.

4.3.5 Age Ranges

The age range for RT-PCR tested patients from the OUH are shown in Figure 4.4.

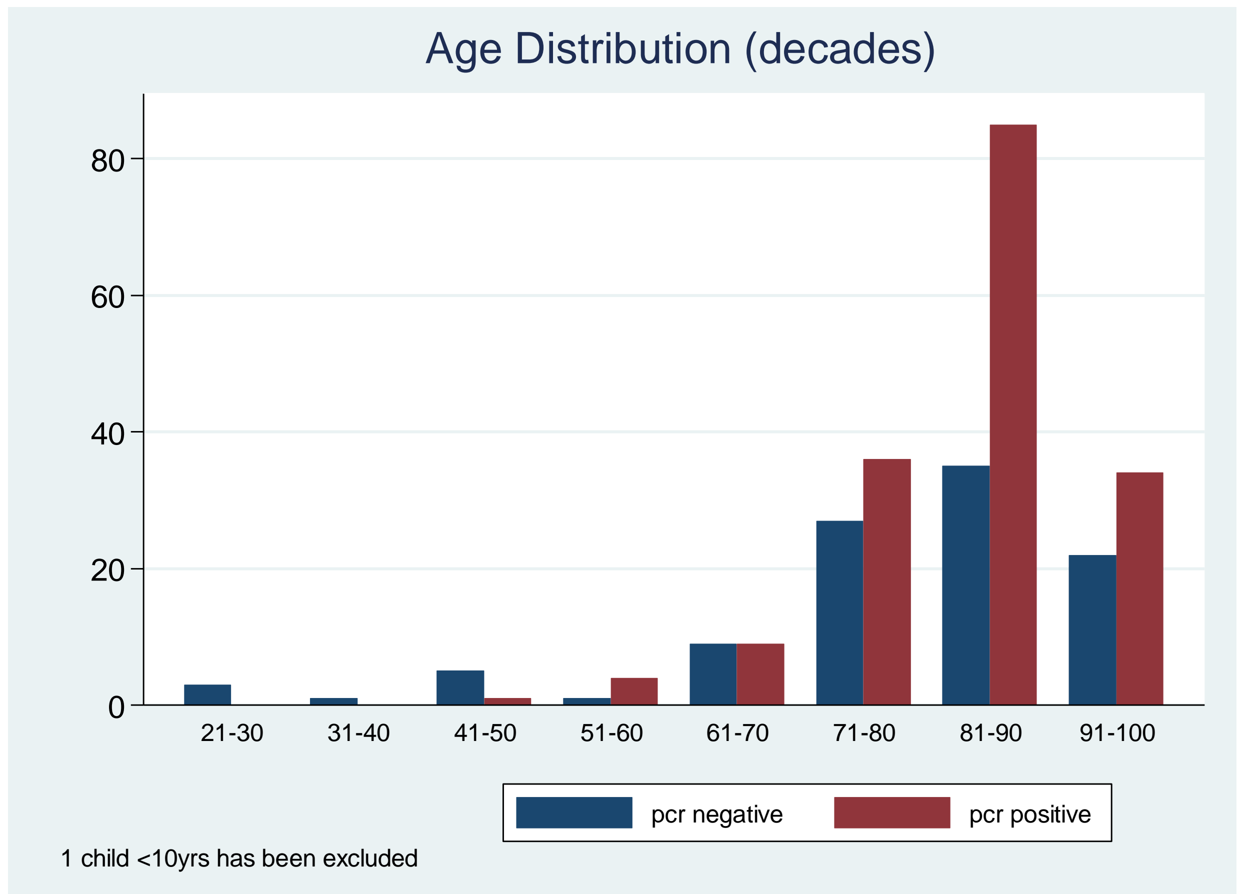


Figure 4.4 Age ranges of RT-PCR tested symptomatic patients between 2009-2013 at the OUH

The mean age amongst all symptomatic patients was 81 years, whilst the mean of all norovirus RT-PCR positives was 83 years. The median age of our cohorts of symptomatic and RT-PCR positive patients is greater than published data (84 years and 85 years, respectively). A recent report by Vega et al., [123] who investigated the epidemiology of norovirus amongst patients within multiple settings found their median age for hospital patients to be lower at 66 years of age.

Statistically there was no significant difference in the age of symptomatic patients who provided a faecal sample and those who did not (Mean age 79.1 with sample vs. 81.2 without sample, two sample t-test p=0.948). Patients who were RT-PCR positive (mean age 81.2 yrs (95%CI 79.6-82.8)) were on average 4.9 yrs (CI 1.7-8.1) older than RT-PCR negative patients (mean age 78.2 yrs (95% CI 75.1-81.2); (t-test P=0.003). There was no

data available on age distributions from all inpatients admitted to OUH during this period to analyse whether this correlates with symptomatic patients. No age information was available from the non-Oxford sites.

4.3.6 Gender

Norovirus affects both genders. The figures for all symptomatic patients within the OUH over the four norovirus seasons show more females were affected (Figure 4.5). 39% (109/273) of patients providing samples were male compared to 33.3% (64/192) of those not providing samples (χ^2=2.1 p=0.15).

Figure 4.5 Histogram of gender ratio of all OUH symptomatic patients over four norovirus seasons

There was no difference in the sex ratio between patients providing RT-PCR positive and negative samples. Proportion of males in RT-PCR negative patients was 38.8% (40/103) compared to 40.6% (69/170) with RT-PCR positive samples; χ^2 0.08, p=0.77)).

Therefore, although more females were symptomatic over the four seasons, this did not affect their sample collection or testing result compared to males. Unfortunately, there was no data from OUH regarding gender distributions during the four seasons from all inpatient admissions to verify if this correlated with the higher proportion of symptomatic female patients.

Gender information was also available for patients supplying samples who were from Brighton and Sussex University Hospitals NHS Trust (Fig 4.6) and Leeds Teaching Hospitals NHS Trust (Fig 4.7).

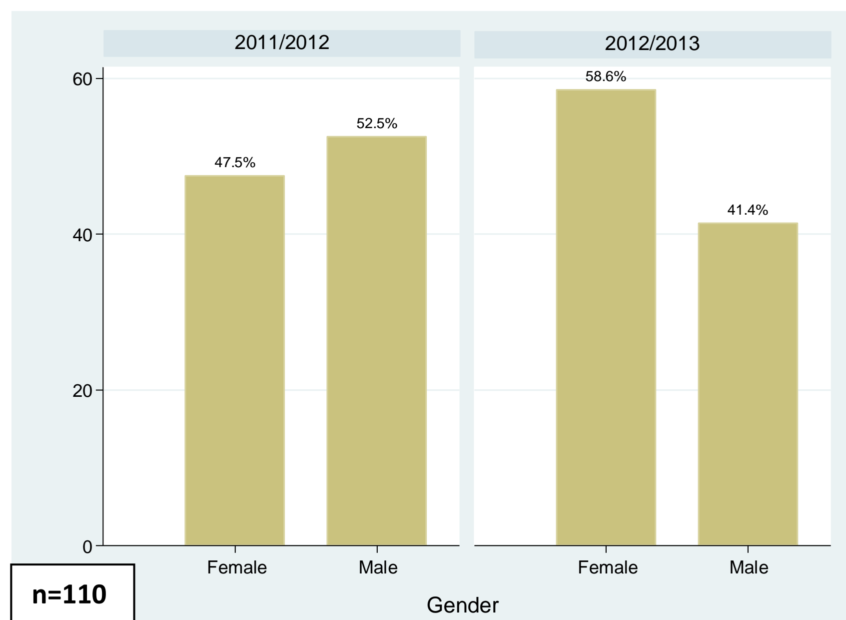


Figure 4.6 Gender ratios from Brighton and Sussex University Hospitals NHS Trust 2011-2013

Figure 4.7 Gender ratios on Leeds Teaching Hospitals NHS Trust 2011-2013

4.3.7 Wards affected

4.3.7.1 OUH

The majority (84%) of symptomatic patients over the four seasons were admitted to general medical wards. Fewer cases (16%) occurred on specialist medical (Gynaecology, Stroke, Chest, Renal and Infectious Diseases) surgical (vascular surgery (6A), Cardiothoracic and cardiothoracic surgery (CT surgery)), and paediatric (Robins) wards. Figure 4.8 shows the distribution of cases among the wards affected by outbreaks. Medical wards are labelled on the x-axis in black, non medical or specialist wards are labelled in red.

4.3.7.2 Brighton and Sussex University Hospitals

A somewhat similar picture is seen in Brighton. Although its medical and elderly care wards did report norovirus outbreaks, there were also cases on specialist wards over the two seasons. This is shown in Figure 4.9. Non medical wards are labelled in red.

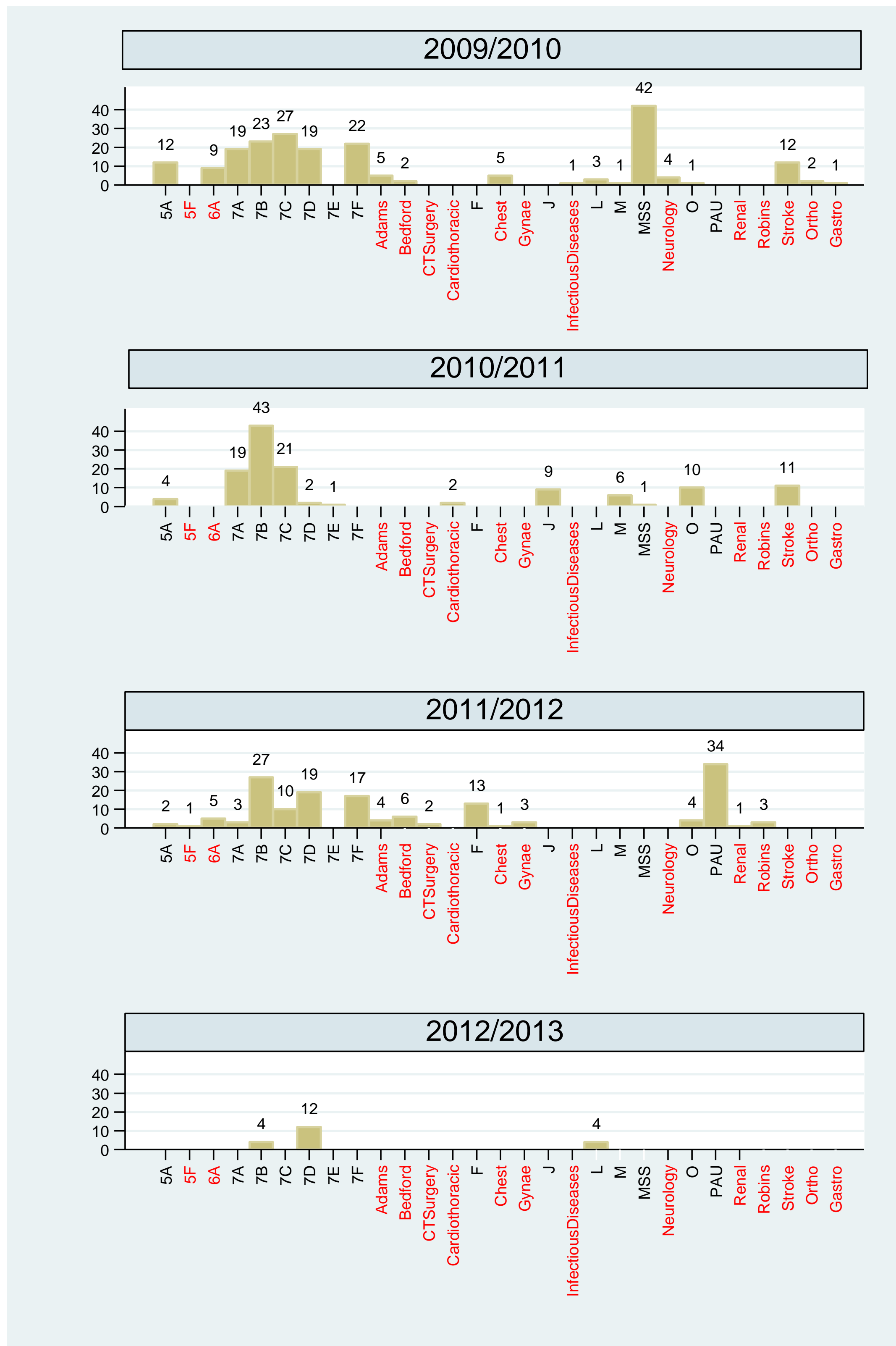


Figure 4.8 Distribution and frequency of all norovirus RT-PCR positive cases over four seasons and their respective wards. MSS (Medical Short Stay) ward became known as PAU (Post Admission Unit) after 2011, which was the largest ward in the OUH with the highest turnover of patients. Medical wards are labelled in black on the x axis, whilst non medical wards are labelled in red.

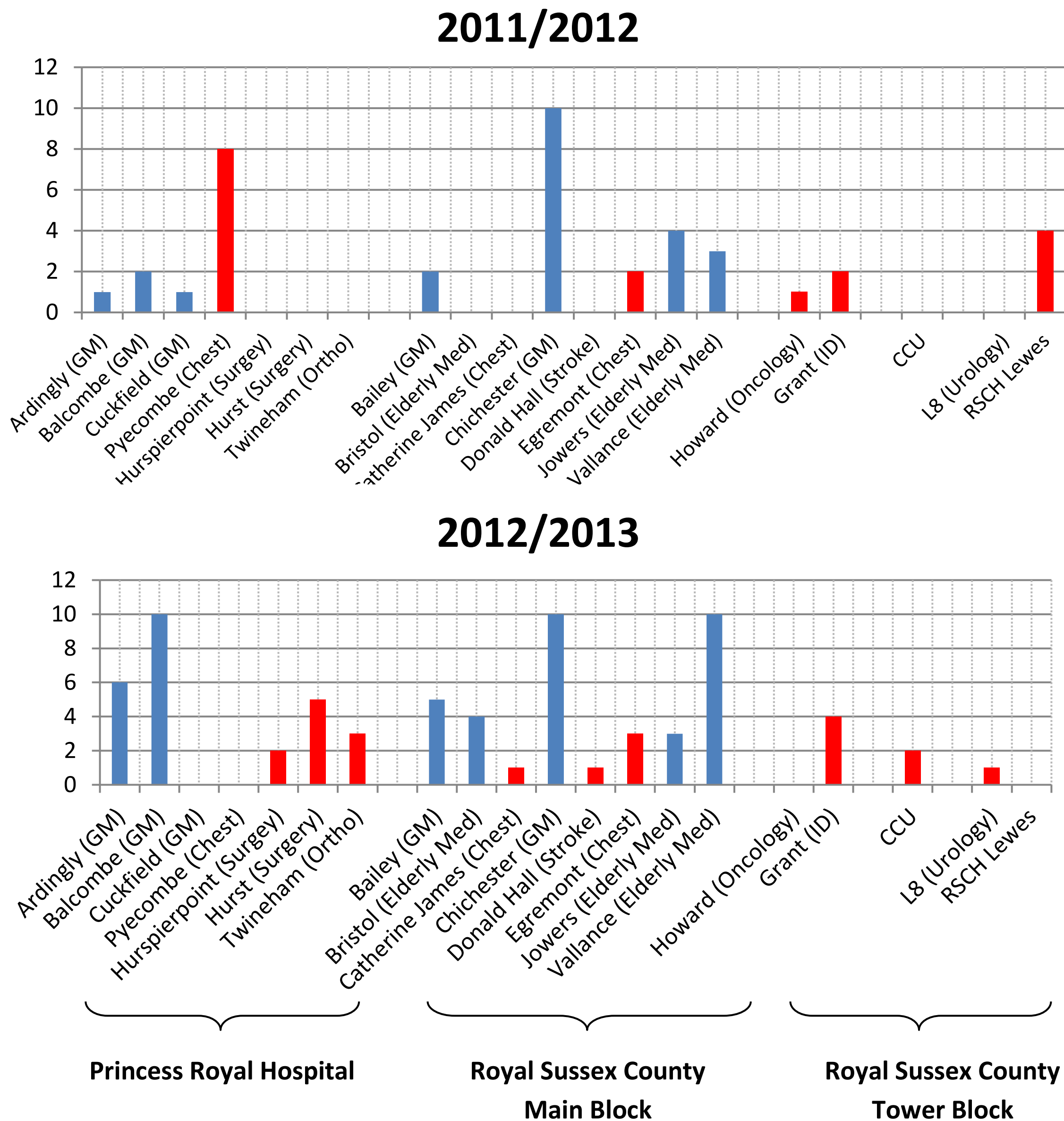


Figure 4.9 Distribution of norovirus RT-PCR positive patients over two seasons (non medical wards are denoted by red histogram bars). Wards have also been segregated by hospital and location.

4.3.7.3Leeds Teaching Hospitals

No information was available on the relationship between specialty and ward in Leeds.

4.3.8 Description of symptoms

Clinical details of vomiting and, or diarrhoea experienced by 469 Oxford patients for their duration of illness were available. Three clinical symptom variables were used in

the analysis; presence of vomiting only, diarrhoea only or presence of both diarrhoea and vomiting. Results for Oxford with RT-PCR positive samples are shown in Fig 4.10, RT-PCR negative patients in Figure 4.11 and patients not providing samples in Fig 4.12. The x-axis indicates the total duration of illness in days. Each horizontal line represents a patient. Each line has been ordered and stacked on top of each other. Missing data is shown in black. The plot shows the most common symptom was 'diarrhoea only' (marked in green). A similar picture is seen with the RT-PCR negative cohort of patients (Fig 4.11). Vomiting, either alone (in orange), or as a combination with diarrhoea (blue), tended to occur earlier in the illness (Odds Ratio for diarrhoea versus no diarrhoea: 0.816 (0.77-0.89) $p<0.001$ for every extra day of illness).

Figure 4.12 illustrates the symptomology in those patients who were unable to provide a stool sample. This shows a predominance of patients who exhibited vomiting only throughout their illness (shown by orange bars) explaining the absence of faecal samples for norovirus testing.

Patients from regions outside Oxford show a similar pattern of symptoms to the Oxford patients. This is illustrated in Figure 4.13 which confirms a similar proportion of symptoms and duration to the OUH.

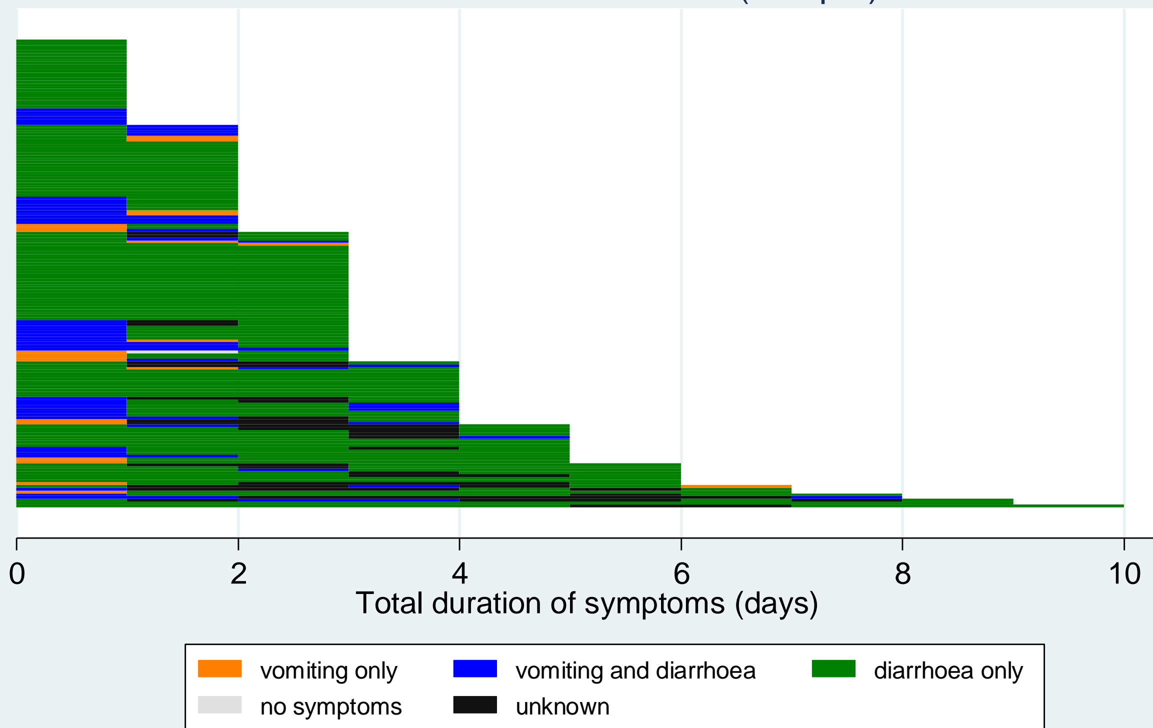


Figure 4.10 Plot of duration of illness in days and symptom description (RT-PCR positive). There are a total of 170 horizontal lines stacked on top of each other which represent individual patients. The length of this line determines their illness duration. Green colours indicate diarrhoea only, orange for vomiting only, whilst blue suggests a combination of diarrhoea and vomiting. Black lines suggest unknown data and grey bars are for no symptoms reported.

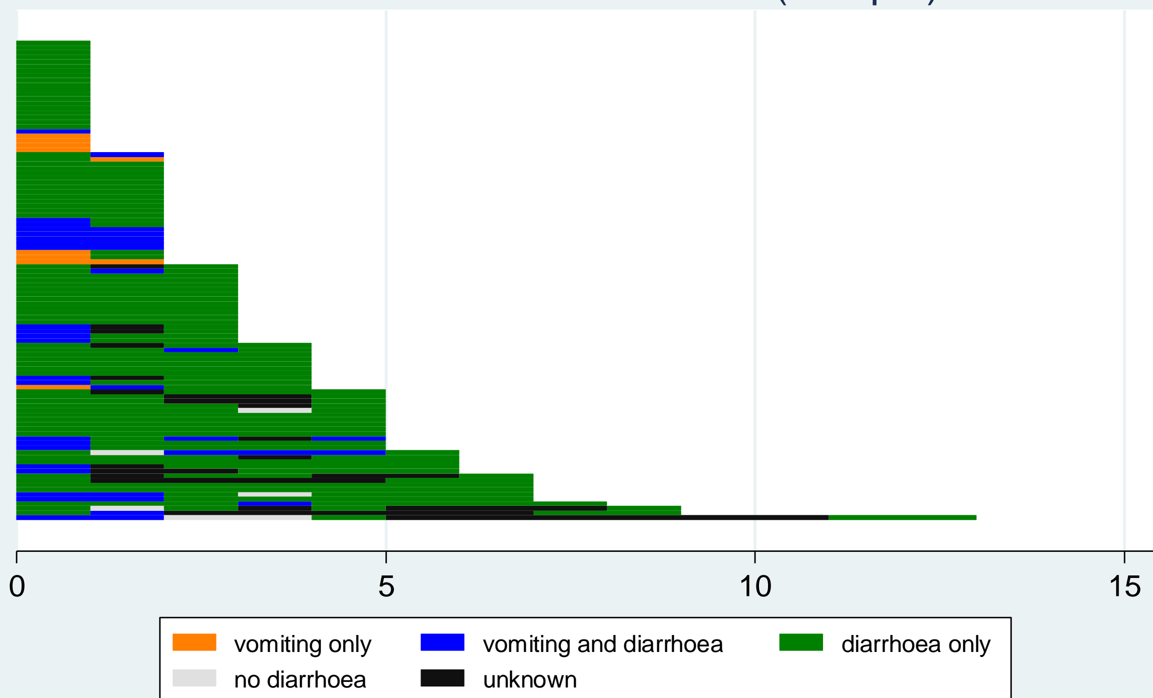


Figure 4.11 Plot of duration of illness in days and symptom description (RT-PCR negative). 103 horizontal lines stacked on top of each other indicating individual patients.

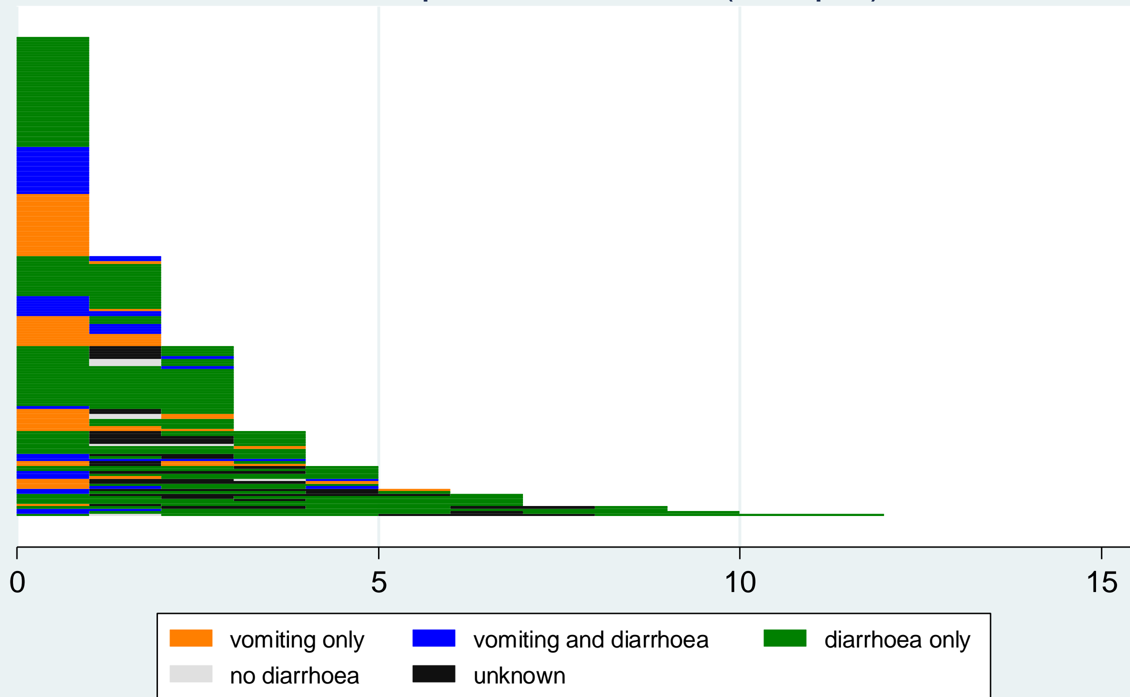


Figure 4.12 Plot of duration of illness in days and symptom description from symptomatic patients with no sample received. There are a total of 161 horizontal lines stacked on top of each other indicating individual patients. Compared to the RT-PCR tested cohorts, more patients exhibited vomiting in combination with diarrhoea or vomiting alone.

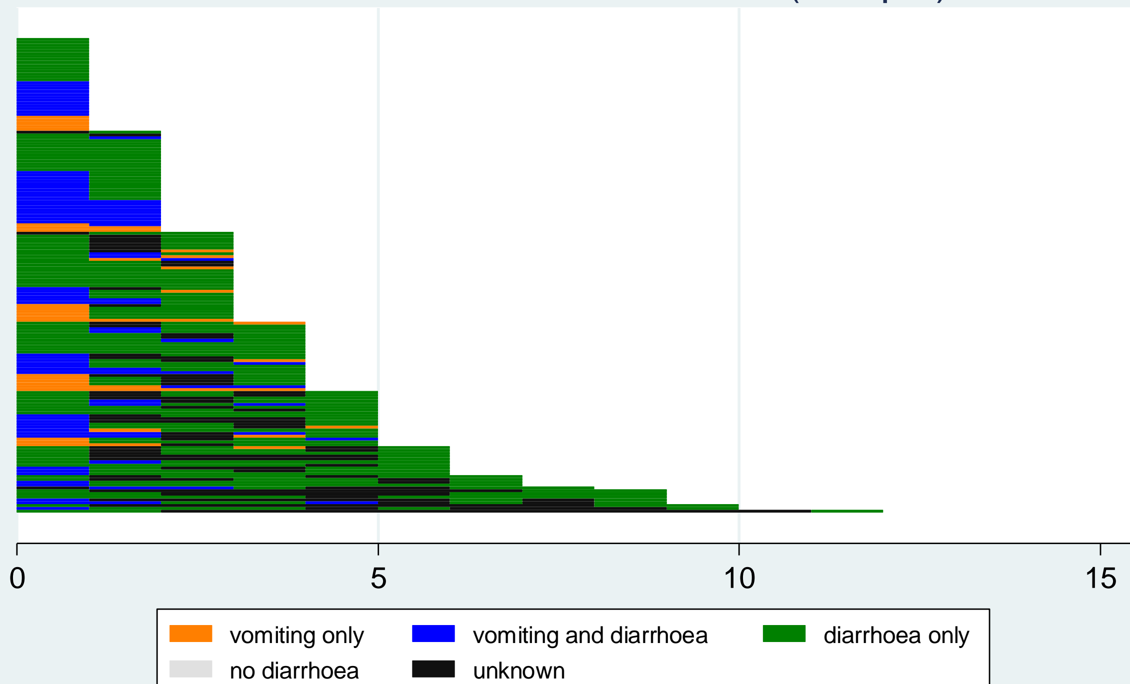


Figure 4.13 Plot of duration of illness in days and symptom description (RT-PCR positive) from non-Oxford samples. A total of 164 horizontal lines stacked on top of each other indicating individual patients. This shows a very similar picture to the Oxford cohort of samples.

The proportion of patients with vomiting (with or without diarrhoea) according to day of illness is shown in Table 4.1. This shows that vomiting occurs earlier in illness in all symptomatic patients.

Day of illness	PCR +	PCR-	No sample sent
1	32.3% (55/179)	29.1% (30/103)	47.4% (91/192)
2	17.3% (24/139)	17.8% (14/79)	20.2% (21/104)
3	6% (6/100)	5.5% (3/55)	11.8% (8/68)
4	13.2% (7/53)	5.2% (2/38)	8.9% (3/34)
5+	4.8% (3/63)	3.1% (2/64)	7.7% (4/52)
Odds Ratio (95% CI)* Change in Proportion vomiting for additional day of illness	0.59 (0.48-0.73) P<0.001	0..51 (0.38-0.69) P<0.001	0.48 (0.30-0.62) P<0.001

Table 4.1 Proportion of patients with vomiting according to days of illness (Oxford only). The denominator value in brackets indicate the total number of symptom episodes dependent on day of illness.

The proportion of diarrhoea for each day of illness is shown in table 4.2.

Day of illness	PCR +	PCR-	No sample sent
1	92.4% (157/170)	92.2% (95/103)	72.4&% (139/192)
2	84.9% (118/139)	82.3% (65/79)	63.4% (66/104)
3	84.0% (84/100)	89.1% (49/55)	73.5% (50/68)
4	79.2% (42/53)	71.1% (27/38)	70.6% (24/34)
5+	77.8% (49/63)	76.5% (49/64)	76.9% (40/52)
Odds Ratio (95% CI)* Change in Proportion vomiting for additional day of illness	0.82 (0.72-0.94) P<0.003	0.81 (0.71-0.91) P<0.001	1.05 (0.94-1.18) P=0.361

Table 4.2 Proportion of patients with diarrhoea according to day of illness (Oxford only). The denominator value in brackets indicate the total number of symptom episodes dependent on day of illness.

This table, which depicts diarrhoea according to day of illness, confirms that irrespective of PCR result or whether a sample was sent or not, diarrhoea was a very common symptom amongst all symptomatic patients in Oxford over the four seasons. Nearly 93% of patients who were RT-PCR positive for norovirus had diarrhoea on the first day of illness.

Table 4.3 summarises the total percentage of symptomatic patients who experienced any vomiting or diarrhoea during their course of illness in Oxford.

	RT-PCR +	RT-PCR -	No sample
Vomiting	41.8% (71/170)	35.9% (37/103)	50.5% (97/192)
Diarrhoea	100% (170/170)	95.2% (98/103)	81.8% (157/192)

Table 4.3 Percentage of patients exhibiting symptoms dependent on RT-PCR result

This shows that all patients who were RT-PCR positive exhibited diarrhoea during their illness. It also shows that only 41.8% of RT-PCR positive patients exhibited symptoms of vomiting during their total illness. This is less than the criteria as suggested by Kaplan and which is used globally to distinguish viral gastroenteritis outbreaks (more than 50% of patients vomiting within outbreaks) [18].

4.3.9 Duration of illness

4.3.9.1 OUH data

Patients who did not provide a stool sample had experienced a mean duration of illness of 2.3 days compared to 3.2 days in those who gave samples; median durations of 2 and 3 days respectively (log rank test for difference of overall duration of symptoms: $p<0.0001$). Information on the duration of illness has been categorised in a number of

ways. Figure 4.14 illustrates the duration of illness from symptomatic patients at the OUH who provided a sample over the past four seasons. Within this symptomatic patient cohort, the median duration of illness was 2 days, mean duration of 2.82 days, with a range between 1 to 12 days.

There was some evidence that the duration of symptoms was shorter after the first season (Odds Ration 0.71 (0.51-0.98) p=0.038). To confirm this, patients with RT-PCR negative samples were analyzed and each season was considered independently.

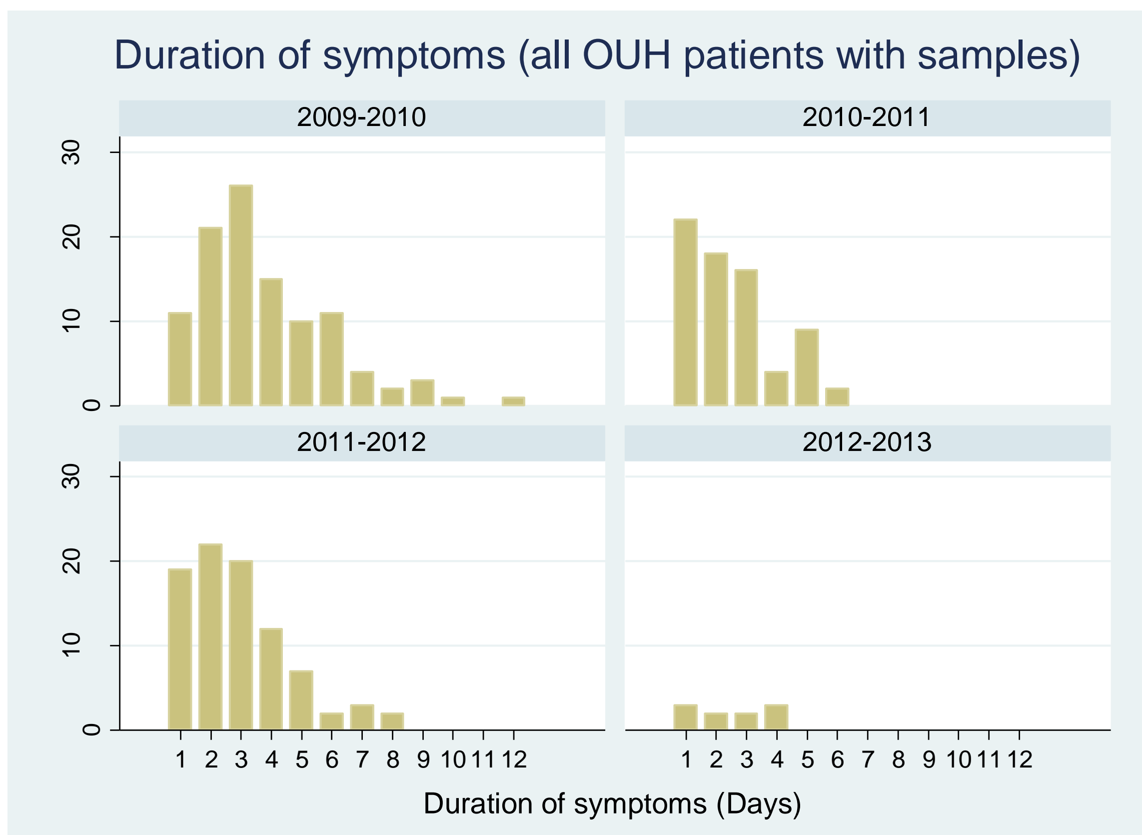


Figure 4.14 Histogram of the number of symptomatic patients who provided a sample and their total duration of symptoms (days) separated by season.

Overall, there was no difference between the symptom durations experienced by patients in seasons 2011-13 but the increased duration of symptoms in 2009-2010 was

seen overall (0.65 (0.50-0.83) p=0.001) and after adjustment for PCR results, gender and age (Global P value for difference between Seasons p=0.003).

Figure 4.15 details the total duration of symptoms for all four seasons when combined and differentiated by RT-PCR result.

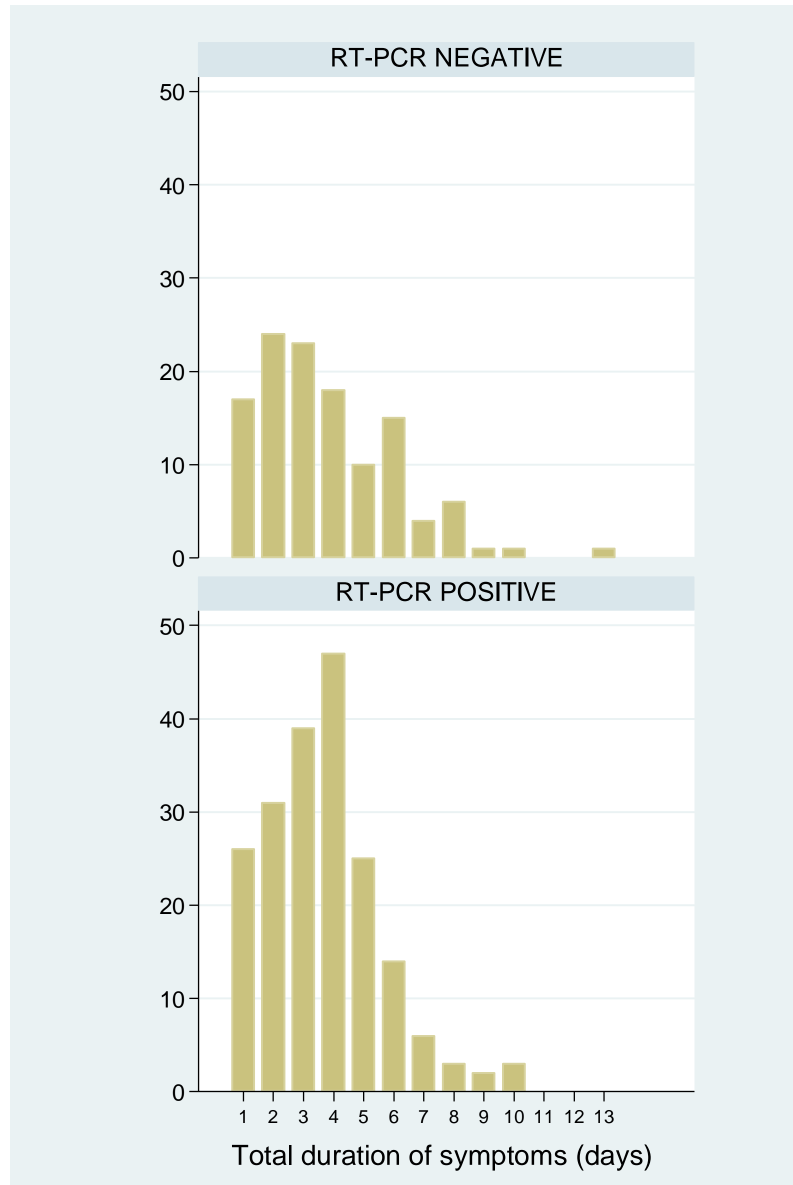


Figure 4.15 Histogram of number of symptomatic patients and their total duration of symptoms (days) separated by RT-PCR result

Amongst patients who provided samples, there was no difference between the duration of symptoms according to RT-PCR result (see Figure 4.15). Mean durations were 3.1

days in RT-PCR positive patients and 3.3 days in RT-PCR-negative patients both with a median duration of 3 days (log rank test for difference of overall duration of symptoms: p= 0.34).

Table 4.4 summarises the duration of illness over the four seasons.

Season	Sample size (n)	Mean (days)	Median (days)	Range (days)	Standard deviation
2009/2010	57	3.61	3	1-10	1.92
2010/2011	42	2.64	3	1-5	1.35
2011/2012	62	2.98	3	1-9	1.82
2012/2013	9	2.78	3	1-5	1.39
Combined	170	3.09	3	1-10	1.77

Table 4.4 Statistical analysis of all RT-PCR positive samples from the four seasons

4.3.9.2 Brighton and Leeds data

4.3.9.2.1 Brighton and Sussex University Hospitals

A total of 110 patients who experienced illness over two seasons were available from Sussex hospitals. Statistics for the two seasons are featured in table 4.5. Figure 4.16 shows the frequency of patients and their respective duration of illness. In summary this confirms that the median duration of illness in Brighton for the two years was 3 days; as observed in Oxford patients over the 4 seasons.

Season	n	Mean (days)	Median (days)	Range (days)	*s.d*
2011/2012	40	3.4	3	1-9	1.86
2012/2013	70	3.6	3	1-12	2.41
Total	110	3.54	3	1-12	2.26

Table 4.5 Statistical analysis of duration of illness

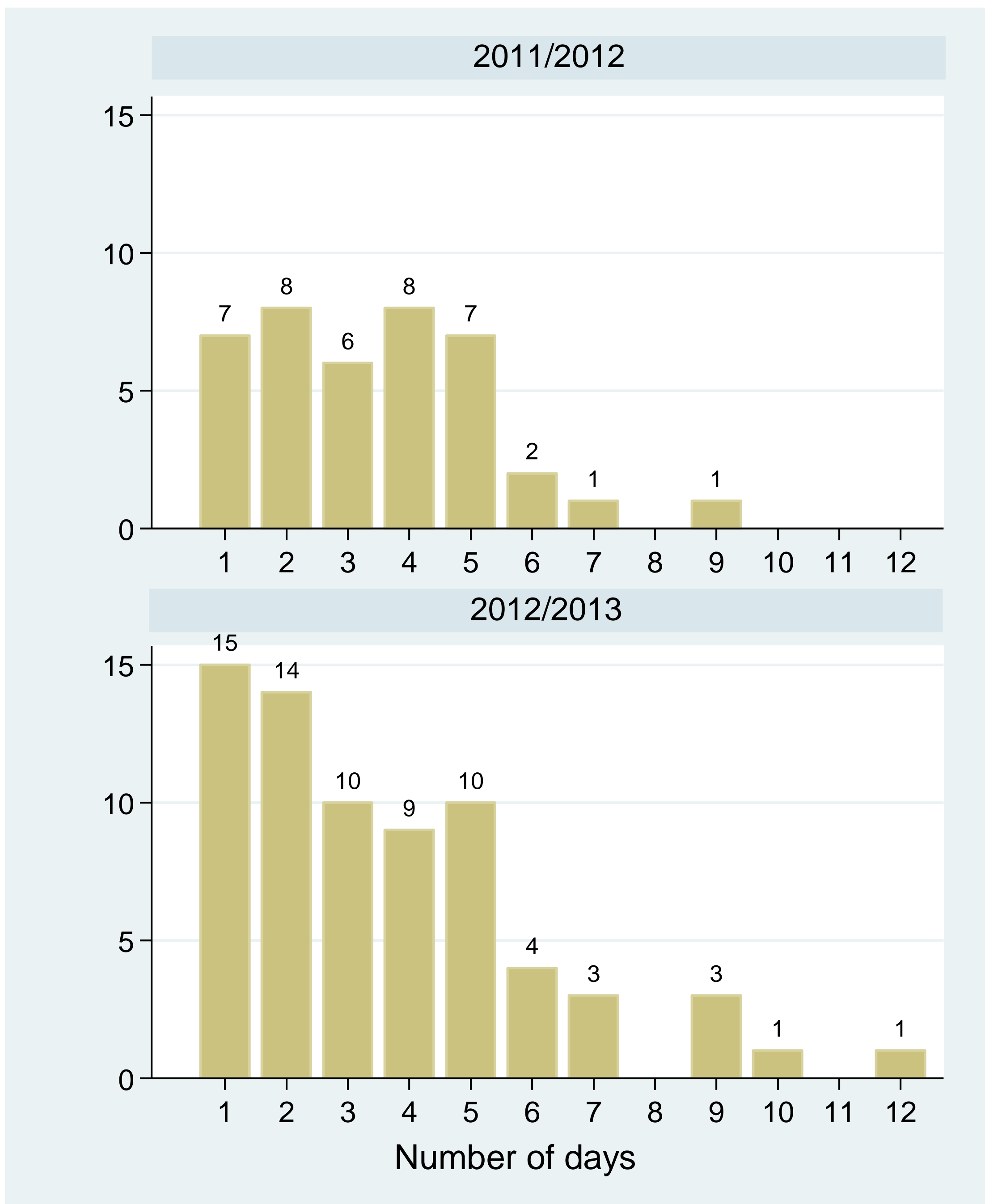


Figure 4.16 Frequency of total duration of symptoms in Brighton per RT-PCR +ve individual

4.3.9.2.2 Leeds Teaching Hospitals

Data were only available for the 2011/2012 season in Leeds. The range of symptomatic days experienced by patients in Leeds was 1-10 days (Figure 4.17). The mean duration of symptoms was 3.1 days and the median was 3 days. Again, this was very similar to data for patients from both Brighton and Oxford.

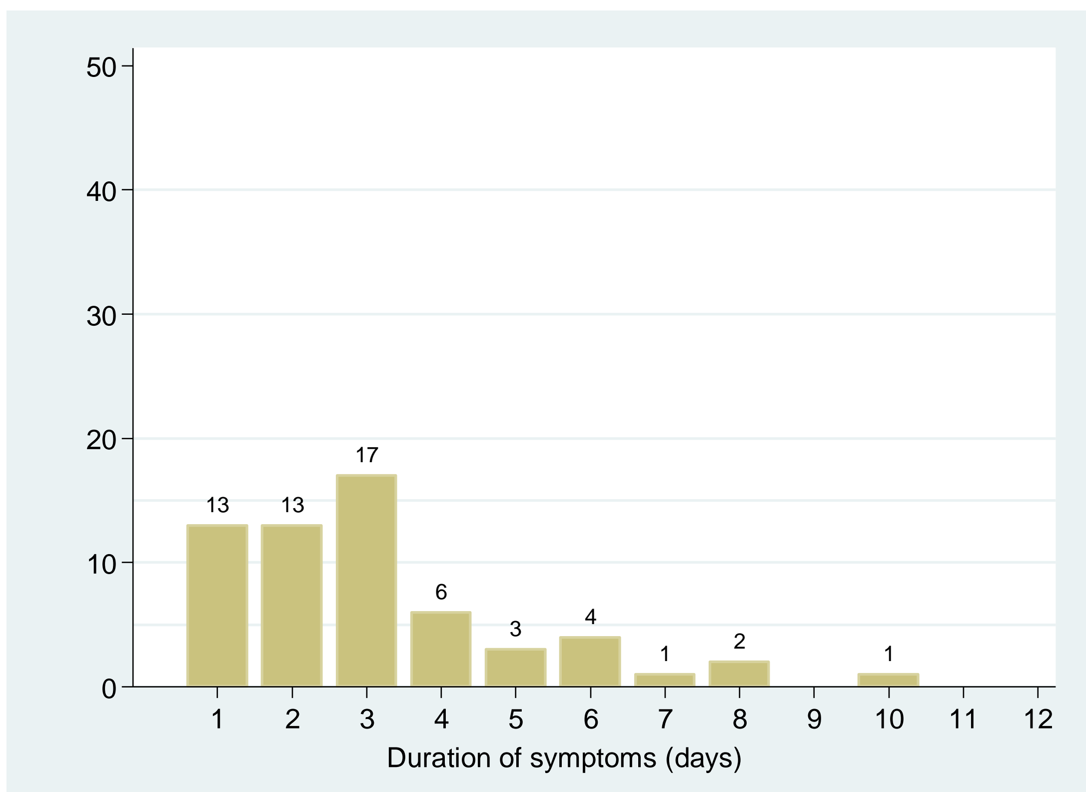


Figure 4.17 Duration of illness in RT-PCR +ve patients in Leeds 2011-2012

4.4 Discussion

The aim of this chapter was to describe the epidemiology of norovirus outbreaks in hospital wards over the past four seasons. In the first instance the data confirmed that norovirus affects all age groups within the healthcare environment, but the elderly population was identified as the most at risk over the duration of the study. The data also confirm that norovirus affects patients on multiple wards within a hospital, but that

the cases are not evenly distributed throughout the hospital. Recent cost analysis for nosocomial gastroenteritis in the UK has estimated the impact to be £635,000 per 1,000 beds, based on revenue lost from closure to new admissions and staff absences [112].

The Oxford data also suggests that norovirus affected females more than males over the past four seasons. This is confirmed with data from both Brighton and Leeds. Reasons for this are unknown. However, the statistical analysis showed that this change in gender ratio has not affected the number of samples received for testing nor its diagnostic RT-PCR result.

Vomiting (either alone or in combination with diarrhoea) occurred in only a minority of patients. This finding differs from one of the criteria used by Kaplan [18] (vomiting in more than half of affected patients) and has been evident in Oxford, Brighton and Leeds over the past four seasons. The use of Kaplan criteria is still advocated by the American CDC and Australian National health. Recent guidelines published by Public Health England do not recommend using these criteria as the Working Party felt that diarrhoea was the predominant feature seen during norovirus outbreaks, and that the use of median and mean incubation time or duration of illness suggested that the criteria can only be used retrospectively [209]. Several volunteer and epidemiological studies have been performed to gain insights into the course of diseases associated with norovirus infection [14, 71, 210]. These studies have been limited to focusing on healthy adult volunteer or at risk communities involved in outbreaks. However, there are, as yet, no studies that have assessed the utility of the Kaplan criteria in healthcare associated outbreaks of norovirus gastroenteritis. The present study supports the change in UK guidelines stating that vomiting may no longer be the predominant clinical feature of

symptomatic norovirus infection, as it once was [81]. It is unclear whether the changing prevalence of vomiting over time reflects the evolution of different norovirus strains [84].

The duration of symptoms, indicated by the epidemiological information, confirmed that the symptomatic infection is relatively short-lived, with mean and median durations of 3 days. The shorter duration of symptoms seen after the 2009/2010 season is unexplained. However it is tempting to speculate that this might be caused by increased immunity in the population.

There are several limitations to this analysis; the overall range in age of all patients admitted within the hospitals over the study period is unknown. This could potentially explain why the average age groups were higher than previously recorded [123]. In addition, the gender ratio of all admitted patient is unknown, hence it is unclear whether more female patients were admitted to hospital overall, which could potentially explain why more females were symptomatically affected than males over the four seasons.

The data presented in this chapter confirm that the Kaplan criteria [153] are not the optimal criteria to use for the early identification of a norovirus ward-based outbreak. Although it appears that the mean/median duration of symptoms do fall within the criteria, it may be inadvisable for infection control specialists to wait for this period of time to elapse, before making decisions aimed controlling an outbreak. Therefore the ability to rapidly confirm norovirus infections in the diagnostic laboratory, and to quickly elucidate transmission routes by employing both epidemiological and genomic data would be of immense practical benefit. Ways in which genomic data, complimented

with epidemiological data can potentially assist our understanding of norovirus transmission dynamics within the healthcare environment will be reviewed and discussed in Chapter 6.

Chapter 5. Description of outbreaks between wards

Chapter objective

To describe the characteristics of hospital ward outbreaks within the OUH between September 2009 to August 2013. To illustrate the number of symptomatic patients who were PCR tested and what proportions were PCR positive for norovirus.

5.1 Background

Hospital outbreaks of norovirus infection are seldom restricted to individual wards [211]. To achieve effective infection control, individual wards and occasionally entire hospitals are closed or restricted to new admissions, and elective surgeries are cancelled to prevent the continued introduction of new susceptibles to the clinical environment, which would likely prolong the outbreak. These measures have a severe impact on the normal running of the hospital [112]. This chapter details the chronological relationship of symptomatic cases within and among wards over the four seasons from 1st September 2009 to 31st August 2013.

In the present study, a norovirus outbreak was defined as two or more cases occurring within a functional unit (section 2.1.3). I have assigned the functional unit as cases meeting the definition within an individual ward. If after seven days, new cases occurred on a ward which previously had symptomatic cases, it was recorded as a new and independent outbreak.

5.2 Epidemiological data from Oxford

5.2.1 2009-2010 season

A total of 210 patients located on 19 wards within the OUH complained of diarrhoea, vomiting or a combination of both during this season. Complete epidemiological data

regarding illness description and duration were available for only 202 patients located in 16 of the wards. The distribution of cases from 15 wards over time is shown in Figure 5.1. The remaining ward (6A) outbreak occurred in late March 2010 and was temporally distinct from the other outbreaks.

Figure 5.1 (overleaf). Distribution of symptomatic cases by ward and time over a 2 month period at the OUH. x axis indicates time, y axis indicates number of cases. Three different sites were involved (JR Building 1, JR Building 2 (which comprises of the Neurology and Orthopaedic departments) and the Churchill Hospital). Each case was defined as such on the first day of recorded symptoms. Laboratory confirmed RT-PCR cases are highlighted in blue, RT-PCR negative cases are in red, whilst cases where no sample was received cases are indicated in duck brown. Confirmed cases were seen in the first 14 days of the two month period which affected all three sites.

2009-2010 OUH Outbreaks

JR Building 1: 5A, 5C, 7A, 7B, 7C, 7D, 7F, MSS, Adams, Bedfor, Stroke

JR Building 2: Neuro

Horton: Ortho

Churchill Hospital: M, Chest

10
5
0

29/12/2009 31/12/2009 02/01/2010 04/01/2010 06/01/2010 08/01/2010 10/01/2010 12/01/2010 14/01/2010 16/01/2010 18/01/2010 20/01/2010 22/01/2010 24/01/2010 26/01/2010 28/01/2010 30/01/2010 01/02/2010 03/02/2010 05/02/2010 07/02/2010 09/02/2010 11/02/2010 13/02/2010 15/02/2010 17/02/2010 19/02/2010

The histogram depicted in Figure 5.1 documented norovirus cases chronologically by ward. All RT-PCR confirmed cases are highlighted in blue. RT-PCR negative cases are highlighted in red, whilst cases lacking a sample are shown in brown. Only the first day of symptoms is represented. Between 28th December 2009 and 14th January 2010, symptomatic cases occurred on multiple wards within the trust. Cases occurred concurrently on neighbouring wards on the same level (exemplified by wards 7A, 7B, 7C, 7D and 7F), but also in wards on different levels, different buildings as well as on a different site (Chest Ward at the Churchill Hospital).

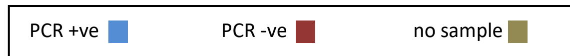
A total of 23 distinct outbreaks involving 202 patients were identified on 16 wards. The mean duration of each outbreak was 7.9 days, median 7 days, and range 2-17 days. Three wards experienced a second outbreak (Wards 5A, 7A, and 7B), whilst two general medical wards (7F and MSS) experienced 3 outbreaks each, within three months of the first outbreak.

5.2.2 2010/2011 season

A total of 14 outbreaks occurred during this season, affecting 129 patients and spanning 12 wards. 104 patients from 8 wards had complete epidemiological data for analysis. Two wards sustained a second outbreak (Wards 7A and 7C), whilst Ward 7B had 3 outbreaks. The mean outbreak duration was 4.9 days, with a median of 4.5 days and a range of 1-10 days. Figure 5.2 represents graphically, the ward outbreaks over time. There were five clusters of temporally (not geographically) clustered cases. Time cluster 1 (Figure 5.2a) in early February 2011, affected four wards at around the same time (Wards 7A, 7B, 7C, and O ward at the Horton Hospital in Banbury). Cluster 2 also affected 4 wards in late March 2011 (Ward 7A, 7B, 7C and Stroke), whilst cluster 3

affected M ward at the Horton in April. Clusters 4 and 5 affected multiple wards in June and July, respectively.

2010-2011 OUH Outbreaks February 2011



JR Building 1: 7A, 7B, 7C, Stroke

JR Building 2: CTW

Horton Hospital: O

Time Cluster 1

Time Cluster 2

Figure 5.2a Distribution of cases in the OUH during February and March 2011. Two distinct temporally related clusters occur during this period (Time Clusters 1 and 2). O ward is at the Horton general Hospital, 20 miles away from the JR and Churchill hospitals. It appears that cases were found simultaneously with cases at the JR. In the absence of genotyping data (ideally highly discriminatory WGS), the possibility that these outbreaks are linked cannot be confirmed.

2010-2011 cont'd

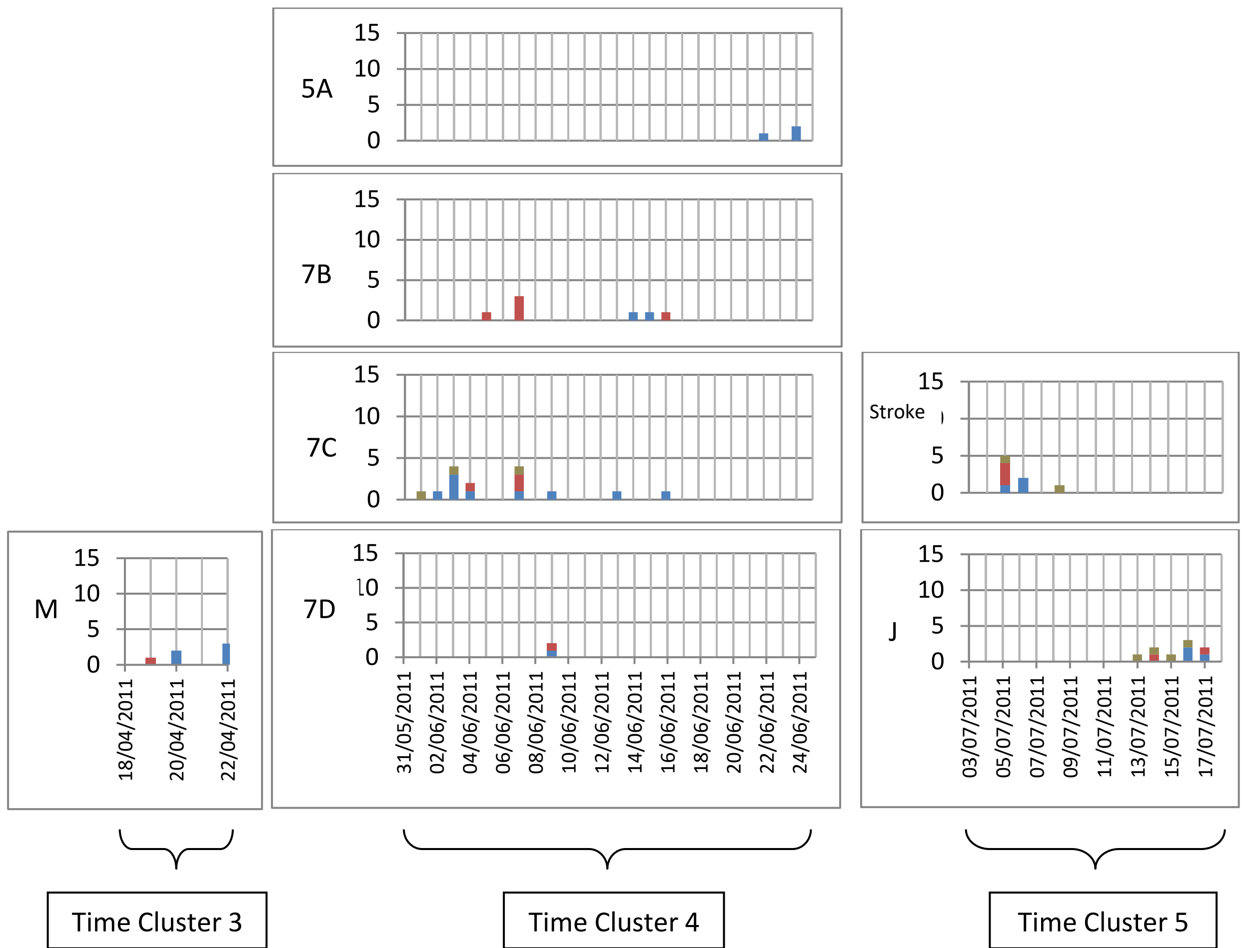


Figure 5.2b Distribution of cases in the OUH between 2010-2011 Clusters 3-5(April to July). Both M (Time cluster 3) and J (Time cluster 5) wards are at the Horton general hospital, 20 miles from the JR.

5.2.3 2011/2012 season

A total of 155 patients had symptoms of gastroenteritis, spanning 18 wards and accounting for 19 outbreaks. Full epidemiological information was available for 145 patients, extending over 16 wards. The mean duration of outbreaks was 7.1 days, median 5 days (range of 2-22 days). Figure 5.3a and 5.3b illustrate the temporal distribution of cases during this season. The majority of cases occurred in February 2012, whilst some isolated confirmed cases occurred at the Horton General Hospital in the summer months (May and July). Three JR wards (Ward 7B, 7C and 7F) experienced two waves of norovirus outbreaks. RT-PCR positive cases were found in neighbouring buildings within the JR as well as the Churchill hospital.

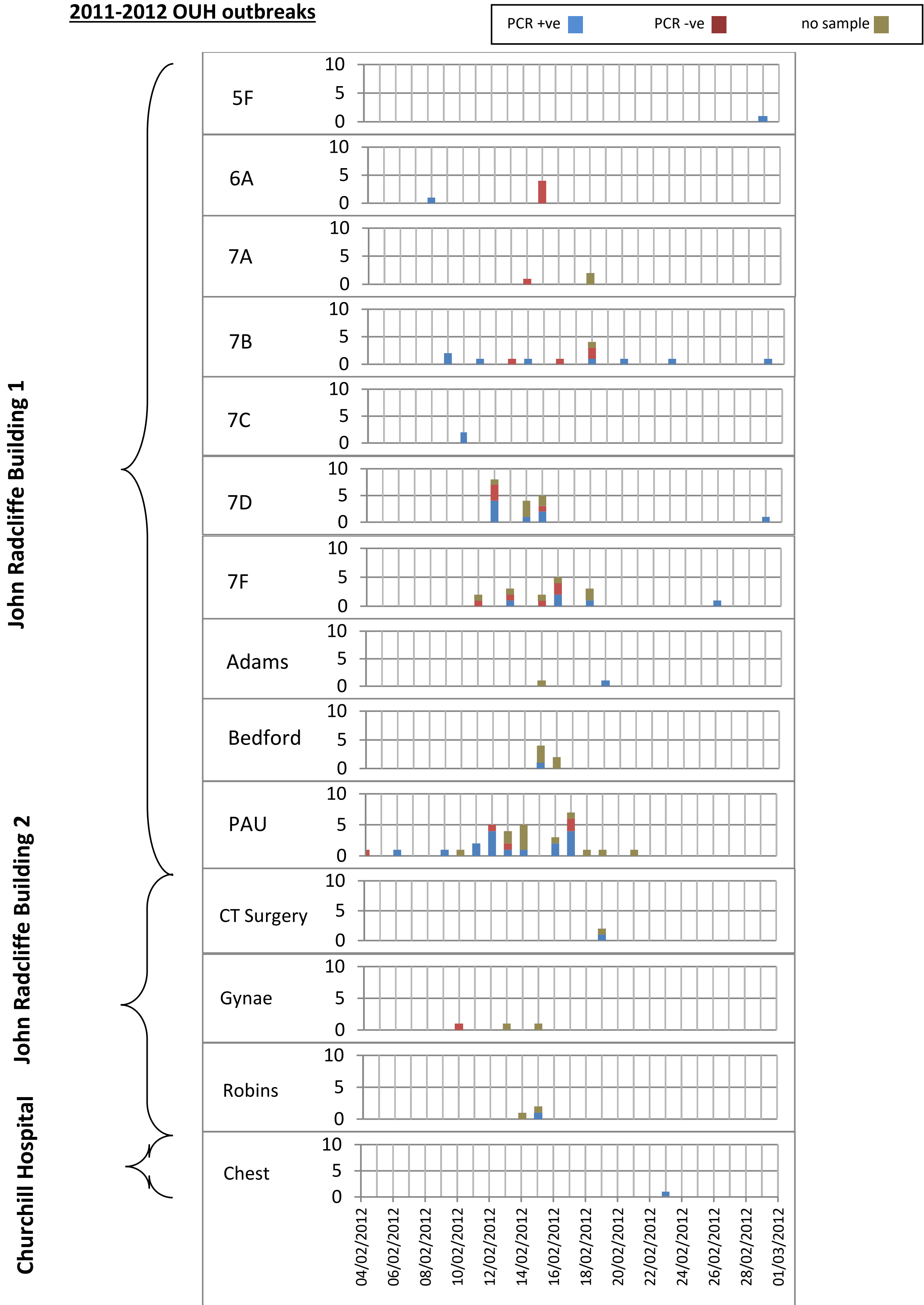


Figure 5.3a Distribution of cases according to ward during 2011-2012

2011-2012 continued

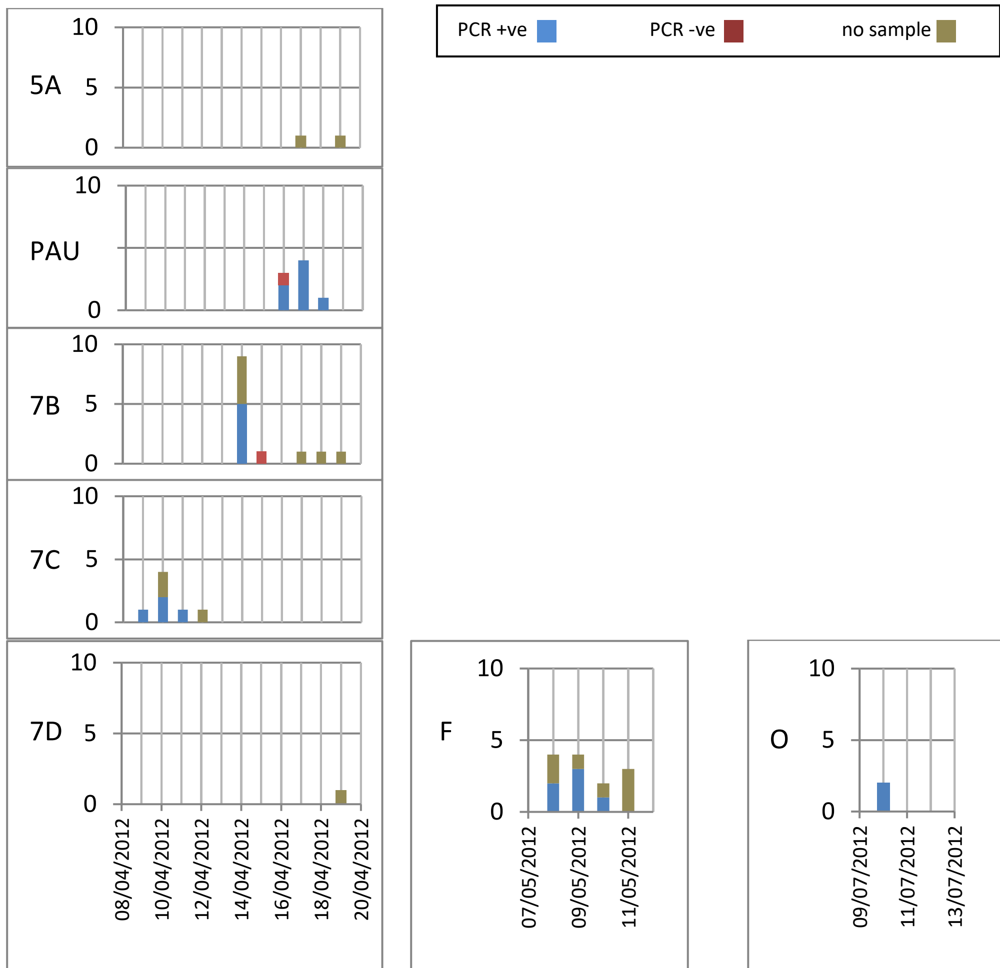


Figure 5.3b Distribution of cases according to ward during 2011-2012

5.2.4 2012/2013 season

Only twenty patients located in three wards had symptoms suggesting norovirus infection, (Figure 5.4). The mean duration of the three outbreaks was 7 days, median 7 days and the range 6-8 days. Of the nine samples that were PCR positive, these covered two wards (Wards 7D and L ward at the Horton Hospital). A total of two confirmed norovirus outbreaks were recorded in 2012/2013.

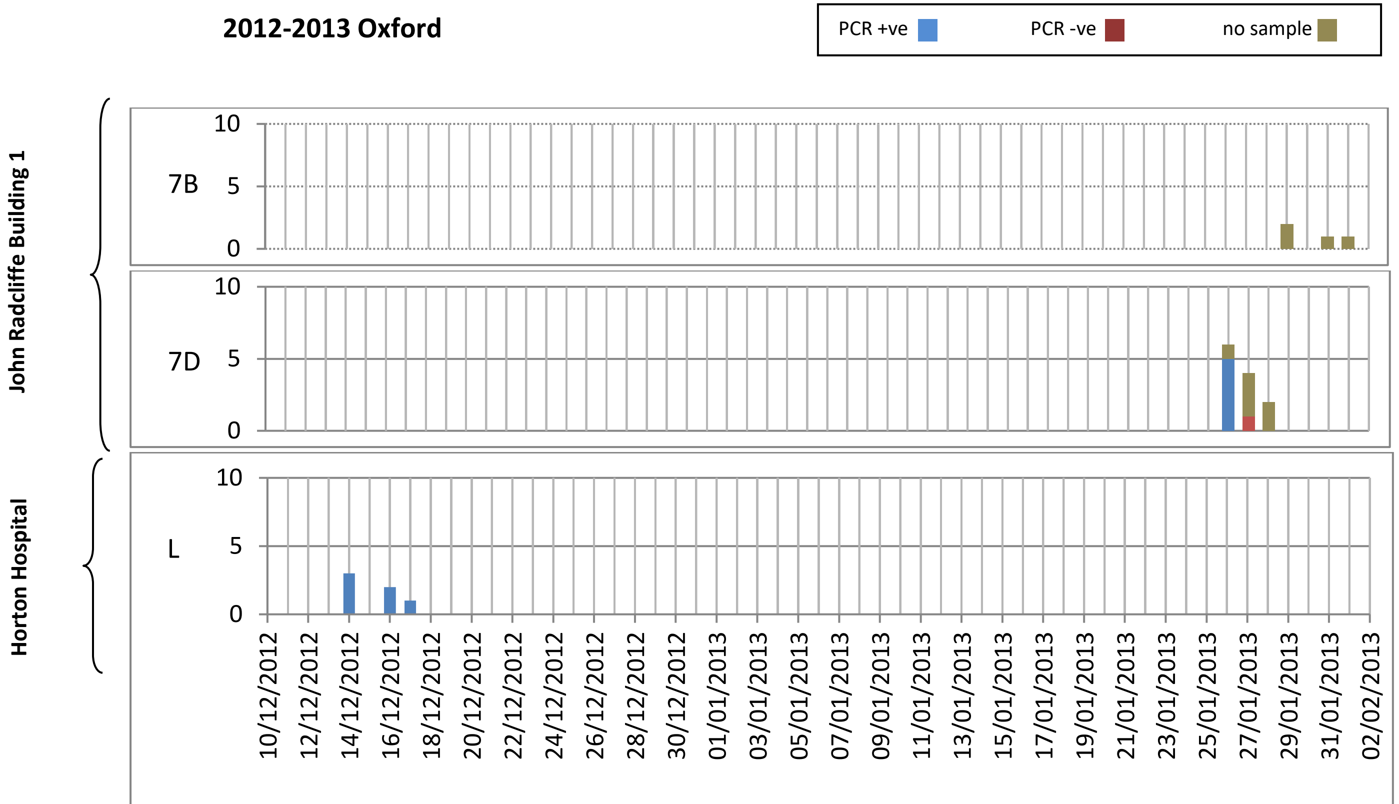


Figure 5.4 Distribution of ward cases in 2012-2013 at the OUH

5.3 Outbreaks in Brighton and Leeds

Similar epidemiological information was collected from the Brighton and Sussex County Hospitals as well as the Leeds Teaching Hospitals NHS Trust. The epidemiological histograms are supplied in Appendix 4. Epidemiological data were only available from patients providing an RT-PCR positive sample, rather than from all symptomatic cases as in Oxford, making analysis of ward-defined outbreaks difficult to analyse from this cohort.

5.4 Discussion

Over the course of the four seasons analysed, multiple wards within the hospitals studied were affected by norovirus. Although weekly national reports of norovirus are

available from Public Health England, descriptions of nosocomial hospital outbreaks on a ward scale are lacking. This chapter has illustrated the intensity and frequency of norovirus outbreaks over four seasons in Oxford.

The OUH infection control team follow recommendations recently published by Public Health England regarding closure of wards during gastroenteritis outbreaks [209]. For the past 5 years, they have utilised a pragmatic, escalatory system of patient isolation using side rooms and cohorting ward bays without compromising patients care. Only rarely have whole wards been closed. This is quite different to previous guidance of traditional complete ward closure [145].

Complimenting epidemiological data with genomic data rapidly is one key step in minimising essential service disruptions and maximising the ability of organisations to deliver care to patients safely and effectively. Current infection control guidelines suggest laboratory confirmation of norovirus infection and the collection of epidemiological data, which takes time and effort. Not all hospitals have in-house laboratory testing available for norovirus. Therefore, usually only a restricted number of suspected samples from outbreaks will be sent to reference laboratories for confirmation. This reduces the ability to retrospectively confirm or refute true infection or transmission in affected ward patients. Outbreaks have usually ended when results become available from the reference laboratory, indicating that testing merely becomes an effective national surveillance tool. In addition, this data does not give any discrimination as to the path of transmission within and between wards.

As illustrated in Chapter 3, whole genome sequencing has the potential to give a detailed resolution of the virus within the patient's faecal sample. Differences between

these genomes can be analysed and transmission directionality may be tracked. The application of WGS (in combination with epidemiological data) to the question of norovirus transmission routes will be described in the next chapter.

Chapter 6 Understanding norovirus transmission during hospital outbreaks by combining genomic and epidemiological data

Chapter objectives

The overarching aim was to understand transmission during norovirus outbreaks. This would be informed by combining two data sources which, prior to the present study, have only been considered independently; firstly viral whole genome sequences and secondly epidemiological (i.e. temporal and geographic) data from infected patients. These data would allow the determination over time of (i) the phylogenetic relationships among viruses collected in a single hospital trust; (ii) the genetic relationship between the hospital viruses and those circulating concurrently in the local community; (iii) the relationships of the geographically localised viruses in (i) and (ii) to those circulating in unconnected clinical environments throughout the UK.

6.1 Background

Previous studies aiming to understand the nosocomial transmission of norovirus have frequently employed an approach in which epidemiological data were combined with nucleotide sequence data derived from only a small part of the genome, typically a capsid or RdRP fragment [1-4]. Epidemiological links among cases were then considered to be 'confirmed' on the basis of short regions of nucleotide sequence identity. Although noroviruses show extreme levels of genetic diversity overall [66, 212], healthcare associated outbreaks were almost always caused by extremely closely related GII.4 viruses [212]. For this reason, studies of norovirus molecular epidemiology which use short genome fragments were compromised by the fact that point mutations or recombination events outside the region sequenced were overlooked. Such studies therefore lacked discriminatory power. Links among cases can be incorrectly

‘confirmed’ when the cases could in fact be epidemiologically independent. Here, the aim was to use whole genome sequencing to enhance the level of discrimination among closely related viruses, in an attempt to improve our understanding of the dynamics of norovirus transmission in the healthcare environment.

The potential for norovirus whole genome sequences to inform our understanding of virus transmission was investigated on three levels; locally, regionally and nationally. The extent of local norovirus transmission was studied within a single UK healthcare environment, which was dynamic in terms of staff and patient movements. The healthcare environment used was the OUH NHS Trust and transmission was investigated within and among wards. A separate investigation of local transmission (in either direction) between this healthcare environment and the community it serves was performed by comparing the norovirus genomes from hospital patients with noroviruses circulating among the Oxfordshire population (~635,000 people, 756 square miles). The possibility that detectable national healthcare-associated transmission links exist was investigated using noroviruses from geographically independent hospitals, predominantly in Leeds, Brighton, East Surrey, and samples from the South East England reference laboratory. The longitudinal nature of this study, spanning four norovirus seasons from September 2009 to August 2013 also provided the possibility of examining norovirus genome evolution over time and concurrent epidemiological changes.

6.2 Virus Genomes Sequenced

The total number of norovirus genome sequences determined successfully using both the Illumina MiSeq and HiSeq platforms is summarised (Figure 6.1). In brief, 424 of the 584 (73%) norovirus positive clinical samples yielded a single contiguous norovirus sequence (Figure 6.1). Of these, 337 (79%) were over 7,500 nt in length (>99% of the genome) and were used to address the aims of the study.

The remaining 160 of the 584 samples (27%) yielded multiple short non-contiguous norovirus sequences. Between one and eleven contigs were identified per sample, with a combined length of 1,833 nt to 6,800 nt. As expected, all but a small minority of genomes sequenced were closely related variants of the predominant healthcare-associated GII.4 strain. Mapping these multiple small contiguous sequences to a reference GII.4 genome, as previously described (chapter 2.7.2 (Genbank GQ845367, Norovirus Hu/GII.4/Orange/NSW001P/2008/AU; Eden et al., 2010)) revealed that the average gap size was 330 nt. These gaps were positioned randomly and hampered comparative genomic analyses, hence the non-contiguous genomes were excluded from further study.

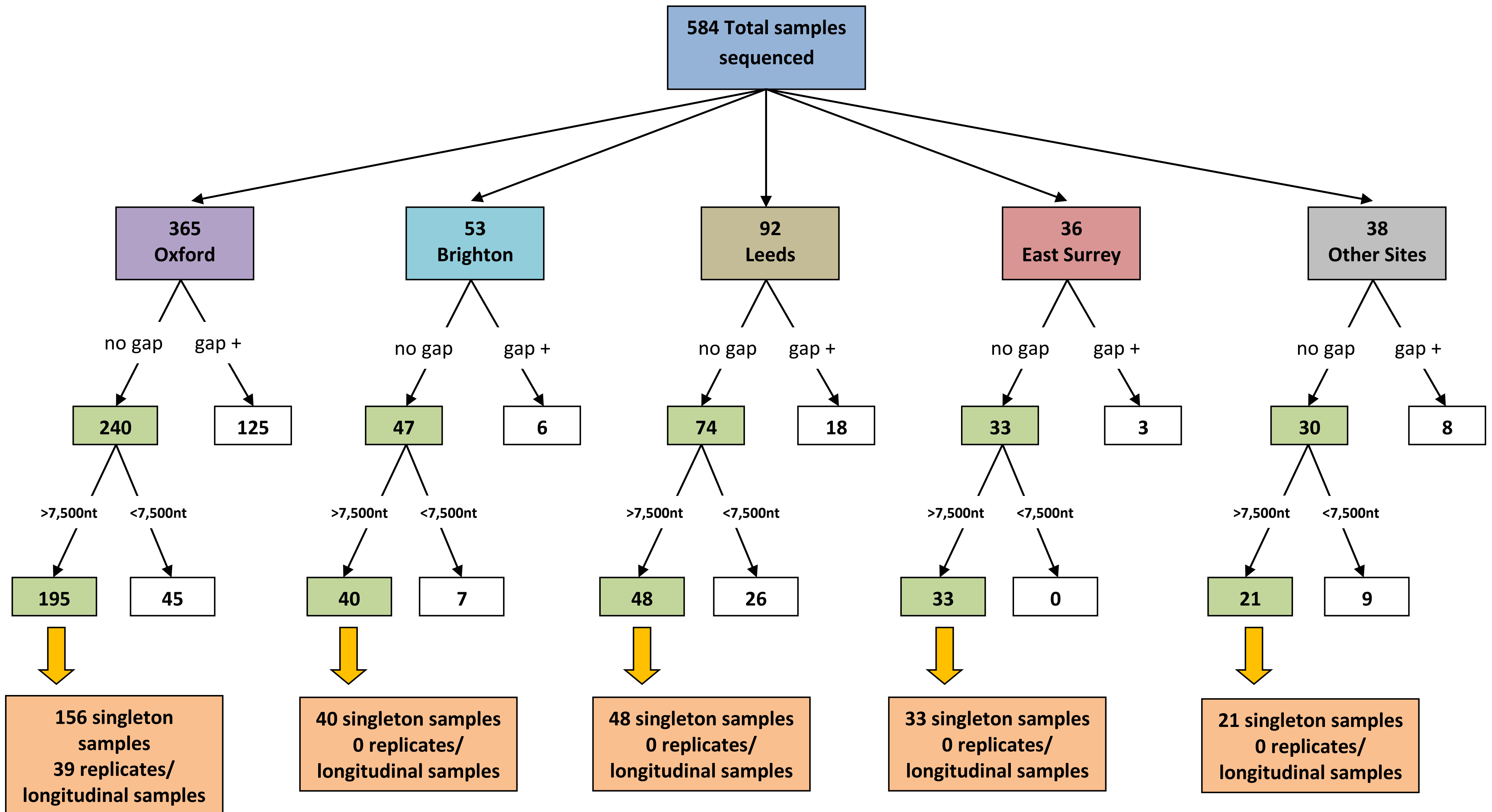


Figure 6.1 Flow diagram summarising the total number of noroviruses sequenced, together with their geographic location. If more than one contig was produced by the de novo assembler, then this genome was excluded from further analysis (indicated by gap+). Only genomes that were longer than 7,500nt (99% of the total genome) were used in the analysis.

6.3 Oxfordshire Norovirus Genomes

A total of 145 hospital outbreak-associated norovirus genomes and 47 community genomes were available from Oxfordshire patients experiencing norovirus infections between September 2009 and August 2013 (Table 6.1). These included thirteen genomes of viruses collected longitudinally from seven symptomatic patients during the course of a single norovirus episode, as well as 19 technical replicates used as a measure of quality control against its "sister" genome sequence. Limited epidemiological data were available for the community norovirus infections. However, the date of collection was used to inform relationships between the hospital and community infections during specific outbreak periods.

Hospital Inpatient Noroviruses				
Season	**Noroviruses representing individual cases (n)**	**Noroviruses collected longitudinally (n) [a]**	**Replicates (n) [b]**	**Total (n)**
2009-2010	22	0	7	**29**
2010-2011	35	0	9	**44**
2011-2012	48	13	3	**64**
2012-2013	8	0	0	**8**
Total	**113**	**13**	**19**	**145**
Community Noroviruses				
Season	**Noroviruses representing individual cases (n)**	**Noroviruses collected longitudinally (n) [a]**	**Replicates (n) [b]**	**Total (n)**
2009-2010	5	0	1	**6**
2010-2011	8	0	0	**8**
2011-2012	24	4	2	**30**
2012-2013	3	0	0	**3**
Total	**40**	**4**	**3**	**47**

Table 6.1 Oxfordshire norovirus genomes by season.

[a] 'Noroviruses collected longitudinally' refers to viruses derived from patients supplying more than one sample during an outbreak.

[b] 'Replicates' refers to virus derived from an individual specimen which underwent sequencing more than once. Only the longest contig produced was included in further analyses.

6.4 Analysis of OUH norovirus genomes

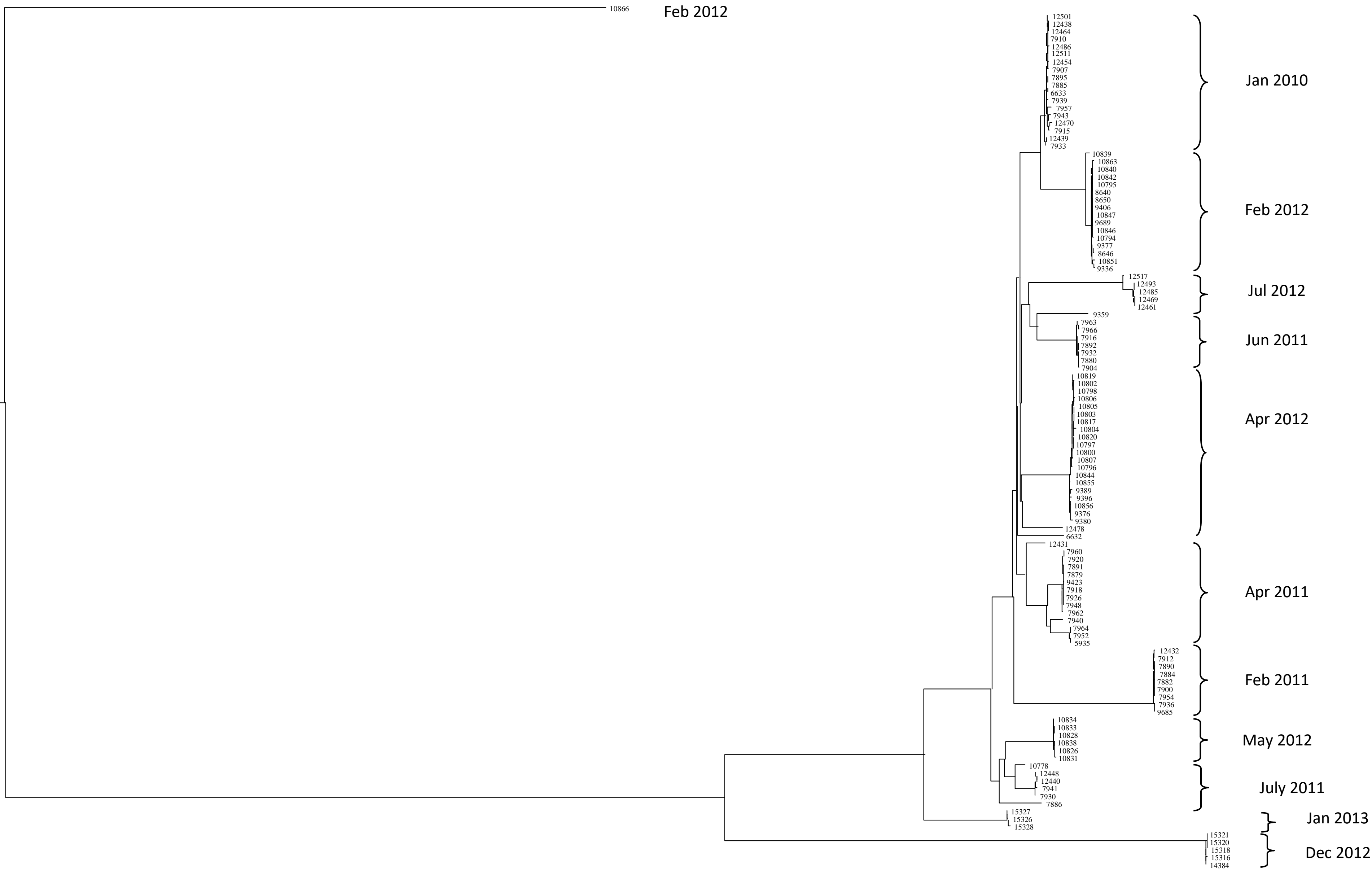
A total of 113 individual norovirus genomes collected over four seasons were used for analysis. The breakdown of genomes by season is indicated in Table 6.2.

Season	Noroviruses representing individual cases (n)
2009-2010	22
2010-2011	35
2011-2012	48
2012-2013	8
Total	**113**

Table 6.2 Table representing number of OUH individual cases used in analysis

The phylogenetic relationship between the 113 genomes was investigated by constructing a maximum likelihood tree (Figure 6.2). Groups, or clusters of closely related genomes were delineated. The tree was annotated with the date each virus had been collected, revealing that the clusters were differentiated by time. Clusters of genomes collected from ward outbreaks which were separated by as little as one month showed high genetic distances, suggesting that re-introductions of previously circulating strains (for instance by persistent environmental contamination) was rare within this cohort. Genetically divergent norovirus clusters were identified that co-circulated, indicating multiple, concurrent independent introductions into the hospital, for example April 2011, February 2012 and April 2012 (Figure 6.2). To facilitate detailed exploration of these data, in the following section the genomes have been analysed by season.

Figure 6.2 (overleaf). Maximum likelihood tree constructed using norovirus genomes from OUH in patients collected between 2009-2013. Date of onset was included on the right of the figure. Phylogenetic clusters of genomes were visibly segregated in time. Within cluster diversity was low, whilst between cluster diversity can be large, despite clusters occurring as little as one month apart.



10866
Feb 2012
12501
12438
12464
7910
12486
12511
12454
7907
7895
7885
6633
7939
7957
7943
12470
7915
12439
7933
Jan 2010
10839
10863
10840
10842
10795
8640
8650
9406
10847
9689
10846
10794
9377
8646
10851
9336
Feb 2012
12517
12493
12485
12469
12461
Jul 2012
9359
7963
7966
7916
7892
7932
7880
7904
Jun 2011
10819
10802
10798
10806
10805
10803
10817
10804
10820
10797
10800
10807
10796
10844
10855
9389
9396
10856
9376
9380
12478
6632
Apr 2012
12431
7960
7920
7891
7879
9423
7918
7926
7948
7962
7940
7964
7952
5935
Apr 2011
12432
7912
7890
7884
7882
7900
7954
7936
9685
Feb 2011
10834
10833
10828
10838
10826
10831
May 2012
10778
12448
12440
7941
7930
7886
July 2011
15327
15326
15328
Jan 2013
15321
15320
15318
15316
14384
Dec 2012
200nt

6.5 Within Season Norovirus Genetic Diversity (OUH 2009-2010)

The nucleotide sequences of 29 hospital outbreak noroviruses were determined. These were derived from 22 unique faecal samples, and the viruses in seven were re-sequenced for quality control. No single nucleotide variants (SNVs) were found among the re-sequenced replicates, confirming the authenticity of the SNVs observed among the 22 viruses. The genetic distances among these 22 viruses ranged from 0 to 197 SNVs (Appendix 5). Eighteen viruses (72%) were within ten SNVs of each other. Three groups of hospital viruses (n=5, 3 and 2) each shared an identical genome sequence specific to their outbreak cluster.

6.5.1 Distribution of genomes according to time clusters

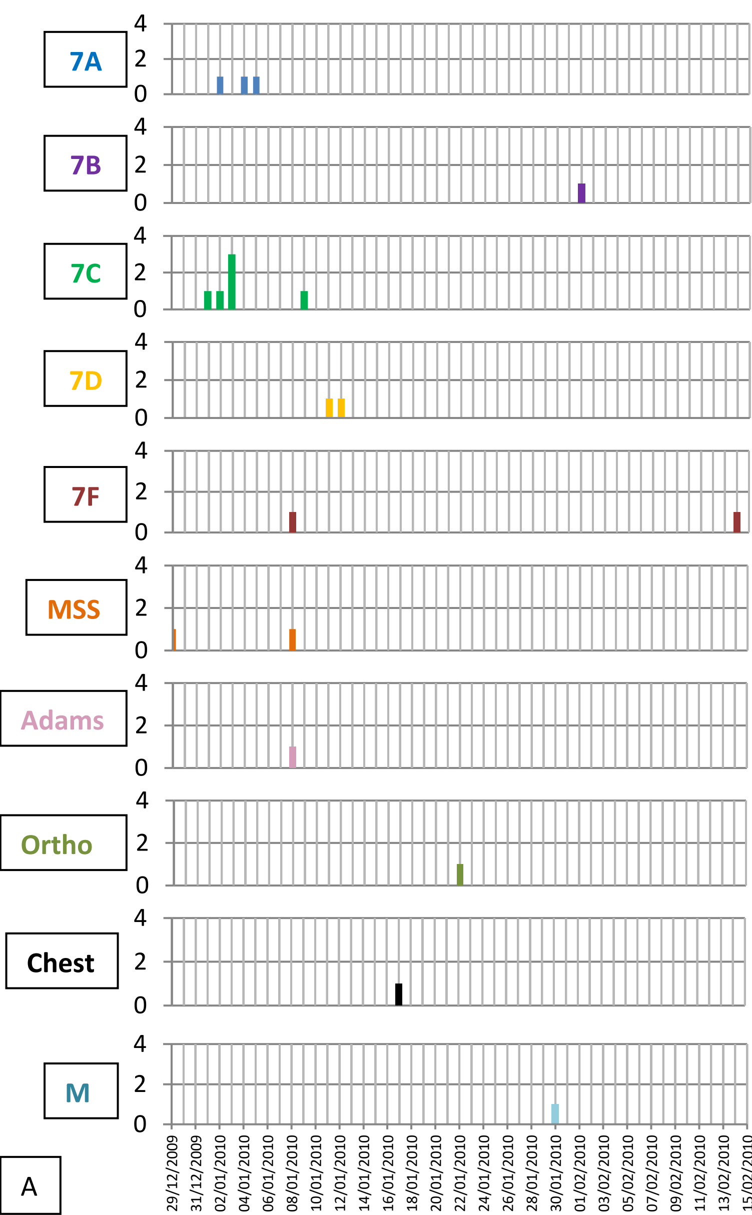
Direct comparison of epidemiological data (patient location [ward], date of sample collection, and number of cases, shown in Figure 6.3a) with the genetic data, in the form of the maximum likelihood phylogenetic tree is shown in Figure 6.3b. The phylogenetic tree was constructed using the maximum likelihood method in the computer program Geneious (Biomatters, New Zealand [213]) and the viewer PhyML [214]). The tree confirmed the three clusters of identical genomes, some of which were derived from viruses infecting symptomatic patients from different wards, during similar periods of time.

An alternative method of representing the genetic relationships between genomes is shown in Figure 6.4. This plot of genetic distances shows information from the maximum likelihood algorithm. Many of the genomes collected within the same ward during an outbreak were genetically identical. A small subset of samples were within 1-2 SNVs from the largest node. More genetically distant genomes are located to the right

of the main node, but were collected from outbreaks up to 2 months after the first sample was collected. These SNV changes correlate within the range of mutation or evolution of norovirus as previously described [89, 92].

The plot of genetic distances (Figure 6.4) also allowed the representation of SNV changes common to multiple viruses, which could be used to track transmission events, and potentially infer the direction of transmission. Viruses sharing common accumulated SNVs within this season were observed on four occasions; on the right of the figure, where four viruses (7915, 12470, 7957 and 7943) are all derived from the larger identical node to the left of the figure. These four viruses all acquired only one common SNV difference from the node, before acquiring further SNVs (n=3 to 6) in the sequence obtained. The four viruses were collected a few weeks after the 'identical node', indicating that further evolutionary intermediate genome(s) may not have been collected and sequenced. Other potential intermediate points are illustrated with a red asterisk (Figure 6.4).

Temporally linked, but more divergent genomes descended from the ‘intermediate node’ but differing from it by >30 SNV were also found (Figure 6.4 red dotted arrow). Two of these genomes (6632 and 12517) were collected within the same ward and at the same time as viruses within the larger node, suggesting the presence of two genetically divergent genomes co-circulating during a ward outbreak.



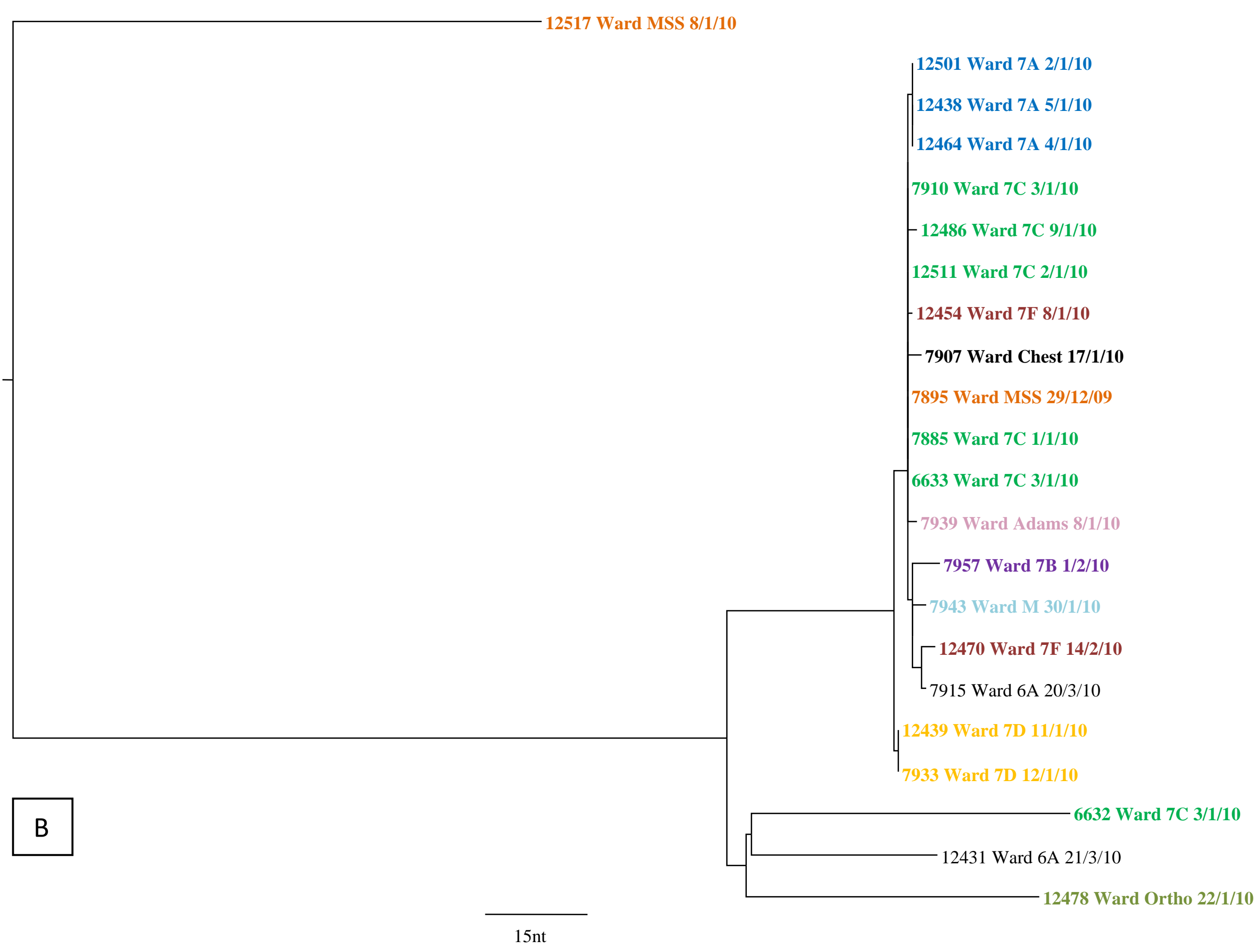


Figure 6.3 (a) Histogram illustrating the location and collection date of sequenced viruses. (b) Phylogenetic maximum likelihood tree of viruses sequenced during the 2009-2010 season. As revealed by the matrix in Appendix 3, there were two clusters of three genomes (Ward 7A, shown in blue) and two genomes (Ward 7D, yellow) which were identical within their group. An additional cluster of 6 genomes (Ward 7C, shown in green) contained four identical genomes but two divergent ones. There was a divergent genome in MSS (sample 12517 shown in brown) which was different to most of the other genomes sequenced during this season.

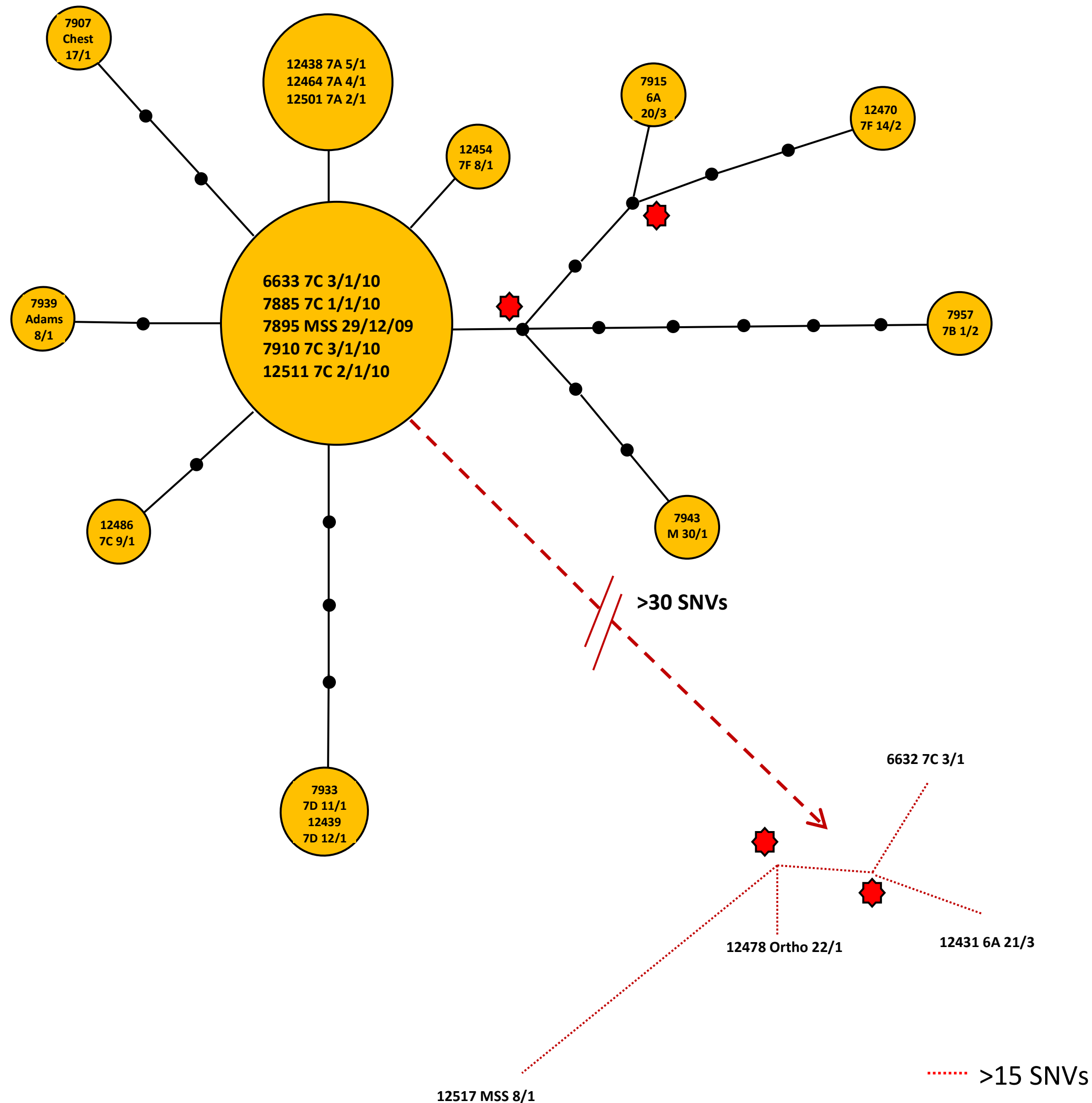


Figure 6.4 Plot of genetic distances among viruses collected between 2009-2010. Each of the yellow circles (nodes) represents a genomic sequence. Black lines separated by a small black dot delineate a difference of one SNV. Black lines without a black dot but ending in a new node signify one SNV difference from the other node. The tree confirmed that the more genetically distant genomes were separated by time, whilst viruses collected within the same period of time were identical. However, five genomes were more than 30 SNVs distant to the central node (bottom right edge, separated by red vertices). The genetic distance between these genomes was more than 15 SNVs. Two viruses of interest (6632, and 12517, in lower left distant tree) were collected from wards at a similar timepoint to those within the larger central node. This genetic distance is indicative of independent introductions. The red asterisks highlight possible intermediate nodes as viruses distal to these have accumulated common SNV(s) before diversifying.

6.5.2 Genetic analysis of viruses within individual wards

Epidemiological data alone identified possible norovirus transmission chains involving cases that shared a common location and date. However, the existence of a transmission link was excluded using genetic data. For example a total of six viruses from patients on Ward 7C at the John Radcliffe Hospital, were sequenced. Figure 6.3b shows that four of these sequences were identical (6633, 7885, 7910, and 12511, all highlighted in green). The clinical samples in which they were contained had been collected within two days of each other (range: 1st January 2010 to 3rd January 2010). Both the epidemiological data and genomes sequence data therefore indicated that these four samples were all part of the same transmission chain. An additional virus (12486, at the top of the tree, in green) was collected 8 days after the first case (7885, 1st January 2010) and was two SNVs from the identical cluster of four. This genome is a putative relative to the original cluster, given that the documented evolutionary clock of norovirus has been estimated to be between 0.6-1.3 substitutions/genome/week [84, 88, 89]. In contrast, the sixth virus (6632, in green and located near the bottom of the phylogenetic tree) was sequenced and collected on the 3rd January 2010. Although it was collected only two days after the first virus, it differed from the identical cluster by 109 SNVs. The patients providing genomes 7910 and 6632 were admitted to neighbouring beds within a four bedded bay on the same ward (Ward 7C). Their genomes differed by 109 SNVs, suggesting that a transmission event did not occur between them. Furthermore, the norovirus genome of patient 6632 was not closely related to any of the viral genomes from symptomatic patients within the 2009/2010

season, suggesting that this patient's virus was not part of an in-hospital onward transmission chain.

Interestingly, sample 7895 (coloured in tan brown and collected on MSS ward) was identical to the cluster of four identical genomes from Ward 7C. This sample was collected up to 3 days prior to the first sample sequenced from Ward 7C and suggests a possible transmission event. Unfortunately, only one further sample from MSS ward was sequenced (12517) and was collected 11 days after sample 7895. This latter sample was very distinct from the original genome (177 SNVs apart). The documented norovirus mutation rate [89, 92], indicates that this was unlikely to be part of a transmission chain.

6.5.3 Distribution of SNVs between closely related genomes

Many studies of norovirus transmission to date have relied upon short sequences from the capsid P2 domain or the region of the ORF1/ORF2 overlap to establish potential transmission links [154]. Figure 6.5 illustrates the location of SNVs corresponding to each of the 22 2009-2010 hospital genomes. The short sequences commonly used in established genotyping approaches are highlighted as regions A, C, and D within the figure. They have been used for outbreak investigations over the last 10 years due to their discriminatory power, which has been considered sufficient to link related cases [215]. However, the higher resolution afforded by whole genome sequencing confirms that SNVs frequently occur outside of these established regions (shown by red dotted circles as examples). Consequently, these data demonstrate that previously accepted genotyping methods would not discriminate these very closely related genomes.

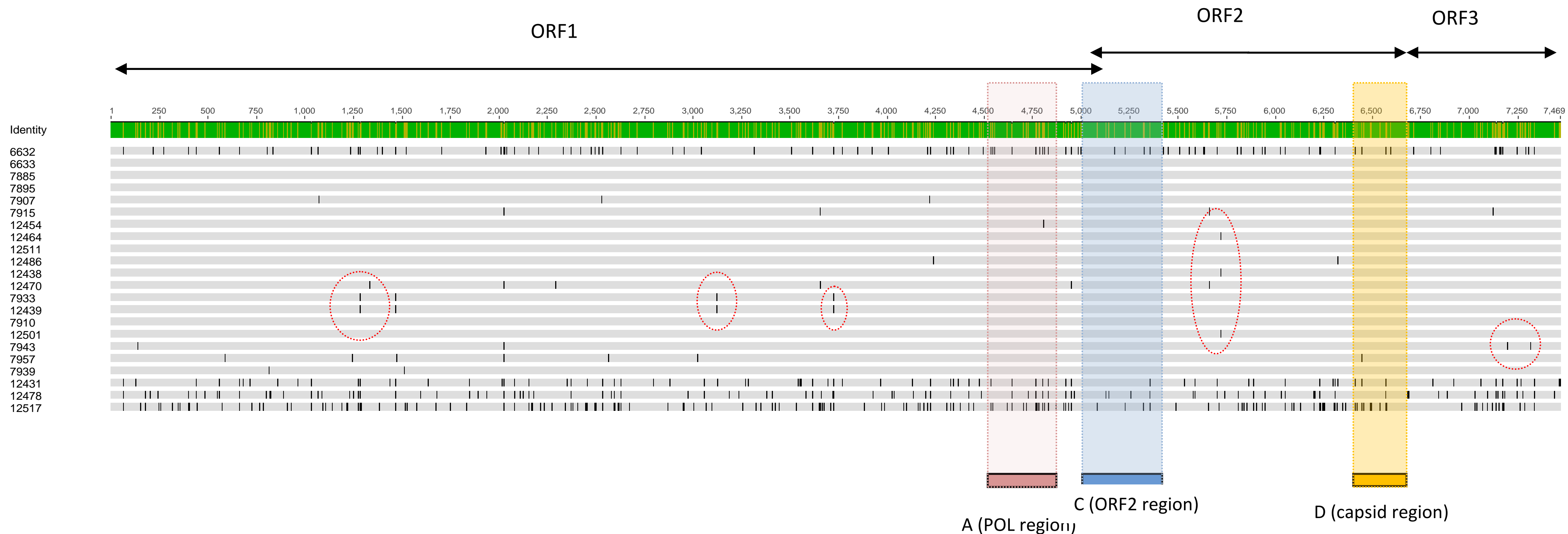


Figure 6.5 Location of SNVs within the genomes of all 2009-2010 isolates. Genomes were aligned, labelled on the left and represented in grey. At the top, the identity bar discriminates between areas which are conserved throughout all the genomes (green) and non conserved (yellow). Grey areas spanning each sample show sites of sequence identity amongst all genomes (the consensus). Vertical black lines indicate locations of SNVs between each of these genomes. SNVs occurred throughout the genome, and there were no mutation hotspots. SNVs occurred outside of both the areas used for genotyping (known as region A covering the POL region (pink box)), region C covering the beginning of ORF2 (blue box), and region D covering the 3' end of the VP1 capsid protein (orange box)). The dotted red rings highlight regions where SNVs occurred outside these 'genotyping' regions. Standard genotyping based on these short fragments would have identified these viruses as 'identical' and inferred the possibility of an epidemiological link.

6.5.4 Within Season Norovirus Genetic Diversity (OUH 2010-2011)

A total of 44 hospital inpatient norovirus genomes (35 unique and 9 replicates) were used in this analysis. Comparison of the 9 replicate pairs confirmed the absence of SNV differences between pairs. The phylogenetic relationships among the 35 viruses were assessed by constructing a maximum likelihood tree (figure 6.6). Five genome clusters involving temporally linked cases accounted for the majority of viruses; only two unique outliers occurred outside the clusters.

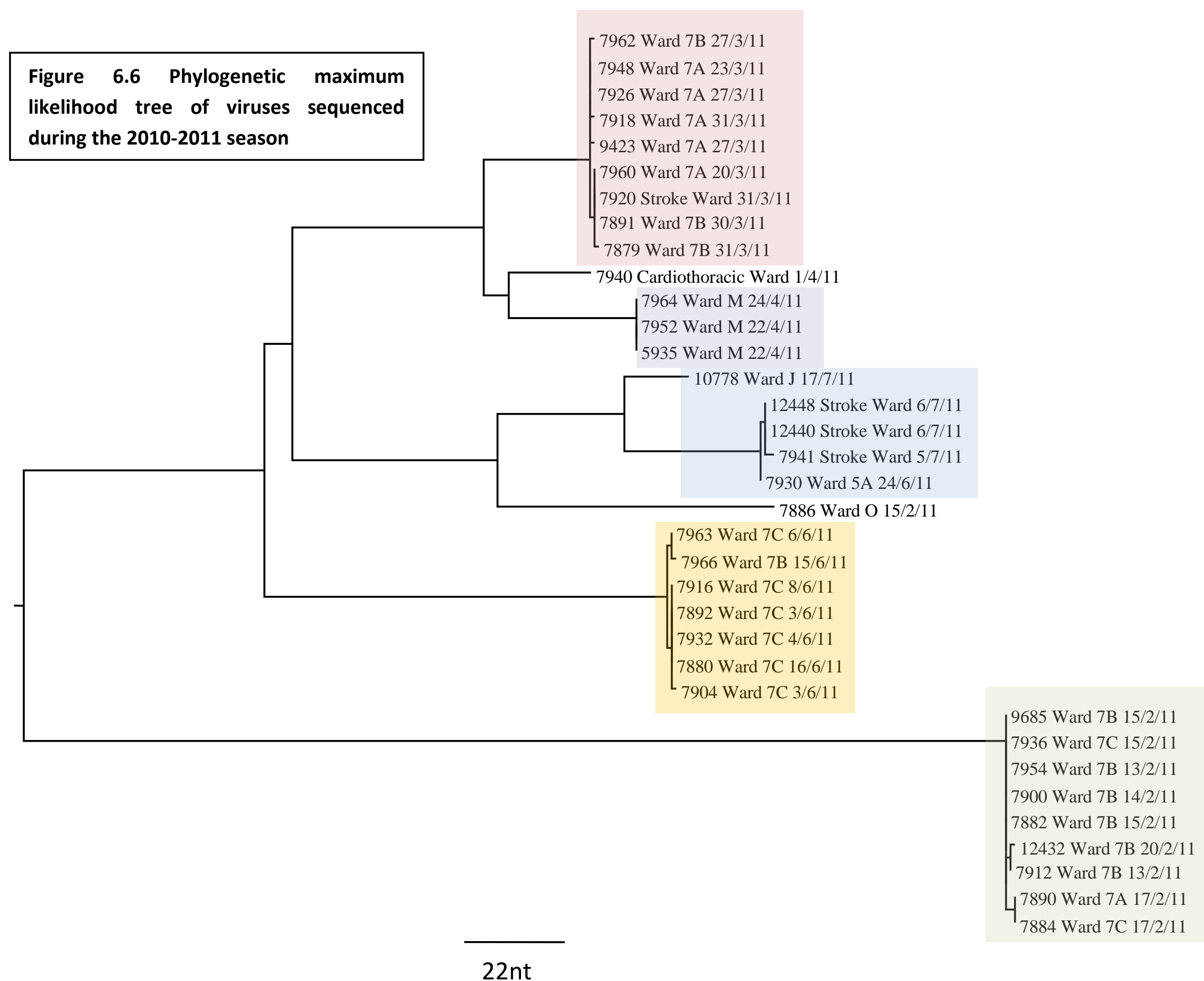


Figure 6.6 Phylogenetic maximum likelihood tree of viruses sequenced during the 2010-2011 season

In order to understand the genetic relationship, separate genetic distance plots are shown accompanied by time histograms summarising epidemiological information (Figs 6.7a-d).

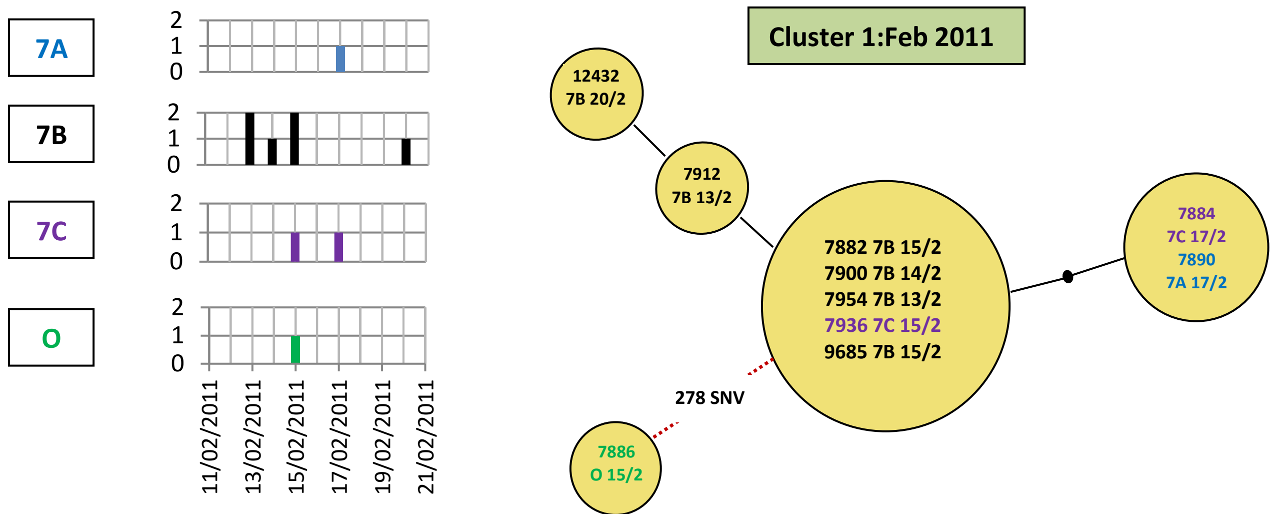


Figure 6.7a Case distribution histogram and corresponding genetic distance plot of viruses collected in February 2011.

Cases occurred on four different wards during February 2011. Three (7A, 7B, and 7C) were adjoining and located on the 7th floor at the John Radcliffe Hospital, whilst the fourth (O ward) was located at the Horton Hospital. The JR viruses contained minimal SNVs, suggesting the possibility of derivation from a common reservoir of infection. The Horton virus (7886, in green) was genetically distinct from the JR variant.

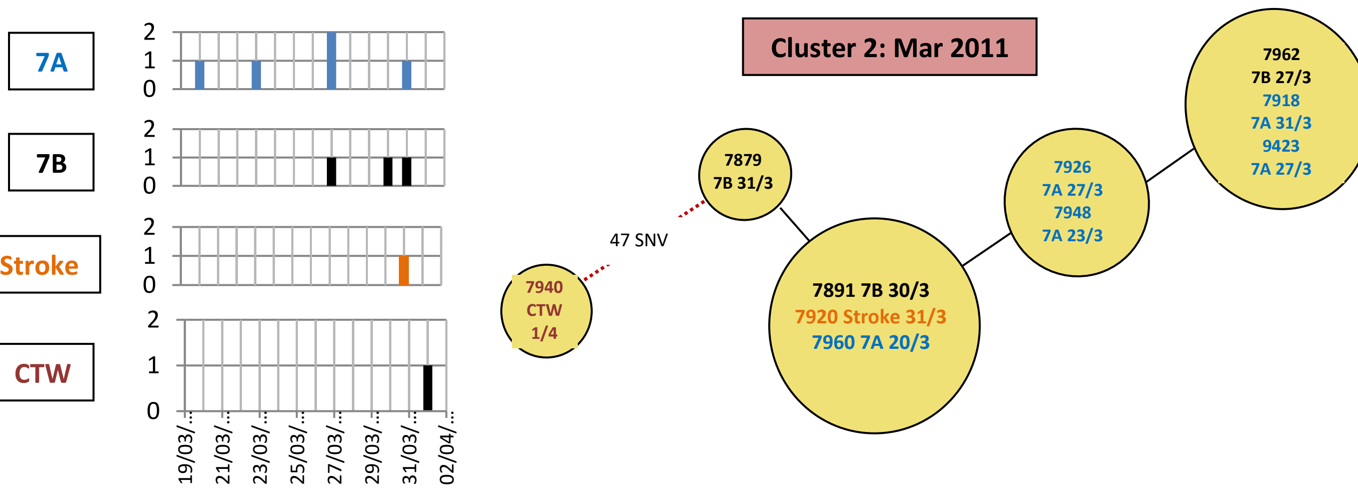


Figure 6.7b Case distribution histogram and corresponding genetic distance plot of viruses collected in March 2011.

The cases occurring during March 2011, were all located within the JR site. In addition to two seventh floor wards, stroke ward (5th floor), experienced a case as did CTW (cardiothoracic ward) which was located in a separate building on the same site. The CTW virus (7940, in maroon) differed significantly from the others (47 SNVs). However, the Stroke ward level 5 virus being was identical to viruses collected from level 7 wards 7A and 7B, although one virus (7960 from Ward 7A) was collected 10 days prior to the other two.

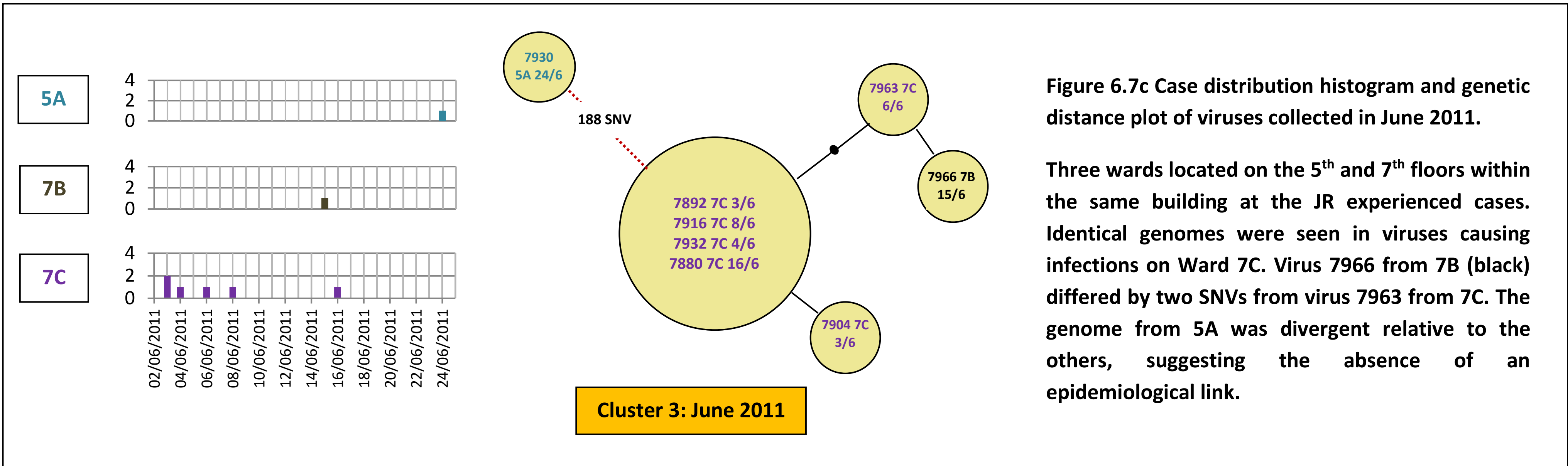


Figure 6.7c Case distribution histogram and genetic distance plot of viruses collected in June 2011.

Three wards located on the 5th and 7th floors within the same building at the JR experienced cases. Identical genomes were seen in viruses causing infections on Ward 7C. Virus 7966 from 7B (black) differed by two SNVs from virus 7963 from 7C. The genome from 5A was divergent relative to the others, suggesting the absence of an epidemiological link.

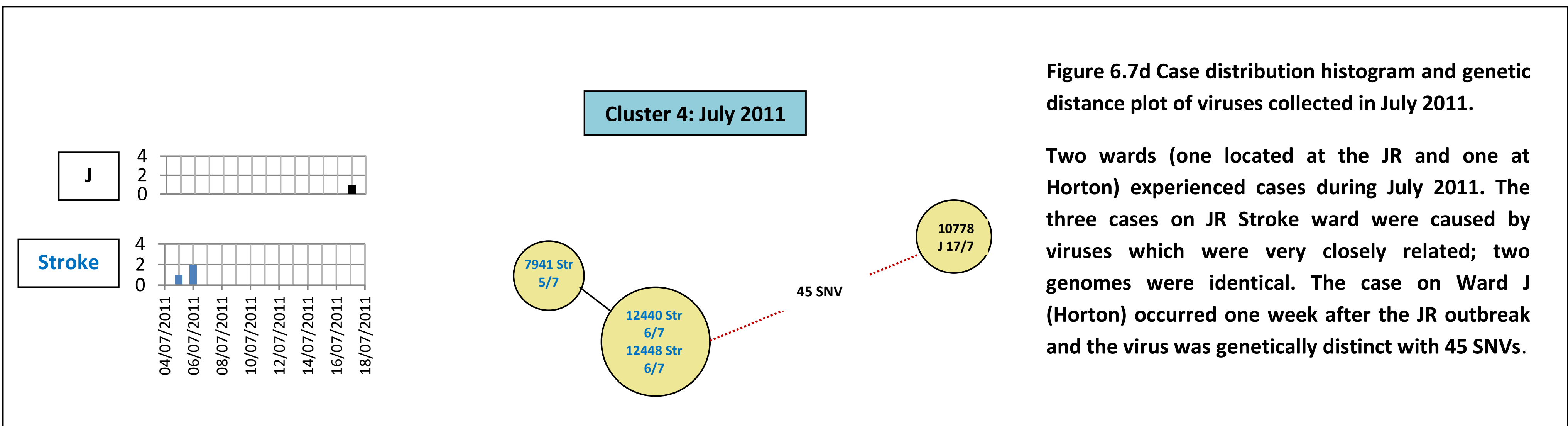


Figure 6.7d Case distribution histogram and genetic distance plot of viruses collected in July 2011.

Two wards (one located at the JR and one at Horton) experienced cases during July 2011. The three cases on JR Stroke ward were caused by viruses which were very closely related; two genomes were identical. The case on Ward J (Horton) occurred one week after the JR outbreak and the virus was genetically distinct with 45 SNVs.

In contrast to the Oxfordshire norovirus season of 2009/2010, during 2010/2011, norovirus cases occurred over four months (February, March, June and July) and formed four distinct temporal clusters (labelled as Clusters 1-4 in Figures 6.7a-d). There were significant SNV differences between clusters. Appendix 6 shows the matrix of SNV differences between the viruses within clusters.

The majority of the viruses sequenced from each temporal cluster differed by a maximum of two SNVs, indicating that the associated cases were potentially epidemiologically related (see Fig 6.7a-d for details). However, in all four temporal clusters individual genomes occurred that were clearly divergent (range 45-278 SNVs). These cases originated in a different hospital within the county or a different floor or different building within a hospital.

Phylogenetic clusters occurring on independent occasions in the same ward differed by large numbers of SNVs, for example, 265 SNVs occurred between outbreaks on Ward 7B in February 2011 and March 2011. This rules out environmental contamination as a reservoir for onward transmission (in this case), since this would cause the genome sequences to be almost identical over time, minimal replication taking place outside the host. Figure 6.8 illustrates the location of SNVs within the viral genomes studied from the 2010/2011 Oxfordshire season. As in 2009/2010, SNVs were distributed throughout the genome.

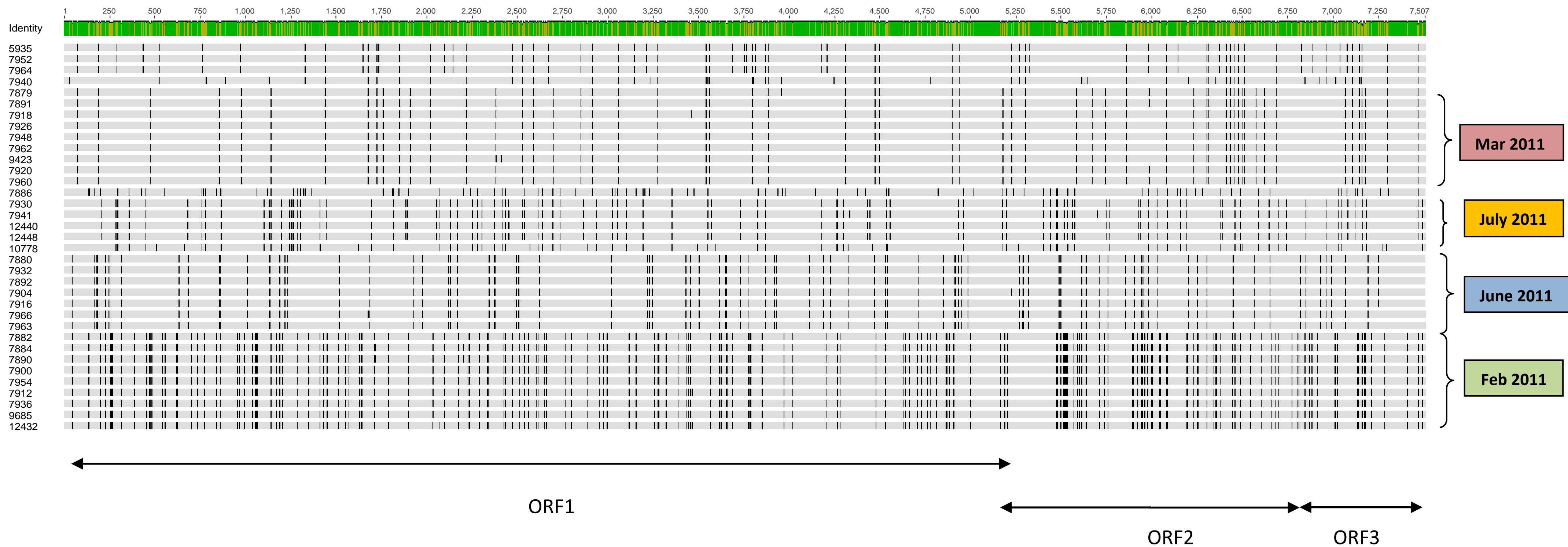


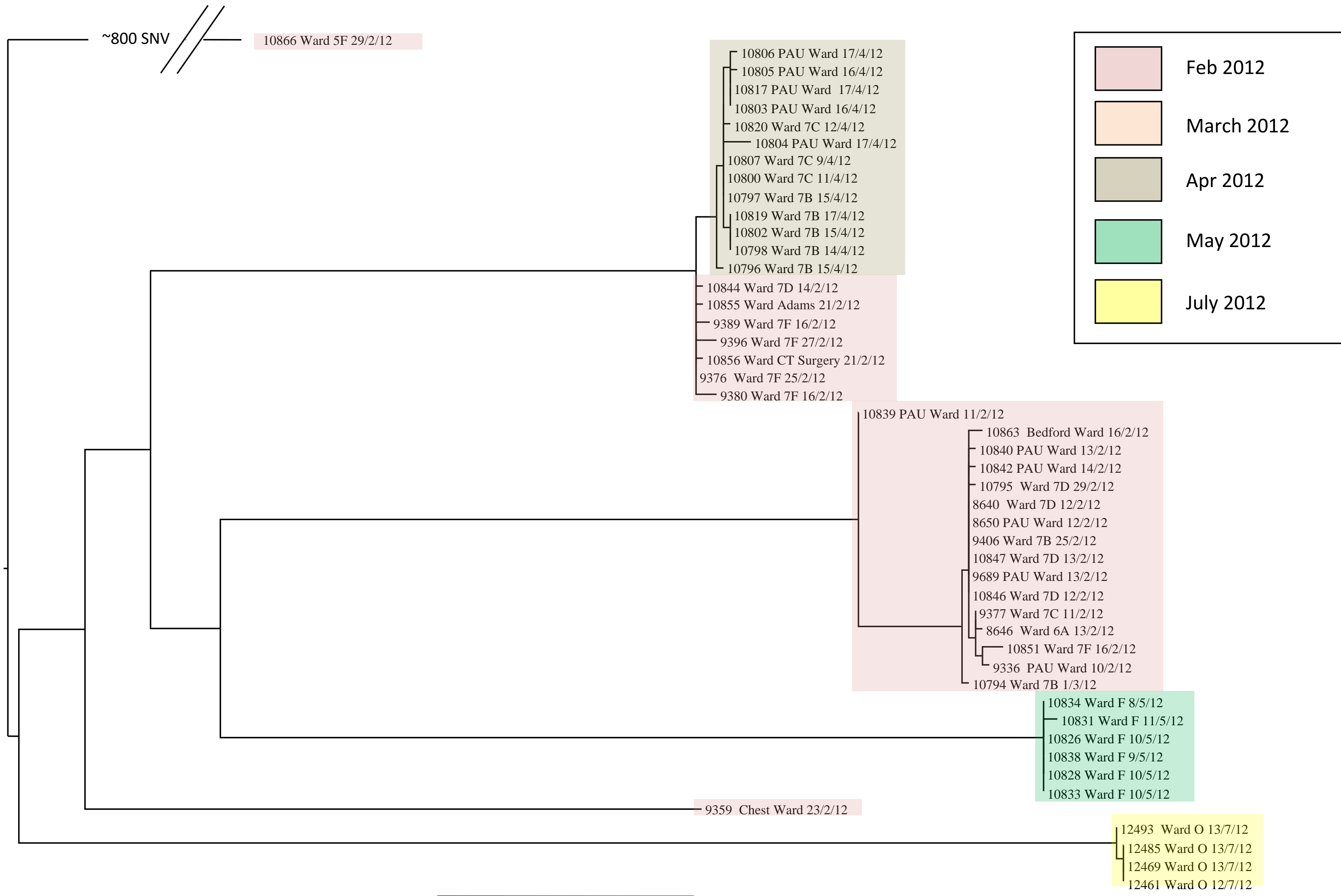
Figure 6.8 Genomic locations of SNVs among the 2010-2011 genomes. As previously illustrated, individual genomes within this season were aligned against each other using the Geneious suite without any reference genome. At the top, the identity bar discriminates between areas of high sequence conservation (green) and areas of low levels of conservation (yellow). Grey areas spanning each genome show sites of identical base call. Black areas, indicate SNVs. A greater number of SNVs were identified than in the 2009/2010 viruses, but again no mutation 'hotspots' were found.

6.5.5 Genomes from the 2011/2012 season

A total of sixty four Oxfordshire hospital norovirus genomes were available for analysis. This comprised forty-eight unique viruses, thirteen genomes collected longitudinally (from four patients) and three replicates from two patients. Comparison of the replicate pairs revealed no SNV differences. The first genome to be collected in each longitudinal series was used in the analysis. Figure 6.9 shows phylogenetic relationship between the genomes.

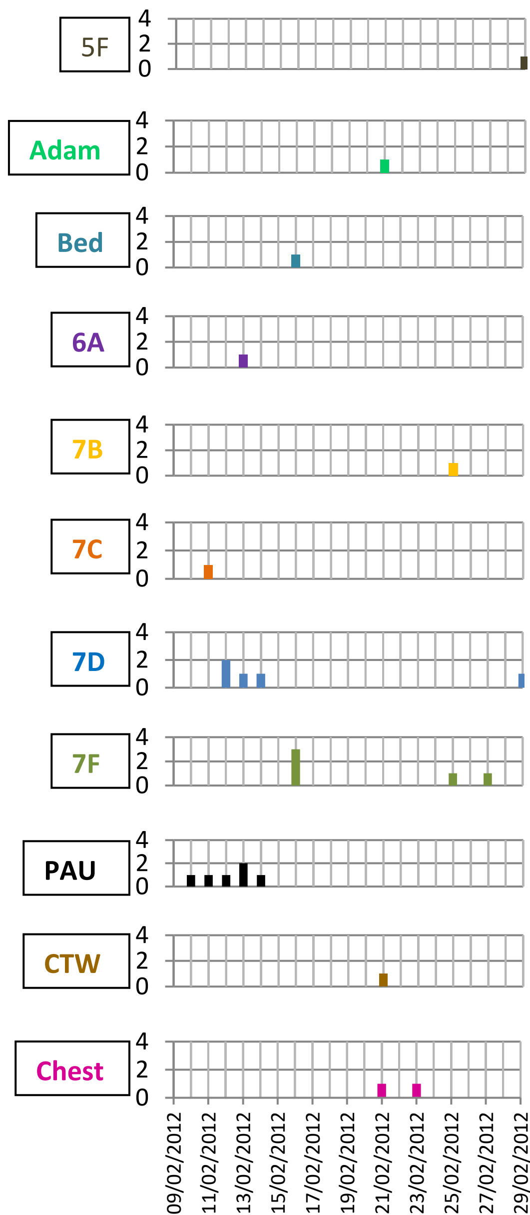
As in the two previous seasons, the genomes clustered temporally and spatially in the majority of cases. A matrix of SNV differences is shown in Appendix 7. The within cluster isolates were almost identical (0-3 SNVs), but there was a large genetic distance between clusters. Spatially, there was evidence of high diversity between clusters within the same county when they occurred close together in time (JR Hospital in April 2011 and Horton Hospital May 2011).

Figure 6.9 (overleaf). Phylogenetic maximum likelihood tree constructed using the WGS of noroviruses collected in Oxfordshire from 2011-2012. All genomes were collected from the JR Hospital apart from the Chest ward (Feb 2011; Churchill Hospital) and Ward F (May 2011; Horton Hospital). Each genome and cluster was assigned a coloured box reflecting the time when symptoms occurred. Of note is the divergence of the two main branches, which are over 1,700 SNVs apart, one of which represents a distinct norovirus genogroup (GII.3). Two pairs of isolates, which shared the same ward and similar time (shown by red (Ward 7F isolates) and green stars (Ward 7D isolates)), were 177SNVs apart, confirming the simultaneous introduction of two virus strains.


~800 SNV
10866 Ward 5F 29/2/12
10806 PAU Ward 17/4/12
10805 PAU Ward 16/4/12
10817 PAU Ward 17/4/12
10803 PAU Ward 16/4/12
10820 Ward 7C 12/4/12
10804 PAU Ward 17/4/12
10807 Ward 7C 9/4/12
10800 Ward 7C 11/4/12
10797 Ward 7B 15/4/12
10819 Ward 7B 17/4/12
10802 Ward 7B 15/4/12
10798 Ward 7B 14/4/12
10796 Ward 7B 15/4/12
10844 Ward 7D 14/2/12
10855 Ward Adams 21/2/12
9389 Ward 7F 16/2/12
9396 Ward 7F 27/2/12
10856 Ward CT Surgery 21/2/12
9376 Ward 7F 25/2/12
9380 Ward 7F 16/2/12
10839 PAU Ward 11/2/12
10863 Bedford Ward 16/2/12
10840 PAU Ward 13/2/12
10842 PAU Ward 14/2/12
10795 Ward 7D 29/2/12
8640 Ward 7D 12/2/12
8650 PAU Ward 12/2/12
9406 Ward 7B 25/2/12
10847 Ward 7D 13/2/12
9689 PAU Ward 13/2/12
10846 Ward 7D 12/2/12
9377 Ward 7C 11/2/12
8646 Ward 6A 13/2/12
10851 Ward 7F 16/2/12
9336 PAU Ward 10/2/12
10794 Ward 7B 1/3/12
10834 Ward F 8/5/12
10831 Ward F 11/5/12
10826 Ward F 10/5/12
10838 Ward F 9/5/12
10828 Ward F 10/5/12
10833 Ward F 10/5/12
9359 Chest Ward 23/2/12
12493 Ward O 13/7/12
12485 Ward O 13/7/12
12469 Ward O 13/7/12
12461 Ward O 12/7/12
40 nt
Feb 2012
March 2012
Apr 2012
May 2012
July 2012

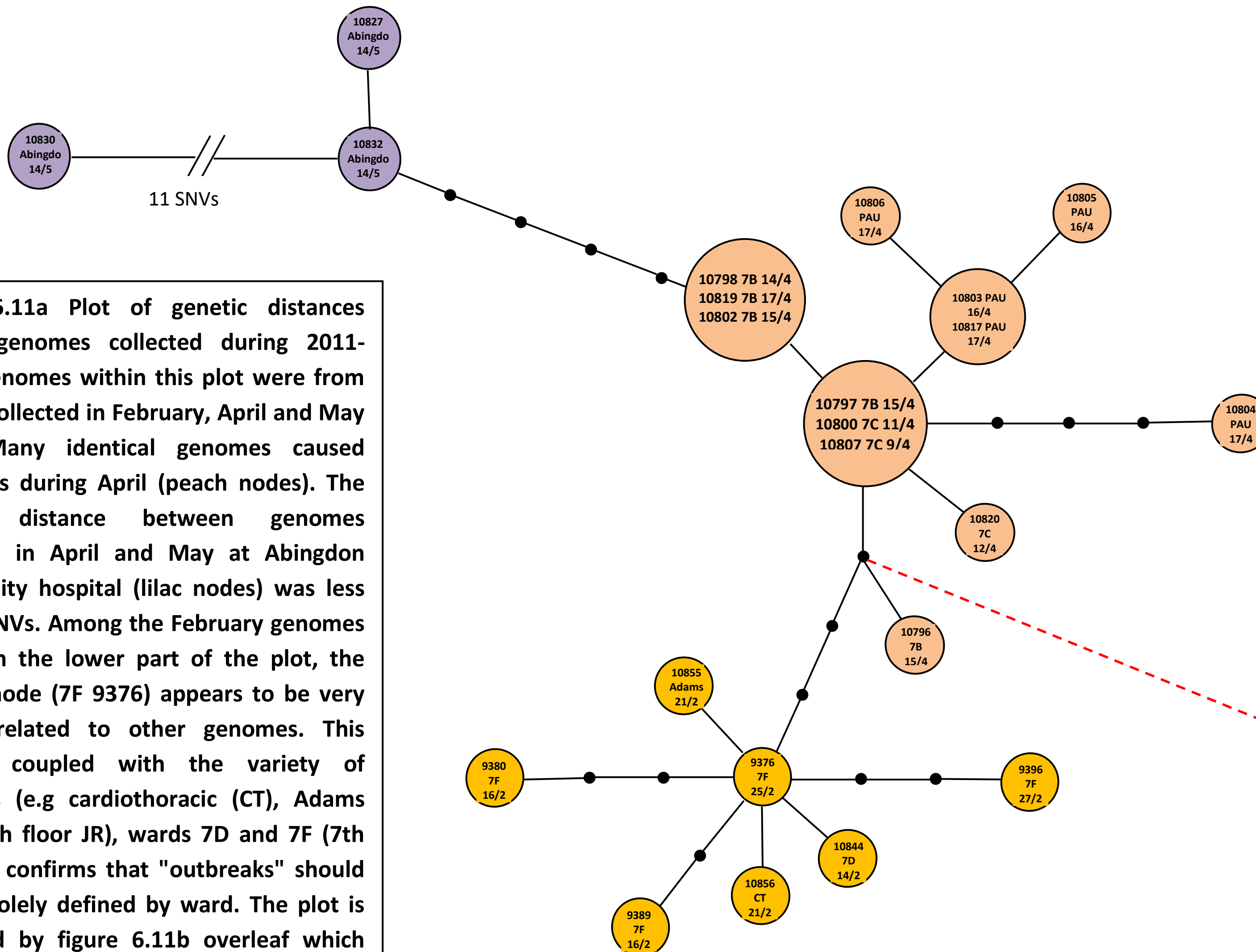
The figure indicated that with the exception of February (pink boxes), phylogenetic clusters of genomes segregated temporally within each month. During February genomes formed four divergent clusters (Figure 6.10). In February 2012, there were two main branches in this tree. In the upper main branch, genomes were within five SNVs of a central identical group of eight genomes. These samples were derived from neighbouring wards (Wards 7B, 7C, 7D, and 7F as well as wards on the sixth floor of the hospital (PAU). The smaller lower branch contained seven genomes, four from Ward 7F, as well one from CT Surgery (different building within the JR site), Adams (different floor) and 7D (an adjoining ward to 7F). The four genomes from 7F were slightly more diverse spanning eight SNVs making it less clear whether or not these were sampled from the same ward based transmission chain. Of particular interest were genomes from both Ward 7D and 7F found within the upper branch, which were separated by 175-180 SNVs, but co-circulated during the same outbreak period, further confirming the presence of multiple virus strains during ward outbreaks. Figure 6.11a and 6.11b are plots of genetic distances for all 2011-2012 genomes.

During April 2012, a single outbreak affecting thirteen patients, caused by viruses separated by five SNVs occurred in the adjacent wards 7B, 7C and PAU (post admission unit). These were labelled by peach nodes in Figure 6.11a. One additional case on PAU differed by four SNVs making its relationship to the others difficult to interpret. In May 2012, there was a completely separate outbreak of six cases on Ward F at the Horton Hospital, all genomically identical and likely to be part of the same ward outbreak (Node in green, Figure 6.11b). Similarly, in July 2012 another independent outbreak of four genetically segregated cases were identified on Ward O at the Horton Hospital, all occurring on the same day (nodes in light blue in Figure 6.11b).



10840 PAU Ward 13/2/12
10795 Ward 7D 29/2/12
10842 PAU Ward 14/2/12
8640 Ward 7D 12/2/12
9689 PAU Ward 13/2/12
10846 Ward 7D 12/2/12
10847 Ward 7D 13/2/12
8650 PAU Ward 12/2/12
9406 Ward 7B 25/2/12
10863 Bedford Ward 16/2/12
10851 Ward 7F 16/2/12
9336 PAU Ward 10/2/12
8646 Ward 6A 13/2/12
9377 Ward 7C 11/2/12
10839 PAU Ward 11/2/12
9359 Chest Ward 23/2/12
10844 Ward 7D 14/2/12
10855 Ward Adams 21/2/12
10856 Ward CT Surgery 21/2/12
9389 Ward 7F 16/2/12
9396 Ward 7F 27/2/12
9380 Ward 7F 16/2/12
~900SNVs
10866 Ward 5F 29/2/12
25 SNVs

Figure 6.10 Maximum likelihood tree constructed using the genomes of viruses circulating in February 2012. Genomes 10866 (Ward 5F) and 9359 (Chest ward) are quite divergent from the two main branches. The two main branches are separated by up to 187 SNVs, despite occurring at a similar time. Also, genomes of viruses collected within the same ward over same period, differed markedly (e.g 9376 from Ward 7F and 10844 from Ward 7D within the lower branch).



To Figure 6.11b

Figure 6.11a Plot of genetic distances among genomes collected during 2011-2012. Genomes within this plot were from viruses collected in February, April and May 2012. Many identical genomes caused infections during April (peach nodes). The genetic distance between genomes collected in April and May at Abingdon Community hospital (lilac nodes) was less than 5 SNVs. Among the February genomes shown in the lower part of the plot, the central node (7F 9376) appears to be very closely related to other genomes. This identity coupled with the variety of locations (e.g cardiothoracic (CT), Adams ward (4th floor JR), wards 7D and 7F (7th floor JR) confirms that "outbreaks" should not be solely defined by ward. The plot is extended by figure 6.11b overleaf which follows >80SNVs from these clusters of sequences.

Abingdo = Abingdon community hospital; Co=community samples; PAU = post admission ward (6th floor JR); 7D, 7F = wards on 7th floor JR; CT = cardiothoracic ward (2nd floor JR)

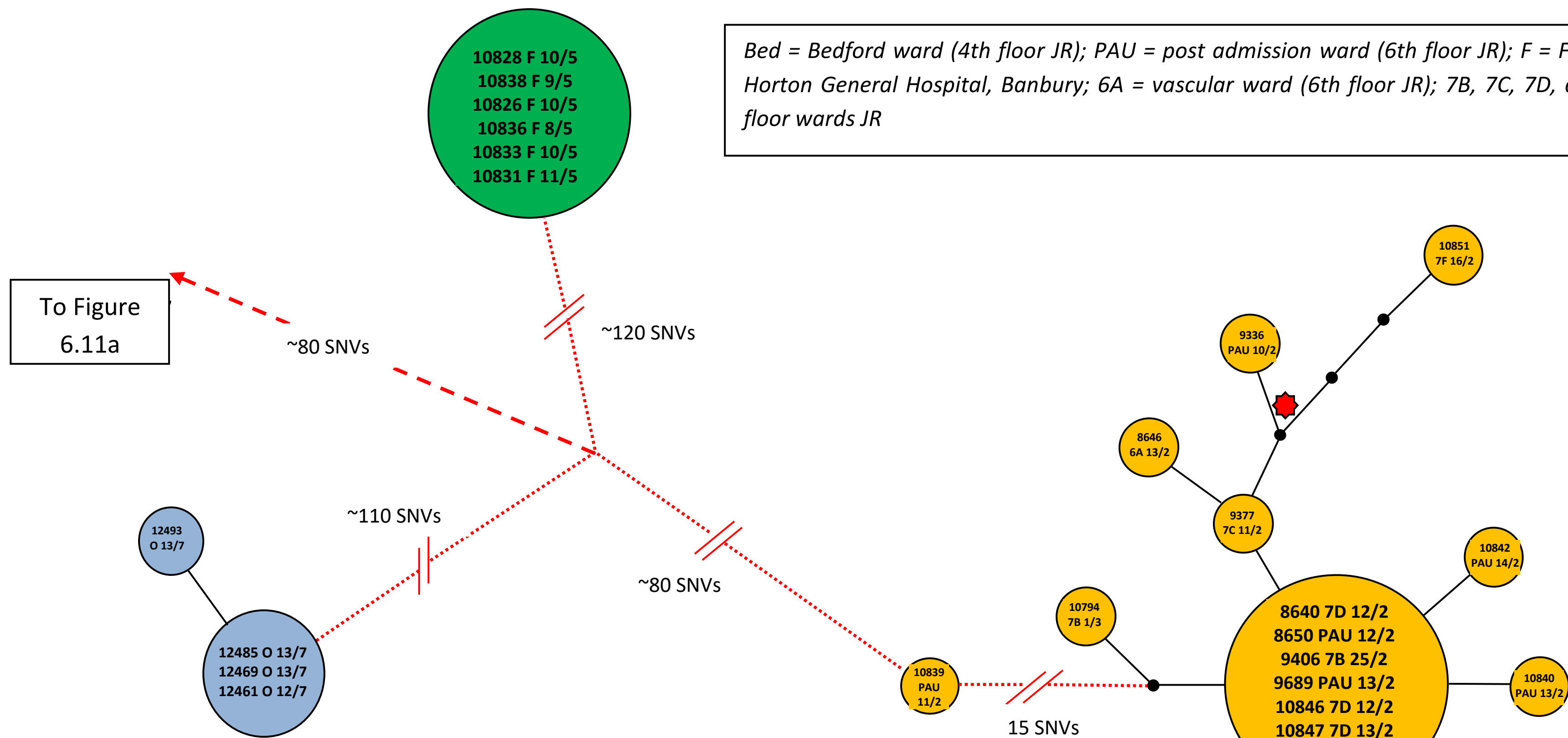


Figure 6.11b Plot of genetic distances among viral genomes from February 2012 (yellow nodes), May 2012 (green nodes) and July 2012 (light blue nodes). This figure is a continuation of Figure 6.11a, which diverges by 160 SNVs at its closest point to the present figure. Genomes recovered from F ward in May 2012 were identical. Many genomes from February 2012 shared identity, but some divergent genomes occurred. By linking with Figure 6.11a, there is evidence that at least 2 divergent strains (of more than 175 SNVs) were circulating concurrently within the 7th floor ward. The red asterisk indicates a location where two genomes distal to this point have the same accumulated intermediary SNV giving rise to possible transmission directionality.

6.5.6 Within Season Norovirus Genetic Diversity (Oxfordshire 2012-2013)

A total of eight genomes were obtained for viruses which caused infections in hospital inpatients during the 2012/2013 season.

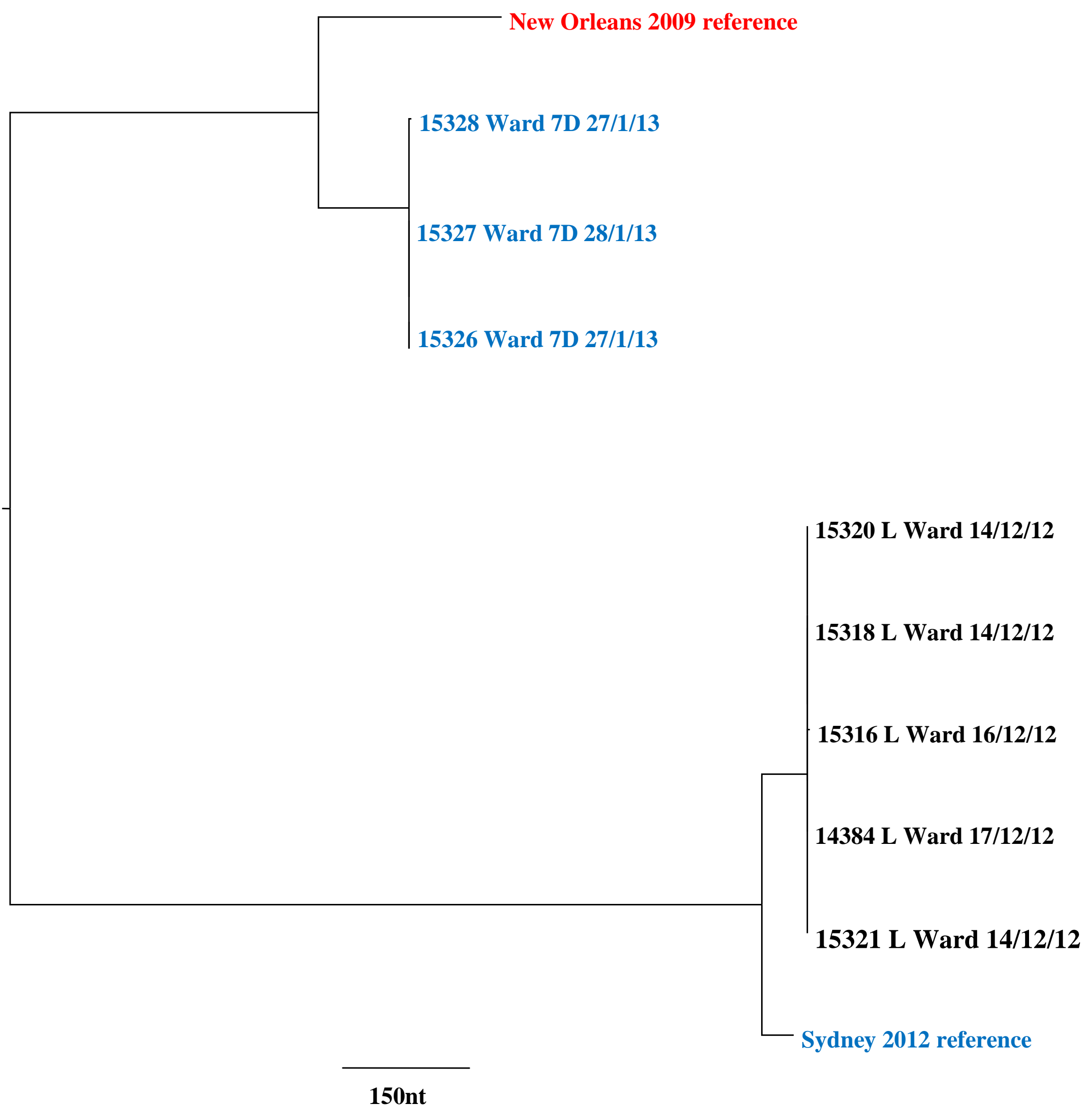


Figure 6.12. Phylogenetic maximum likelihood tree constructed for 2012-2013 genomes from Oxfordshire. Both the New Orleans 2009 and Sydney 2012 references have been included in this tree.

A ML tree revealed two phylogenetic clusters. The viruses within each cluster were geographically linked: Ward 7D (JR Hospital) highlighted in blue, and L Ward (Horton

hospital) highlighted in black. The two clusters occurred around one month apart. The matrix of SNV differences is shown in appendix 8. Once again, the genomes within each cluster were genetically very closely related (range 0 to 3 SNVs, and 1 to 4 SNVs). Between cluster variation was large (range 763-765 SNVs), indicating significant divergence.

It is noteworthy that during this season a novel emergent norovirus strain, designated GII.4 Sydney 2012 [43], was reported worldwide and achieved high prevalence, at least equalling the previously circulating GII4 strain (but not replacing it). The Sydney 2012 variant is believed to be a recombinant GII.4 virus, comprising sequences designated GII.e Sydney 2012 (ORF1) and GII.4 Sydney 2012 (ORF2 and 3) [67]. Subsequent reports have described other recombinant viruses as comprising the older GII.4 New Orleans 2009 virus RNA (ORF1) and GII.4 Sydney (ORF2 and 3) [206].

Although the Oxford norovirus 2012/2013 collection was relatively small (n=8), it was still important to determine whether the GII4 Sydney 2012 variant was present. Figure 6.13 compares the Oxford 2012/2013 and GII4 Sydney 2012 sequences across the genome.

The pattern of nucleotide sequence variation across the Oxford genomes, the reference New Orleans 2009 (Genbank accession number JN595867.1) and Sydney 2012 (Genbank accession number JX459908.1) genomes was interesting (Figures 6.13 and 6.14). It showed that across the alignment there was a different SNV frequency between the different ORFs, specifically ORF1 (many SNVs) versus ORF2 and ORF3 (fewer SNVs). The ML trees constructed for each of the three ORFs (Figure 6.14) lacked congruence. Both

these lines of evidence strongly support the hypothesis that recombination has taken place between GII4 viruses, as described elsewhere [67], and that this mechanism may underlie the emergence of the novel Sydney variant [43, 160, 216]. This analysis confirmed that five of the Oxford 2012-13 viruses (orange asterisks) were closely related to the Sydney 2012 strain.

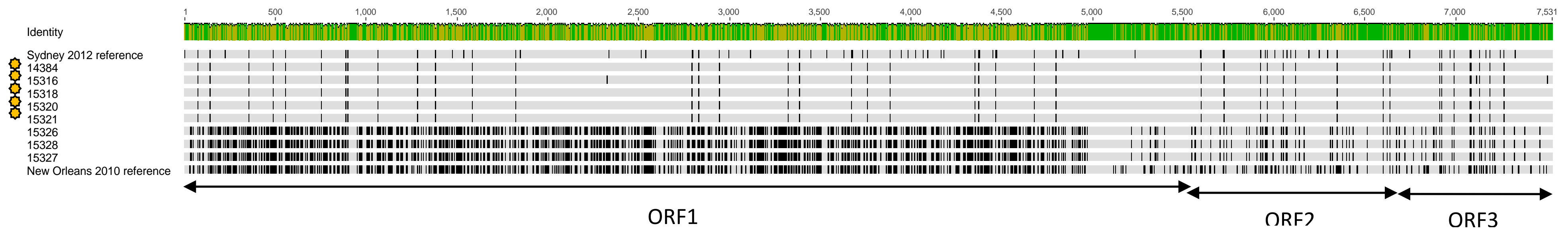


Figure 6.13. Location of differences between sequences during 2012/2013 season. Of particular interest are the top five sequences (denoted with orange asterisks) which are very different to the three genomes below it, particularly within ORF1 where there are highly dense areas of variation (shown as black lines). Both the Sydney 2012 and New Orleans 2009 reference genomes have been included.

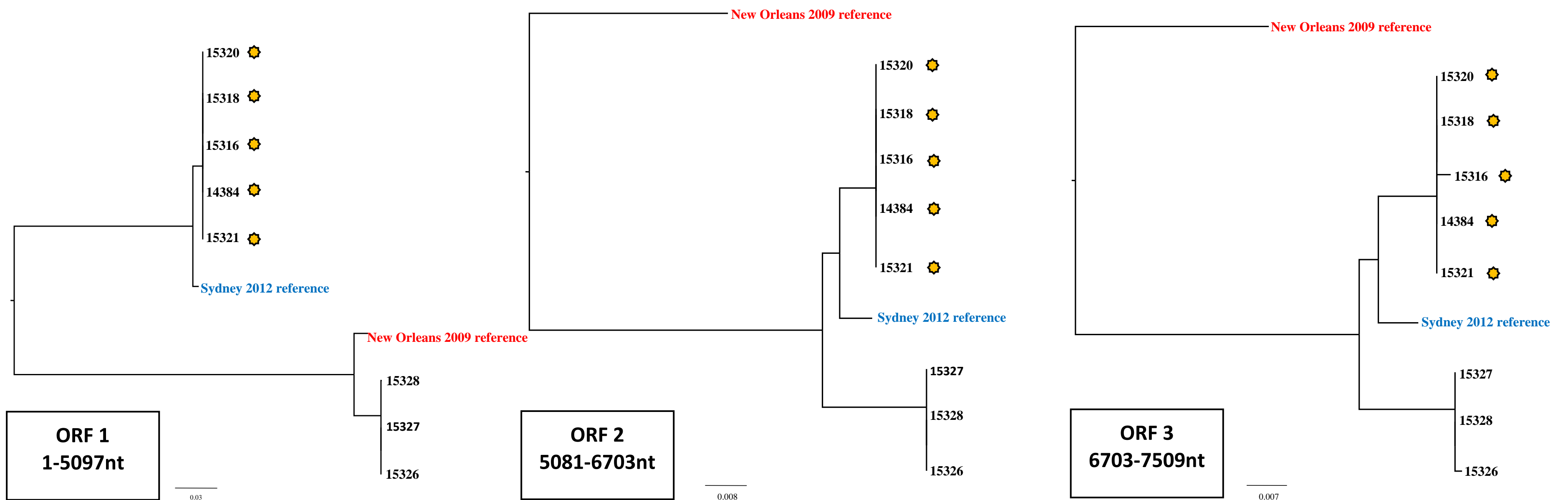


Figure 6.14. Phylogenetic maximum likelihood trees of all 2012-2013 sequences split by the 3 open reading frames. Both New Orleans 2009 (red label) and Sydney 2012 (blue label) references have been inserted within the trees for reference association. Asterisked sequences are derived from Figure 6.14 above.

6.6 Community noroviruses and their relationship to viruses causing hospital outbreaks

The second aim of the study was to determine whether norovirus WGS can be used to identify possible transmission links between the healthcare environment and the local community. To achieve this aim, the phylogenetic relationships between noroviruses causing Oxfordshire hospital outbreaks and those circulating in the region it serves were investigated. This is important because information on the frequency of virus introductions from the community versus internal spread within the healthcare environment would be useful to inform approaches to infection control. Community noroviruses are known to exhibit significantly greater phylogenetic diversity overall compared to hospital viruses [217, 218]. However, the GII.4 variant which predominates in the hospital environment is also present in the community [123, 217, 219].

A total of 40 noroviruses were collected and sequenced from cases which occurred in the community served by ORH and the Horton Hospitals (population size 650,000) between September 2009 and September 2013. These viruses were present in stool specimens submitted by GPs to the OUH Clinical Microbiology Laboratory. Five of these noroviruses were identified as genotype GII.3, and were therefore excluded from further analysis. A maximum likelihood tree was constructed using the genomes of the remaining 35 GII4 community noroviruses (Figure 6.15). This confirmed that like the hospital norovirus genomes described above, and the community noroviruses did not overlap phylogenetically between seasons. Furthermore, community viruses that clustered genetically were also temporally related, apart from one virus collected in December 2011 (12496) which was related to genomes of viruses collected in January 2010.

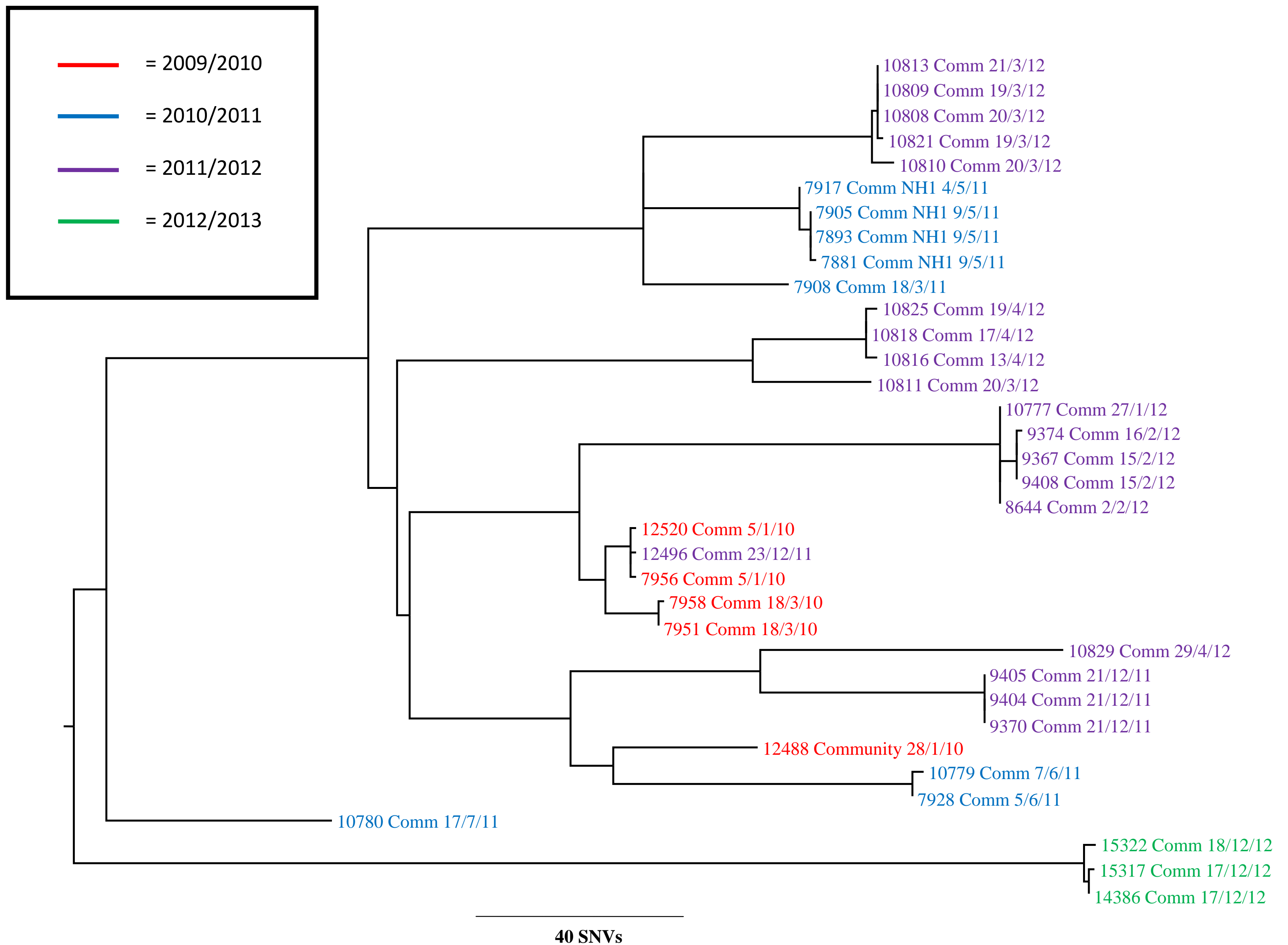


Figure 6.15 Maximum likelihood tree of all 35 community samples from Oxfordshire collected between 2009-2013 (GII.3 viruses have been excluded. NH=nursing home

The relationship between healthcare-associated and community norovirus genomes can be best studied phylogenetically, using the maximum likelihood method. A ML tree was therefore built for each of the Norovirus seasons for which community and hospital viruses were available. This is shown for the 2009/2010 norovirus season below (Figure 6.16).

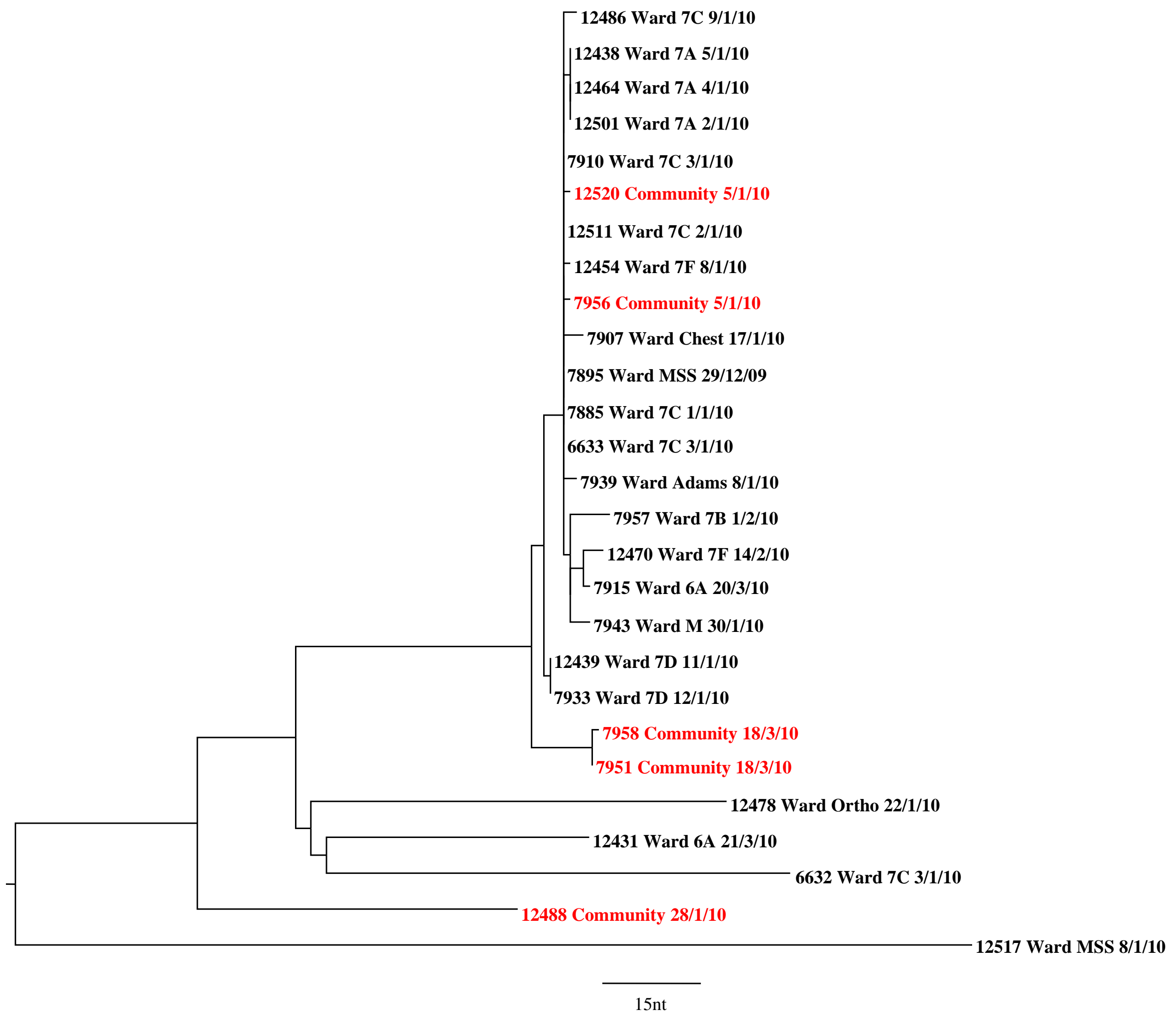


Figure 6.16 Relationship between hospital outbreak and community genomes during the 2009-2010 norovirus season. Community genomes are highlighted in red.

Of the five community norovirus genomes sequenced, three were phylogenetically distinct from the concurrent hospital outbreak genomes. Two of the community genomes (7956 and 12520, both shown in red) were within 1 SNV of the hospital

outbreak genomes. This type of relationship was repeated among hospital and community genomes collected the following season (2010-2011) (figure 6.17).

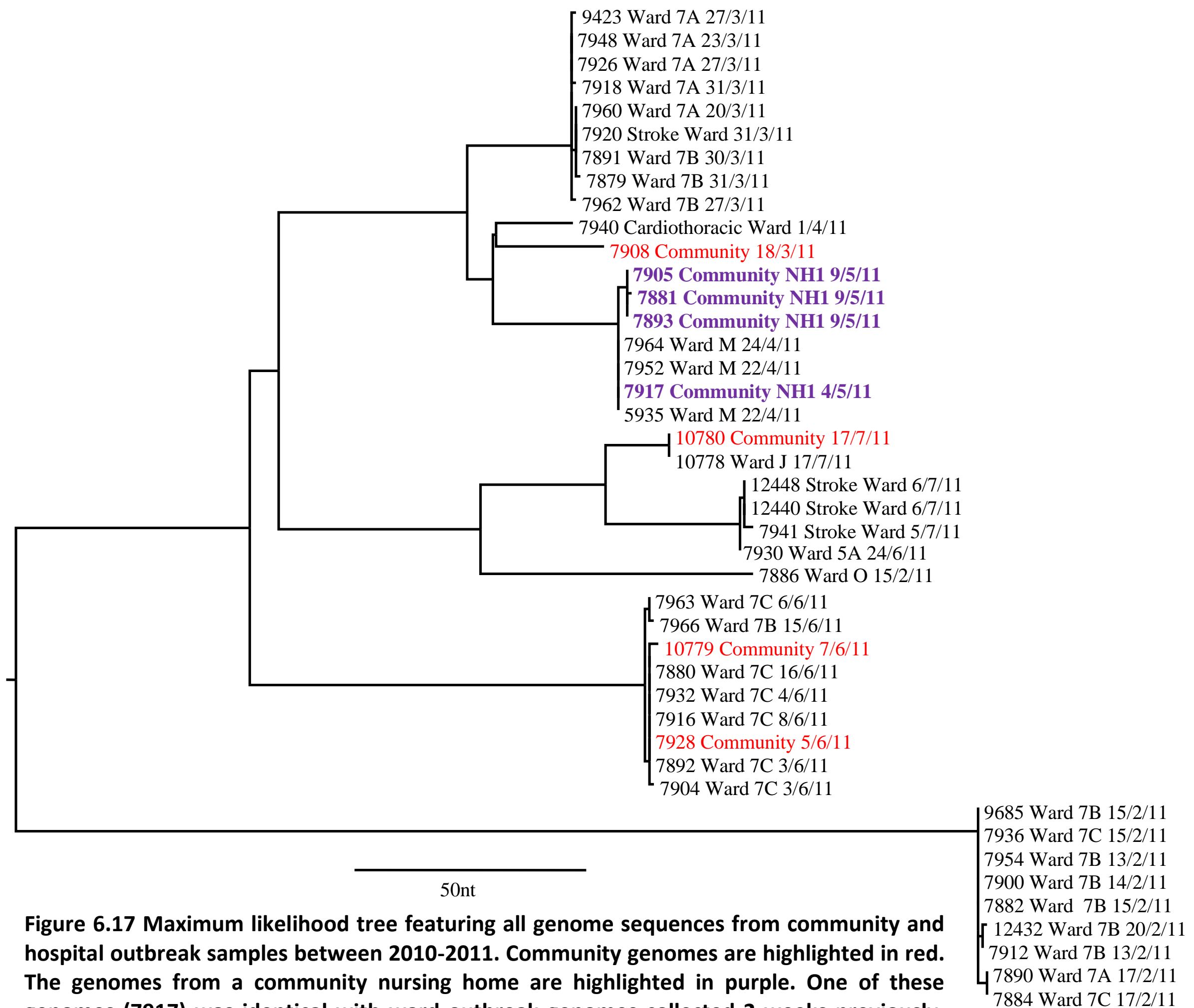


Figure 6.17 Maximum likelihood tree featuring all genome sequences from community and hospital outbreak samples between 2010-2011. Community genomes are highlighted in red. The genomes from a community nursing home are highlighted in purple. One of these genomes (7917) was identical with ward outbreak genomes collected 2 weeks previously, suggesting possible onward transmission.

In this figure, two of the community genomes (10780 and 7928, both in red) share identical genomes with temporally linked hospital genomes. The sequence identity observed between some community and hospital inpatient norovirus genomes provides an indication that frequent transmission occurs between the community and healthcare environments, rather than the two settings harbouring independent pools of viruses. Furthermore, one genome from the community nursing home (7917, in purple) was

identical to genomes collected on a ward outbreak nearly two weeks previously. The patient who provided this sample had previously been admitted on the same ward where others had been infected and sequenced (Ward M). This suggests onward transmission from ward to community.

Similarities during the 2011-2012 season between community and hospital outbreak genomes are demonstrated in Figure 6.18. Once again, the tree confirms the presence of temporally linked hospital and community cases which shared genomes with minimal differences. During both February and April 2012, there were six community norovirus genomes sequenced which were identical to temporally related hospital norovirus genomes. However, temporally linked but divergent noroviruses were also identified during this season.

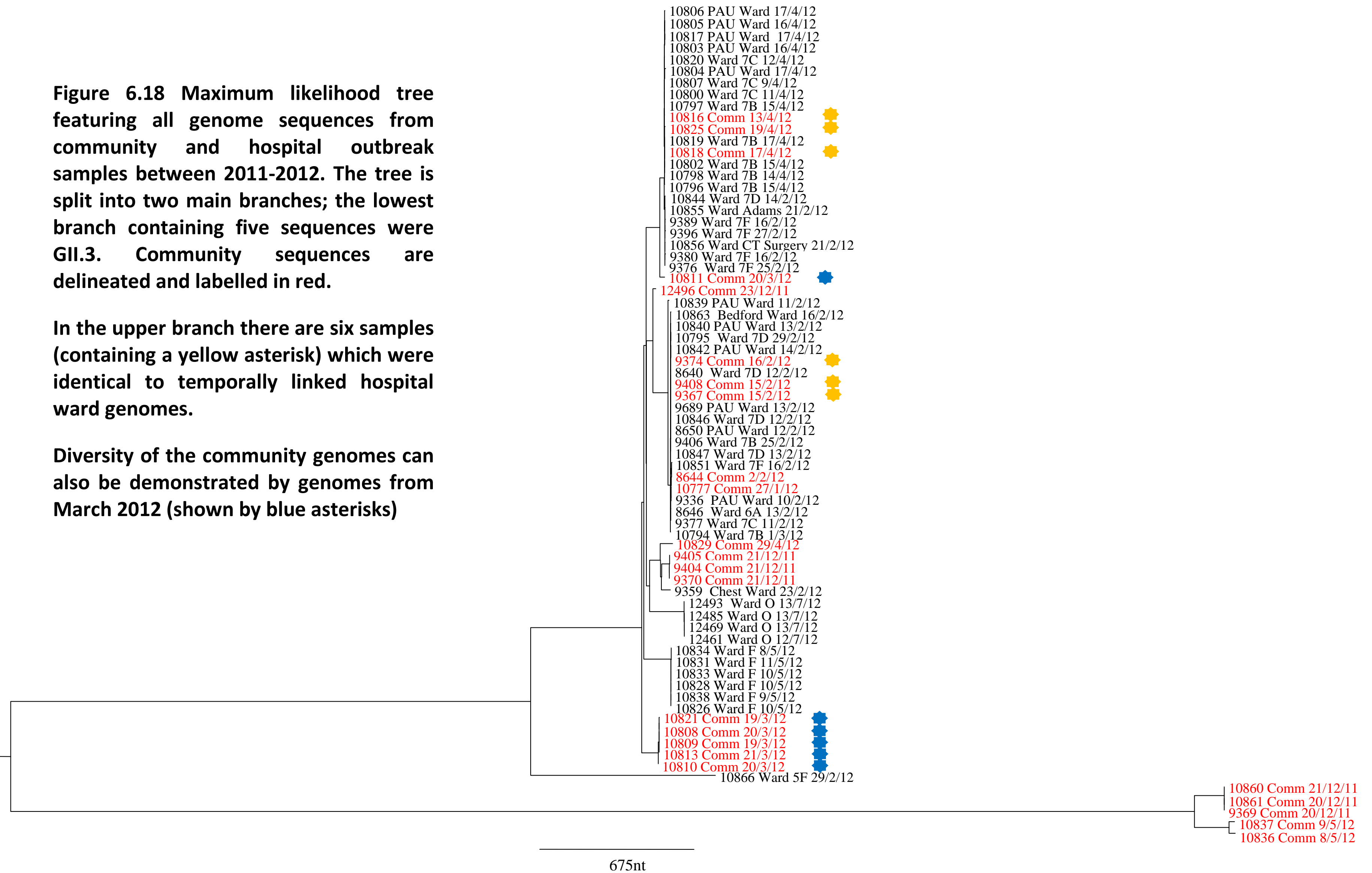


Figure 6.18 Maximum likelihood tree featuring all genome sequences from community and hospital outbreak samples between 2011-2012. The tree is split into two main branches; the lowest branch containing five sequences were GII.3. Community sequences are delineated and labelled in red.

In the upper branch there are six samples (containing a yellow asterisk) which were identical to temporally linked hospital ward genomes.

Diversity of the community genomes can also be demonstrated by genomes from March 2012 (shown by blue asterisks)

Community genomes from 2012-2013 season were also available. Unlike previous seasons, these community samples were divergent from temporally linked hospital outbreak genomes. A maximum likelihood tree of these samples are shown in Figure 6.19 below.

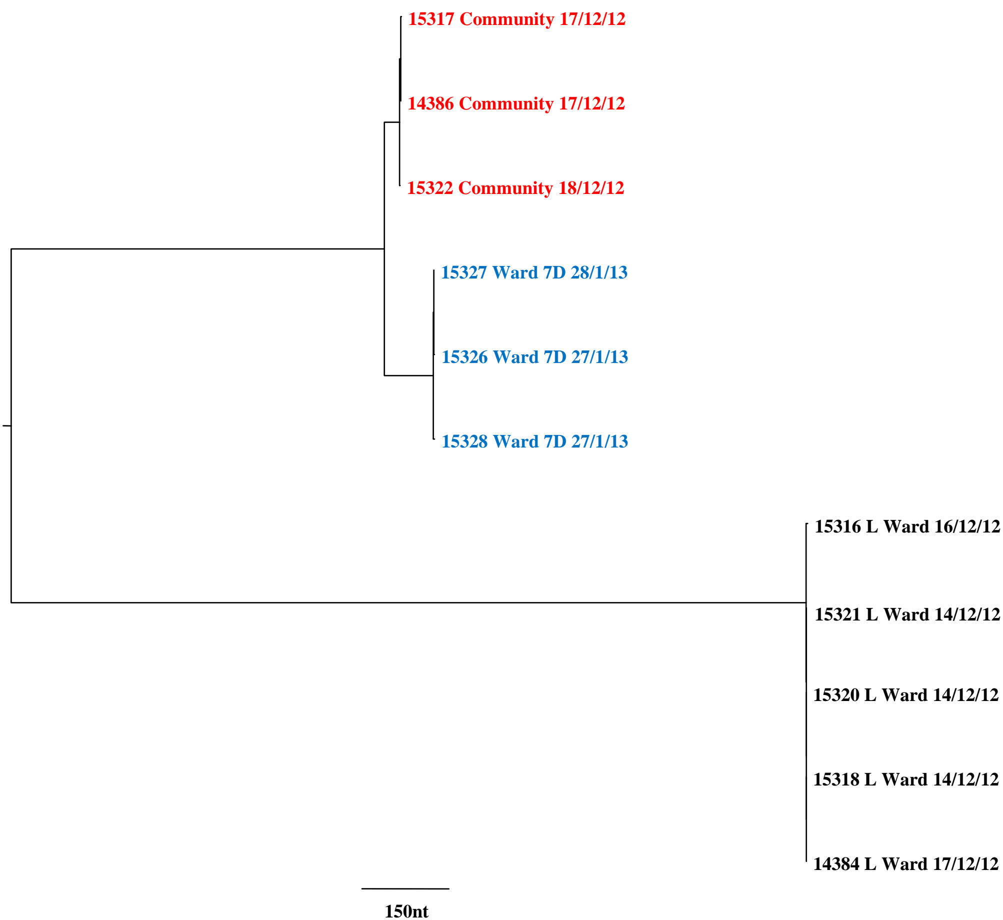


Figure 6.19. Phylogenetic maximum likelihood of 2012-2013 genomes within Oxfordshire. Community genomes are highlighted in red, whilst temporally linked (December 2012) hospital ward samples (highlighted in black) were highly divergent from their community genomes.

The information gained from the community genomes and temporally linked hospital outbreak genomes confirms that potential transmission links are present and that

identical strains of norovirus were found in the community and hospital concurrently. More divergent noroviruses were also found in the community.

6.7 Genomic diversity between geographical sites

Noroviruses collected within hospital outbreaks from other geographical UK sites were sequenced. Phylogenetic analysis of these data would facilitate the third aim; an investigation of the relationship between norovirus variants circulating in Oxfordshire, and those present in the healthcare system nationally. This would allow questions such as the extent of rapid national norovirus healthcare-associated transmission to be investigated. If virtually identical norovirus genomes were found in multiple geographic locations, the possibility of national waves of norovirus transmission within the healthcare system would be indicated. The phylogenetic relationships of the noroviruses within each of the additional locations further confirmed the findings reported earlier in this chapter with respect to the within-Oxfordshire healthcare-associated noroviruses. To avoid repetition, these 'within locality' analyses have been supplied in the supplementary information. .

For analytical purposes, and to avoid duplication all identical genomes within temporally linked outbreak clusters were excluded. National and regional representatives were included (n = 140; from eleven UK sites), and this is represented by a maximum likelihood phylogenetic tree (Figure 6.20). Both the Sydney 2012 and New Orleans 2009 reference genomes were included in this tree.

The tree was divided into two main branches which contained the two reference genomes. A SNV matrix of all genomes (n=62) sequenced has been included in Appendix 9.

Within this subset of 62 genomes, more than 27% (2,004 out of 7,489 sites) of the total norovirus genome were involved in variant nucleotide changes over the four seasons. Genomes sequenced from the same period of time from different geographic locations were different and diverse (Range 4 - 845 SNVs). These findings strongly suggest that the evolution and transmission of norovirus was occurring most rapidly within the regions, and circulating "local" reservoirs were the source of mutations or evolution. Transmission outside the immediate locality was apparently less frequent, but must have occurred because the newly emergent Sydney variant was identified within the data set in eleven locations.

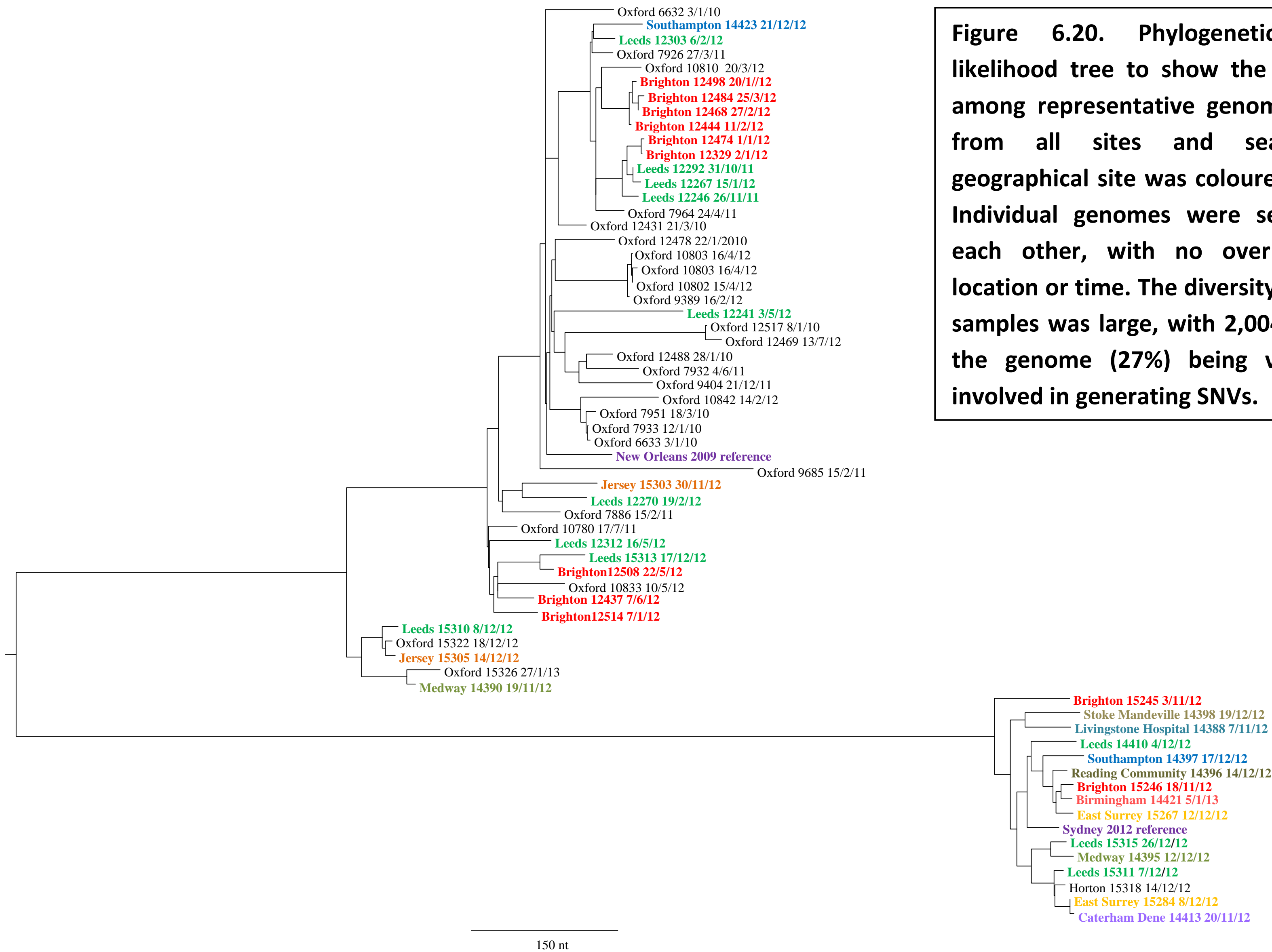


Figure 6.20. Phylogenetic maximum likelihood tree to show the relationships among representative genome sequences from all sites and seasons. Each geographical site was coloured separately. Individual genomes were separate from each other, with no overlap between location or time. The diversity within these samples was large, with 2,004 sites within the genome (27%) being variable sites involved in generating SNVs.

6.8 Overall relationships among local, regional and national noroviruses

To summarise graphically the overall relationships among genomes locally, regional and nationally, a pair-wise comparison of each genomes with is closest relative was performed (Figure 6.21). This was demonstrated both within the same and between seasons. This plot confirmed that, within wards, the majority of genomes were identical. Despite it being a minimum distance plot, the appearance of a divergent genome (i.e. more than 20 SNVs) was seen on one occasion within the ward, confirming a new introduction. Sequences between wards during the same season depict a slightly different pattern; although some genomes were identical, there were sequences with some diversity of up to 16 SNVs. In addition, more divergent genomes (>20 SNVs) were found. As discussed previously, pair-wise comparison of genomes between cities (i.e. between Brighton, Leeds and Oxford) revealed minimal identity within the same season, confirming that, geographically, norovirus did not "travel" frequently. Furthermore, genomes from different seasons were genetically divergent, confirming that they continued to evolve throughout the study period and excluding persistent environmental contamination as a major transmission route.

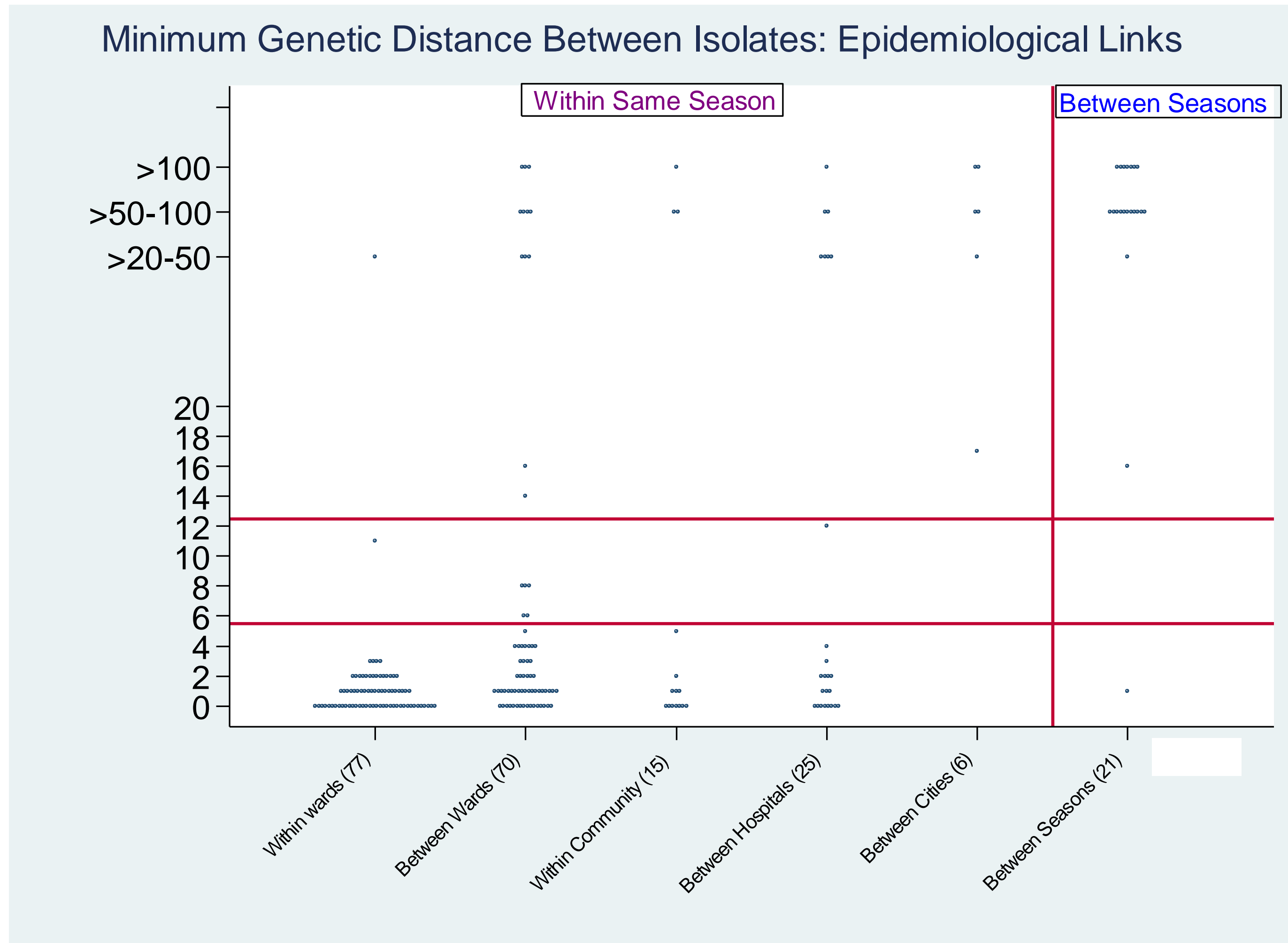


Figure 6.21 Minimum genetic distance between genomes. This plot was divided between genomes within the same ward, which were generally identical or very closely related (less than 5 SNVs). Genomes from different wards also demonstrated similarity, but divergent genomes were also present (>20 SNVs). Genomes from different cities were divergent, as well as those from different seasons.

6.9 Discussion

Current genotyping tools for norovirus, which utilise sequencing a number of small regions along the genome [66, 220, 221], have been shown to omit a number of significant SNVs along the genome outside of these regions. This confirms that whole genome sequencing offers more discrimination between closely related genomes.

In this chapter, a strict filtering criteria was used and only near full genomes of norovirus with uninterrupted gaps used for analyses [160]. As newer bioinformatics assembly packages become available with enhanced filtering criteria, many more genomes may be successfully assembled for comparison in the future.

Three fundamental aims have been described. Firstly, within wards during an outbreak, the majority of genomes were found to be identical. However, the presence of divergent genomes within a ward outbreak confirmed that separate introductions of noroviruses occurred. The demonstration of multiple introductions during ward outbreaks has important implications for hospital infection control. Concerns about poor and inadequate deep cleaning may be answered with data from whole genome sequencing confirming divergent new introductions of norovirus from the community [222]. Furthermore, there was no evidence that a similar strain of norovirus persisted within the hospital temporally, refuting that environmental contamination due to poor deep cleaning or ongoing carriage occurred. In real time, this detail can help the infection control professional with knowledge to respond quickly in cohorting and isolating symptomatic patients.

For my second aim, genomes from the community were analysed and compared with temporally related hospital outbreak samples. Some genomes were clearly divergent

from those in hospital, but a small minority had near or identical genomes to hospital genomes. This suggests that community noroviruses may have contributed to hospital ward outbreaks, and vice versa.

Thirdly, the analyses have been extended to other geographical locations, where similar transmission dynamics were observed within ward outbreaks; identical sequences within ward outbreaks, and the presence of diverse genomes within the same ward, suggesting multiple introductions of virus into the hospital. Comparison of hospital genomes from geographically unrelated regions showed no relationship both spatially and temporally. The data suggested that norovirus transmission was geographically localised with the virus possibly evolving locally.

An incidental finding from the genomes sequenced from the last season 2012-2013 confirmed the presence of a new recombinant strain of GII.4 norovirus, known as Sydney 2012, as reported recently in Australia, Denmark and Scotland [43, 206, 223]. The genomic data also confirmed the presence of recombinants with the previous strain, New Orleans 2009, with recombination occurring between the genome's open reading frames. This further illustrated the need for whole genomic sequencing, especially in light of recent reports that more recombination hotspots may be present within the genome [67]. Standard genotypic sequencing of small isolated regions may not give enough resolution to identify these recombinant events [67].

Although the sample numbers described in this thesis were small, they served as a good example of what had happened within three major hospital NHS trusts over the last four years (Oxford, Leeds and Brighton). The robustness of this approach and the wealth

of information gathered will benefit both the healthcare professionals during outbreaks and scientists in unravelling the evolution of norovirus.

Chapter 7. Final conclusions

Hospital norovirus outbreaks pose a large burden for healthcare [112]. New insights on the descriptive epidemiology of outbreaks have been described. Furthermore, I have developed a novel method of determining norovirus whole genome sequences and employing them in analyses to understand transmission dynamics within hospital outbreaks.

7.1 Whole genome sequences of norovirus can be determined using next generation sequencing technology

Two major components were pre-requisite for the production of whole genome sequences of norovirus: a robust and optimised laboratory protocol to extract and sequence norovirus RNA from faecal material, followed by a reliable and compatible bioinformatic package to generate whole genomic sequences for comparative analysis. Three sequencing approaches and a range of bioinformatics packages were evaluated to generate the norovirus genomes [168, 196, 198, 200, 201]. The relative performance of these methods and bioinformatic packages have been described, and the optimal method identified from the initial pilot study was scaled up to determine norovirus whole genomes sequences from a large number of clinical samples [159, 160]. This protocol was also adapted successfully for the determination of whole genome sequences of another positive sense RNA virus, Hepatitis C virus [159].

A total of 337 norovirus whole genomes, from eleven sites within the United Kingdom, were assembled using the method described. These were used in analyses to elucidate transmission dynamics within hospital outbreaks.

7.2 Epidemiological data question the reliability of the Kaplan criteria in outbreak identification

The Kaplan criteria have been adopted in many countries since they have been considered to provide a sensitive and specific method for systematically identifying viral gastroenteritis outbreaks [18, 153]. Although the duration of symptoms observed in our outbreak(s) was within the range specified by Kaplan, the data collected from Oxford, Leeds and Brighton over the past four seasons confirmed that vomiting occurred less frequently in all symptomatic patients (42% of all norovirus RT-PCR positive patients in Oxford versus >50% as described by Kaplan). The use of the term "winter vomiting virus" may no longer be appropriate to describe norovirus.

7.3 WGS provides enhanced discrimination among norovirus strains

Studies which aim to use norovirus sequences for epidemiological purposes currently determine the sequence of up to five small regions along the norovirus genome [66, 220]. Whole genome sequences revealed multiple single nucleotide variants missed by the current approach [154]. Consequently, these data demonstrated that previously accepted genotyping methods were insufficiently discriminatory to resolve transmission reliably.

7.4 Norovirus genomes and epidemiological data can be combined to characterise virus transmission events

Combining norovirus whole genome sequences with detailed epidemiological information provided sufficient resolution to discriminate between the majority of viruses studied. The major findings were as follows:

- Viruses sharing identical genomes were used to rule in likely transmission events, especially when supporting epidemiological data were available. As a consequence, the data were also used to rule out transmission events, (despite being suggested by epidemiological data) when viruses with divergent genomes infected patients in adjacent bays or beds within the same ward.
- Persistence of environmental contamination leading to later infection in the ward was not seen, as no secondary infections occurred with strains sharing identity temporally.
- Transmission between hospital and community, in either direction, was suggested by complete or near identity of viral strains (<2 SNVs).
- Comparison of norovirus genomes from regions spanning the UK revealed divergent norovirus variants (still of the predominant GII.4 strain) within each location, reflecting the predominance of local strains causing outbreaks. This was even apparent when a novel Sydney variant dispersed across the UK and indicates strong geographical structuring of the data.
- Norovirus evolution occurred across the four seasons studied, and a new recombinant strain of the virus emerged in the UK as described in other countries [67, 206, 224].

7.5 Limitations of the study

7.5.1 Epidemiological Data

In this thesis, I attempted to collect as much epidemiological information as feasible. With the help of colleagues in infection control, I was able to capture detailed clinical and demographic information from symptomatic patients. However, additional information regarding their symptom history prior to arrival on wards was not

systematically obtained, therefore possible transmission whilst they were asymptomatic before movement to the ward was lacking. Using a linked electronic database, it may be possible in future to investigate the contribution to transmission by patients at any stage during their pathway through the hospital. In the case of community noroviruses, there was no additional epidemiological information on the spatial relationship between the genomes to allow detailed epidemiological investigations to be performed.

7.5.2 Missing norovirus sequences – due to technical limitations

A minority of samples (160 out of 584 (27%)) that underwent sequencing did not yield a complete genome sequence. The criteria for analysis was strict and many partially sequenced samples were excluded from analyses (87 out of 584 (15%)). Newer sequencing technology has just been released for research testing. As described in chapter 1, this relies on third generation single molecule sequencing, which promises to reduce costs for reagents and instrumentation [225]. Although the best available assembly program was used, updates and new programs are rapidly becoming available; these use different algorithms for analysis [226, 227], minimise the need for expensive computing power, and can potentially run from a desktop computer [228]. This will be explored in the future with a view to re-assembling all samples, to determine whether more full length genomes can be produced.

7.5.3 Missing norovirus sequences – due to missing clinical samples

An area that remains to be explored is the contribution made to transmission by symptomatic and asymptomatic hospital staff, and transient ward visitors including other healthcare professionals and relatives. I have not been able to request the participation of all ward-linked individuals during these outbreaks, and it remains an

important issue to address in the future. Ethical approval will be required to collect faecal samples from all patients (symptomatic and asymptomatic) as well as all healthcare professionals working within the affected ward.

7.6 Translational benefits and future directions

The information presented in this thesis provide the clinician or infection control practitioner an insight into norovirus outbreaks within and between hospitals in the UK over four seasons. The prospect of having whole genome sequencing optimised as a direct diagnostic test is feasible. With improvements in technology and reduced costs and labour, most samples will yield full and closed genomes for which outbreaks can be inferred or rejected, thus aiding rapid management of norovirus outbreaks. Regional and national scale analysis of such data can contribute to ongoing surveillance.

With the changing characteristics of norovirus, as evident in this thesis, it may become more difficult for the healthcare professional to determine on symptomatic grounds alone whether diarrhoeal episodes are most likely to be viral, bacterial or non infectious. The use of metagenomics in whole genome sequencing, as discussed in chapter 1, has the ability to detect and confirm the presence of pathogens within patients' samples. One possibility is that bacterial, fungal, or viral pathogens could be directly detected from clinical samples, thus replacing current microbiological approaches such as culture, microscopy or the agar plate [158, 229, 230].

The sequencing method described in this thesis required no *a priori* knowledge of pathogens genome sequneces, hence the RNA extracted can be generically used with "tailored" bioinformatic packages to discover the viral pathogen(s) of interest. This

creates the possibility for novel viral pathogens to be discovered as a consequence of routine analyses in the near future.

Lastly, this thesis has opened avenues for future work; the methods for whole genome sequencing norovirus can provide data to further appreciate the viruses' evolving characteristics. These will include studies to determine both the antigenicity and immunity between different variants of norovirus and the host; fundamental elements that may alleviate some of the current challenges in norovirus vaccine and antiviral research [49].

By collating a library of whole genome sequences from national and international centres, the wealth of information gathered may provide the grounds for a greater understanding for future surveillance, as well as vaccine development [44, 231].

References

1. Bryce, J., et al., *WHO estimates of the causes of death in children.* Lancet, 2005. **365**(9465): p. 1147-52.
2. Lozano, R., et al., *Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010.* Lancet, 2012. **380**(9859): p. 2095-128.
3. King, C.K., et al., *Managing acute gastroenteritis among children: oral rehydration, maintenance, and nutritional therapy.* MMWR Recomm Rep, 2003. **52**(RR-16): p. 1-16.
4. Kotloff, K.L., et al., *Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study.* Lancet, 2013. **382**(9888): p. 209-22.
5. Tam, C.C., et al., *Changes in causes of acute gastroenteritis in the United Kingdom over 15 years: microbiologic findings from 2 prospective, population-based studies of infectious intestinal disease.* Clin Infect Dis, 2012. **54**(9): p. 1275-86.
6. Chiba, S., et al., *An outbreak of gastroenteritis associated with calicivirus in an infant home.* J Med Virol, 1979. **4**(4): p. 249-54.
7. Green, K.Y., et al., *Taxonomy of the caliciviruses.* J Infect Dis, 2000. **181 Suppl 2**: p. S322-30.
8. Patel, M.M., et al., *Systematic literature review of role of noroviruses in sporadic gastroenteritis.* Emerg Infect Dis, 2008. **14**(8): p. 1224-31.
9. Dolin, R., *Perspective: Noroviruses - Challenges to control.* New England Journal of Medicine, 2007. **357**(11): p. 1072-1073.
10. Goodgame, R., *Norovirus gastroenteritis.* Curr Gastroenterol Rep, 2006. **8**(5): p. 401-8.
11. Koopmans, M., *Progress in understanding norovirus epidemiology.* Curr Opin Infect Dis, 2008. **21**(5): p. 544-52.
12. Kapikian, A.Z., et al., *Visualization by immune electron microscopy of a 27-nm particle associated with acute infectious nonbacterial gastroenteritis.* J Virol, 1972. **10**(5): p. 1075-81.
13. International Committee on Taxonomy of Viruses. and A.M.Q. King, *Virus taxonomy : classification and nomenclature of viruses : ninth report of the International Committee on Taxonomy of Viruses*. 2012, Amsterdam ; London: Elsevier Academic Press. x, 1327 p.
14. Dolin, R., et al., *Transmission of acute infectious nonbacterial gastroenteritis to volunteers by oral administration of stool filtrates.* J Infect Dis, 1971. **123**(3): p. 307-12.
15. Thornhill, T.S., et al., *Detection by Immune Electron-Microscopy of 26-Nm to 27-Nm Viruslike Particles Associated with 2 Family Outbreaks of Gastroenteritis.* Journal of Infectious Diseases, 1977. **135**(1): p. 20-27.
16. Greenberg, H.B., et al., *Solid-phase microtiter radioimmunoassay for detection of the Norwalk strain of acute nonbacterial, epidemic gastroenteritis virus and its antibodies.* J Med Virol, 1978. **2**(2): p. 97-108.
17. Blacklow, N.R. and G. Cukor, *Viral gastroenteritis.* N Engl J Med, 1981. **304**(7): p. 397-406.
18. Kaplan, J.E., et al., *Epidemiology of Norwalk gastroenteritis and the role of Norwalk virus in outbreaks of acute nonbacterial gastroenteritis.* Ann Intern Med, 1982. **96**(6 Pt 1): p. 756-61.
19. Herrmann, J.E., N.A. Nowak, and N.R. Blacklow, *Detection of Norwalk virus in stools by enzyme immunoassay.* J Med Virol, 1985. **17**(2): p. 127-33.
20. Ho, M.S., et al., *Viral gastroenteritis aboard a cruise ship.* Lancet, 1989. **2**(8669): p. 961-5.

21. Xi, J.N., et al., *Norwalk virus genome cloning and characterization.* Science, 1990. **250**(4987): p. 1580-3.
22. Jiang, X., et al., *Detection of Norwalk virus in stool by polymerase chain reaction.* J Clin Microbiol, 1992. **30**(10): p. 2529-34.
23. Jiang, X., et al., *Expression, self-assembly, and antigenicity of the Norwalk virus capsid protein.* J Virol, 1992. **66**(11): p. 6527-32.
24. Dingle, K.E., et al., *Human enteric Caliciviridae: the complete genome sequence and expression of virus-like particles from a genetic group II small round structured virus.* J Gen Virol, 1995. **76 (Pt 9)**: p. 2349-55.
25. Jiang, X., et al., *Sequence and genomic organization of Norwalk virus.* Virology, 1993. **195**(1): p. 51-61.
26. Hardy, M.E. and M.K. Estes, *Completion of the Norwalk virus genome sequence.* Virus Genes, 1996. **12**(3): p. 287-90.
27. Green, J., et al., *Broadly reactive reverse transcriptase polymerase chain reaction for the diagnosis of SRSV-associated gastroenteritis.* J Med Virol, 1995. **47**(4): p. 392-8.
28. Vinje, J. and M.P. Koopmans, *Molecular detection and epidemiology of small round-structured viruses in outbreaks of gastroenteritis in the Netherlands.* J Infect Dis, 1996. **174**(3): p. 610-5.
29. Fankhauser, R.L., et al., *Molecular epidemiology of "Norwalk-like viruses" in outbreaks of gastroenteritis in the United States.* J Infect Dis, 1998. **178**(6): p. 1571-8.
30. Green, J., et al., *The role of environmental contamination with small round structured viruses in a hospital outbreak investigated by reverse-transcriptase polymerase chain reaction assay.* J Hosp Infect, 1998. **39**(1): p. 39-45.
31. Daniels, N.A., et al., *A foodborne outbreak of gastroenteritis associated with Norwalk-like viruses: first molecular traceback to deli sandwiches contaminated during preparation.* J Infect Dis, 2000. **181**(4): p. 1467-70.
32. Fankhauser, R.L., et al., *Epidemiologic and molecular trends of "Norwalk-like viruses" associated with outbreaks of gastroenteritis in the United States.* J Infect Dis, 2002. **186**(1): p. 1-7.
33. Glass, R.I., et al., *The epidemiology of enteric caliciviruses from humans: a reassessment using new diagnostics.* J Infect Dis, 2000. **181 Suppl 2**: p. S254-61.
34. Cramer, E.H., et al., *Outbreaks of gastroenteritis associated with noroviruses on cruise ships - United States, 2002 (Reprinted from MMWR, vol 51, pg 1112-1115, 2002).* Jama-Journal of the American Medical Association, 2003. **289**(2): p. 167-169.
35. Lopman, B., et al., *Increase in viral gastroenteritis outbreaks in Europe and epidemic spread of new norovirus variant.* Lancet, 2004. **363**(9410): p. 682-8.
36. Lopman, B.A., et al., *Viral gastroenteritis outbreaks in Europe, 1995-2000.* Emerg Infect Dis, 2003. **9**(1): p. 90-6.
37. Kageyama, T., et al., *Broadly reactive and highly sensitive assay for Norwalk-like viruses based on real-time quantitative reverse transcription-PCR.* Journal of Clinical Microbiology, 2003. **41**(4): p. 1548-1557.
38. Duizer, E., et al., *Laboratory efforts to cultivate noroviruses.* J Gen Virol, 2004. **85**(Pt 1): p. 79-87.
39. Gallimore, C.I., et al., *Asymptomatic and symptomatic excretion of noroviruses during a hospital outbreak of gastroenteritis.* J Clin Microbiol, 2004. **42**(5): p. 2271-4.
40. Fauquet, C.M. and D. Fargette, *International Committee on Taxonomy of Viruses and the 3,142 unassigned species.* Virol J, 2005. **2**: p. 64.
41. Bull, R.A., et al., *Emergence of a new norovirus genotype II.4 variant associated with global outbreaks of gastroenteritis.* J Clin Microbiol, 2006. **44**(2): p. 327-33.
42. Teunis, P.F., et al., *Norwalk virus: how infectious is it?* J Med Virol, 2008. **80**(8): p. 1468-76.

43. van Beek, J., et al., *Indications for worldwide increased norovirus activity associated with emergence of a new variant of genotype II.4, late 2012.* Eurosurveillance, 2013. **18**(1): p. 8-9.
44. Debbink, K., L.C. Lindesmith, and R.S. Baric, *The State of Norovirus Vaccines.* Clin Infect Dis, 2014.
45. Zheng, D.P., et al., *Norovirus classification and proposed strain nomenclature.* Virology, 2006. **346**(2): p. 312-23.
46. Belliot, G., et al., *In vitro proteolytic processing of the MD145 norovirus ORF1 nonstructural polyprotein yields stable precursors and products similar to those detected in calicivirus-infected cells.* Journal of Virology, 2003. **77**(20): p. 10957-10974.
47. Prasad, B.V.V., et al., *X-ray crystallographic structure of the Norwalk virus capsid.* Science, 1999. **286**(5438): p. 287-290.
48. Glass, P.J., et al., *Norwalk virus open reading frame 3 encodes a minor structural protein.* Journal of Virology, 2000. **74**(14): p. 6581-6591.
49. Donaldson, E.F., et al., *Viral shape-shifting: norovirus evasion of the human immune system.* Nat Rev Microbiol, 2010. **8**(3): p. 231-41.
50. Vongpunsawad, S., B.V. Venkataram Prasad, and M.K. Estes, *Norwalk Virus Minor Capsid Protein VP2 Associates within the VP1 Shell Domain.* J Virol, 2013. **87**(9): p. 4818-25.
51. Prasad, B.V., et al., *Three-dimensional structure of baculovirus-expressed Norwalk virus capsids.* J Virol, 1994. **68**(8): p. 5117-25.
52. Prasad, B.V., et al., *X-ray crystallographic structure of the Norwalk virus capsid.* Science, 1999. **286**(5438): p. 287-90.
53. Bertolotti-Ciarlet, A., et al., *Structural requirements for the assembly of Norwalk virus-like particles.* J Virol, 2002. **76**(8): p. 4044-55.
54. Tan, M., et al., *Mutations within the P2 domain of norovirus capsid affect binding to human histo-blood group antigens: evidence for a binding pocket.* J Virol, 2003. **77**(23): p. 12562-71.
55. Bu, W., et al., *Structural basis for the receptor binding specificity of Norwalk virus.* J Virol, 2008. **82**(11): p. 5340-7.
56. Bok, K. and K.Y. Green, *Norovirus gastroenteritis in immunocompromised patients.* N Engl J Med, 2012. **367**(22): p. 2126-32.
57. Weir, M.L., A. Klip, and W.S. Trimble, *Identification of a human homologue of the vesicle-associated membrane protein (VAMP)-associated protein of 33 kDa (VAP-33): a broadly expressed protein that binds to VAMP.* Biochem J, 1998. **333 (Pt 2)**: p. 247-51.
58. Pfister, T. and E. Wimmer, *Polypeptide p41 of a Norwalk-like virus is a nucleic acid-independent nucleoside triphosphatase.* J Virol, 2001. **75**(4): p. 1611-9.
59. Doedens, J.R., T.H. Giddings, Jr., and K. Kirkegaard, *Inhibition of endoplasmic reticulum-to-Golgi traffic by poliovirus protein 3A: genetic and ultrastructural analysis.* J Virol, 1997. **71**(12): p. 9054-64.
60. Burroughs, J.N. and F. Brown, *Presence of a covalently linked protein on calicivirus RNA.* J Gen Virol, 1978. **41**(2): p. 443-6.
61. Herbert, T.P., I. Brierley, and T.D. Brown, *Identification of a protein linked to the genomic and subgenomic mRNAs of feline calicivirus and its role in translation.* J Gen Virol, 1997. **78 (Pt 5)**: p. 1033-40.
62. Daughenbaugh, K.F., C.E. Wobus, and M.E. Hardy, *VPg of murine norovirus binds translation initiation factors in infected cells.* Virol J, 2006. **3**: p. 33.
63. Nakamura, K., et al., *A norovirus protease structure provides insights into active and substrate binding site integrity.* Journal of Virology, 2005. **79**(21): p. 13685-13693.
64. Hogbom, M., et al., *The active form of the norovirus RNA-dependent RNA polymerase is a homodimer with cooperative activity.* J Gen Virol, 2009. **90**(Pt 2): p. 281-91.
65. Lindesmith, L.C., et al., *Mechanisms of GII.4 norovirus persistence in human populations.* PLoS Med, 2008. **5**(2): p. e31.

66. Kroneman, A., et al., *Proposal for a unified norovirus nomenclature and genotyping.* Arch Virol, 2013. **158**(10): p. 2059-68.
67. Eden, J.S., et al., *Recombination within the pandemic norovirus GII.4 lineage.* J Virol, 2013. **87**(11): p. 6270-82.
68. Cheetham, S., et al., *Pathogenesis of a genogroup II human norovirus in gnotobiotic pigs.* J Virol, 2006. **80**(21): p. 10372-81.
69. Bok, K., et al., *Chimpanzees as an animal model for human norovirus infection and vaccine development.* Proc Natl Acad Sci U S A, 2011. **108**(1): p. 325-30.
70. Hutson, A.M., et al., *Norwalk virus infection and disease is associated with ABO histo-blood group type.* J Infect Dis, 2002. **185**(9): p. 1335-7.
71. Graham, D.Y., et al., *Norwalk virus infection of volunteers: new insights based on improved assays.* J Infect Dis, 1994. **170**(1): p. 34-43.
72. Lindesmith, L., et al., *Human susceptibility and resistance to Norwalk virus infection.* Nat Med, 2003. **9**(5): p. 548-53.
73. Frenck, R., et al., *Predicting susceptibility to norovirus GII.4 by use of a challenge model involving humans.* J Infect Dis, 2012. **206**(9): p. 1386-93.
74. Lindesmith, L., et al., *Cellular and humoral immunity following Snow Mountain virus challenge.* J Virol, 2005. **79**(5): p. 2900-9.
75. Siebenga, J.J., et al., *Epochal evolution of GGII.4 norovirus capsid proteins from 1995 to 2006.* Journal of Virology, 2007. **81**(18): p. 9932-9941.
76. Donaldson, E.F., et al., *Norovirus pathogenesis: mechanisms of persistence and immune evasion in human populations.* Immunol Rev, 2008. **225**: p. 190-211.
77. Velasquez, L.S., et al., *Intranasal delivery of Norwalk virus-like particles formulated in an in situ gelling, dry powder vaccine.* Vaccine, 2011. **29**(32): p. 5221-31.
78. Atmar, R.L., et al., *Norovirus vaccine against experimental human Norwalk Virus illness.* N Engl J Med, 2011. **365**(23): p. 2178-87.
79. Huang, P., et al., *Noroviruses bind to human ABO, Lewis, and secretor histo-blood group antigens: identification of 4 distinct strain-specific patterns.* J Infect Dis, 2003. **188**(1): p. 19-31.
80. Jones, M.K., et al., *Enteric bacteria promote human and mouse norovirus infection of B cells.* Science, 2014. **346**(6210): p. 755-9.
81. Patel, M.M., et al., *Noroviruses: a comprehensive review.* J Clin Virol, 2009. **44**(1): p. 1-8.
82. Mounts, A.W., et al., *Cold weather seasonality of gastroenteritis associated with Norwalk-like viruses.* J Infect Dis, 2000. **181 Suppl 2**: p. S284-7.
83. Siebenga, J.J., et al., *Norovirus illness is a global problem: emergence and spread of norovirus GII.4 variants, 2001-2007.* J Infect Dis, 2009. **200**(5): p. 802-12.
84. Bok, K., et al., *Evolutionary dynamics of GII.4 noroviruses over a 34-year period.* J Virol, 2009. **83**(22): p. 11890-901.
85. Jiang, X., et al., *Characterization of a novel human calicivirus that may be a naturally occurring recombinant.* Arch Virol, 1999. **144**(12): p. 2377-87.
86. Drake, J.W. and J.J. Holland, *Mutation rates among RNA viruses.* Proceedings of the National Academy of Sciences of the United States of America, 1999. **96**(24): p. 13910-13913.
87. Roossinck, M.J., *Mechanisms of plant virus evolution.* Annu Rev Phytopathol, 1997. **35**: p. 191-209.
88. Siebenga, J.J., et al., *Phylodynamic Reconstruction Reveals Norovirus GII. 4 Epidemic Expansions and their Molecular Determinants.* PLoS Pathog, 2010. **6**(5).
89. Bull, R.A., et al., *Rapid evolution of pandemic noroviruses of the GII.4 lineage.* PLoS Pathog, 2010. **6**(3): p. e1000831.
90. Domingo, E., et al., *Quasispecies dynamics and RNA virus extinction.* Virus Res, 2005. **107**(2): p. 129-39.

91. Pfeiffer, J.K. and K. Kirkegaard, *Increased fidelity reduces poliovirus fitness and virulence under selective pressure in mice.* PLoS Pathog, 2005. **1**(2): p. 102-110.
92. Bull, R.A., et al., *Contribution of Intra- and Interhost Dynamics to Norovirus Evolution.* Journal of Virology, 2012. **86**(6): p. 3219-3229.
93. Bergstrom, C.T., P. McElhany, and L.A. Real, *Transmission bottlenecks as determinants of virulence in rapidly evolving pathogens.* Proc Natl Acad Sci U S A, 1999. **96**(9): p. 5095-100.
94. Bull, R.A., et al., *Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection.* PLoS Pathog, 2011. **7**(9): p. e1002243.
95. Fischer, W., et al., *Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing.* PLoS One, 2010. **5**(8): p. e12303.
96. Keele, B.F., *Identifying and characterizing recently transmitted viruses.* Current Opinion in Hiv and Aids, 2010. **5**(4): p. 327-334.
97. Siebenga, J.J., et al., *High prevalence of prolonged norovirus shedding and illness among hospitalized patients: a model for in vivo molecular evolution.* J Infect Dis, 2008. **198**(7): p. 994-1001.
98. Bull, R.A., et al., *Norovirus recombination in ORF1/ORF2 overlap.* Emerg Infect Dis, 2005. **11**(7): p. 1079-85.
99. Han, G.Z. and M. Worobey, *Homologous recombination in negative sense RNA viruses.* Viruses, 2011. **3**(8): p. 1358-73.
100. Nayak, M.K., et al., *A new variant of Norovirus GII.4/2007 and inter-genotype recombinant strains of NVGII causing acute watery diarrhoea among children in Kolkata, India.* J Clin Virol, 2009. **45**(3): p. 223-9.
101. Eden, J.S., et al., *Norovirus GII.4 variant 2006b caused epidemics of acute gastroenteritis in Australia during 2007 and 2008.* J Clin Virol, 2010. **49**(4): p. 265-71.
102. Waters, A., S. Coughlan, and W.W. Hall, *Characterisation of a novel recombination event in the norovirus polymerase gene.* Virology, 2007. **363**(1): p. 11-4.
103. Ambert-Balay, K., et al., *Characterization of new recombinant noroviruses.* J Clin Microbiol, 2005. **43**(10): p. 5179-86.
104. Rohayem, J., J. Munch, and A. Rethwilm, *Evidence of recombination in the norovirus capsid gene.* J Virol, 2005. **79**(8): p. 4977-90.
105. Lopman, B.A., D.W. Brown, and M. Koopmans, *Human caliciviruses in Europe.* J Clin Virol, 2002. **24**(3): p. 137-60.
106. Schmid, D., et al., *Outbreak of norovirus infection associated with contaminated flood water, Salzburg, 2005.* Euro Surveill, 2005. **10**(6): p. E050616 3.
107. Glass, R.I., U.D. Parashar, and M.K. Estes, *Norovirus gastroenteritis.* N Engl J Med, 2009. **361**(18): p. 1776-85.
108. Rockx, B., et al., *Natural history of human calicivirus infection: a prospective cohort study.* Clin Infect Dis, 2002. **35**(3): p. 246-53.
109. Atmar, R.L., et al., *Norwalk virus shedding after experimental human infection.* Emerg Infect Dis, 2008. **14**(10): p. 1553-7.
110. Turcios-Ruiz, R.M., et al., *Outbreak of necrotizing enterocolitis caused by norovirus in a neonatal intensive care unit.* J Pediatr, 2008. **153**(3): p. 339-44.
111. Sai, L., et al., *Epidemiology and clinical features of rotavirus and norovirus infection among children in Ji'nan, China.* Virol J, 2013. **10**: p. 302.
112. Lopman, B.A., et al., *Epidemiology and cost of nosocomial gastroenteritis, Avon, England, 2002-2003.* Emerg Infect Dis, 2004. **10**(10): p. 1827-34.
113. Fretz, R., et al., *Outbreaks of gastroenteritis due to infections with Norovirus in Switzerland, 2001-2003.* Epidemiol Infect, 2005. **133**(3): p. 429-37.
114. Morotti, R.A., et al., *Calicivirus infection in pediatric small intestine transplant recipients: pathological considerations.* Hum Pathol, 2004. **35**(10): p. 1236-40.

115. Westhoff, T.H., et al., *Chronic norovirus infection in renal transplant recipients.* Nephrol Dial Transplant, 2009. **24**(3): p. 1051-3.
116. Troeger, H., et al., *Structural and functional changes of the duodenum in human norovirus infection.* Gut, 2009. **58**(8): p. 1070-7.
117. Dolin, R., et al., *Viral gastroenteritis induced by the Hawaii agent. Jejunal histopathology and serologic response.* Am J Med, 1975. **59**(6): p. 761-8.
118. Medici, M.C., et al., *Norovirus RNA in the blood of a child with gastroenteritis and convulsions--A case report.* J Clin Virol, 2010. **48**(2): p. 147-9.
119. Schreiber, D.S., N.R. Blacklow, and J.S. Trier, *The mucosal lesion of the proximal small intestine in acute infectious nonbacterial gastroenteritis.* N Engl J Med, 1973. **288**(25): p. 1318-23.
120. Meeroff, J.C., et al., *Abnormal gastric motor function in viral gastroenteritis.* Ann Intern Med, 1980. **92**(3): p. 370-3.
121. Agus, S.G., et al., *Acute infectious nonbacterial gastroenteritis: intestinal histopathology. Histologic and enzymatic alterations during illness produced by the Norwalk agent in man.* Ann Intern Med, 1973. **79**(1): p. 18-25.
122. Mathijs, E., et al., *A review of known and hypothetical transmission routes for noroviruses.* Food Environ Virol, 2012. **4**(4): p. 131-52.
123. Vega, E., et al., *Genotypic and Epidemiologic Trends of Norovirus Outbreaks in the United States, 2009-2013.* J Clin Microbiol, 2013.
124. Lopman, B.A., et al., *Two epidemiologic patterns of norovirus outbreaks: surveillance in England and wales, 1992-2000.* Emerg Infect Dis, 2003. **9**(1): p. 71-7.
125. Siebenga, J.J., et al., *Gastroenteritis caused by norovirus GGII.4, The Netherlands, 1994-2005.* Emerg Infect Dis, 2007. **13**(1): p. 144-6.
126. Gallimore, C.I., et al., *Environmental monitoring for gastroenteric viruses in a pediatric primary immunodeficiency unit.* J Clin Microbiol, 2006. **44**(2): p. 395-9.
127. Gallimore, C.I., et al., *Contamination of the hospital environment with gastroenteric viruses: comparison of two pediatric wards over a winter season.* J Clin Microbiol, 2008. **46**(9): p. 3112-5.
128. Diggs, R., et al., *Norovirus outbreak in an elementary school district of Columbia, February 2007 (Reprinted from MMWR, vol 56, pg 1340-1343, 2008).* Jama-Journal of the American Medical Association, 2008. **299**(6): p. 627-630.
129. Bosch, A., *Human enteric viruses in the water environment: a minireview.* Int Microbiol, 1998. **1**(3): p. 191-6.
130. Borchardt, M.A., et al., *Viruses in nondisinfected drinking water from municipal wells and community incidence of acute gastrointestinal illness.* Environ Health Perspect, 2012. **120**(9): p. 1272-9.
131. Sartorius, B., et al., *Outbreak of norovirus in Vastra Gotaland associated with recreational activities at two lakes during August 2004.* Scand J Infect Dis, 2007. **39**(4): p. 323-31.
132. Yoder, J.S., et al., *Surveillance for waterborne-disease outbreaks associated with recreational water--United States, 2001-2002.* MMWR Surveill Summ, 2004. **53**(8): p. 1-22.
133. Dziuban, E.J., et al., *Surveillance for waterborne disease and outbreaks associated with recreational water--United States, 2003-2004.* MMWR Surveill Summ, 2006. **55**(12): p. 1-30.
134. Silverman, A.I., et al., *Quantification of human norovirus GII, human adenovirus, and fecal indicator organisms in wastewater used for irrigation in Accra, Ghana.* J Water Health, 2013. **11**(3): p. 473-88.
135. Schaeffer, J., et al., *Norovirus contamination on French marketed oysters.* Int J Food Microbiol, 2013. **166**(2): p. 244-8.

136. Le Guyader, F.S., et al., *Detection of multiple noroviruses associated with an international gastroenteritis outbreak linked to oyster consumption.* Journal of Clinical Microbiology, 2006. **44**(11): p. 3878-3882.
137. Kingsley, D.H., G.K. Meade, and G.P. Richards, *Detection of both hepatitis A virus and Norwalk-like virus in imported clams associated with food-borne illness.* Applied and Environmental Microbiology, 2002. **68**(8): p. 3914-3918.
138. Wang, J. and Z. Deng, *Detection and forecasting of oyster norovirus outbreaks: recent advances and future perspectives.* Mar Environ Res, 2012. **80**: p. 62-9.
139. Made, D., et al., *Detection and Typing of Norovirus from Frozen Strawberries Involved in a Large-Scale Gastroenteritis Outbreak in Germany.* Food Environ Virol, 2013.
140. Lee, H.M., et al., *Rapid detection of norovirus from fresh lettuce using immunomagnetic separation and a quantum dots assay.* J Food Prot, 2013. **76**(4): p. 707-11.
141. Koser, C.U., et al., *Whole-genome sequencing for rapid susceptibility testing of M. tuberculosis.* N Engl J Med, 2013. **369**(3): p. 290-2.
142. Reuter, S., et al., *A pilot study of rapid whole-genome sequencing for the investigation of a Legionella outbreak.* BMJ Open, 2013. **3**(1).
143. Torok, M.E., et al., *Rapid whole-genome sequencing for investigation of a suspected tuberculosis outbreak.* J Clin Microbiol, 2013. **51**(2): p. 611-4.
144. Eyre, D.W., et al., *A pilot study of rapid benchtop sequencing of Staphylococcus aureus and Clostridium difficile for outbreak detection and surveillance.* BMJ Open, 2012. **2**(3).
145. Chadwick, P.R., et al., *Management of hospital outbreaks of gastro-enteritis due to small round structured viruses.* Journal of Hospital Infection, 2000. **45**(1): p. 1-10.
146. Illingworth, E., et al., *Is closure of entire wards necessary to control norovirus outbreaks in hospital? Comparing the effectiveness of two infection control strategies.* Journal of Hospital Infection, 2011. **79**(1): p. 32-7.
147. Harris, J.P., B.A. Lopman, and S.J. O'Brien, *Infection control measures for norovirus: a systematic review of outbreaks in semi-enclosed settings.* J Hosp Infect, 2010. **74**(1): p. 1-9.
148. Weber, D.J., et al., *Role of hospital surfaces in the transmission of emerging health care-associated pathogens: norovirus, Clostridium difficile, and Acinetobacter species.* Am J Infect Control, 2010. **38**(5 Suppl 1): p. S25-33.
149. Moe, C.L., *Preventing norovirus transmission: how should we handle food handlers?* Clin Infect Dis, 2009. **48**(1): p. 38-40.
150. Aoki, Y., et al., *Duration of norovirus excretion and the longitudinal course of viral load in norovirus-infected elderly patients.* J Hosp Infect, 2010. **75**(1): p. 42-6.
151. Rahamat-Langendoen, J.C., et al., *Rapid detection of a norovirus pseudo-outbreak by using real-time sequence based information.* J Clin Virol, 2013. **58**(1): p. 245-8.
152. Kaplan, J.E., et al., *The frequency of a Norwalk-like pattern of illness in outbreaks of acute gastroenteritis.* Am J Public Health, 1982. **72**(12): p. 1329-32.
153. Turcios, R.M., et al., *Reevaluation of epidemiological criteria for identifying outbreaks of acute gastroenteritis due to norovirus: United States, 1998-2000.* Clinical Infectious Diseases, 2006. **42**(7): p. 964-969.
154. Xerry, J., et al., *Transmission events within outbreaks of gastroenteritis determined through analysis of nucleotide sequences of the P2 domain of genogroup II noroviruses.* J Clin Microbiol, 2008. **46**(3): p. 947-53.
155. Bruggink, L.D., et al., *Evaluation of the RIDA((R))QUICK immunochromatographic norovirus detection assay using specimens from Australian gastroenteritis incidents.* J Virol Methods, 2011. **173**(1): p. 121-6.
156. Lee, B.Y., et al., *Economic impact of outbreaks of norovirus infection in hospitals.* Infect Control Hosp Epidemiol, 2011. **32**(2): p. 191-3.

157. Walker, T.M., et al., *Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study.* Lancet Infectious Diseases, 2013. **13**(2): p. 137-146.
158. Didelot, X., et al., *Transforming clinical microbiology with bacterial genome sequencing.* Nat Rev Genet, 2012. **13**(9): p. 601-12.
159. Batty, E.M., et al., *A Modified RNA-Seq Approach for Whole Genome Sequencing of RNA Viruses from Faecal and Blood Samples.* PLoS One, 2013. **8**(6): p. e66129.
160. Wong, T.H., et al., *Whole genome sequencing and de novo assembly identifies Sydney-like variant noroviruses and recombinants during the winter 2012/2013 outbreak in England.* Virol J, 2013. **10**(1): p. 335.
161. Sanger, F., *The Croonian Lecture, 1975. Nucleotide sequences in DNA.* Proc R Soc Lond B Biol Sci, 1975. **191**(1104): p. 317-33.
162. Sanger, F. and A.R. Coulson, *A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.* J Mol Biol, 1975. **94**(3): p. 441-8.
163. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
164. Maxam, A.M. and W. Gilbert, *A new method for sequencing DNA.* Proc Natl Acad Sci U S A, 1977. **74**(2): p. 560-4.
165. Smith, L.M., et al., *Fluorescence detection in automated DNA sequence analysis.* Nature, 1986. **321**(6071): p. 674-9.
166. Ansorge, W., et al., *Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis.* Nucleic Acids Res, 1987. **15**(11): p. 4593-602.
167. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.
168. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.
169. Loman, N.J., et al., *Performance comparison of benchtop high-throughput sequencing platforms.* Nat Biotechnol, 2012. **30**(5): p. 434-9.
170. Hoper, D., B. Hoffmann, and M. Beer, *Simple, sensitive, and swift sequencing of complete H5N1 avian influenza virus genomes.* J Clin Microbiol, 2009. **47**(3): p. 674-9.
171. Woolhouse, M.E., et al., *Temporal trends in the discovery of human viruses.* Proc Biol Sci, 2008. **275**(1647): p. 2111-5.
172. Ksiazek, T.G., et al., *A novel coronavirus associated with severe acute respiratory syndrome.* N Engl J Med, 2003. **348**(20): p. 1953-66.
173. Drosten, C., et al., *Severe acute respiratory syndrome: identification of the etiological agent.* Trends Mol Med, 2003. **9**(8): p. 325-7.
174. Rohde, H., et al., *Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4.* N Engl J Med, 2011. **365**(8): p. 718-24.
175. Victoria, J.G., et al., *Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis.* J Virol, 2009. **83**(9): p. 4642-51.
176. Steinhauer, D.A., E. Domingo, and J.J. Holland, *Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase.* Gene, 1992. **122**(2): p. 281-8.
177. Eriksson, N., et al., *Viral population estimation using pyrosequencing.* PLoS Comput Biol, 2008. **4**(4): p. e1000074.
178. Homs, M., et al., *Ultra-deep pyrosequencing analysis of the hepatitis B virus preCore region and main catalytic motif of the viral polymerase in the same viral genome.* Nucleic Acids Res, 2011. **39**(19): p. 8457-71.
179. Hiraga, N., et al., *Rapid emergence of telaprevir resistant hepatitis C virus strain from wildtype clone in vivo.* Hepatology, 2011. **54**(3): p. 781-8.
180. Varghese, V., et al., *Minority variants associated with transmitted and acquired HIV-1 nonnucleoside reverse transcriptase inhibitor resistance: implications for the use of*

second-generation nonnucleoside reverse transcriptase inhibitors. J Acquir Immune Defic Syndr, 2009. **52**(3): p. 309-15.
181. Tapparel, C., et al., *Rhinovirus genome variation during chronic upper and lower respiratory tract infections.* PLoS One, 2011. **6**(6): p. e21163.
182. Willerth, S.M., et al., *Development of a low bias method for characterizing viral populations using next generation sequencing technology.* PLoS One, 2010. **5**(10): p. e13564.
183. Bimber, B.N., et al., *Whole-genome characterization of human and simian immunodeficiency virus intrahost diversity by ultradeep pyrosequencing.* J Virol, 2010. **84**(22): p. 12087-92.
184. Eyre, D.W., et al., *A pilot study of rapid benchtop sequencing of Staphylococcus aureus and Clostridium difficile for outbreak detection and surveillance.* Bmj Open, 2012. **2**(3).
185. Chin, C.S., et al., *The origin of the Haitian cholera outbreak strain.* N Engl J Med, 2011. **364**(1): p. 33-42.
186. Daly, G.M., et al., *A viral discovery methodology for clinical biopsy samples utilising massively parallel next generation sequencing.* PLoS One, 2011. **6**(12): p. e28879.
187. Nakamura, S., et al., *Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach.* PLoS One, 2009. **4**(1): p. e4219.
188. Bull, R.A. and P.A. White, *Mechanisms of GII.4 norovirus evolution.* Trends Microbiol, 2011. **19**(5): p. 233-40.
189. Lauck, M., et al., *Analysis of hepatitis C virus intrahost diversity across the coding region by ultradeep pyrosequencing.* J Virol, 2012. **86**(7): p. 3952-60.
190. Mamanova, L., et al., *Target-enrichment strategies for next-generation sequencing.* Nat Methods, 2010. **7**(2): p. 111-8.
191. Depledge, D.P., et al., *Specific capture and whole-genome sequencing of viruses from clinical samples.* PLoS One, 2011. **6**(11): p. e27805.
192. Vennema, H., E. de Bruin, and M. Koopmans, *Rational optimization of generic primers used for Norwalk-like virus detection by reverse transcriptase polymerase chain reaction.* J Clin Virol, 2002. **25**(2): p. 233-5.
193. Vinje, J., et al., *International collaborative study to compare reverse transcriptase PCR assays for detection and genotyping of noroviruses.* J Clin Microbiol, 2003. **41**(4): p. 1423-33.
194. Bonfield, J.K., K. Smith, and R. Staden, *A new DNA sequence assembly program.* Nucleic Acids Res, 1995. **23**(24): p. 4992-9.
195. Mendez, II, et al., *A comparative analysis of freon substitutes in the purification of reovirus and calicivirus.* J Virol Methods, 2000. **90**(1): p. 59-67.
196. Lennon, N.J., et al., *A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454.* Genome Biol, 2010. **11**(2): p. R15.
197. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads*. 2011. Vol. 17. 2011.
198. Zerbino, D.R. and E. Birney, *Velvet: Algorithms for de novo short read assembly using de Bruijn graphs.* Genome Research, 2008. **18**(5): p. 821-829.
199. Myers, E.W., et al., *A whole-genome assembly of Drosophila.* Science, 2000. **287**(5461): p. 2196-2204.
200. Hernandez, D., et al., *De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer.* Genome Research, 2008. **18**(5): p. 802-809.
201. Yang, X., et al., *De novo assembly of highly diverse viral populations.* Bmc Genomics, 2012. **13**.
202. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.

203. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
204. Katoh, K. and D.M. Standley, *MAFFT multiple sequence alignment software version 7: improvements in performance and usability.* Mol Biol Evol, 2013. **30**(4): p. 772-80.
205. Drummond, A.J., et al., *Bayesian phylogenetics with BEAUti and the BEAST 1.7.* Mol Biol Evol, 2012. **29**(8): p. 1969-73.
206. Fonager, J., S. Barzinci, and T.K. Fischer, *Emergence of a new recombinant Sydney 2012 norovirus variant in Denmark, 26 December 2012 to 22 March 2013.* Euro Surveill, 2013. **18**(25).
207. Haustein, T., et al., *Hospital admissions due to norovirus in adult and elderly patients in England.* Clin Infect Dis, 2009. **49**(12): p. 1890-2.
208. Lopman, B.A., et al., *A summertime peak of "winter vomiting disease": surveillance of noroviruses in England and Wales, 1995 to 2002.* BMC Public Health, 2003. **3**: p. 13.
209. England, P.H., *Guidelines for the management of norovirus outbreaks in acute and community health and social care settings*.
210. Thornhill, T.S., et al., *Pattern of shedding of the Norwalk particle in stools during experimentally induced gastroenteritis in volunteers as determined by immune electron microscopy.* J Infect Dis, 1975. **132**(1): p. 28-34.
211. Harris, J.P., et al., *The development of web-based surveillance provides new insights into the burden of norovirus outbreaks in hospitals in England.* Epidemiol Infect, 2013: p. 1-9.
212. Gallimore, C.I., et al., *Inter-seasonal diversity of norovirus genotypes: emergence and selection of virus variants.* Arch Virol, 2007. **152**(7): p. 1295-303.
213. Kearse, M., et al., *Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data.* Bioinformatics, 2012. **28**(12): p. 1647-9.
214. Guindon, S., et al., *PhyML: Fast and Accurate Phylogeny Reconstruction by Maximum Likelihood.* Infection Genetics and Evolution, 2009. **9**(3): p. 384-385.
215. Lopman, B., H. Vennema, and E. Kohli, *Increase in viral gastroenteritis outbreaks in Europe and epidemic spread of new norovirus variant. (vol 363, pg 682, 2004).* Lancet, 2004. **364**(9435): p. 664-664.
216. Allen, D.J., et al., *Emergence of the GII-4 Norovirus Sydney2012 strain in England, winter 2012-2013.* PLoS One, 2014. **9**(2): p. e88978.
217. Bruggink, L. and J. Marshall, *The relationship between health care and nonhealth care norovirus outbreak settings and norovirus genotype in Victoria, Australia, 2002-2005.* J Microbiol Immunol Infect, 2011. **44**(4): p. 241-6.
218. Bruggink, L.D., et al., *Molecular and epidemiological features of gastroenteritis outbreaks involving genogroup I norovirus in Victoria, Australia, 2002-2010.* J Med Virol, 2012. **84**(9): p. 1437-48.
219. Blanton, L.H., et al., *Molecular and epidemiologic trends of caliciviruses associated with outbreaks of acute gastroenteritis in the United States, 2000-2004.* J Infect Dis, 2006. **193**(3): p. 413-21.
220. Vinje, J., R.A. Hamidjaja, and M.D. Sobsey, *Development and application of a capsid VP1 (region D) based reverse transcription PCR assay for genotyping of genogroup I and II noroviruses.* J Virol Methods, 2004. **116**(2): p. 109-17.
221. Sukhrie, F.H., et al., *Using molecular epidemiology to trace transmission of nosocomial norovirus infection.* J Clin Microbiol, 2011. **49**(2): p. 602-6.
222. Nenonen, N.P., et al., *Norovirus GII.4 detection in environmental samples from patient rooms during nosocomial outbreaks.* J Clin Microbiol, 2014. **52**(7): p. 2352-8.
223. Bennett, S., et al., *Increased norovirus activity in Scotland in 2012 is associated with the emergence of a new norovirus GII. 4 variant.* Eurosurveillance, 2013. **18**(2): p. 22-23.

224. Fonager, J., L.S. Hindbaek, and T.K. Fischer, *Rapid emergence and antigenic diversification of the norovirus 2012 Sydney variant in Denmark, October to December, 2012.* Eurosurveillance, 2013. **18**(9): p. 2-5.
225. Mikheyev, A.S. and M.M. Tin, *A first look at the Oxford Nanopore MinION sequencer.* Mol Ecol Resour, 2014.
226. Kajitani, R., et al., *Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads.* Genome Res, 2014. **24**(8): p. 1384-95.
227. Bradnam, K.R., et al., *Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species.* Gigascience, 2013. **2**(1): p. 10.
228. Kleftogiannis, D., P. Kalnis, and V.B. Bajic, *Comparing memory-efficient genome assemblers on stand-alone and cloud infrastructures.* PLoS One, 2013. **8**(9): p. e75505.
229. Lipkin, W.I., *Microbe hunting.* Microbiol Mol Biol Rev, 2010. **74**(3): p. 363-77.
230. Loman, N.J., et al., *High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity.* Nat Rev Microbiol, 2012. **10**(9): p. 599-606.
231. Atmar, R.L. and M.K. Estes, *Norovirus vaccine development: next steps.* Expert Rev Vaccines, 2012. **11**(9): p. 1023-5.

Supplementary information

Genome sequences from other geographical locations

S1 Brighton

A total of 40 norovirus genomes from 2 seasons (2011-2012: 30 genomes; 2012-2013: 10 genomes) were generated to full sequence for analysis. All genomes generated were submitted into the BLAST database, confirming that all genomes were of the GII.4 genotype. There were no replicates nor longitudinal samples from this cohort.

S1.1 2011-2012 genomes

Figure S1 is a phylogenetic maximum likelihood tree depicting the sequences from 2011-2012. The genomes separate into clusters within time and location. Similar to that seen in the OUH, the "within - cluster" SNV diversity was small (less than 6 SNV) but "between-cluster" diversity was large (up to 226 SNVs; Appendix 10).

Three distinct clusters developed within the hospital wards over the same month (January 2012, highlighted Clusters 1-3 with the rose coloured boxes). Clusters 1 and 2 were only ~4 SNVs apart, whilst cluster 3 was more diverse at more than 86 SNVs. The samples sequenced in both clusters 1 and 2 were collected from Princess Royal Hospital, whilst samples from cluster 3 were collected from wards within the Royal Sussex and County hospital. Both hospitals were separated geographically by 15 miles. This gives anecdotal evidence that two different noroviruses were circulating within the same time period in two geographically locations within the same county.

In addition, genomes collected from Chichester Ward in January 2012 (highlighted in red) reinforce the notion that two norovirus genomes circulating within a single ward

during an outbreak. Samples 12483 and 12491 share the same genome and were collected within 24 hours of each other, sample 12475 collected 2 days earlier was 175 SNVs apart.

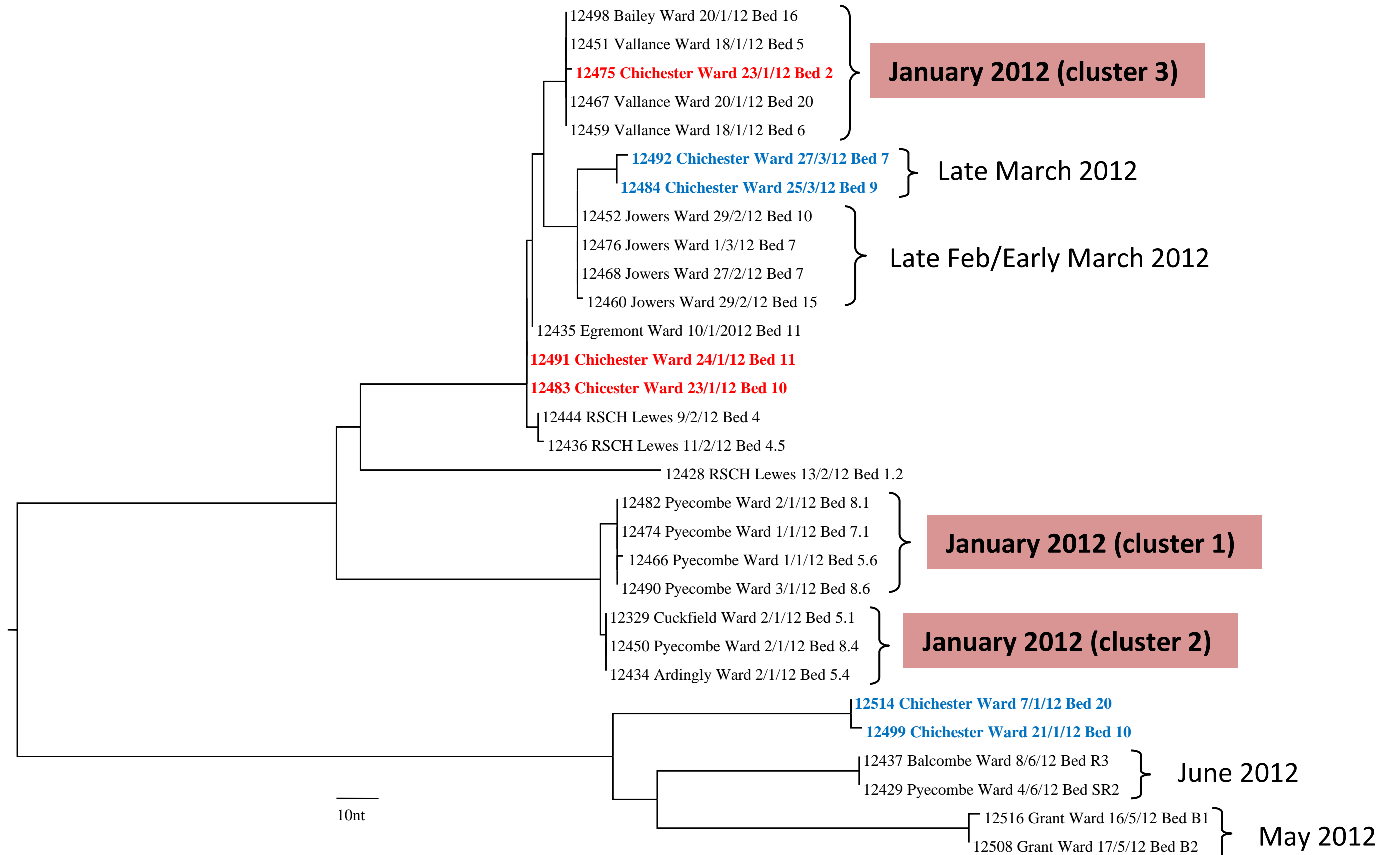


Figure S1 Phylogenetic maximum likelihood trees of all 2011-2012 sequences from Brighton

Furthermore, genomes sequenced from RT-PCR positive patients from the same ward at different times during the season differ markedly in SNVs, indicating that a new "wave" or strain had been the cause of infection. Examples include isolates from Chichester Ward (12475, 12492, 12484, 12491, 12483, 12514 and 12499, marked in blue and red). Had the sequences been identical over a long period of time, the possibility of surface contamination may have been a distinct possibility, given that the virus cannot evolve or mutate on contaminated surfaces. A plot of genetic distances of the upper branch of genomes is illustrated in Figure S2.

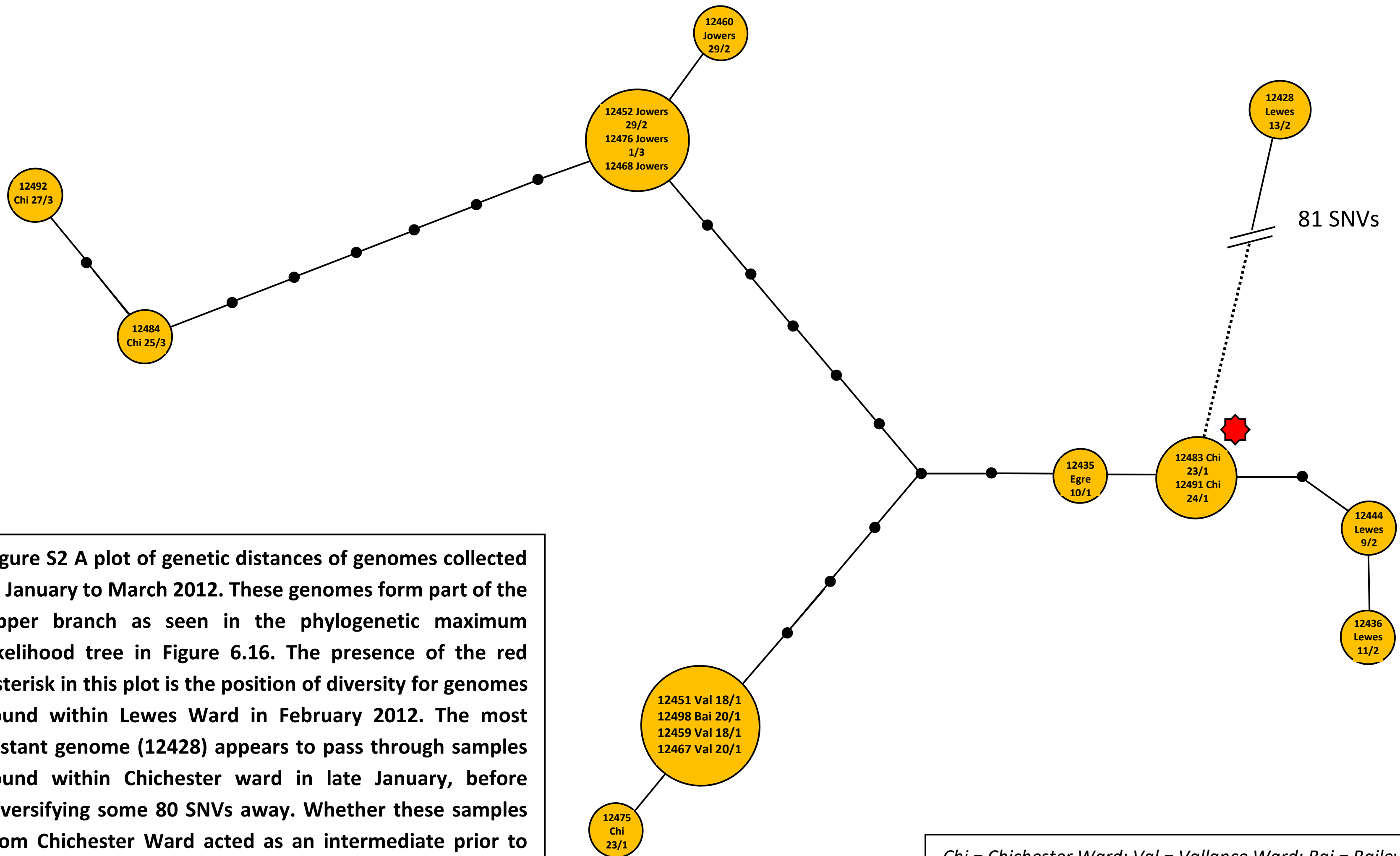


Figure S2 A plot of genetic distances of genomes collected in January to March 2012. These genomes form part of the upper branch as seen in the phylogenetic maximum likelihood tree in Figure 6.16. The presence of the red asterisk in this plot is the position of diversity for genomes found within Lewes Ward in February 2012. The most distant genome (12428) appears to pass through samples found within Chichester ward in late January, before diversifying some 80 SNVs away. Whether these samples from Chichester Ward acted as an intermediate prior to diversifying is unknown, but it is clearly a very different genome to that found within the same ward during the same time (samples 12444 and 12436)

Chi = Chichester Ward; Val = Vallance Ward; Bai = Bailey Ward; Egre = Egremont Ward

S1.2 2012-2013 genomes

A total of ten sequences were available for analysis from viruses originating during this season. Figure S3 is a phylogenetic maximum likelihood tree depicting the sequences and their relationship to each other. Given the presence of the GII.4 Sydney strains as described in Oxford, the tree also includes the reference genomes from Genbank of both the older strain (GII.4 New Orleans 2009) as well as the latest strain (GII.4 Sydney 2012), these are highlighted in green on the tree.

The sequences were all collected from symptomatic patients within a small period of time in November 2012. The tree shows that the genomes produced two main branches. None of the sequences bear any similarity to the New Orleans 2009 strain (sequences are between 811-829 SNVs apart; table S1) and all appear closer to the new Sydney 2012 strain (range 73-137 SNVs).

	15239	15247	15248	15246	Sydney 2012 reference	15240	15241	15243	15244	15245	15242	New Orleans 2009 reference
15239		2	2	96	91	145	146	146	146	146	146	819
15247	2		0	94	89	143	144	144	144	144	144	819
15248	2	0		94	89	143	144	144	144	144	144	819
15246	96	94	94		73	146	147	147	147	147	147	812
Sydney 2012 reference	91	89	89	73		136	137	137	137	137	137	811
15240	145	143	143	146	136		1	1	1	1	1	825
15241	146	144	144	147	137	1		0	0	0	0	824
15243	146	144	144	147	137	1	0		0	0	0	824
15244	146	144	144	147	137	1	0	0		0	0	824
15245	146	144	144	147	137	1	0	0	0		0	824
15242	146	144	144	147	137	1	0	0	0	0		824
New Orleans 2009 reference	819	819	819	812	811	825	824	824	824	824	824	

Table S1. Matrix of the Brighton 2012-2013 genomes

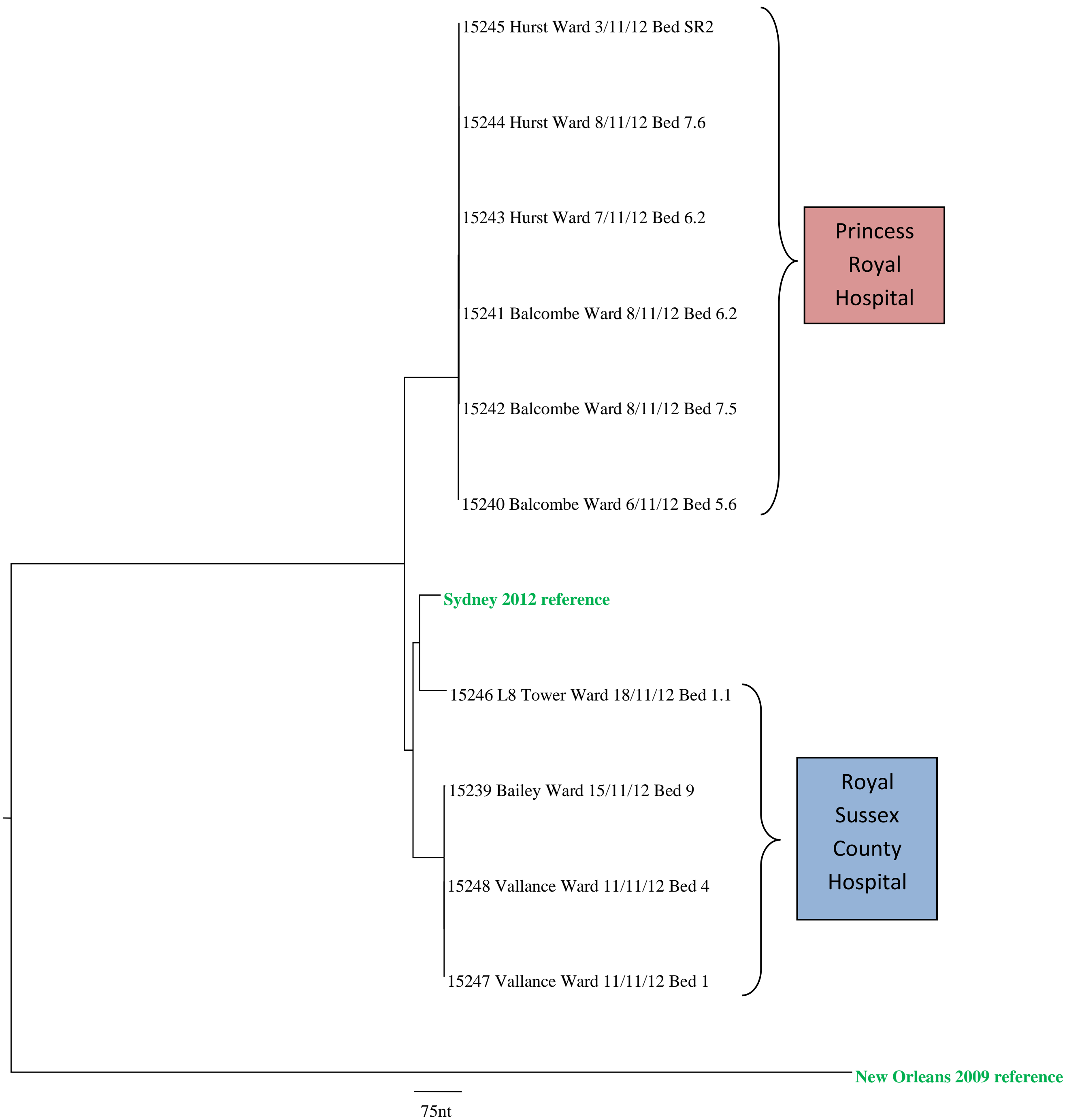


Figure S3 Phylogenetic tree depicting the sequences from Brighton between 2012-2013. All sequences share similarities to the Sydney 2012 strain compared to the New Orleans 2009 strain. The genomes within the two main branches are separated by their geographical locations.

S2 Leeds

A total of 46 *de novo* assembled genomes were available for analysis from two seasons: 2011-2012 (38 samples) and 2012-2013 (8 samples). No replicates or longitudinal samples were sequenced in this cohort.

Figure S4 below is a phylogenetic tree depicting all the full genomes from both seasons. On inspection there are two main branches. The reference genomes from New Orleans 2009 and Sydney 2012 have been added to this tree and the two branches are separated by the two reference genomes. There were approximately 800 SNVs between the two reference genomes.

In summary the 2011/2012 sequences had closer homology to the New Orleans 2009 reference. This was similar to what was described in both Oxford and Brighton. For the 2012/2013, six out of eight sequences clustered within the Sydney 2012 branch, whilst the remaining two sequences (highlighted with a pink coloured box in Figure S4) clustered between the New Orleans 2009 and Sydney 2012 branches.

In order to investigate whether recombination may have occurred, Figure S5 shows a maximum likelihood phylogenetic tree of ten samples, chosen to consist of the eight sequences from 2012/2013 season, as well as two samples that cluster outside of the main 2011/2012 cluster (12270 and 12312). In Figure S4 these are the ten samples distal to the black arrow. This shows that one sample, 15310 (sampled in December 2012) to have an ORF1 located closer to New Orleans 2009, but closer to Sydney 2012 in both ORF2 and ORF3, reaffirming recombination in Leeds as well as that seen Oxford from chapter 6.

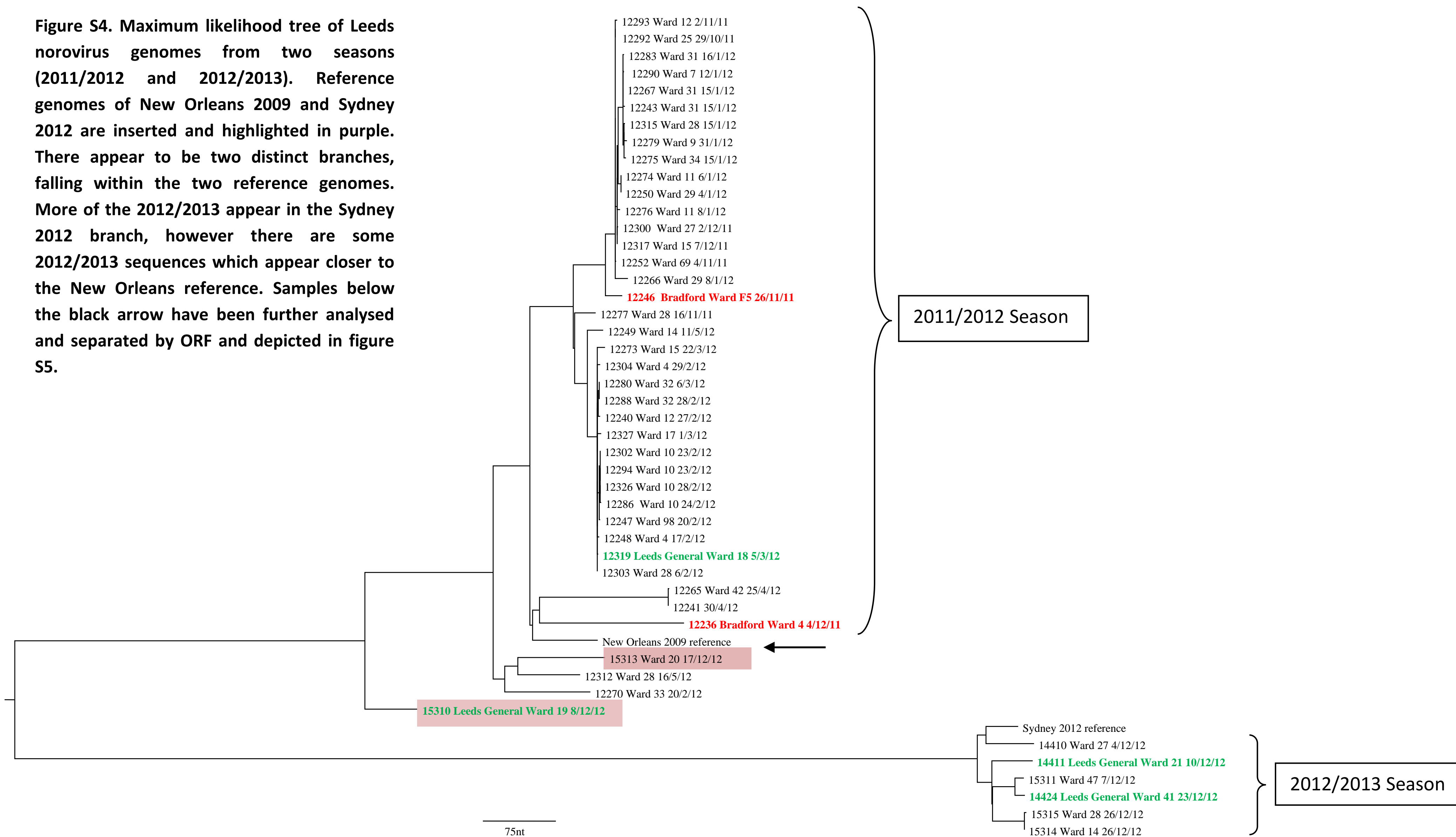


Figure S4. Maximum likelihood tree of Leeds norovirus genomes from two seasons (2011/2012 and 2012/2013). Reference genomes of New Orleans 2009 and Sydney 2012 are inserted and highlighted in purple. There appear to be two distinct branches, falling within the two reference genomes. More of the 2012/2013 appear in the Sydney 2012 branch, however there are some 2012/2013 sequences which appear closer to the New Orleans reference. Samples below the black arrow have been further analysed and separated by ORF and depicted in figure S5.

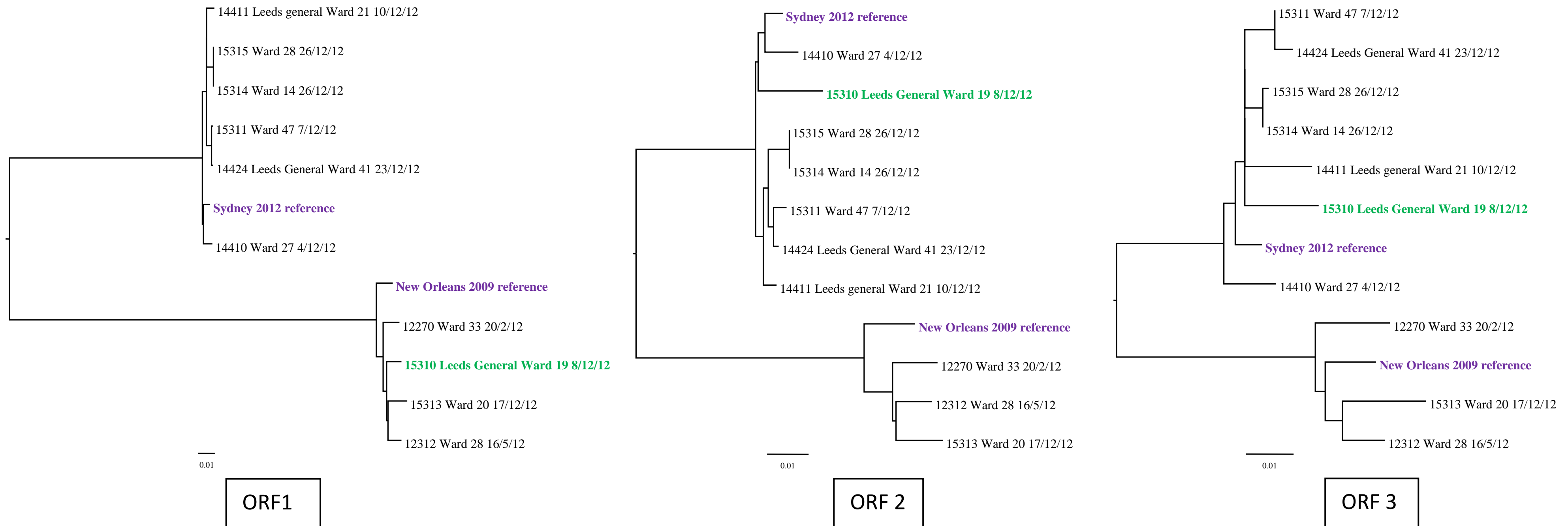


Figure S5 Maximum likelihood trees of ten samples separated by open reading frame (ORF). The reference strains from Genbank (Sydney 2012 and New Orleans 2009) have been coloured in Purple. There are two main branches, separated by the two strains. Of the ten samples that were analysed here, one sample (15310, highlighted in green) falls within the New Orleans reference in ORF1, but falls within the Sydney 2012 within both ORF2 and ORF3, confirming recombination.

S3 East Surrey Hospital and other sites

A total of 33 genomes were available for analysis from East Surrey hospitals. The relationship of these genomes is shown in the phylogenetic tree (figure S6). Unfortunately full epidemiological data was not available from all East Surrey Hospital samples, so ward locations are missing in some genomes.

Figure S6 Maximum likelihood tree of the East Surrey Hospitals sequences 2012-2013

The tree displays two main branches, separated by the two different reference strains (New Orleans 2009 and Sydney 2012, as depicted in purple). Only one sample was

located closest to the New Orleans 2009 branch whilst the other 32 sequences were closest to Sydney 2012.

Within the larger branch containing 32 sequences, two clusters have formed; of samples taken from Leigh Ward during an outbreak between 8-10th December 2012, three samples were sequenced (14412, 15268 and 15284, all highlighted in red) on Figure S6. One sample (14412) was 100 SNVs apart from the other two samples. Like Oxford, Brighton and Leeds, this indicates the presence of two diverse norovirus strains circulating within the same ward during the same period of time.

An additional 21 sequences were available for analysis. These were derived from faecal samples obtained at the following sites: Southampton hospital, Jersey General Hospital, Medway, Birmingham, Stoke Mandeville and two community hospitals. A maximum likelihood phylogenetic tree is shown in Figure S7 below.

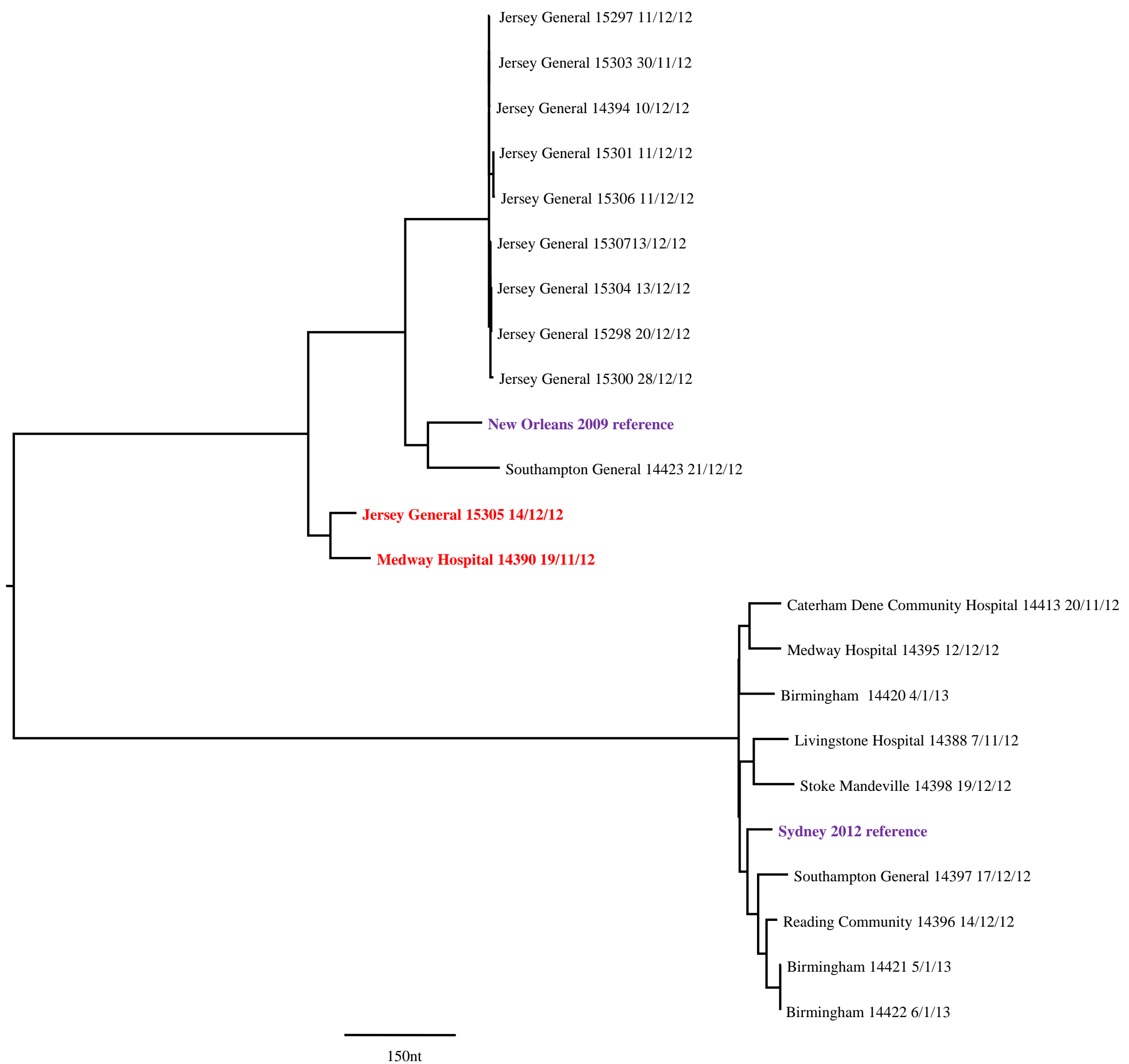


Figure S7 Maximum likelihood tree of all outliers sequences 2012-2013. Reference genomes from Sydney 2012 and New Orleans 2009 have been included. Two genomes highlighted in red display large genomic differences from the other genomes.

The samples from Jersey General Hospital were similar despite being collected up to 30 days apart (Range: 0-30 days; 0-14 SNVs). However, between national locations, there were no similarities at all; the nearest partner being 32 SNVs apart.

Of particular interest in this subset were two samples which lie within the tree between the two reference genomes. These two samples (15305 and 14390, highlighted in red on Figure S7) were collected from two UK locations (Jersey General Hospital and Medway hospital) independently during outbreak periods when other samples within the same hospital were also collected and sequenced. Genomically, these two samples were distinctly different from genomes within the same locality. In addition, both genomes were distinct from both the reference genomes in the maximum likelihood phylogenetic tree. As previously described, the open reading frames were divided to investigate whether recombination occurred. Figure S8 shows that both these samples showed evidence of recombination; with characteristic New Orleans 2009 like sequences in ORF1 whilst both ORF2 and ORF3 showed similarity to Sydney 2012.

This confirms the presence of intermediate recombination displaying partial genomes of the previous and current strain. This phenomena has also been recently described [206]. Whether these intermediate recombinants are part of the driving force for full genome recombination and evolution is, as yet, unknown.

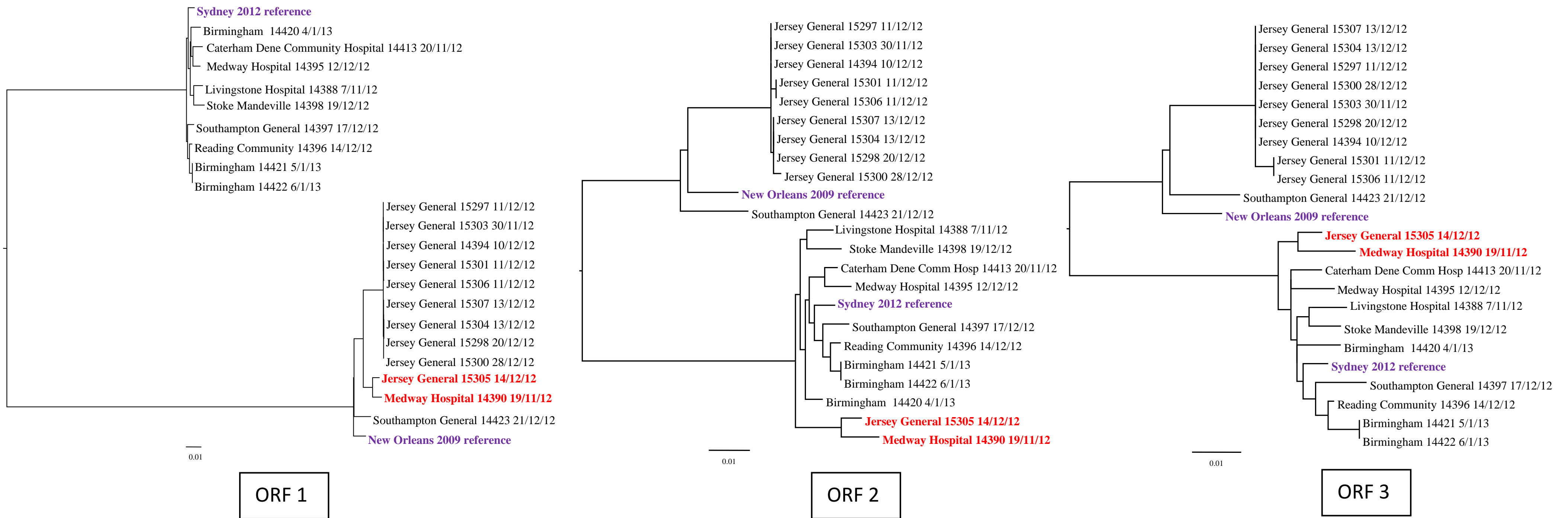


Figure S8 Maximum likelihood trees of norovirus genomes from viruses originating in Jersey, Medway, Southampton, Birmingham, Reading, Stoke Mandeville and other community-based sites within England, separated by open reading frame (ORF). Reference genomes from New Orleans 2009 and Sydney 2012 have been included in all three trees (highlighted in purple). Changes in tree topology across the ORFs illustrates recombination; the two red highlighted genomes share similarity within the New Orleans 2009 reference branch within ORF1, whilst they are completely different in both ORF2 and ORF3 with similarity to Sydney 2012, suggesting recombination of both the older (New Orleans 2009) and newer (Sydney 2012) strains.

Appendices

Appendix 1. Details of primers used for producing the seven overlapping amplicon fragments.

Primer	Polarity	Sequence (5'-3')	Binding site (nt)
PCR1F	+	GTGAATGAAGATGGCGTCTAAC	1
PCR1R	-	GGGAARAACCAYTTCATGAC	1160
INT1F	+	CTCCCCAGTACCTCATCTC	537
INT1R	-	CCCAACAACAAAAGGCATAGC	620
PCR-2F	+	ATGTTGTCATCTGCTGCATCC	1085
PCR-2R	-	CACATTGCCCTTGCCATCAG	2307
INT2F	+	GAACTCGCTGACACTTGTCC	1685
INT2R	-	TGATAATGATGACATCGCTGTC	1751
PCR3F	+	GCTCAAGCAAGCACTCAAG	2231
PCR3R	-	AGGGACAGCCACAATCAC	3438
INT3F	+	AGAAGAGTACCTTCAGGACAG	2758
INT3R	-	CTCCTCCTCTTCACAGAAGTC	2825
PCR4F	+	CCCATGCAACCATGAAAATCC	3336
PCR4R	-	AGAGGGTGAGAAGCCAGTG	4508
INT4F	+	TGAGAAATGGTCGTTCGCAC	3874
INT4R	-	AGCTTGCCTGTGAAGGACTC	3971
PCR5F	+	TGGACGTGGGTGACTTCAC	4425
PCR5R	-	AGAAACTGTGAAGACATCGTCC	5660
INT5F	+	CAAGAGCCAATGTTCAGATGG	5003
INT5R	-	GTCACTCGACGCCATCTTC	5087
PCR6F	+	CTGTGTTGATCCCCTTACCCG	5542
PCR6R	-	GACGTCAGATGCCAATCCAG	6726
INT6F	+	CCCTCTAGGAACTCCAGATTTC	6041
INT6R	-	CTCCCAGTGTACACTGTAGC	6129
PCR7F	+	GTCAACCAGTTCTACACACTTGC	6645
PCR7R	-	NVTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	7559
INT7F	+	AGAGTTAAACAGGCAATGCTC	6947
INT7R	-	CTATTTTGATCCTCAACCCAGC	7302

Internal primers have been included with the abbreviation INT followed by number. All primers listed were designed in house. *F= forward; R= reverse*

Appendix 2. List of all 477 norovirus genomes downloaded from NCBI for bioinformatic processing

Accession	Description	SeqLength
AF093797.1	Norwalk virus, complete genome.	7598
AF504671.2	Norwalk virus strain Vietnam 026, complete genome.	7540
M87661.2	Norwalk virus nonstructural polyprotein, 58 kd capsid protein, and orf3 genes, complete cds.	7654
NC_001959.2	Norwalk virus, complete genome.	7654
L07418.1	Southampton virus capsid protein, polyprotein, and ORF genes, complete cds.	7708
AY134748.1	Snow Mountain virus, complete genome.	7537
X86557.1	Lordsdale virus complete genome.	7555
AB039774.1	Norwalk-like virus genomic RNA, complete genome, isolate:SzUG1.	7700
AB039775.1	Norwalk-like virus genomic RNA, complete genome, isolate:Saitama U1.	7521
AB039776.1	Norwalk-like virus genomic RNA, complete genome, isolate:Saitama U3.	7564
AB039777.1	Norwalk-like virus genomic RNA, complete genome, isolate:Saitama U4.	7564
AB039778.1	Norwalk-like virus genomic RNA, complete genome, isolate:Saitama U16.	7546
AB039779.1	Norwalk-like virus genomic RNA, complete genome, isolate:Saitama U17.	7546
AB039780.1	Norwalk-like virus genomic RNA, complete genome, isolate:Saitama U25.	7524
AB039781.1	Norwalk-like virus genomic RNA, complete genome, isolate:Saitama U18.	7541
AB039782.1	Norwalk-like virus genomic RNA, complete genome, isolate:Saitama U201.	7541
AB081723.1	Norwalk-like virus genomic RNA, complete genome, strain:WUG1.	7688
AB045603.2	Norwalk-like virus genomic RNA, complete genome, specimen_voucher: Gifu'96.	7521
AY126474.2	Norwalk-like virus isolate Bo/Dumfries/94/UK, complete genome.	7311
AB083780.1	Norwalk-like virus genomic RNA, complete genome, strain: YURI.	7560
AB083781.1	Norwalk-like virus genomic RNA, complete genome, strain: YURI 32073.	7535
AB084071.1	Norwalk-like virus genomic RNA, complete genome, strain: GIFU'99.	7562
AB126320.1	Norwalk-like virus genomic RNA, complete genome, strain: SW/NV/swine43/JP.	7537
AB042808.1	Chiba virus genomic RNA, complete genome.	7697
AY772730.1	Norovirus Hu/NLV/GII/Neustrelitz260/2000/DE from Germany, complete	7579

	genome.	
AY032605.1	Human calicivirus Hu/NLV/GII/MD145-12/1987/US, complete genome.	7556
AY228235.2	Murine norovirus 1 isolate Mu/NoV/GV/MNV1/2002/USA, complete genome.	7382
DQ285629.1	Murine norovirus 1 clone CW1, complete genome.	7382
NC_008311.1	Murine norovirus 1, complete genome.	7382
EF014462.1	Murine norovirus 1 clone CW3, complete genome.	7382
EU004654.1	Murine norovirus 1 clone CW2, complete genome.	7382
EU004655.1	Murine norovirus 1 clone CW4, complete genome.	7382
EU004656.1	Murine norovirus 1 clone CW5, complete genome.	7382
EU004657.1	Murine norovirus 1 clone CW6, complete genome.	7382
EU004658.1	Murine norovirus 1 clone CW7, complete genome.	7382
EU004659.1	Murine norovirus 1 clone CW8, complete genome.	7382
EU004660.1	Murine norovirus 1 clone CW9, complete genome.	7382
EU004661.1	Murine norovirus 1 clone CW10, complete genome.	7382
EU004662.1	Murine norovirus 1 clone CW11, complete genome.	7382
AY237415.2	Human calicivirus strain Mc37, complete genome.	7541
AY485642.1	Human calicivirus NLV/GII/Langen1061/2002/DE, complete genome.	7558
AY581254.1	Human calicivirus Hu/NLV/Oxford/B5S22/2003/UK, complete genome.	7558
AY587983.1	Norovirus Hu/NLV/Oxford/B4S2/2002/UK, complete genome.	7558
AY587984.1	Norovirus Hu/NLV/Oxford/B4S5/2002/UK, complete genome.	7558
AY587985.1	Norovirus Hu/NLV/Oxford/B4S6/2002/UK, complete genome.	7558
AY587986.1	Norovirus Hu/NLV/Oxford/B4S4/2002/UK, complete genome.	7558
AY587987.1	Norovirus Hu/NLV/Oxford/B4S7/2002/UK, complete genome.	7558
AY587988.1	Norovirus Hu/NLV/Oxford/B4S1/2002/UK, complete genome.	7558
AY587989.1	Norovirus Hu/NLV/Oxford/B2S16/2002/UK, complete genome.	7558
AB187514.1	Norovirus Hu/GI/Otofuke/1979/JP genomic RNA, complete genome.	7746
AY741811.1	Norovirus Hu/NLV/Dresden174/pUS-NorII/1997/GE, complete genome.	7555
AY823305.2	Norovirus swine/GII/OH-QW125/03/US, complete genome.	7612
DQ078814.2	Norovirus Hu/GII.4/Hunter504D/04O/AU, complete genome.	7559
AB220921.1	Norovirus Hu/Chiba/04-1050/2005/JP genomic RNA, complete genome.	7559
AF097917.5	Norovirus Bo/Newbury2/1976/UK, complete genome.	7311
FJ446720.1	Murine norovirus 3 strain K4, complete genome.	7384
FJ446719.1	Murine norovirus 4 strain S18, complete genome.	7383

DQ911368.1	Murine norovirus strain Berlin/04/06/DE, complete genome.	7390
EF531290.1	Murine norovirus strain Berlin/05/06/DE, complete genome.	7391
EF531291.1	Murine norovirus strain Berlin/06/06/DE, complete genome.	7383
AB435514.1	Murine norovirus genomic RNA, complete genome, strain: S7-P2.	7382
AB435515.1	Murine norovirus genomic RNA, complete genome, strain: S7-PP3.	7382
HQ317203.1	Murine norovirus strain Guangzhou/K162/09/CHN, complete genome.	7380
AB601769.1	Murine norovirus genomic RNA, complete genome, strain: MT30-2.	7383
JX048594.1	Murine norovirus strain KHU-1, complete genome.	7402
DQ369797.1	Norovirus Hu/Guangzhou/NVgz01/CHN, complete genome.	7558
DQ456824.1	Norovirus Hu/MK04/2004/JP, complete genome.	7536
DQ658413.1	Norovirus Hu/GII.4/MD-2004/2004/US, complete genome.	7558
EF187497.2	Norovirus Hu/GII.4/Kenepuru/NZ327/2006/NZL, complete genome.	7559
EF650480.1	Murine norovirus 5 polyprotein gene, partial cds; and capsid protein gene, complete cds.	6596
EF650481.1	Murine norovirus 6 polyprotein gene, partial cds; and capsid protein gene, complete cds.	6610
EU004663.1	Murine norovirus GV/WU11/2005/USA, complete genome.	7390
EU004664.1	Murine norovirus GV/WU12/2005/USA, complete genome.	7384
EU004665.1	Murine norovirus GV/WU20/2005/USA, complete genome.	7383
EU004666.1	Murine norovirus GV/WU21/2005/USA, complete genome.	7380
EU004667.1	Murine norovirus GV/WU22/2005/USA, complete genome.	7383
EU004668.1	Murine norovirus GV/WU23/2005/USA, complete genome.	7383
EU004669.1	Murine norovirus GV/WU24/2005/USA, complete genome.	7383
EU004670.1	Murine norovirus GV/WU25/2005/USA, complete genome.	7383
EU004671.1	Murine norovirus GV/WU26/2005/USA, complete genome.	7384
EU004672.1	Murine norovirus GV/CR1/2005/USA, complete genome.	7382
EU004673.1	Murine norovirus GV/CR3/2005/USA, complete genome.	7383
EU004674.1	Murine norovirus GV/CR4/2005/USA, complete genome.	7382
EU004675.1	Murine norovirus GV/CR5/2005/USA, complete genome.	7382
EU004676.1	Murine norovirus GV/CR6/2005/USA, complete genome.	7383
JQ237823.1	Murine norovirus GV/CR6/2005/USA clone STL1, complete genome.	7383
EU004677.1	Murine norovirus GV/CR7/2005/USA, complete genome.	7380
EU004678.1	Murine norovirus GV/CR10/2005/USA, complete genome.	7386

EU004679.1	Murine norovirus GV/CR11/2005/USA, complete genome.	7386
EU004680.1	Murine norovirus GV/CR13/2005/USA, complete genome.	7386
EU004681.1	Murine norovirus GV/CR15/2005/USA, complete genome.	7389
EU004682.1	Murine norovirus GV/CR17/2005/USA, complete genome.	7389
EU004683.1	Murine norovirus GV/CR18/2005/DEU, complete genome.	7381
EF684915.2	Norovirus Hu/GII.4/Shellharbour/NSW696T/2006/AUS, complete genome.	7560
AB365435.1	Norovirus Hu/Texas/TCH04-577/2004/US genomic RNA, complete genome.	7627
EU310927.1	Norovirus Hu/Houston/TCH186/2002/US, complete genome.	7559
AB447427.1	Norovirus Hu/GII-4/Hokkaido1/2006/JP genomic RNA, complete genome.	7511
AB447428.1	Norovirus Hu/GII-4/Hokkaido2/2006/JP genomic RNA, complete genome.	7511
AB447429.1	Norovirus Hu/GII-4/Hokkaido3/2006/JP genomic RNA, complete genome.	7511
AB447430.1	Norovirus Hu/GII-4/Hokkaido4/2006/JP genomic RNA, complete genome.	7511
AB447431.1	Norovirus Hu/GII-4/Hokkaido5/2006/JP genomic RNA, complete genome.	7511
AB447432.1	Norovirus Hu/GII-4/Aomori1/2006/JP genomic RNA, complete genome.	7502
AB447433.1	Norovirus Hu/GII-4/Aomori2/2006/JP genomic RNA, complete genome.	7502
AB447434.1	Norovirus Hu/GII-4/Aomori4/2006/JP genomic RNA, complete genome.	7511
AB447435.1	Norovirus Hu/GII-4/Aomori5/2006/JP genomic RNA, complete genome.	7511
AB447436.1	Norovirus Hu/GII-4/Akita1/2006/JP genomic RNA, complete genome.	7511
AB447437.1	Norovirus Hu/GII-4/Akita2/2006/JP genomic RNA, complete genome.	7511
AB447438.1	Norovirus Hu/GII-4/Akita4/2006/JP genomic RNA, complete genome.	7511
AB447439.1	Norovirus Hu/GII-4/Akita5/2006/JP genomic RNA, complete genome.	7511
AB447440.1	Norovirus Hu/GII-4/Miyagi2/2006/JP genomic RNA, complete genome.	7511
AB447441.1	Norovirus Hu/GII-4/Miyagi4/2006/JP genomic RNA, complete genome.	7511
AB447442.1	Norovirus Hu/GII-4/Miyagi5/2006/JP genomic RNA, complete genome.	7511
AB447443.1	Norovirus Hu/GII-4/Toyama1/2006/JP genomic RNA, complete genome.	7511
AB447444.1	Norovirus Hu/GII-4/Toyama4/2006/JP genomic RNA, complete genome.	7511
AB447445.1	Norovirus Hu/GII-4/Toyama5/2006/JP genomic RNA, complete genome.	7511
AB447446.1	Norovirus Hu/GII-4/Aichi3/2006/JP genomic RNA, complete genome.	7511
AB447447.1	Norovirus Hu/GII-4/Aichi4/2006/JP genomic RNA, complete genome.	7511
AB447448.1	Norovirus Hu/GII-4/Sakai2/2006/JP genomic RNA, complete genome.	7511
AB447449.1	Norovirus Hu/GII-4/Sakai3/2006/JP genomic RNA, complete genome.	7511
AB447450.1	Norovirus Hu/GII-4/Sakai4/2006/JP genomic RNA, complete genome.	7511
AB447451.1	Norovirus Hu/GII-4/Hiroshima1/2006/JP genomic RNA, complete genome.	7511

AB447452.1	Norovirus Hu/GII-4/Hiroshima2/2006/JP genomic RNA, complete genome.	7511
AB447453.1	Norovirus Hu/GII-4/Ehime1/2006/JP genomic RNA, complete genome.	7511
AB447454.1	Norovirus Hu/GII-4/Ehime2/2006/JP genomic RNA, complete genome.	7511
AB447455.1	Norovirus Hu/GII-4/Ehime5/2006/JP genomic RNA, complete genome.	7511
AB447456.1	Norovirus Hu/GII-4/Saga1/2006/JP genomic RNA, complete genome.	7511
AB447457.1	Norovirus Hu/GII-4/Saga4/2006/JP genomic RNA, complete genome.	7511
AB447458.1	Norovirus Hu/GII-4/Saga5/2006/JP genomic RNA, complete genome.	7511
AB447459.1	Norovirus Hu/GII-4/Kumamoto1/2006/JP genomic RNA, complete genome.	7511
AB447460.1	Norovirus Hu/GII-4/Kumamoto2/2006/JP genomic RNA, complete genome.	7511
AB447461.1	Norovirus Hu/GII-4/Kumamoto3/2006/JP genomic RNA, complete genome.	7511
AB447462.1	Norovirus Hu/GII-4/Kumamoto4/2006/JP genomic RNA, complete genome.	7511
AB447463.1	Norovirus Hu/GII-4/Kumamoto5/2006/JP genomic RNA, complete genome.	7511
EU854589.1	Norovirus mouse/Hannover1/2007/DEU, complete genome.	7383
EU794907.1	Norovirus Bo/GIII/B309/2003/BEL polyprotein, capsid, and VP2 genes, complete cds.	7317
EU921344.2	Norovirus Hu/Pune/PC15/2006/India, complete genome.	7559
EU921388.2	Norovirus Hu/Pune/PC51/2007/India, complete genome.	7559
EU921389.2	Norovirus Hu/Pune/PC52/2007/India, complete genome.	7547
FJ514242.1	Norovirus Hu/GII-4/CUK-3/2008/KR, complete genome.	7559
FJ515294.1	Norovirus Hu/GI.2/Leuven/2003/BEL, complete genome.	7699
FJ537135.1	Norovirus Hu/GII.4/CHDC2094/1974/US, complete genome.	7576
FJ537136.1	Norovirus Hu/GII.4/CHDC3967/1988/US, complete genome.	7580
FJ537137.1	Norovirus Hu/GII.4/CHDC4108/1987/US, complete genome.	7580
FJ537138.1	Norovirus Hu/GII.4/CHDC4871/1977/US, complete genome.	7580
FJ537134.1	Norovirus Hu/GII.4/CHDC5191/1974/US, complete genome.	7580
FJ692500.1	Norovirus dog/GVI.1/HKU_Ca026F/2007/HKG ORF1 polyprotein, VP1 capsid protein, and VP2 minor structural protein genes, complete cds.	7637
FJ692501.1	Norovirus dog/GVI.1/HKU_Ca035F/2007/HKG ORF1 polyprotein, VP1 capsid protein, and VP2 minor structural protein genes, complete cds.	7637
GQ845024.2	Norovirus Hu/GII.4/Rathmines/NSW287R/2007/AUS, complete genome.	7562
GQ845367.2	Norovirus Hu/GII.4/Orange/NSW001P/2008/AU, complete genome.	7559
GQ845370.2	Norovirus Hu/GII.g-GII.12/StGeorge/NSW199U/2008/AU, complete genome.	7523
GQ845366.2	Norovirus Hu/GII.4/Westmead/NSW3639/2008/AUS, complete genome.	7560

GQ845369.3	Norovirus Hu/GII.4/Armidale/NSW390I/2008/AU, complete genome.	7560
GQ845368.2	Norovirus Hu/GII.4/Sutherland/NSW505G/2007/AUS, complete genome.	7559
GU017903.2	Norovirus Hu/8533/Maizuru/08/JPN, complete genome.	7529
GU325839.2	Norovirus Hu/GII.4/HS194/2009/US, complete genome.	7560
AB541190.1	Norovirus Hu/GII-4/2007a_ORF2-3/JP genes for polyprotein, VP1, VP2, partial and complete cds.	2751
AB541192.1	Norovirus Hu/GII-4/2007b_ORF2-3/JP genes for polyprotein, VP1, VP2, partial and complete cds.	2748
AB541195.1	Norovirus Hu/GII-4/2008a_ORF2-3/JP genes for polyprotein, VP1, VP2, partial and complete cds.	2751
AB541198.1	Norovirus Hu/GII-4/2008b_ORF2-3/JP genes for polyprotein, VP1, VP2, partial and complete cds.	2751
AB541201.1	Norovirus Hu/GII-4/Aichi1/2007/JP genomic RNA, complete genome.	7509
AB541202.1	Norovirus Hu/GII-4/Aichi1/2008/JP genomic RNA, complete genome.	7509
AB541203.1	Norovirus Hu/GII-4/Aichi2/2007/JP genomic RNA, complete genome.	7509
AB541204.1	Norovirus Hu/GII-4/Aichi2/2008/JP genomic RNA, complete genome.	7509
AB541205.1	Norovirus Hu/GII-4/Aichi3/2007/JP genomic RNA, complete genome.	7509
AB541206.1	Norovirus Hu/GII-4/Aichi3/2008/JP genomic RNA, complete genome.	7509
AB541207.1	Norovirus Hu/GII-4/Aichi4/2007/JP genomic RNA, complete genome.	7509
AB541208.1	Norovirus Hu/GII-4/Aichi4/2008/JP genomic RNA, complete genome.	7509
AB541209.1	Norovirus Hu/GII-4/Aichi5/2007/JP genomic RNA, complete genome.	7509
AB541210.1	Norovirus Hu/GII-4/Aichi5/2008/JP genomic RNA, complete genome.	7509
AB541211.1	Norovirus Hu/GII-4/Akita1/2007/JP genomic RNA, complete genome.	7509
AB541212.1	Norovirus Hu/GII-4/Akita1/2008/JP genomic RNA, nearly complete genome.	7404
AB541213.1	Norovirus Hu/GII-4/Akita2/2008/JP genomic RNA, complete genome.	7509
AB541214.1	Norovirus Hu/GII-4/Akita3/2007/JP genomic RNA, complete genome.	7509
AB541215.1	Norovirus Hu/GII-4/Akita3/2008/JP genomic RNA, complete genome.	7509
AB541216.1	Norovirus Hu/GII-4/Akita4/2008/JP genomic RNA, complete genome.	7509
AB541217.1	Norovirus Hu/GII-4/Akita5/2007/JP genomic RNA, complete genome.	7509
AB541218.1	Norovirus Hu/GII-4/Aomori1/2007/JP genomic RNA, complete genome.	7509
AB541219.1	Norovirus Hu/GII-4/Aomori1/2008/JP genomic RNA, complete genome.	7509
AB541220.1	Norovirus Hu/GII-4/Aomori2/2007/JP genomic RNA, complete genome.	7509
AB541221.1	Norovirus Hu/GII-4/Aomori2/2008/JP genomic RNA, nearly complete genome.	7404

AB541222.1	Norovirus Hu/GII-4/Aomori3/2007/JP genomic RNA, complete genome.	7509
AB541223.1	Norovirus Hu/GII-4/Aomori3/2008/JP genomic RNA, complete genome.	7509
AB541224.1	Norovirus Hu/GII-4/Aomori4/2007/JP genomic RNA, complete genome.	7509
AB541225.1	Norovirus Hu/GII-4/Aomori4/2008/JP genomic RNA, complete genome.	7509
AB541226.1	Norovirus Hu/GII-4/Aomori5/2007/JP genomic RNA, complete genome.	7509
AB541227.1	Norovirus Hu/GII-4/Aomori5/2008/JP genomic RNA, complete genome.	7509
AB541228.1	Norovirus Hu/GII-4/Chiba1/2007/JP genomic RNA, complete genome.	7509
AB541229.1	Norovirus Hu/GII-4/Chiba1/2008/JP genomic RNA, complete genome.	7509
AB541230.1	Norovirus Hu/GII-4/Chiba2/2007/JP genomic RNA, complete genome.	7509
AB541231.1	Norovirus Hu/GII-4/Chiba2/2008/JP genomic RNA, complete genome.	7509
AB541232.1	Norovirus Hu/GII-4/Chiba4/2007/JP genomic RNA, complete genome.	7509
AB541233.1	Norovirus Hu/GII-4/Chiba4/2008/JP genomic RNA, complete genome.	7497
AB541234.1	Norovirus Hu/GII-4/Chiba5/2007/JP genomic RNA, complete genome.	7509
AB541235.1	Norovirus Hu/GII-4/Chiba5/2008/JP genomic RNA, complete genome.	7509
AB541236.1	Norovirus Hu/GII-4/Ehime1/2007/JP genomic RNA, complete genome.	7509
AB541237.1	Norovirus Hu/GII-4/Ehime1/2008/JP genomic RNA, complete genome.	7509
AB541238.1	Norovirus Hu/GII-4/Ehime2/2007/JP genomic RNA, complete genome.	7509
AB541239.1	Norovirus Hu/GII-4/Ehime3/2007/JP genomic RNA, complete genome.	7509
AB541240.1	Norovirus Hu/GII-4/Ehime3/2008/JP genomic RNA, complete genome.	7509
AB541241.1	Norovirus Hu/GII-4/Ehime4/2007/JP genomic RNA, complete genome.	7509
AB541242.1	Norovirus Hu/GII-4/Ehime4/2008/JP genomic RNA, complete genome.	7509
AB541243.1	Norovirus Hu/GII-4/Ehime5/2008/JP genomic RNA, complete genome.	7509
AB541244.1	Norovirus Hu/GII-4/Fukui1/2008/JP genomic RNA, complete genome.	7509
AB541245.1	Norovirus Hu/GII-4/Fukui2/2007/JP genomic RNA, complete genome.	7509
AB541246.1	Norovirus Hu/GII-4/Fukui2/2008/JP genomic RNA, complete genome.	7509
AB541247.1	Norovirus Hu/GII-4/Fukui4/2007/JP genomic RNA, complete genome.	7509
AB541248.1	Norovirus Hu/GII-4/Fukui4/2008/JP genomic RNA, complete genome.	7509
AB541249.1	Norovirus Hu/GII-4/Fukui5/2007/JP genomic RNA, complete genome.	7509
AB541250.1	Norovirus Hu/GII-4/Fukui5/2008/JP genomic RNA, complete genome.	7509
AB541251.1	Norovirus Hu/GII-4/Hiroshima1/2007/JP genomic RNA, complete genome.	7509
AB541252.1	Norovirus Hu/GII-4/Hiroshima1/2008/JP genomic RNA, complete genome.	7509
AB541253.1	Norovirus Hu/GII-4/Hiroshima2/2007/JP genomic RNA, complete genome.	7509
AB541254.1	Norovirus Hu/GII-4/Hiroshima2/2008/JP genomic RNA, complete genome.	7509

AB541255.1	Norovirus Hu/GII-4/Hiroshima3/2007/JP genomic RNA, complete genome.	7509
AB541256.1	Norovirus Hu/GII-4/Hiroshima3/2008/JP genomic RNA, complete genome.	7509
AB541257.1	Norovirus Hu/GII-4/Hiroshima4/2007/JP genomic RNA, complete genome.	7509
AB541258.1	Norovirus Hu/GII-4/Hiroshima4/2008/JP genomic RNA, complete genome.	7509
AB541259.1	Norovirus Hu/GII-4/Hiroshima5/2008/JP genomic RNA, complete genome.	7509
AB541260.1	Norovirus Hu/GII-4/Hokkaido1/2007/JP genomic RNA, complete genome.	7509
AB541261.1	Norovirus Hu/GII-4/Hokkaido1/2008/JP genomic RNA, complete genome.	7509
AB541262.1	Norovirus Hu/GII-4/Hokkaido2/2007/JP genomic RNA, complete genome.	7509
AB541263.1	Norovirus Hu/GII-4/Hokkaido2/2008/JP genomic RNA, complete genome.	7509
AB541264.1	Norovirus Hu/GII-4/Hokkaido3/2008/JP genomic RNA, complete genome.	7509
AB541265.1	Norovirus Hu/GII-4/Hokkaido4/2007/JP genomic RNA, complete genome.	7509
AB541266.1	Norovirus Hu/GII-4/Hokkaido4/2008/JP genomic RNA, complete genome.	7509
AB541267.1	Norovirus Hu/GII-4/Hokkaido5/2007/JP genomic RNA, complete genome.	7509
AB541268.1	Norovirus Hu/GII-4/Hokkaido5/2008/JP genomic RNA, complete genome.	7509
AB541269.1	Norovirus Hu/GII-4/Iwate1/2008/JP genomic RNA, complete genome.	7509
AB541270.1	Norovirus Hu/GII-4/Iwate2/2007/JP genomic RNA, complete genome.	7509
AB541271.1	Norovirus Hu/GII-4/Iwate3/2007/JP genomic RNA, complete genome.	7509
AB541272.1	Norovirus Hu/GII-4/Iwate3/2008/JP genomic RNA, complete genome.	7509
AB541273.1	Norovirus Hu/GII-4/Iwate4/2007/JP genomic RNA, complete genome.	7509
AB541274.1	Norovirus Hu/GII-4/Iwate4/2008/JP genomic RNA, complete genome.	7509
AB541275.1	Norovirus Hu/GII-4/Iwate5/2007/JP genomic RNA, complete genome.	7506
AB541276.1	Norovirus Hu/GII-4/Iwate5/2008/JP genomic RNA, complete genome.	7509
AB541277.1	Norovirus Hu/GII-4/Kumamoto1/2007/JP genomic RNA, complete genome.	7509
AB541278.1	Norovirus Hu/GII-4/Kumamoto2/2007/JP genomic RNA, complete genome.	7509
AB541279.1	Norovirus Hu/GII-4/Kumamoto3/2007/JP genomic RNA, complete genome.	7509
AB541280.1	Norovirus Hu/GII-4/Kumamoto4/2007/JP genomic RNA, complete genome.	7509
AB541281.1	Norovirus Hu/GII-4/Miyagi1/2007/JP genomic RNA, complete genome.	7509
AB541282.1	Norovirus Hu/GII-4/Miyagi2/2007/JP genomic RNA, complete genome.	7509
AB541283.1	Norovirus Hu/GII-4/Miyagi3/2007/JP genomic RNA, complete genome.	7509
AB541284.1	Norovirus Hu/GII-4/Miyagi5/2008/JP genomic RNA, complete genome.	7509
AB541285.1	Norovirus Hu/GII-4/Miyazaki10/2008/JP genomic RNA, complete genome.	7509
AB541286.1	Norovirus Hu/GII-4/Miyazaki12/2008/JP genomic RNA, complete genome.	7509
AB541287.1	Norovirus Hu/GII-4/Miyazaki13/2008/JP genomic RNA, complete genome.	7509

AB541288.1	Norovirus Hu/GII-4/Miyazaki1/2007/JP genomic RNA, complete genome.	7509
AB541289.1	Norovirus Hu/GII-4/Miyazaki1/2008/JP genomic RNA, complete genome.	7509
AB541290.1	Norovirus Hu/GII-4/Miyazaki2/2007/JP genomic RNA, complete genome.	7509
AB541291.1	Norovirus Hu/GII-4/Miyazaki2/2008/JP genomic RNA, complete genome.	7509
AB541292.1	Norovirus Hu/GII-4/Miyazaki3/2007/JP genomic RNA, complete genome.	7509
AB541293.1	Norovirus Hu/GII-4/Miyazaki3/2008/JP genomic RNA, complete genome.	7509
AB541294.1	Norovirus Hu/GII-4/Miyazaki4/2007/JP genomic RNA, complete genome.	7509
AB541295.1	Norovirus Hu/GII-4/Miyazaki4/2008/JP genomic RNA, complete genome.	7509
AB541296.1	Norovirus Hu/GII-4/Miyazaki5/2007/JP genomic RNA, complete genome.	7509
AB541297.1	Norovirus Hu/GII-4/Miyazaki6/2008/JP genomic RNA, complete genome.	7509
AB541298.1	Norovirus Hu/GII-4/Miyazaki7/2008/JP genomic RNA, complete genome.	7509
AB541299.1	Norovirus Hu/GII-4/Miyazaki8/2008/JP genomic RNA, complete genome.	7509
AB541300.1	Norovirus Hu/GII-4/Miyazaki9/2008/JP genomic RNA, nearly complete genome.	7263
AB541301.1	Norovirus Hu/GII-4/Nagano1/2007/JP genomic RNA, complete genome.	7509
AB541302.1	Norovirus Hu/GII-4/Nagano1/2008/JP genomic RNA, complete genome.	7509
AB541303.1	Norovirus Hu/GII-4/Nagano2/2007/JP genomic RNA, complete genome.	7509
AB541304.1	Norovirus Hu/GII-4/Nagano2/2008/JP genomic RNA, complete genome.	7509
AB541305.1	Norovirus Hu/GII-4/Nagano3/2007/JP genomic RNA, complete genome.	7512
AB541306.1	Norovirus Hu/GII-4/Nagano3/2008/JP genomic RNA, complete genome.	7509
AB541307.1	Norovirus Hu/GII-4/Nagano4/2008/JP genomic RNA, complete genome.	7509
AB541308.1	Norovirus Hu/GII-4/Nagano5/2007/JP genomic RNA, complete genome.	7509
AB541309.1	Norovirus Hu/GII-4/Niigata1/2007/JP genomic RNA, complete genome.	7509
AB541310.1	Norovirus Hu/GII-4/Niigata1/2008/JP genomic RNA, complete genome.	7509
AB541311.1	Norovirus Hu/GII-4/Niigata2/2007/JP genomic RNA, complete genome.	7509
AB541312.1	Norovirus Hu/GII-4/Niigata2/2008/JP genomic RNA, complete genome.	7509
AB541313.1	Norovirus Hu/GII-4/Niigata3/2007/JP genomic RNA, complete genome.	7509
AB541314.1	Norovirus Hu/GII-4/Niigata3/2008/JP genomic RNA, complete genome.	7509
AB541315.1	Norovirus Hu/GII-4/Niigata4/2007/JP genomic RNA, complete genome.	7509
AB541316.1	Norovirus Hu/GII-4/Niigata4/2008/JP genomic RNA, complete genome.	7509
AB541317.1	Norovirus Hu/GII-4/Niigata5/2007/JP genomic RNA, complete genome.	7509
AB541318.1	Norovirus Hu/GII-4/Niigata5/2008/JP genomic RNA, complete genome.	7509
AB541319.1	Norovirus Hu/GII-4/Osaka1/2007/JP genomic RNA, complete genome.	7509

AB541320.1	Norovirus Hu/GII-4/Osaka1/2008/JP genomic RNA, complete genome.	7509
AB541321.1	Norovirus Hu/GII-4/Osaka2/2007/JP genomic RNA, complete genome.	7509
AB541322.1	Norovirus Hu/GII-4/Osaka2/2008/JP genomic RNA, complete genome.	7509
AB541323.1	Norovirus Hu/GII-4/Osaka3/2007/JP genomic RNA, complete genome.	7509
AB541324.1	Norovirus Hu/GII-4/Osaka3/2008/JP genomic RNA, complete genome.	7509
AB541325.1	Norovirus Hu/GII-4/Osaka4/2007/JP genomic RNA, complete genome.	7509
AB541326.1	Norovirus Hu/GII-4/Osaka4/2008/JP genomic RNA, complete genome.	7509
AB541327.1	Norovirus Hu/GII-4/Osaka5/2007/JP genomic RNA, complete genome.	7509
AB541328.1	Norovirus Hu/GII-4/Osaka5/2008/JP genomic RNA, complete genome.	7509
AB541329.1	Norovirus Hu/GII-4/Osaka6/2008/JP genomic RNA, complete genome.	7509
AB541330.1	Norovirus Hu/GII-4/Saga1/2007/JP genomic RNA, complete genome.	7509
AB541331.1	Norovirus Hu/GII-4/Saga1/2008/JP genomic RNA, complete genome.	7509
AB541332.1	Norovirus Hu/GII-4/Saga2/2007/JP genomic RNA, complete genome.	7509
AB541333.1	Norovirus Hu/GII-4/Saga2/2008/JP genomic RNA, complete genome.	7509
AB541334.1	Norovirus Hu/GII-4/Saga3/2008/JP genomic RNA, complete genome.	7509
AB541335.1	Norovirus Hu/GII-4/Saga4/2007/JP genomic RNA, complete genome.	7509
AB541336.1	Norovirus Hu/GII-4/Saga4/2008/JP genomic RNA, complete genome.	7509
AB541337.1	Norovirus Hu/GII-4/Saga5/2007/JP genomic RNA, complete genome.	7509
AB541338.1	Norovirus Hu/GII-4/Saga5/2008/JP genomic RNA, complete genome.	7509
AB541339.1	Norovirus Hu/GII-4/Sakai1/2007/JP genomic RNA, complete genome.	7509
AB541340.1	Norovirus Hu/GII-4/Sakai1/2008/JP genomic RNA, complete genome.	7509
AB541341.1	Norovirus Hu/GII-4/Sakai2/2007/JP genomic RNA, complete genome.	7509
AB541342.1	Norovirus Hu/GII-4/Sakai3/2007/JP genomic RNA, complete genome.	7509
AB541343.1	Norovirus Hu/GII-4/Sakai3/2008/JP genomic RNA, complete genome.	7509
AB541344.1	Norovirus Hu/GII-4/Sakai4/2007/JP genomic RNA, complete genome.	7509
AB541345.1	Norovirus Hu/GII-4/Sakai4/2008/JP genomic RNA, complete genome.	7509
AB541346.1	Norovirus Hu/GII-4/Shimane1/2007/JP genomic RNA, complete genome.	7509
AB541347.1	Norovirus Hu/GII-4/Shimane2/2007/JP genomic RNA, complete genome.	7509
AB541348.1	Norovirus Hu/GII-4/Shimane2/2008/JP genomic RNA, nearly complete genome.	7401
AB541349.1	Norovirus Hu/GII-4/Shimane3/2007/JP genomic RNA, complete genome.	7509
AB541350.1	Norovirus Hu/GII-4/Shimane3/2008/JP genomic RNA, complete genome.	7509
AB541351.1	Norovirus Hu/GII-4/Shimane4/2007/JP genomic RNA, complete genome.	7509

AB541352.1	Norovirus Hu/GII-4/Shimane5/2007/JP genomic RNA, complete genome.	7509
AB541353.1	Norovirus Hu/GII-4/Shimane5/2008/JP genomic RNA, complete genome.	7509
AB541354.1	Norovirus Hu/GII-4/Toyama1/2007/JP genomic RNA, complete genome.	7509
AB541355.1	Norovirus Hu/GII-4/Toyama2/2007/JP genomic RNA, complete genome.	7509
AB541356.1	Norovirus Hu/GII-4/Toyama2/2008/JP genomic RNA, complete genome.	7509
AB541357.1	Norovirus Hu/GII-4/Toyama3/2007/JP genomic RNA, complete genome.	7512
AB541358.1	Norovirus Hu/GII-4/Toyama3/2008/JP genomic RNA, complete genome.	7509
AB541359.1	Norovirus Hu/GII-4/Toyama4/2007/JP genomic RNA, complete genome.	7509
AB541360.1	Norovirus Hu/GII-4/Toyama4/2008/JP genomic RNA, complete genome.	7509
AB541361.1	Norovirus Hu/GII-4/Toyama5/2007/JP genomic RNA, complete genome.	7509
AB541362.1	Norovirus Hu/GII-4/Toyama5/2008/JP genomic RNA, complete genome.	7509
AB543808.1	Norovirus Hu/GII-4/FUMI/2010/JP genomic RNA, complete genome.	7509
GU445325.2	Norovirus Hu/GII.4/New Orleans1805/2009/USA, complete genome.	7559
GU594162.1	Norovirus Hu/GII/8610/Saga/2008/JPN, complete genome.	7529
GU991353.1	Norovirus Hu/GII/Shanghai/SH2/2008/CHN, complete genome.	7555
GU991354.1	Norovirus Hu/Shanghai/SH5/2009/CHN, complete genome.	7511
GU991355.1	Norovirus Hu/Shanghai/SH312/2009/CHN, complete genome.	7544
GU980585.1	Norovirus Hu/GII.3/CBNU1/2006/KOR, complete genome.	7577
HQ009513.1	Norovirus Hu/GII.4/JB-15/KOR/2008, complete genome.	7558
HM748971.2	Norovirus Hu/GII.4/Beecroft/NSW305P/2009/AUS, complete genome.	7560
HM748972.2	Norovirus Hu/GII.4/Teralba/NSW881Z/2009/AUS, complete genome.	7559
HM748973.2	Norovirus Hu/GII.4/Turramurra/NSW892U/2009/AUS, complete genome.	7560
HQ449728.1	Norovirus Hu/GII.12/HS210/2010/USA, complete genome.	7523
JF320644.1	Murine norovirus GV/NIH-2409/2005/USA, complete genome.	7383
JF320645.1	Murine norovirus GV/NIH-2410/2005/USA, complete genome.	7383
JF320646.1	Murine norovirus GV/NIH-2411/2005/USA, complete genome.	7383
JF320647.1	Murine norovirus GV/NIH-2747/2005/USA, complete genome.	7383
JF320648.1	Murine norovirus GV/NIH-2750/2005/USA, complete genome.	7383
JF320649.1	Murine norovirus GV/NIH-4421/2005/USA, complete genome.	7383
JF320650.1	Murine norovirus GV/NIH-4428/2005/USA, complete genome.	7383
JF320651.1	Murine norovirus GV/NIH-4431/2005/USA, complete genome.	7383
JF320652.1	Murine norovirus GV/NIH-A114/2006/USA, complete genome.	7382
JF320653.1	Murine norovirus GV/NIH-D220/2007/USA, complete genome.	7382

HQ664990.1	Norovirus Hu/GII.12/HS206/2010/USA, complete genome.	7523
HQ392821.1	Norovirus pig/GII/Ch6/China/2009, complete genome.	7548
JN595867.1	Norovirus Hu/GII.4/New Orleans/2010/USA, complete genome.	7559
JF781268.1	Norovirus cat/GIV.2/CU081210E/USA/2010, complete genome.	7839
JQ388274.1	Norovirus Hu/GI.6/Kingston/ACT160D/2010/AU, complete genome.	7691
JQ911594.1	Norovirus Hu/GI/10360/2010/VNM, complete genome.	7696
JQ911595.1	Norovirus Hu/GII/10002/2009/VNM, complete genome.	7511
JQ911596.1	Norovirus Hu/GII/10003/2009/VNM, complete genome.	7511
JQ911597.1	Norovirus Hu/GII/10012/2009/VNM, complete genome.	7510
JQ911598.1	Norovirus Hu/GII/10037/2009/VNM, complete genome.	7511
JQ613552.2	Norovirus Hu/GII.4/NSW123B/2010/AU, complete genome.	7559
JQ613570.1	Norovirus Hu/GII.4/Rockdale/NSW006D/2009/AU, complete genome.	7559
JQ613567.1	Norovirus Hu/GIV.1/LakeMacquarie/NSW268O/2010/AU, complete genome.	7527
JQ613569.1	Norovirus Hu/GII.g-GII.12/Gunnedah/NSW895P/2010/AU, complete genome.	7523
JQ613568.1	Norovirus Hu/GII.g-GII.12/Wahroonga/NSW004P/2009/AU, complete genome.	7523
JQ622197.1	Norovirus Hu/GII-4/CBNU2/2007/KR, complete genome.	7583
JX023286.1	Norovirus Hu/GII.4/CHDC5191/1974/USA, complete genome.	7549
JX023285.1	Norovirus Hu/GI.1/8FIIa/1968/USA, complete genome.	7633
JN400599.1	Norovirus Hu/GII-4/CGMH01/2006/TW, complete genome.	7509
JN400600.1	Norovirus Hu/GII-4/CGMH02/2006/TW, complete genome.	7509
JN400601.1	Norovirus Hu/GII-4/CGMH03/2006/TW, complete genome.	7509
JN400602.1	Norovirus Hu/GII-4/CGMH04/2006/TW, complete genome.	7509
JN400603.1	Norovirus Hu/GII-4/CGMH05/2006/TW, complete genome.	7509
JN400604.1	Norovirus Hu/GII-4/CGMH06/2006/TW, complete genome.	7509
JN400605.1	Norovirus Hu/GII-4/CGMH07/2006/TW, complete genome.	7509
JN400606.1	Norovirus Hu/GII-4/CGMH08/2006/TW, complete genome.	7509
JN400607.1	Norovirus Hu/GII-4/CGMH09/2006/TW, complete genome.	7509
JN400608.1	Norovirus Hu/GII-4/CGMH10/2006/TW, complete genome.	7509
JN400609.1	Norovirus Hu/GII-4/CGMH11/2006/TW, complete genome.	7509
JN400610.1	Norovirus Hu/GII-4/CGMH12/2007/TW, complete genome.	7509
JN400611.1	Norovirus Hu/GII-4/CGMH13/2007/TW, complete genome.	7509
JN400612.1	Norovirus Hu/GII-4/CGMH14/2007/TW, complete genome.	7509
JN400613.1	Norovirus Hu/GII-4/CGMH15/2007/TW, complete genome.	7509

JN400614.1	Norovirus Hu/GII-4/CGMH16/2007/TW, complete genome.	7509
JN400615.1	Norovirus Hu/GII-4/CGMH17/2007/TW, complete genome.	7509
JN400616.1	Norovirus Hu/GII-4/CGMH18/2008/TW, complete genome.	7509
JN400617.1	Norovirus Hu/GII-4/CGMH19/2009/TW, complete genome.	7509
JN400618.1	Norovirus Hu/GII-4/CGMH20/2009/TW, complete genome.	7476
JN400619.1	Norovirus Hu/GII-4/CGMH21/2010/TW, complete genome.	7500
JN400620.1	Norovirus Hu/GII-4/CGMH22/2010/TW, complete genome.	7509
JN400621.1	Norovirus Hu/GII-4/CGMH23/2010/TW, complete genome.	7509
JN400622.1	Norovirus Hu/GII-4/CGMH24/2010/TW, complete genome.	7509
JN400623.1	Norovirus Hu/GII-4/CGMH25/2010/TW, complete genome.	7509
JN400624.1	Norovirus Hu/GII-4/CGMH26/2010/TW, complete genome.	7509
JN400625.1	Norovirus Hu/GII-4/CGMH27/2010/TW, complete genome.	7509
JN400626.1	Norovirus Hu/GII-4/CGMH28/2010/TW, complete genome.	7509
JQ798158.1	Norovirus Hu/GII.4/5M/USA/2004, complete genome.	7558
JX126913.1	Norovirus Hu/GII.4/Ohio/7G/2012/USA, complete genome.	7559
JX126912.1	Norovirus Hu/GII.4/Ohio/7I/2012/USA, complete genome.	7558
JX459902.1	Norovirus Hu/GII.4/Berowra/NSW767L/2012/AU, complete genome.	7559
JX459901.1	Norovirus Hu/GII.4/Caringbah/NSW409G/2011/AU, complete genome.	7559
JX459904.1	Norovirus Hu/GII.4/Doonside/NSW536I/2011/AU, complete genome.	7559
JX459903.1	Norovirus Hu/GII.4/Jannali/NSW774M/2011/AU, complete genome.	7559
JX459906.1	Norovirus Hu/GII.4/Miranda/NSW850K/2011/AU, complete genome.	7560
JX459900.1	Norovirus Hu/GII.4/Randwick/NSW882J/2011/AU, complete genome.	7558
JX459905.1	Norovirus Hu/GII.4/Randwick/NSW938K/2011/AU, complete genome.	7560
JX459908.1	Norovirus Hu/GII.4/Sydney/NSW0514/2012/AU, complete genome.	7564
JX459907.1	Norovirus Hu/GII.4/Woonona/NSW3309/2012/AU, complete genome.	7560
JX486101.1	Norovirus Rn/GV/HKU_CT2/HKG/2011 polyprotein gene, partial cds; and VP1 capsid protein and VP2 minor structural protein genes, complete cds.	7528
JX486102.1	Norovirus Rn/GV/HKU_KT/HKG/2012 polyprotein, VP1 capsid protein, and VP2 minor structural protein genes, complete cds.	7542
JX846924.1	Norovirus Hu/NLV/GII.3/Norwalk/HK71/1978/CHN, complete genome.	7527
JX846927.1	Norovirus Hu/NLV/GII.9/Norwalk/CHDC4073/1984/USA, complete genome.	7534
KC175342.1	Norovirus Hu/Norwalk/10034/2009/VNM, complete genome.	7511
KC175343.1	Norovirus Hu/Norwalk/10051/2009/VNM, complete genome.	7511

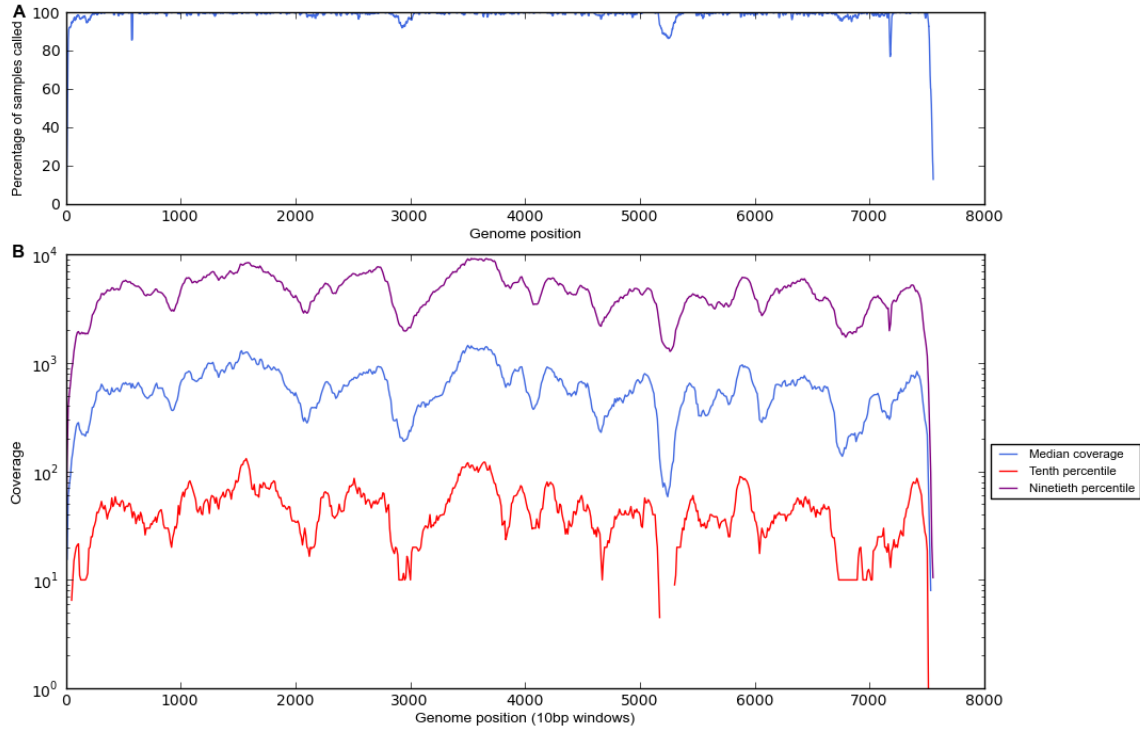
KC175344.1	Norovirus Hu/Norwalk/10054/2009/VNM, complete genome.	7511
KC175345.1	Norovirus Hu/Norwalk/10062/2009/VNM, complete genome.	7511
KC175346.1	Norovirus Hu/Norwalk/10074/2009/VNM, complete genome.	7511
KC175347.1	Norovirus Hu/Norwalk/10075/2009/VNM, complete genome.	7511
KC175348.1	Norovirus Hu/Norwalk/10078/2009/VNM, complete genome.	7511
KC175349.1	Norovirus Hu/Norwalk/10079/2009/VNM, complete genome.	7511
KC175350.1	Norovirus Hu/Norwalk/10110/2009/VNM, complete genome.	7511
KC175351.1	Norovirus Hu/Norwalk/10114/2009/VNM, complete genome.	7511
KC175352.1	Norovirus Hu/Norwalk/10116/2009/VNM, complete genome.	7511
KC175353.1	Norovirus Hu/Norwalk/10129/2009/VNM, complete genome.	7511
KC175354.1	Norovirus Hu/Norwalk/10136/2009/VNM, complete genome.	7511
KC175355.1	Norovirus Hu/Norwalk/10137/2009/VNM, complete genome.	7511
KC175356.1	Norovirus Hu/Norwalk/10145/2009/VNM, complete genome.	7511
KC175357.1	Norovirus Hu/Norwalk/10148/2009/VNM, complete genome.	7511
KC175358.1	Norovirus Hu/Norwalk/10158/2009/VNM, complete genome.	7511
KC175359.1	Norovirus Hu/Norwalk/10160/2009/VNM, complete genome.	7511
KC175360.1	Norovirus Hu/Norwalk/10162/2009/VNM, complete genome.	7511
KC175361.1	Norovirus Hu/Norwalk/10163/2009/VNM, complete genome.	7511
KC175362.1	Norovirus Hu/Norwalk/10169/2009/VNM, complete genome.	7511
KC175363.1	Norovirus Hu/Norwalk/10173/2009/VNM, complete genome.	7511
KC175364.1	Norovirus Hu/Norwalk/10176/2009/VNM, complete genome.	7511
KC175365.1	Norovirus Hu/Norwalk/10177/2009/VNM, complete genome.	7511
KC175366.1	Norovirus Hu/Norwalk/10179/2009/VNM, complete genome.	7511
KC175367.1	Norovirus Hu/Norwalk/10182/2009/VNM, complete genome.	7511
KC175368.1	Norovirus Hu/Norwalk/10183/2009/VNM, complete genome.	7511
KC175369.1	Norovirus Hu/Norwalk/10194/2009/VNM, complete genome.	7511
KC175370.1	Norovirus Hu/Norwalk/10199/2009/VNM, complete genome.	7511
KC175371.1	Norovirus Hu/Norwalk/10203/2009/VNM, complete genome.	7511
KC175372.1	Norovirus Hu/Norwalk/10204/2009/VNM, complete genome.	7511
KC175373.1	Norovirus Hu/Norwalk/10222/2009/VNM, complete genome.	7511
KC175374.1	Norovirus Hu/Norwalk/10223/2009/VNM, complete genome.	7511
KC175375.1	Norovirus Hu/Norwalk/10235/2009/VNM, complete genome.	7511
KC175376.1	Norovirus Hu/Norwalk/10236/2009/VNM, complete genome.	7511

KC175377.1	Norovirus Hu/Norwalk/10238/2009/VNM, complete genome.	7511
KC175378.1	Norovirus Hu/Norwalk/10247/2009/VNM, complete genome.	7511
KC175379.1	Norovirus Hu/Norwalk/10255/2009/VNM, complete genome.	7511
KC175380.1	Norovirus Hu/Norwalk/10285/2010/VNM, complete genome.	7511
KC175381.1	Norovirus Hu/Norwalk/10296/2010/VNM, complete genome.	7511
KC175382.1	Norovirus Hu/Norwalk/10313/2010/VNM, complete genome.	7511
KC175383.1	Norovirus Hu/Norwalk/10325/2010/VNM, complete genome.	7511
KC175384.1	Norovirus Hu/Norwalk/10328/2010/VNM, complete genome.	7511
KC175385.1	Norovirus Hu/Norwalk/10368/2010/VNM, complete genome.	7511
KC175386.1	Norovirus Hu/Norwalk/10378/2010/VNM, complete genome.	7511
KC175387.1	Norovirus Hu/Norwalk/10386/2010/VNM, complete genome.	7511
KC175388.1	Norovirus Hu/Norwalk/20008/2009/VNM, complete genome.	7511
KC175389.1	Norovirus Hu/Norwalk/20010/2009/VNM, complete genome.	7511
KC175390.1	Norovirus Hu/Norwalk/20014/2009/VNM, complete genome.	7511
KC175391.1	Norovirus Hu/Norwalk/20016/2009/VNM, complete genome.	7511
KC175392.1	Norovirus Hu/Norwalk/20019/2009/VNM, complete genome.	7511
KC175393.1	Norovirus Hu/Norwalk/20033/2009/VNM, complete genome.	7511
KC175394.1	Norovirus Hu/Norwalk/20035/2009/VNM, complete genome.	7511
KC175395.1	Norovirus Hu/Norwalk/20044/2009/VNM, complete genome.	7511
KC175396.1	Norovirus Hu/Norwalk/20047/2009/VNM, complete genome.	7511
KC175397.1	Norovirus Hu/Norwalk/20066/2009/VNM, complete genome.	7511
KC175398.1	Norovirus Hu/Norwalk/20067/2009/VNM, complete genome.	7511
KC175399.1	Norovirus Hu/Norwalk/20069/2009/VNM, complete genome.	7511
KC175400.1	Norovirus Hu/Norwalk/20092/2009/VNM, complete genome.	7511
KC175401.1	Norovirus Hu/Norwalk/20093/2009/VNM, complete genome.	7511
KC175402.1	Norovirus Hu/Norwalk/20094/2009/VNM, complete genome.	7511
KC175403.1	Norovirus Hu/Norwalk/20118/2009/VNM, complete genome.	7511
KC175404.1	Norovirus Hu/Norwalk/20122/2009/VNM, complete genome.	7511
KC175405.1	Norovirus Hu/Norwalk/20123/2009/VNM, complete genome.	7511
KC175406.1	Norovirus Hu/Norwalk/20128/2009/VNM, complete genome.	7511
KC175407.1	Norovirus Hu/Norwalk/20135/2009/VNM, complete genome.	7511
KC175408.1	Norovirus Hu/Norwalk/20139/2009/VNM, complete genome.	7511
KC175409.1	Norovirus Hu/Norwalk/20140/2009/VNM, complete genome.	7511

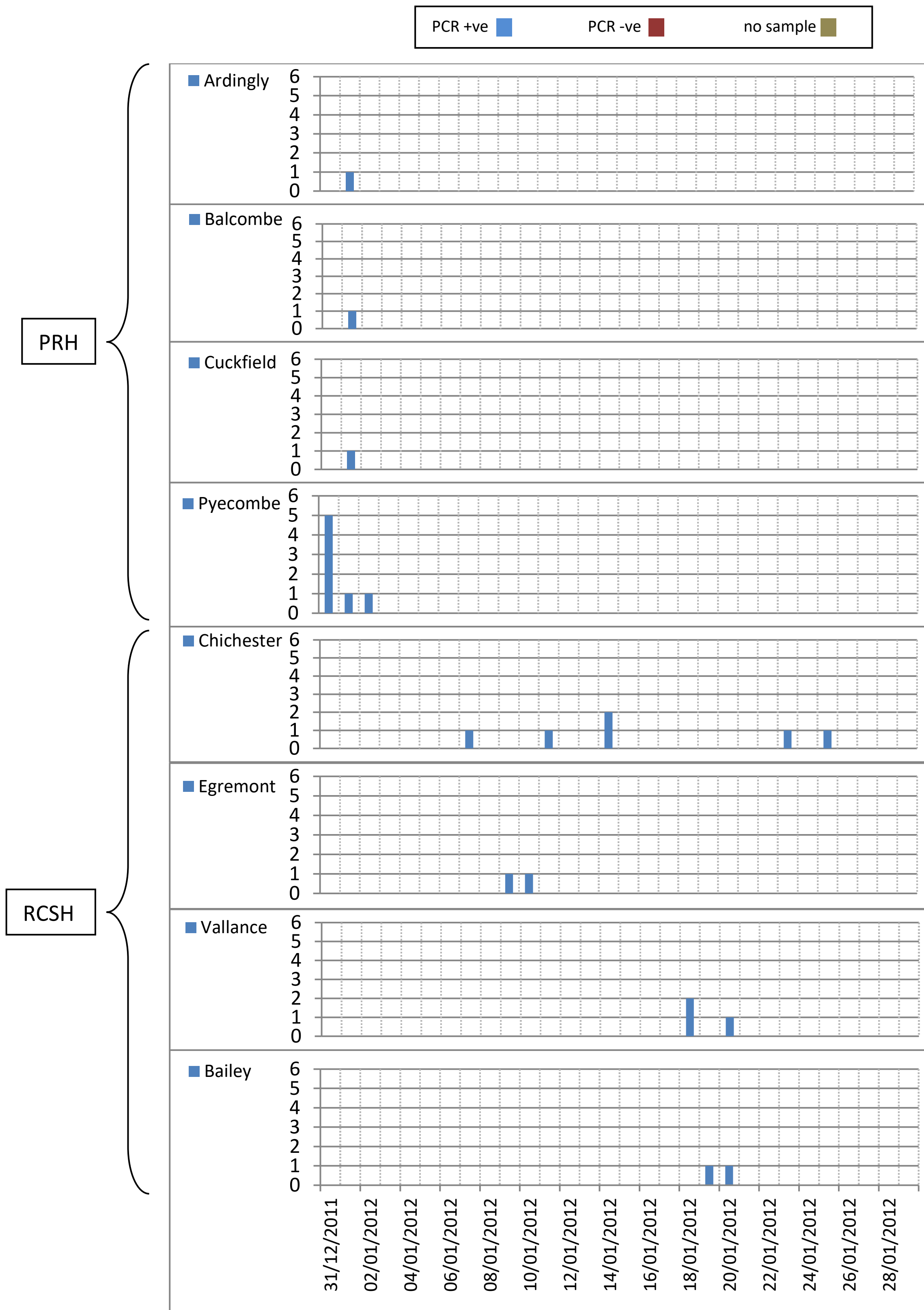
KC175410.1	Norovirus Hu/Norwalk/20142/2009/VNM, complete genome.	7511
KC175323.1	Norovirus Hu/GII.4/Hong Kong/CUHK3630/2012/CHN, complete genome.	7559
KC013592.1	Norovirus Hu/GII.4/HS191/2004/USA, complete genome.	7556

Appendix 3. Coverage and percentage plots of 61 HiSeq sequences.

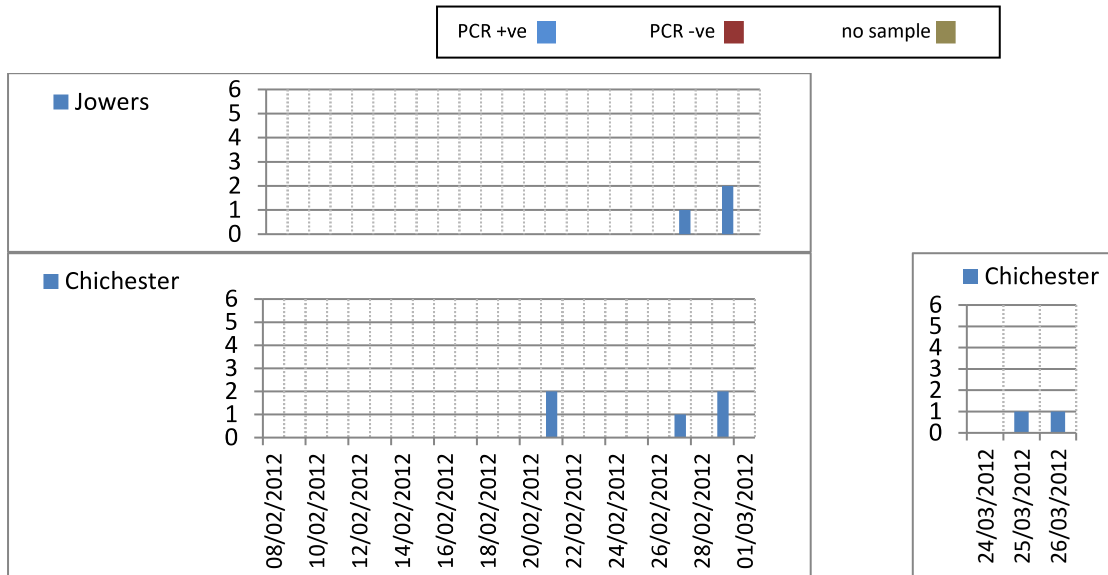
A – percentage of the 61 Norovirus HiSeq samples which could be called at each position in the genome, averaged across 5bp windows. B – the median, tenth percentile, and ninetieth percentile of the coverage across the genome for the 61 HiSeq samples. Note: 10th percentile is 0 where not shown.



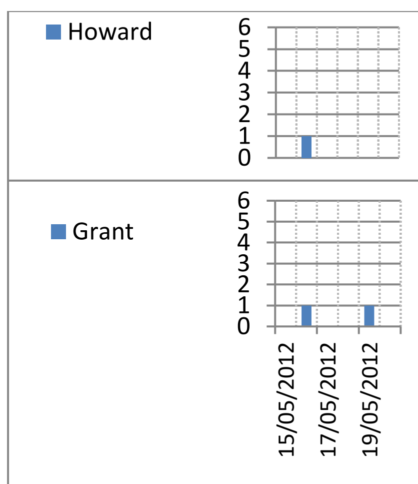
Appendix 4. Epidemiological histograms for outbreaks in Brighton and Leeds



Distribution of symptomatic cases within the Brighton and Sussex county hospitals, which are divided into two main sites: Princess Royal Hospital and Royal Sussex County Hospital. The distance between these two sites is 15 miles.



Distribution of cases in February 2012 and March 2012 (Brighton). Cases appear to recur in Chichester ward, seen firstly in January, then February and then March.



Distribution of cases in May 2012 (Brighton)

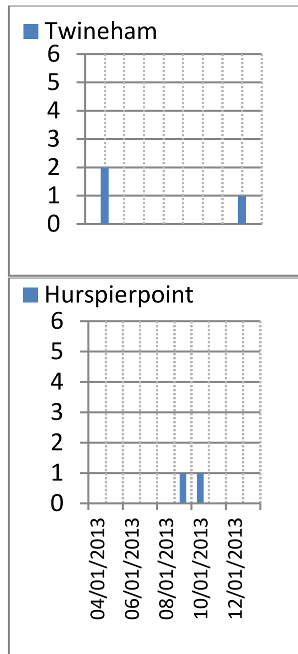
PRH
RSCH

Balcombe
Hurst
CCU
Vallance
Bailey
L8 Tower
Chichester
Bristol
Jowers

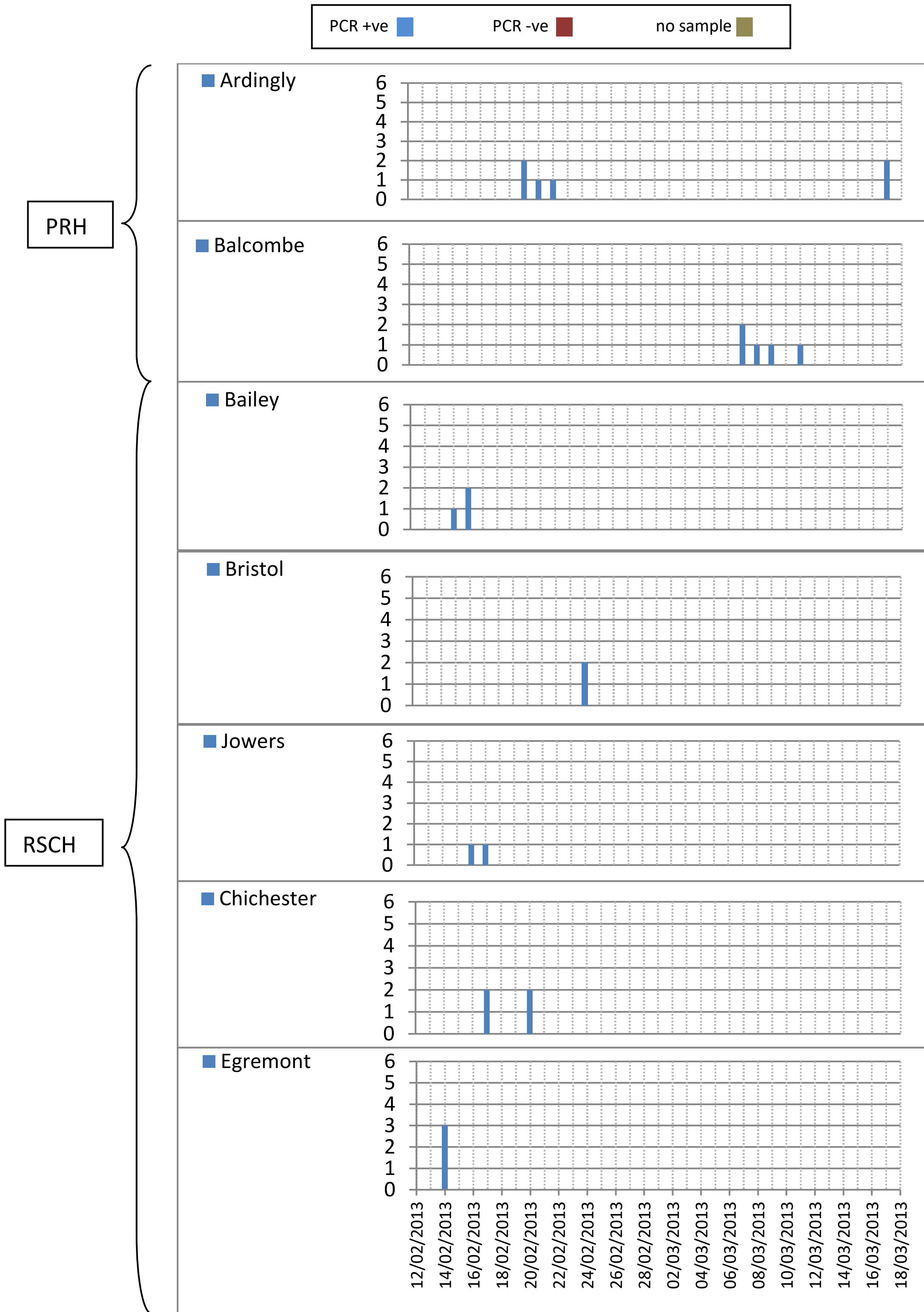
6
5
4
3
2
1
0

01/11/2012
03/11/2012
05/11/2012
07/11/2012
09/11/2012
11/11/2012
13/11/2012
15/11/2012
17/11/2012
19/11/2012
21/11/2012
23/11/2012
25/11/2012
27/11/2012
29/11/2012
01/12/2012
03/12/2012
05/12/2012
07/12/2012
09/12/2012

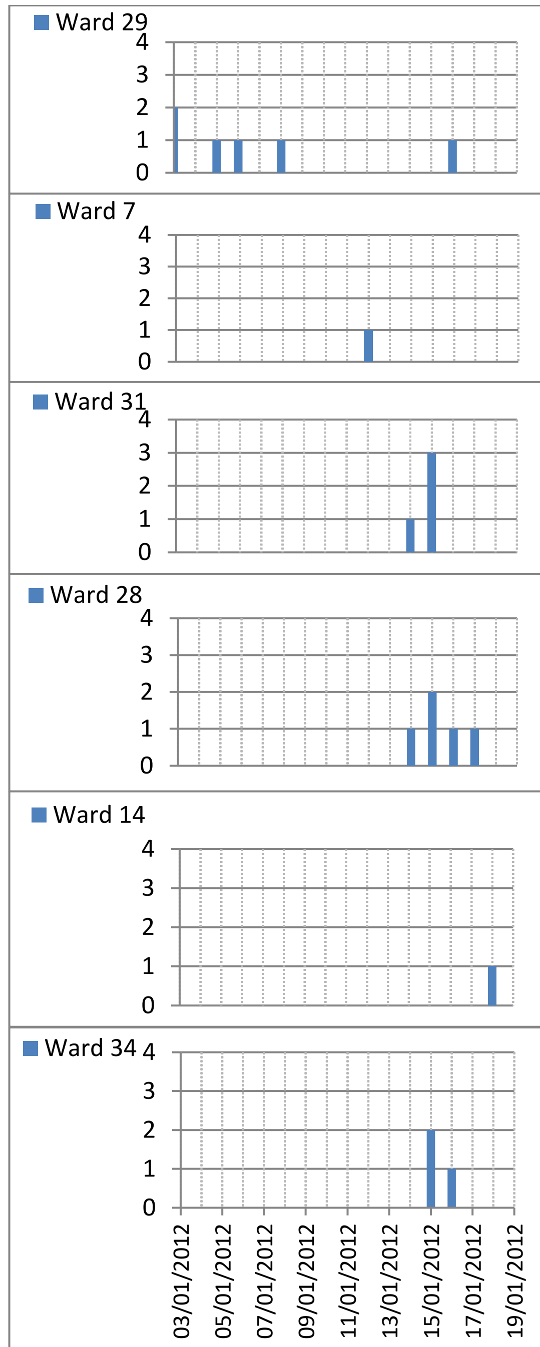
Ward distribution of cases during November and December 2012 (Brighton). Note that a few wards only had sporadic cases although they have been included to illustrate when they occurred relative to the other cases.



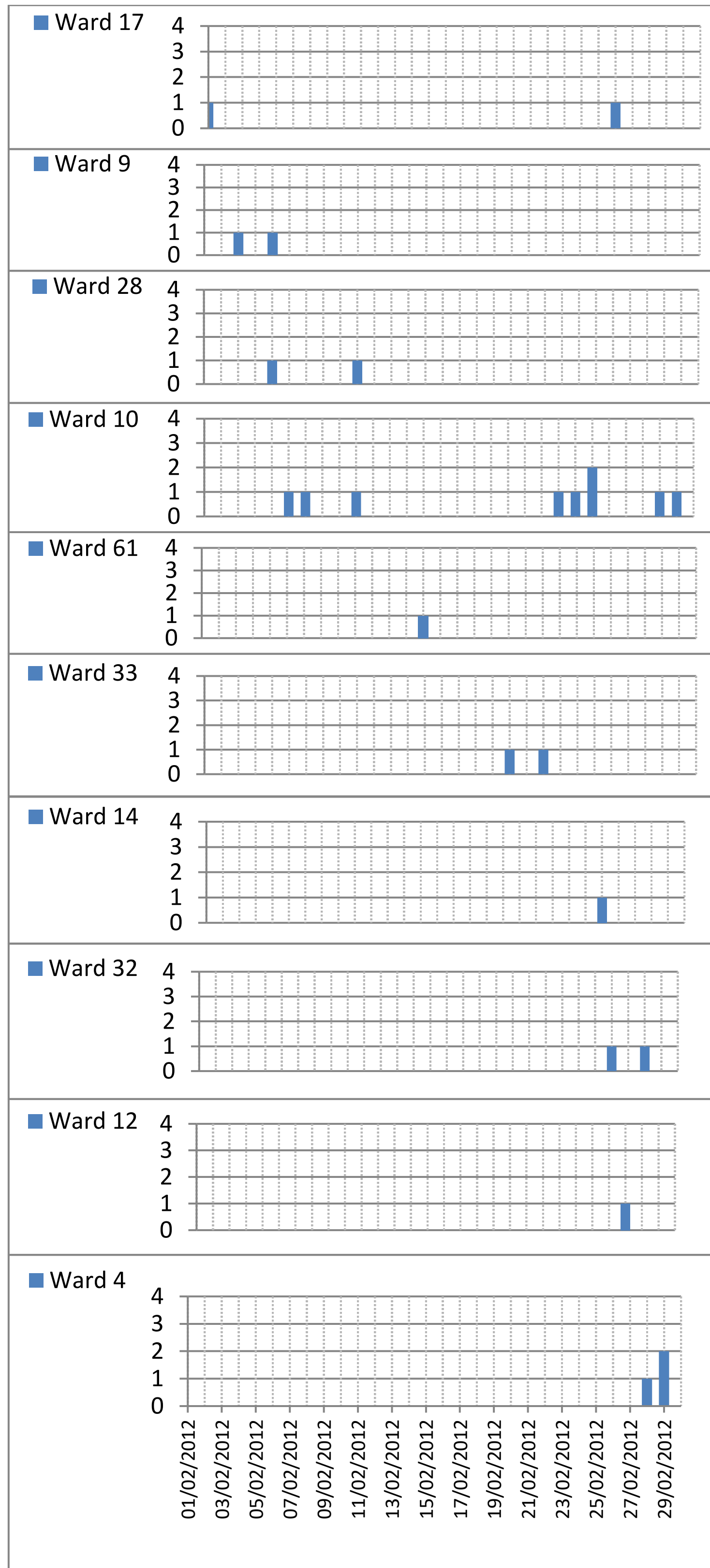
Distribution of cases in January 2013 (Brighton)



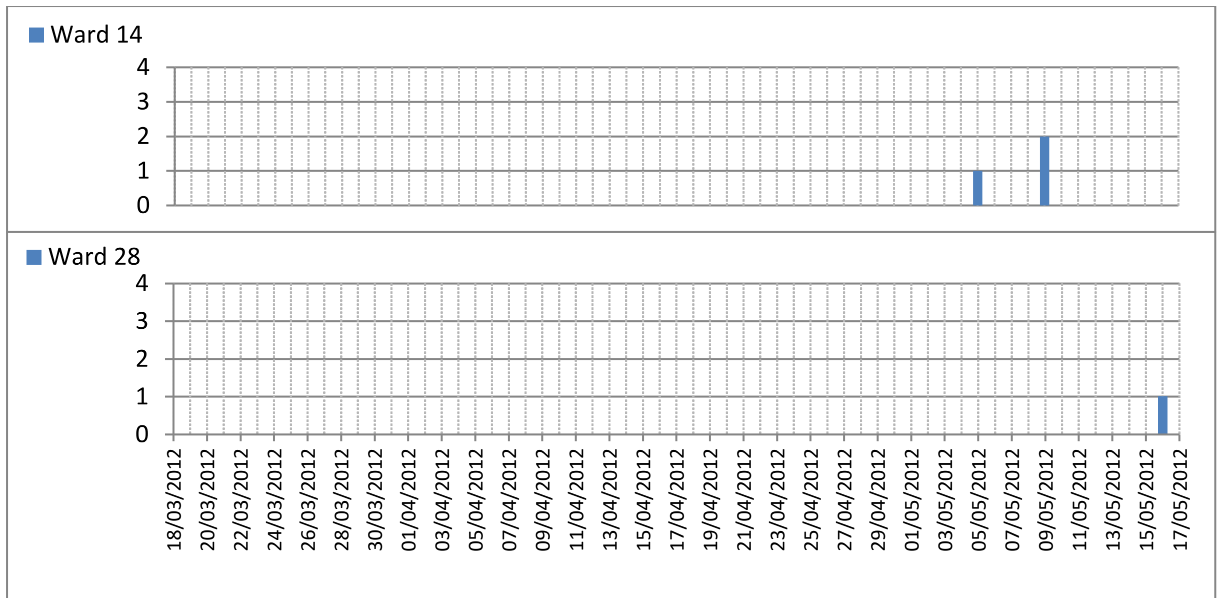
Distribution of ward cases between February - March 2013 (Brighton)



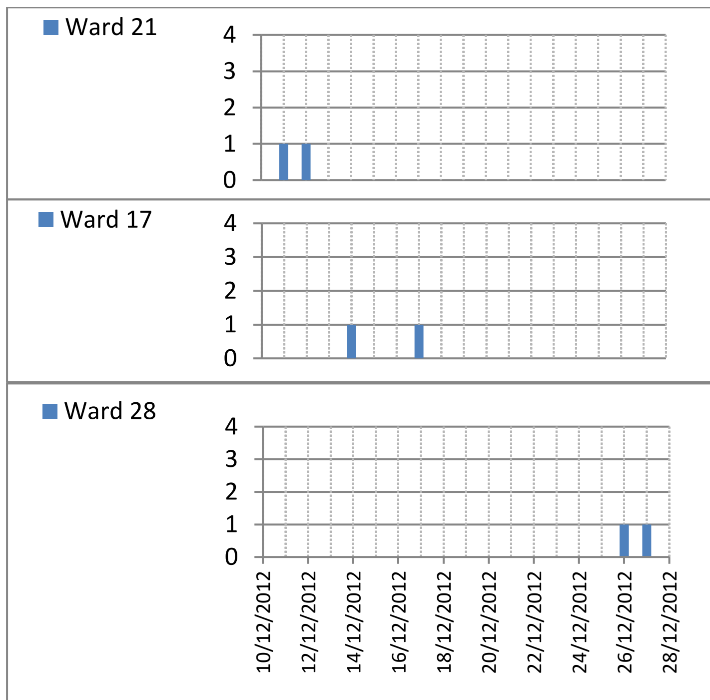
Distribution of ward cases between January 2012 (Leeds)



Distribution of ward cases within February 2012 (Leeds)



Distribution of ward cases within April 2012 (Leeds)



Distribution of ward cases within December 2012 (Leeds)

Appendix 5. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples season 2009-2010.

	6632	6633	7885	7895	7907	12454	12501	12511	7933	12439	7910	12464	7943	7915	12438	12470	12486	7957	7939	12431	12478	12517
6632		109	109	109	112	110	110	109	107	107	109	110	111	111	110	111	111	112	111	105	128	197
6633	109		0	0	3	1	1	0	4	4	0	1	4	4	1	6	2	7	2	81	101	177
7885	109	0		0	3	1	1	0	4	4	0	1	4	4	1	6	2	7	2	81	101	177
7895	109	0	0		3	1	1	0	4	4	0	1	4	4	1	6	2	7	2	81	101	177
7907	112	3	3	3		4	4	3	7	7	3	4	7	7	4	9	5	10	5	84	104	180
12454	110	1	1	1	4		2	1	5	5	1	2	5	5	2	7	3	8	3	82	102	178
12501	110	1	1	1	4	2		1	5	5	1	0	5	5	0	7	3	8	3	82	102	178
12511	109	0	0	0	3	1	1		4	4	0	1	4	4	1	6	2	7	2	81	101	177
7933	107	4	4	4	7	5	5	4		0	4	5	8	8	5	10	6	11	6	79	99	175
12439	107	4	4	4	7	5	5	4	0		4	5	8	8	5	10	6	11	6	79	99	175
7910	109	0	0	0	3	1	1	0	4	4		1	4	4	1	6	2	7	2	81	101	177
12464	110	1	1	1	4	2	0	1	5	5	1		5	5	0	7	3	8	3	82	102	178
7943	111	4	4	4	7	5	5	4	8	8	4	5		6	5	8	6	9	6	83	103	179
7915	111	4	4	4	7	5	5	4	8	8	4	5	6		5	4	6	9	6	83	103	179
12438	110	1	1	1	4	2	0	1	5	5	1	0	5	5		7	3	8	3	82	102	178
12470	111	6	6	6	9	7	7	6	10	10	6	7	8	4	7		8	11	8	83	103	179
12486	111	2	2	2	5	3	3	2	6	6	2	3	6	6	3	8		9	4	83	103	179
7957	112	7	7	7	10	8	8	7	11	11	7	8	9	9	8	11	9		9	84	106	180
7939	111	2	2	2	5	3	3	2	6	6	2	3	6	6	3	8	4	9		83	103	179
12431	105	81	81	81	84	82	82	81	79	79	81	82	83	83	82	83	83	84	83		100	185
12478	128	101	101	101	104	102	102	101	99	99	101	102	103	103	102	103	103	106	103	100		197
12517	197	177	177	177	180	178	178	177	175	175	177	178	179	179	178	179	179	180	179	185	197	

Appendix 6. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples season 2010-2011

	5935	7952	7964	7940	7879	7891	7918	7926	7948	7962	9423	7920	7960	7886	7930	7941	12440	12448	10778	7880	7932	7892	7904	7916	7966	7963	7882	7884	7890	7900	7954	7912	7936	9685	12432
5935		0	0	46	59	58	58	57	57	58	58	58	58	171	170	173	171	171	154	161	161	161	160	161	161	161	270	272	272	270	270	271	270	270	272
7952	0		0	46	59	58	58	57	57	58	58	58	58	171	170	173	171	171	154	161	161	161	160	161	161	161	270	272	272	270	270	271	270	270	272
7964	0	0		46	59	58	58	57	57	58	58	58	58	171	170	173	171	171	154	161	161	161	160	161	161	161	270	272	272	270	270	271	270	270	272
7940	46	46	46		47	48	48	47	47	48	48	48	48	165	162	165	163	163	148	149	149	149	148	149	149	149	262	264	264	262	262	263	262	262	264
7879	59	59	59	47		1	3	2	2	3	3	1	1	162	163	166	164	164	145	152	152	152	151	152	152	152	265	267	267	265	265	266	265	265	267
7891	58	58	58	48	1		2	1	1	2	2	0	0	161	162	165	163	163	144	151	151	151	150	151	151	151	264	266	266	264	264	265	264	264	266
7918	58	58	58	48	3	2		1	1	2	2	2	2	161	162	165	163	163	144	151	151	151	150	151	151	151	264	266	266	264	264	265	264	264	266
7926	57	57	57	47	2	1	1		0	1	1	1	1	160	161	164	162	162	143	150	150	150	149	150	150	150	263	265	265	263	263	264	263	263	265
7948	57	57	57	47	2	1	1	0		1	1	1	1	160	161	164	162	162	143	150	150	150	149	150	150	150	263	265	265	263	263	264	263	263	265
7962	58	58	58	48	3	2	2	1	1		2	2	2	161	162	165	163	163	144	151	151	151	150	151	151	151	262	264	264	262	262	263	262	262	264
9423	58	58	58	48	3	2	2	1	1	2		2	2	161	162	165	163	163	144	151	151	151	150	151	151	151	264	266	266	264	264	265	264	264	266
7920	58	58	58	48	1	0	2	1	1	2	2		0	161	162	165	163	163	144	151	151	151	150	151	151	151	264	266	266	264	264	265	264	264	266
7960	58	58	58	48	1	0	2	1	1	2	2	0		161	162	165	163	163	144	151	151	151	150	151	151	151	264	266	266	264	264	265	264	264	266
7886	171	171	171	165	162	161	161	160	160	161	161	161	161		114	117	115	115	101	188	188	188	189	188	189	188	278	278	278	278	278	279	278	278	280
7930	170	170	170	162	163	162	162	161	161	162	162	162	162	114		3	1	1	44	188	188	188	189	188	189	188	282	284	284	282	282	283	282	282	284
7941	173	173	173	165	166	165	165	164	164	165	165	165	165	117	3		2	2	47	191	191	191	192	191	192	191	285	287	287	285	285	286	285	285	287
12440	171	171	171	163	164	163	163	162	162	163	163	163	163	115	1	2		0	45	189	189	189	190	189	190	189	283	285	285	283	283	284	283	283	285
12448	171	171	171	163	164	163	163	162	162	163	163	163	163	115	1	2	0		45	189	189	189	190	189	190	189	283	285	285	283	283	284	283	283	285
10778	154	154	154	148	145	144	144	143	143	144	144	144	144	101	44	47	45	45		171	171	171	172	171	172	171	270	272	272	270	270	271	270	270	272
7880	161	161	161	149	152	151	151	150	150	151	151	151	151	188	188	191	189	189	171		0	0	1	0	3	2	275	277	277	275	275	276	275	275	277
7932	161	161	161	149	152	151	151	150	150	151	151	151	151	188	188	191	189	189	171	0		0	1	0	3	2	275	277	277	275	275	276	275	275	277
7892	161	161	161	149	152	151	151	150	150	151	151	151	151	188	188	191	189	189	171	0	0		1	0	3	2	275	277	277	275	275	276	275	275	277
7904	160	160	160	148	151	150	150	149	149	150	150	150	150	189	189	192	190	190	172	1	1	1		1	4	3	276	278	278	276	276	277	276	276	278
7916	161	161	161	149	152	151	151	150	150	151	151	151	151	188	188	191	189	189	171	0	0	0	1		3	2	275	277	277	275	275	276	275	275	277
7966	161	161	161	149	152	151	151	150	150	151	151	151	151	189	189	192	190	190	172	3	3	3	4	3		1	276	278	278	276	276	277	276	276	278
7963	161	161	161	149	152	151	151	150	150	151	151	151	151	188	188	191	189	189	171	2	2	2	3	2	1		275	277	277	275	275	276	275	275	277
7882	270	270	270	262	265	264	264	263	263	262	264	264	264	278	282	285	283	283	270	275	275	275	276	275	276	275		2	2	0	0	1	0	0	2
7884	272	272	272	264	267	266	266	265	265	264	266	266	266	278	284	287	285	285	272	277	277	277	278	277	278	277	2		0	2	2	3	2	2	4
7890	272	272	272	264	267	266	266	265	265	264	266	266	266	278	284	287	285	285	272	277	277	277	278	277	278	277	2	0		2	2	3	2	2	4
7900	270	270	270	262	265	264	264	263	263	262	264	264	264	278	282	285	283	283	270	275	275	275	276	275	276	275	0	2	2		0	1	0	0	2
7954	270	270	270	262	265	264	264	263	263	262	264	264	264	278	282	285	283	283	270	275	275	275	276	275	276	275	0	2	2	0		1	0	0	2
7912	271	271	271	263	266	265	265	264	264	263	265	265	265	279	283	286	284	284	271	276	276	276	277	276	277	276	1	3	3	1	1		1	1	1
7936	270	270	270	262	265	264	264	263	263	262	264	264	264	278	282	285	283	283	270	275	275	275	276	275	276	275	0	2	2	0	0	1		0	2
9685	270	270	270	262	265	264	264	263	263	262	264	264	264	278	282	285	283	283	270	275	275	275	276	275	276	275	0	2	2	0	0	1	0		2
12432	272	272	272	264	267	266	266	265	265	264	266	266	266	280	284	287	285	285	272	277	277	277	278	277	278	277	2	4	4	2	2	1	2	2	

Appendix 7. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples season 2011-2012

	8640	8646	8650	9406	10847	9689	10794	10795	10842	10846	10840	10863	10851	9336	9377	10839	9376	9380	9396	10856	9389	10796	10798	10802	10855	10797	10800	10807
8640		2	0	0	0	0	2	1	1	0	1	2	5	3	1	16	176	179	179	177	178	179	181	181	177	180	180	180
8646	2		2	2	2	2	4	3	3	2	3	4	5	3	1	18	176	179	179	177	178	179	181	181	177	180	180	180
8650	0	2		0	0	0	2	1	1	0	1	2	5	3	1	16	176	179	179	177	178	179	181	181	177	180	180	180
9406	0	2	0		0	0	2	1	1	0	1	2	5	3	1	16	176	179	179	177	178	179	181	181	177	180	180	180
10847	0	2	0	0		0	2	1	1	0	1	2	5	3	1	16	176	179	179	177	178	179	181	181	177	180	180	180
9689	0	2	0	0	0		2	1	1	0	1	2	5	3	1	16	176	179	179	177	178	179	181	181	177	180	180	180
10794	2	4	2	2	2	2		3	3	2	3	4	7	5	3	16	176	179	179	177	178	179	181	181	177	180	180	180
10795	1	3	1	1	1	1	3		2	1	2	3	6	4	2	17	177	180	180	178	179	180	182	182	178	181	181	181
10842	1	3	1	1	1	1	3	2		1	2	3	6	4	2	17	177	180	180	178	179	180	182	182	178	181	181	181
10846	0	2	0	0	0	0	2	1	1		1	2	5	3	1	16	176	179	179	177	178	179	181	181	177	180	180	180
10840	1	3	1	1	1	1	3	2	2	1		3	6	4	2	17	177	180	180	178	179	180	182	182	178	181	181	181
10863	2	4	2	2	2	2	4	3	3	2	3		7	5	3	18	176	179	179	177	178	179	181	181	177	180	180	180
10851	5	5	5	5	5	5	7	6	6	5	6	7		4	4	21	175	178	178	176	177	178	180	180	176	179	179	179
9336	3	3	3	3	3	3	5	4	4	3	4	5	4		2	19	175	178	178	176	177	178	180	180	176	179	179	179
9377	1	1	1	1	1	1	3	2	2	1	2	3	4	2		17	175	178	178	176	177	178	180	180	176	179	179	179
10839	16	18	16	16	16	16	16	17	17	16	17	18	21	19	17		164	167	167	165	166	167	169	169	165	168	168	168
9376	176	176	176	176	176	176	176	177	177	176	177	176	175	175	175	164		3	3	1	2	4	5	5	1	4	4	4
9380	179	179	179	179	179	179	179	180	180	179	180	179	178	178	178	167	3		6	4	5	7	8	8	4	7	7	7
9396	179	179	179	179	179	179	179	180	180	179	180	179	178	178	178	167	3	6		4	5	7	8	8	4	7	7	7
10856	177	177	177	177	177	177	177	178	178	177	178	177	176	176	176	165	1	4	4		3	5	6	6	2	5	5	5
9389	178	178	178	178	178	178	178	179	179	178	179	178	177	177	177	166	2	5	5	3		6	7	7	3	6	6	6
10796	179	179	179	179	179	179	179	180	180	179	180	179	178	178	178	167	4	7	7	5	6		3	3	5	2	2	2
10798	181	181	181	181	181	181	181	182	182	181	182	181	180	180	180	169	5	8	8	6	7	3		0	6	1	1	1
10802	181	181	181	181	181	181	181	182	182	181	182	181	180	180	180	169	5	8	8	6	7	3	0		6	1	1	1
10855	177	177	177	177	177	177	177	178	178	177	178	177	176	176	176	165	1	4	4	2	3	5	6	6		5	5	5
10797	180	180	180	180	180	180	180	181	181	180	181	180	179	179	179	168	4	7	7	5	6	2	1	1	5		0	0
10800	180	180	180	180	180	180	180	181	181	180	181	180	179	179	179	168	4	7	7	5	6	2	1	1	5	0		0
10807	180	180	180	180	180	180	180	181	181	180	181	180	179	179	179	168	4	7	7	5	6	2	1	1	5	0	0	
10844	177	177	177	177	177	177	177	178	178	177	178	177	176	176	176	165	1	4	4	2	3	5	6	6	2	5	5	5
10803	181	181	181	181	181	181	181	182	182	181	182	181	180	180	180	169	5	8	8	6	7	3	2	2	6	1	1	1
10817	181	181	181	181	181	181	181	182	182	181	182	181	180	180	180	169	5	8	8	6	7	3	2	2	6	1	1	1
10804	184	184	184	184	184	184	184	185	185	184	185	184	183	183	183	172	8	11	11	9	10	6	5	5	9	4	4	4
10820	179	179	179	179	179	179	179	180	180	179	180	179	178	178	178	167	5	8	8	6	7	3	2	2	6	1	1	1
10805	182	182	182	182	182	182	182	183	183	182	183	182	181	181	181	170	6	9	9	7	8	4	3	3	7	2	2	2
10806	182	182	182	182	182	182	182	183	183	182	183	182	181	181	181	170	6	9	9	7	8	4	3	3	7	2	2	2
10819	181	181	181	181	181	181	181	182	182	181	182	181	180	180	180	169	5	8	8	6	7	3	0	0	6	1	1	1
9359	195	195	195	195	195	195	195	196	196	195	196	195	194	194	194	186	165	168	168	166	167	168	170	170	166	169	169	169
10826	212	212	212	212	212	212	212	213	213	212	213	212	213	213	211	196	195	198	198	196	197	196	198	198	196	197	197	197
10838	212	212	212	212	212	212	212	213	213	212	213	212	213	213	211	196	195	198	198	196	197	196	198	198	196	197	197	197
10828	212	212	212	212	212	212	212	213	213	212	213	212	213	213	211	196	195	198	198	196	197	196	198	198	196	197	197	197
10833	212	212	212	212	212	212	212	213	213	212	213	212	213	213	211	196	195	198	198	196	197	196	198	198	196	197	197	197
10831	214	214	214	214	214	214	214	215	215	214	215	214	215	215	213	198	197	200	200	198	199	198	200	200	198	199	199	199
10834	212	212	212	212	212	212	212	213	213	212	213	212	213	213	211	196	195	198	198	196	197	196	198	198	196	197	197	197
12461	255	255	255	255	255	255	255	256	256	255	256	255	254	254	254	250	227	228	230	228	229	228	230	230	228	229	229	229
12469	255	255	255	255	255	255	255	256	256	255	256	255	254	254	254	250	227	228	230	228	229	228	230	230	228	229	229	229
12485	255	255	255	255	255	255	255	256	256	255	256	255	254	254	254	250	227	228	230	228	229	228	230	230	228	229	229	229
12493	254	254	254	254	254	254	254	255	255	254	255	254	253	253	253	249	226	227	229	227	228	227	229	229	227	228	228	228
10866	950	950	950	950	950	950	949	951	951	950	949	950	947	949	949	950	958	959	961	959	958	959	959	959	959	960	960	960

Appendix 7. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples season 2011-2012 (cont'd)

	10844	10803	10817	10804	10820	10805	10806	10819	9359	10826	10838	10828	10833	10831	10834	12461	12469	12485	12493	10866
8640	177	181	181	184	179	182	182	181	195	212	212	212	212	214	212	255	255	255	254	950
8646	177	181	181	184	179	182	182	181	195	212	212	212	212	214	212	255	255	255	254	950
8650	177	181	181	184	179	182	182	181	195	212	212	212	212	214	212	255	255	255	254	950
9406	177	181	181	184	179	182	182	181	195	212	212	212	212	214	212	255	255	255	254	950
10847	177	181	181	184	179	182	182	181	195	212	212	212	212	214	212	255	255	255	254	950
9689	177	181	181	184	179	182	182	181	195	212	212	212	212	214	212	255	255	255	254	950
10794	177	181	181	184	179	182	182	181	195	212	212	212	212	214	212	255	255	255	254	949
10795	178	182	182	185	180	183	183	182	196	213	213	213	213	215	213	256	256	256	255	951
10842	178	182	182	185	180	183	183	182	196	213	213	213	213	215	213	256	256	256	255	951
10846	177	181	181	184	179	182	182	181	195	212	212	212	212	214	212	255	255	255	254	950
10840	178	182	182	185	180	183	183	182	196	213	213	213	213	215	213	256	256	256	255	949
10863	177	181	181	184	179	182	182	181	195	212	212	212	212	214	212	255	255	255	254	950
10851	176	180	180	183	178	181	181	180	194	213	213	213	213	215	213	254	254	254	253	947
9336	176	180	180	183	178	181	181	180	194	213	213	213	213	215	213	254	254	254	253	949
9377	176	180	180	183	178	181	181	180	194	211	211	211	211	213	211	254	254	254	253	949
10839	165	169	169	172	167	170	170	169	186	196	196	196	196	198	196	250	250	250	249	950
9376	1	5	5	8	5	6	6	5	165	195	195	195	195	197	195	227	227	227	226	958
9380	4	8	8	11	8	9	9	8	168	198	198	198	198	200	198	228	228	228	227	959
9396	4	8	8	11	8	9	9	8	168	198	198	198	198	200	198	230	230	230	229	961
10856	2	6	6	9	6	7	7	6	166	196	196	196	196	198	196	228	228	228	227	959
9389	3	7	7	10	7	8	8	7	167	197	197	197	197	199	197	229	229	229	228	958
10796	5	3	3	6	3	4	4	3	168	196	196	196	196	198	196	228	228	228	227	959
10798	6	2	2	5	2	3	3	0	170	198	198	198	198	200	198	230	230	230	229	959
10802	6	2	2	5	2	3	3	0	170	198	198	198	198	200	198	230	230	230	229	959
10855	2	6	6	9	6	7	7	6	166	196	196	196	196	198	196	228	228	228	227	959
10797	5	1	1	4	1	2	2	1	169	197	197	197	197	199	197	229	229	229	228	960
10800	5	1	1	4	1	2	2	1	169	197	197	197	197	199	197	229	229	229	228	960
10807	5	1	1	4	1	2	2	1	169	197	197	197	197	199	197	229	229	229	228	960
10844		6	6	9	6	7	7	6	166	194	194	194	194	196	194	228	228	228	227	959
10803	6		0	5	2	1	1	2	170	198	198	198	198	200	198	230	230	230	229	961
10817	6	0		5	2	1	1	2	170	198	198	198	198	200	198	230	230	230	229	961
10804	9	5	5		5	6	6	5	173	201	201	201	201	203	201	233	233	233	232	964
10820	6	2	2	5		3	3	2	170	198	198	198	198	200	198	230	230	230	229	959
10805	7	1	1	6	3		2	3	171	199	199	199	199	201	199	231	231	231	230	962
10806	7	1	1	6	3	2		3	171	199	199	199	199	201	199	231	231	231	230	960
10819	6	2	2	5	2	3	3		170	198	198	198	198	200	198	230	230	230	229	959
9359	166	170	170	173	170	171	171	170		208	208	208	208	210	208	228	228	228	227	961
10826	194	198	198	201	198	199	199	198	208		0	0	0	2	0	256	256	256	255	955
10838	194	198	198	201	198	199	199	198	208	0		0	0	2	0	256	256	256	255	955
10828	194	198	198	201	198	199	199	198	208	0	0		0	2	0	256	256	256	255	955
10833	194	198	198	201	198	199	199	198	208	0	0	0		2	0	256	256	256	255	955
10831	196	200	200	203	200	201	201	200	210	2	2	2	2		2	256	256	256	255	955
10834	194	198	198	201	198	199	199	198	208	0	0	0	0	2		256	256	256	255	955
12461	228	230	230	233	230	231	231	230	228	256	256	256	256	256	256		0	0	1	962
12469	228	230	230	233	230	231	231	230	228	256	256	256	256	256	256	0		0	1	962
12485	228	230	230	233	230	231	231	230	228	256	256	256	256	256	256	0	0		1	962
12493	227	229	229	232	229	230	230	229	227	255	255	255	255	255	255	1	1	1		961
10866	959	961	961	964	959	962	960	959	961	955	955	955	955	955	955	962	962	962	961	

Appendix 8. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples season 2012-2013

	14384	**15316**	**15318**	**15320**	**15321**	**15326**	**15328**	**15327**
14384		3	0	0	0	764	765	763
15316	3		3	3	3	765	766	764
15318	0	3		0	0	764	765	763
15320	0	3	0		0	764	765	763
15321	0	3	0	0		764	765	763
15326	764	765	764	764	764		4	1
15328	765	766	765	765	765	4		3
15327	763	764	763	763	763	1	3	

Appendix 9. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples

	9389	10802	10832	10803	9404	7932	12488	6633	7933	7951	12431	10842	7926	12303	7964	12246	12267	12292	12329	12474	12444	12468	12484	12498	10810	14423	12478	6632	7886	10780	12437	12514
9389		7	12	7	169	153	132	118	116	120	120	179	145	149	152	166	167	161	172	172	161	196	173	166	162	168	128	145	182	163	177	182
10802	7		5	2	172	156	135	121	119	123	123	182	148	152	155	169	170	164	175	175	164	199	176	169	165	171	131	146	185	166	182	185
10832	12	5		7	177	161	140	126	124	128	128	187	153	157	160	174	175	169	180	180	169	204	181	174	170	176	136	151	188	171	187	188
10803	7	2	7		172	156	135	121	119	123	123	182	148	152	155	169	170	164	175	175	164	199	176	169	165	171	131	148	185	166	182	185
9404	169	172	177	172		130	101	135	133	137	142	195	164	168	177	191	188	182	193	195	184	220	197	189	184	191	154	162	195	178	190	194
7932	153	156	161	156	130		77	120	118	122	124	174	150	152	161	175	172	168	179	181	162	198	175	167	168	172	138	150	188	170	186	188
12488	132	135	140	135	101	77		96	94	98	105	156	133	133	144	158	157	151	162	164	149	185	162	154	149	160	119	127	169	147	163	165
6633	118	121	126	121	135	120	96		4	14	82	85	108	112	121	135	134	128	139	141	122	158	135	127	126	133	102	109	144	128	144	145
7933	116	119	124	119	133	118	94	4		12	80	83	106	110	119	133	132	126	137	139	120	156	133	125	126	131	100	107	142	126	142	143
7951	120	123	128	123	137	122	98	14	12		86	91	112	116	125	137	138	132	143	145	126	162	139	131	132	137	104	113	147	131	147	149
12431	120	123	128	123	142	124	105	82	80	86		141	82	84	93	107	106	100	111	113	98	134	111	103	102	109	102	106	142	122	138	135
10842	179	182	187	182	195	174	156	85	83	91	141		162	166	177	189	186	182	193	195	178	214	191	183	184	187	163	172	203	185	203	204
7926	145	148	153	148	164	150	133	108	106	112	82	162		44	57	71	72	64	73	75	62	100	77	69	66	71	131	127	160	142	158	155
12303	149	152	157	152	168	152	133	112	110	116	84	166	44		59	73	72	66	75	77	64	100	77	69	72	69	133	133	164	148	164	161
7964	152	155	160	155	177	161	144	121	119	125	93	177	57	59		70	73	65	74	76	63	97	74	68	71	86	142	138	171	153	169	166
12246	166	169	174	169	191	175	158	135	133	137	107	189	71	73	70		35	27	36	38	73	109	86	78	83	98	150	154	187	169	175	180
12267	167	170	175	170	188	172	157	134	132	138	106	186	72	72	73	35		8	29	31	80	115	93	85	88	99	153	152	186	168	176	179
12292	161	164	169	164	182	168	151	128	126	132	100	182	64	66	65	27	8		21	23	72	108	85	77	80	91	147	147	180	162	170	173
12329	172	175	180	175	193	179	162	139	137	143	111	193	73	75	74	36	29	21		4	81	117	94	86	91	100	156	158	189	171	181	184
12474	172	175	180	175	195	181	164	141	139	145	113	195	75	77	76	38	31	23	4		83	119	96	88	93	102	158	160	191	173	183	186
12444	161	164	169	164	184	162	149	122	120	126	98	178	62	64	63	73	80	72	81	83		40	17	9	68	91	147	147	174	160	176	171
12468	196	199	204	199	220	198	185	158	156	162	134	214	100	100	97	109	115	108	117	119	40		35	39	103	127	181	183	209	194	207	205
12484	173	176	181	176	197	175	162	135	133	139	111	191	77	77	74	86	93	85	94	96	17	35		16	81	102	158	158	187	173	187	184
12498	166	169	174	169	189	167	154	127	125	131	103	183	69	69	68	78	85	77	86	88	9	39	16		73	96	150	152	177	165	181	178
10810	162	165	170	165	184	168	149	126	126	132	102	184	66	72	71	83	88	80	91	93	68	103	81	73		99	150	141	170	149	168	169
14423	168	171	176	171	191	172	160	133	131	137	109	187	71	69	86	98	99	91	100	102	91	127	102	96	99		156	158	183	171	187	184

Appendix 9. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples (cont'd)

	12312	12508	15313	10833	12270	15303	12241	12469	12517	14390	15326	15305	15322	15310	9685	14388	14398	14395	15315	14413	15284	15311	15318	14396	14421	15246	15267	14397	14410	15245
9389	190	196	216	197	205	213	194	229	214	291	309	275	273	273	267	828	835	821	818	815	815	814	818	808	811	813	803	816	823	824
10802	193	199	219	198	206	216	197	230	215	294	312	278	276	276	270	831	838	824	821	818	818	817	821	811	814	816	806	819	826	827
10832	198	204	224	203	211	221	202	235	220	297	315	281	279	279	275	833	840	826	823	820	820	819	823	813	816	818	808	821	828	829
10803	193	199	219	198	206	216	197	230	215	292	310	276	274	274	270	829	836	822	819	816	816	815	819	809	812	814	804	817	824	825
9404	200	211	229	208	220	226	212	224	213	296	314	284	282	282	275	823	828	815	812	809	809	810	814	807	809	809	805	816	815	822
7932	205	205	225	200	210	225	200	210	197	295	313	281	279	279	274	825	828	816	813	810	811	814	814	809	813	813	807	818	819	818
12488	182	182	200	181	191	196	175	189	176	270	290	257	255	255	258	810	815	802	799	796	796	797	801	794	798	798	792	805	804	803
6633	162	153	173	160	167	175	168	193	178	261	277	246	244	244	237	812	816	802	799	798	798	796	801	791	792	792	788	797	798	805
7933	160	151	171	158	165	173	166	191	176	259	277	244	242	242	235	811	815	801	798	797	797	795	800	790	791	791	787	796	797	804
7951	165	156	174	163	169	178	172	193	180	265	283	250	248	248	239	815	821	807	804	803	803	801	806	796	797	797	793	802	803	810
12431	153	153	179	157	167	177	168	202	187	259	277	244	242	244	240	801	811	796	793	791	791	790	794	786	787	787	783	792	791	797
10842	220	214	228	213	222	228	220	256	241	311	325	298	296	296	293	829	835	819	818	815	815	817	820	813	813	813	809	818	815	822
7926	174	173	194	178	185	189	186	220	205	276	294	261	259	259	262	820	828	815	812	808	808	809	813	804	807	806	801	810	812	815
12303	179	181	201	182	187	199	192	225	210	275	293	261	259	259	260	807	817	804	801	797	797	800	802	793	796	795	790	799	801	805
7964	183	184	203	184	196	202	200	231	216	281	299	269	268	268	269	816	828	811	808	804	804	801	809	804	805	806	801	810	812	813
12246	199	200	221	203	207	218	208	247	234	292	312	283	279	281	285	816	824	815	812	808	808	809	813	800	805	804	803	810	814	813
12267	198	199	220	200	210	215	205	244	229	290	312	279	277	277	280	814	822	813	810	806	806	809	811	798	803	804	801	808	814	809
12292	192	193	214	194	204	209	201	240	225	288	308	279	277	277	280	814	822	813	810	806	806	807	811	798	803	804	801	808	812	809
12329	201	200	220	205	211	218	212	251	236	297	317	290	288	286	291	824	832	821	818	816	816	817	821	808	813	814	811	818	822	817
12474	203	202	222	207	213	220	214	253	238	299	319	292	290	288	291	826	834	823	820	818	818	819	823	810	815	816	813	820	824	819
12444	190	191	212	194	195	205	203	237	222	284	302	275	271	273	274	816	826	813	812	810	810	809	811	804	805	806	801	810	812	817
12468	224	223	245	228	231	240	237	269	256	318	336	309	305	306	305	844	852	841	840	838	838	837	839	832	833	834	828	837	840	845
12484	203	202	223	207	208	218	214	246	233	297	315	288	284	286	285	825	833	822	821	819	819	818	820	813	814	815	810	819	821	826
12498	195	196	217	198	201	209	206	242	227	289	307	280	276	278	278	818	828	815	814	812	812	811	813	806	807	808	803	812	814	819
10810	183	185	208	190	191	197	205	238	223	279	297	267	265	267	269	817	826	816	813	807	807	808	812	807	808	810	806	811	813	820
14423	198	202	224	205	212	218	214	242	227	297	317	281	277	281	279	823	833	824	819	817	817	820	820	811	812	813	808	818	823	824

Appendix 9. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples (cont'd)

	9389	10802	10832	10803	9404	7932	12488	6633	7933	7951	12431	10842	7926	12303	7964	12246	12267	12292	12329	12474	12444	12468	12484	12498	10810	14423	12478	6632	7886	10780	12437	12514
12478	128	131	136	131	154	138	119	102	100	104	102	163	131	133	142	150	153	147	156	158	147	181	158	150	150	156		129	167	148	166	167
6632	145	146	151	148	162	150	127	109	107	113	106	172	127	133	138	154	152	147	158	160	147	183	158	152	141	158	129		169	148	162	165
7886	182	185	188	185	195	188	169	144	142	147	142	203	160	164	171	187	186	180	189	191	174	209	187	177	170	183	167	169		100	114	115
10780	163	166	171	166	178	170	147	128	126	131	122	185	142	148	153	169	168	162	171	173	160	194	173	165	149	171	148	148	100		74	77
12437	177	182	187	182	190	186	163	144	142	147	138	203	158	164	169	175	176	170	181	183	176	207	187	181	168	187	166	162	114	74		83
12514	182	185	188	185	194	188	165	145	143	149	135	204	155	161	166	180	179	173	184	186	171	205	184	178	169	184	167	165	115	77	83	
12312	190	193	198	193	200	205	182	162	160	165	153	220	174	179	183	199	198	192	201	203	190	224	203	195	183	198	181	176	127	89	99	108
12508	196	199	204	199	211	205	182	153	151	156	153	214	173	181	184	200	199	193	200	202	191	223	202	196	185	202	179	179	127	91	89	102
15313	216	219	224	219	229	225	200	173	171	174	179	228	194	201	203	221	220	214	220	222	212	245	223	217	208	224	201	199	155	121	119	126
10833	197	198	203	198	208	200	181	160	158	163	157	213	178	182	184	203	200	194	205	207	194	228	207	198	190	205	184	184	136	100	100	111
12270	205	206	211	206	220	210	191	167	165	169	167	222	185	187	196	207	210	204	211	213	195	231	208	201	191	212	192	191	141	133	147	150
15303	213	216	221	216	226	225	196	175	173	178	177	228	189	199	202	218	215	209	218	220	205	240	218	209	197	218	201	201	143	135	151	155
12241	194	197	202	197	212	200	175	168	166	172	168	220	186	192	200	208	205	201	212	214	203	237	214	206	205	214	176	192	222	204	216	223
12469	229	230	235	230	224	210	189	193	191	193	202	256	220	225	231	247	244	240	251	253	237	269	246	242	238	242	210	213	254	233	237	245
12517	214	215	220	215	213	197	176	178	176	180	187	241	205	210	216	234	229	225	236	238	222	256	233	227	223	227	199	198	239	218	226	232
14390	291	294	297	292	296	295	270	261	259	265	259	311	276	275	281	292	290	288	297	299	284	318	297	289	279	297	286	288	261	219	234	237
15326	309	312	315	310	314	313	290	277	277	283	277	325	294	293	299	312	312	308	317	319	302	336	315	307	297	317	304	306	275	237	254	251
15305	275	278	281	276	284	281	257	246	244	250	244	298	261	261	269	283	279	279	290	292	275	309	288	280	267	281	271	273	241	204	217	220
15322	273	276	279	274	282	279	255	244	242	248	242	296	259	259	268	279	277	277	288	290	271	305	284	276	265	277	269	271	239	202	215	218
15310	273	276	279	274	282	279	255	244	242	248	244	296	259	259	268	281	277	277	286	288	273	306	286	278	267	281	269	270	239	204	217	220
9685	267	270	275	270	275	274	258	237	235	239	240	293	262	260	269	285	280	280	291	291	274	305	285	278	269	279	261	264	277	270	284	275
14388	828	831	833	829	823	825	810	812	811	815	801	829	820	807	816	816	814	814	824	826	816	844	825	818	817	823	835	819	846	832	829	845
14398	835	838	840	836	828	828	815	816	815	821	811	835	828	817	828	824	822	822	832	834	826	852	833	828	826	833	846	821	855	835	832	846
14395	821	824	826	822	815	816	802	802	801	807	796	819	815	804	811	815	813	813	821	823	813	841	822	815	816	824	829	810	838	824	819	831
15315	818	821	823	819	812	813	799	799	798	804	793	818	812	801	808	812	810	810	818	820	812	840	821	814	813	819	825	809	834	820	815	831
14413	815	818	820	816	809	810	796	798	797	803	791	815	808	797	804	808	806	806	816	818	810	838	819	812	807	817	823	805	832	814	811	829
15284	815	818	820	816	809	811	796	798	797	803	791	815	808	797	804	808	806	806	816	818	810	838	819	812	807	817	823	805	832	814	811	829
15311	814	817	819	815	810	814	797	796	795	801	790	817	809	800	801	809	809	807	817	819	809	837	818	811	808	820	821	800	833	815	812	830
15318	818	821	823	819	814	814	801	801	800	806	794	820	813	802	809	813	811	811	821	823	811	839	820	813	812	820	826	808	837	819	816	834

Appendix 9. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples (cont'd)

	9389	10802	10832	10803	9404	7932	12488	6633	7933	7951	12431	10842	7926	12303	7964	12246	12267	12292	12329	12474	12444	12468	12484	12498	10810	14423	12478	6632	7886	10780	12437	12514
14396	808	811	813	809	807	809	794	791	790	796	786	813	804	793	804	800	798	798	808	810	804	832	813	806	807	811	820	804	831	817	814	828
14421	811	814	816	812	809	813	798	792	791	797	787	813	807	796	805	805	803	803	813	815	805	833	814	807	808	812	819	806	835	818	815	831
15246	813	816	818	814	809	813	798	792	791	797	787	813	806	795	806	804	804	804	814	816	806	834	815	808	810	813	815	806	835	820	815	831
15267	803	806	808	804	805	807	792	788	787	793	783	809	801	790	801	803	801	801	811	813	801	828	810	803	806	808	817	803	831	816	815	827
14397	816	819	821	817	816	818	805	797	796	802	792	818	810	799	810	810	808	808	818	820	810	837	819	812	811	818	826	808	837	823	816	832
14410	823	826	828	824	815	819	804	798	797	803	791	815	812	801	812	814	814	812	822	824	812	840	821	814	813	823	825	811	839	822	825	835
15245	824	827	829	825	822	818	803	805	804	810	797	822	815	805	813	813	809	809	817	819	817	845	826	819	820	824	833	817	844	823	820	836

Appendix 9. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples (cont'd)

	12312	12508	15313	10833	12270	15303	12241	12469	12517	14390	15326	15305	15322	15310	9685	14388	14398	14395	15315	14413	15284	15311	15318	14396	14421	15246	15267	14397	14410	15245
12478	181	179	201	184	192	201	176	210	199	286	304	271	269	269	261	835	846	829	825	823	823	821	826	820	819	815	817	826	825	833
6632	176	179	199	184	191	201	192	213	198	288	306	273	271	270	264	819	821	810	809	805	805	800	808	804	806	806	803	808	811	817
7886	127	127	155	136	141	143	222	254	239	261	275	241	239	239	277	846	855	838	834	832	832	833	837	831	835	835	831	837	839	844
10780	89	91	121	100	133	135	204	233	218	219	237	204	202	204	270	832	835	824	820	814	814	815	819	817	818	820	816	823	822	823
12437	99	89	119	100	147	151	216	237	226	234	254	217	215	217	284	829	832	819	815	811	811	812	816	814	815	815	815	816	825	820
12514	108	102	126	111	150	155	223	245	232	237	251	220	218	220	275	845	846	831	831	829	829	830	834	828	831	831	827	832	835	836
12312		120	143	127	158	163	233	263	248	250	268	233	231	233	292	849	848	837	833	829	829	832	834	826	829	829	827	833	836	837
12508	120		58	121	158	163	235	245	234	246	266	225	223	225	289	827	831	824	820	816	816	817	821	813	814	818	812	817	824	820
15313	143	58		147	183	188	251	268	255	262	280	253	249	253	317	838	844	834	830	827	827	830	832	824	827	829	825	834	837	827
10833	127	121	147		166	177	235	256	241	254	272	243	241	243	303	856	859	847	842	840	840	843	847	837	840	840	836	843	846	845
12270	158	158	183	166		140	247	265	250	283	299	260	258	260	299	847	854	837	835	831	831	834	838	832	833	833	828	840	839	837
15303	163	163	188	177	140		251	274	259	288	304	271	269	269	307	837	849	831	826	826	826	825	829	826	827	827	823	832	833	833
12241	233	235	251	235	247	251		262	247	332	352	319	315	319	308	832	836	829	822	817	817	822	826	812	819	817	815	824	820	830
12469	263	245	268	256	265	274	262		17	344	364	330	328	328	321	838	853	839	838	828	828	831	835	831	834	832	830	839	839	844
12517	248	234	255	241	250	259	247	17		333	353	315	313	313	312	826	842	829	826	818	818	821	825	819	822	820	818	827	825	834
14390	250	246	262	254	283	288	332	344	333		42	86	83	89	349	762	769	755	758	752	752	757	755	746	749	753	745	753	765	763
15326	268	266	280	272	299	304	352	364	353	42		110	107	113	369	773	780	764	767	761	761	766	764	757	760	762	756	764	772	776
15305	233	225	253	243	260	271	319	330	315	86	110		17	29	338	754	759	748	748	740	740	745	743	734	739	741	733	741	755	755
15322	231	223	249	241	258	269	315	328	313	83	107	17		26	334	750	755	744	744	736	736	741	739	730	735	737	729	737	751	751
15310	233	225	253	243	260	269	319	328	313	89	113	29	26		334	747	755	743	743	736	736	740	738	729	734	736	728	736	750	748
9685	292	289	317	303	299	307	308	321	312	349	369	338	334	334		824	831	812	810	808	808	811	809	807	810	814	806	818	822	813
14388	849	827	838	856	847	837	832	838	826	762	773	754	750	747	824		97	118	112	114	110	109	109	111	116	117	117	123	121	148
14398	848	831	844	859	854	849	836	853	842	769	780	759	755	755	831	97		124	120	125	121	116	118	118	123	124	124	130	130	159
14395	837	824	834	847	837	831	829	839	829	755	764	748	744	743	812	118	124		38	81	78	73	73	95	104	103	105	117	111	150
15315	833	820	830	842	835	826	822	838	826	758	767	748	744	743	810	112	120	38		75	72	65	67	87	96	95	97	109	105	144
14413	829	816	827	840	831	826	817	828	818	752	761	740	736	736	808	114	125	81	75		4	29	29	95	100	101	103	115	109	148
15284	829	816	827	840	831	826	817	828	818	752	761	740	736	736	808	110	121	78	72	4		25	25	91	96	97	99	111	105	144
15311	832	817	830	843	834	825	822	831	821	757	766	745	741	740	811	109	116	73	65	29	25		20	88	92	93	96	108	102	141
15318	834	821	832	847	838	829	826	835	825	755	764	743	739	738	809	109	118	73	67	29	25	20		88	93	92	96	108	104	143

Appendix 9. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples (cont'd)

	12312	12508	15313	10833	12270	15303	12241	12469	12517	14390	15326	15305	15322	15310	9685	14388	14398	14395	15315	14413	15284	15311	15318	14396	14421	15246	15267	14397	14410	15245
14396	826	813	824	837	832	826	812	831	819	746	757	734	730	729	807	111	118	95	87	95	91	88	88		33	34	34	64	80	137
14421	829	814	827	840	833	827	819	834	822	749	760	739	735	734	810	116	123	104	96	100	96	92	93	33		23	33	69	81	141
15246	829	818	829	840	833	827	817	832	820	753	762	741	737	736	814	117	124	103	95	101	97	93	92	34	23		34	70	80	146
15267	827	812	825	836	828	823	815	830	818	745	756	733	729	728	806	117	124	105	97	103	99	96	96	34	33	34		70	84	147
14397	833	817	834	843	840	832	824	839	827	753	764	741	737	736	818	123	130	117	109	115	111	108	108	64	69	70	70		92	159
14410	836	824	837	846	839	833	820	839	825	765	772	755	751	750	822	121	130	111	105	109	105	102	104	80	81	80	84	92		152
15245	837	820	827	845	837	833	830	844	834	763	776	755	751	748	813	148	159	150	144	148	144	141	143	137	141	146	147	159	152	

Appendix 10. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples season 2011-2012 (Brighton)

	12428	12434	12450	12490	12329	12466	12474	12482	12436	12444	12483	12491	12435	12467	12475	12451	12498	12452	12459	12460	12468	12476	12484	12492
12428		102	102	104	102	105	104	104	84	83	81	81	82	88	89	88	88	90	88	91	119	90	97	99
12434	102		0	4	0	5	4	4	82	81	79	79	80	86	87	86	86	88	86	89	117	88	95	97
12450	102	0		4	0	5	4	4	82	81	79	79	80	86	87	86	86	88	86	89	117	88	95	97
12490	104	4	4		4	1	0	0	84	83	81	81	82	88	89	88	88	90	88	91	119	90	97	99
12329	102	0	0	4		5	4	4	82	81	79	79	80	86	87	86	86	88	86	89	117	88	95	97
12466	105	5	5	1	5		1	1	85	84	82	82	83	89	90	89	89	91	89	92	120	91	98	100
12474	104	4	4	0	4	1		0	84	83	81	81	82	88	89	88	88	90	88	91	119	90	97	99
12482	104	4	4	0	4	1	0		84	83	81	81	82	88	89	88	88	90	88	91	119	90	97	99
12436	84	82	82	84	82	85	84	84		1	3	3	4	10	11	10	10	12	10	13	41	12	19	21
12444	83	81	81	83	81	84	83	83	1		2	2	3	9	10	9	9	11	9	12	40	11	18	20
12483	81	79	79	81	79	82	81	81	3	2		0	1	7	8	7	7	9	7	10	38	9	16	18
12491	81	79	79	81	79	82	81	81	3	2	0		1	7	8	7	7	9	7	10	38	9	16	18
12435	82	80	80	82	80	83	82	82	4	3	1	1		6	7	6	6	8	6	9	37	8	15	17
12467	88	86	86	88	86	89	88	88	10	9	7	7	6		1	0	0	10	0	11	39	10	17	19
12475	89	87	87	89	87	90	89	89	11	10	8	8	7	1		1	1	11	1	12	40	11	18	20
12451	88	86	86	88	86	89	88	88	10	9	7	7	6	0	1		0	10	0	11	39	10	17	19
12498	88	86	86	88	86	89	88	88	10	9	7	7	6	0	1	0		10	0	11	39	10	17	19
12452	90	88	88	90	88	91	90	90	12	11	9	9	8	10	11	10	10		10	1	29	0	7	9
12459	88	86	86	88	86	89	88	88	10	9	7	7	6	0	1	0	0	10		11	39	10	17	19
12460	91	89	89	91	89	92	91	91	13	12	10	10	9	11	12	11	11	1	11		30	1	8	10
12468	119	117	117	119	117	120	119	119	41	40	38	38	37	39	40	39	39	29	39	30		29	36	38
12476	90	88	88	90	88	91	90	90	12	11	9	9	8	10	11	10	10	0	10	1	29		7	9
12484	97	95	95	97	95	98	97	97	19	18	16	16	15	17	18	17	17	7	17	8	36	7		2
12492	99	97	97	99	97	100	99	99	21	20	18	18	17	19	20	19	19	9	19	10	38	9	2	
12429	194	182	182	184	182	185	184	184	178	177	175	175	176	182	183	182	182	182	182	183	208	182	189	191
12437	194	182	182	184	182	185	184	184	178	177	175	175	176	182	183	182	182	182	182	183	208	182	189	191
12499	192	188	188	190	188	191	190	190	176	175	175	175	176	182	183	182	182	182	182	183	209	182	189	191
12514	190	186	186	188	186	189	188	188	174	173	173	173	174	180	181	180	180	180	180	181	207	180	187	189
12508	207	201	201	203	201	204	203	203	193	192	190	190	191	197	198	197	197	197	197	198	224	197	204	206
12516	209	203	203	205	203	206	205	205	195	194	192	192	193	199	200	199	199	199	199	200	226	199	206	208

Appendix 10. Matrix of single nucleotide variants (SNVs) between hospital outbreak samples season 2011-2012 (cont'd)

	12429	12437	12499	12514	12508	12516
12428	194	194	192	190	207	209
12434	182	182	188	186	201	203
12450	182	182	188	186	201	203
12490	184	184	190	188	203	205
12329	182	182	188	186	201	203
12466	185	185	191	189	204	206
12474	184	184	190	188	203	205
12482	184	184	190	188	203	205
12436	178	178	176	174	193	195
12444	177	177	175	173	192	194
12483	175	175	175	173	190	192
12491	175	175	175	173	190	192
12435	176	176	176	174	191	193
12467	182	182	182	180	197	199
12475	183	183	183	181	198	200
12451	182	182	182	180	197	199
12498	182	182	182	180	197	199
12452	182	182	182	180	197	199
12459	182	182	182	180	197	199
12460	183	183	183	181	198	200
12468	208	208	209	207	224	226
12476	182	182	182	180	197	199
12484	189	189	189	187	204	206
12492	191	191	191	189	206	208
12429		0	86	84	89	91
12437	0		86	84	89	91
12499	86	86		2	105	107
12514	84	84	2		103	105
12508	89	89	105	103		2
12516	91	91	107	105	2	